Semantic techniques for successful related-document retrieval from speech-recognized news items

THESIS under supervision of Laura Hollink (VU) and Corné Versloot (TNO) submitted in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE in ARTIFICIAL INTELLIGENCE

Astrid van Aggelen

Abstract.

This study explores different semantic techniques for the task of linking speech-recognized Dutch news videos to related documents. The approach taken is similarity-based retrieval. Due to frequent mistakes of the automatic speech recognition (ASR), simple term matching leads to precision and recall issues. Adopting the bag-of-words vector space model, we formulate a set of semantically motivated term reduction and weighting methods to reduce the effect of ASR errors. These techniques exploit Dutch Wikipedia's entities for term selection and weighting. Second, adopting a topic-based language modeling approach, we propose a document representation method as well as a query reformulation method based on Latent Dirichlet Allocation (LDA). We experimentally validate our evaluation approach and evaluate the proposed methods. The results suggest the potential of LDA to find topically broadly related documents for ASR materials, and demonstrate the success of Wikipedia as a background vocabulary for query rewriting in speech-recognized documents.

Acknowledgements.

I would like to thank my supervisors Laura Hollink and Corné Versloot for their assistance and guidance with the development of this thesis. Also, I would like to thank my second supervisor at TNO, Rianne Kaptein, for her advice, Martijn Spitters for his programming assistance, and all other colleagues at TNO, who were always willing to help me.

Astrid van Aggelen
Amsterdam, Netherlands
November 18, 2013

Table of contents

List of tables

List of figures

List of Algorithms

Chapter 1

Introduction

Since decades, providers of written and audiovisual media have archived their materials, and tools and algorithms are being developed to exploit and structure ever-growing data sources. Nowadays, not only do media services aim to make data available; a growing trend exists towards integrating different sources of information such as text, image and video. One party that is actively researching and developing multimedia applications is the department of Media and Networking Services of TNO, where this thesis project was carried out as an internship.

In multimedia applications, automatic speech recognition (ASR) techniques play a large role. By instantaneously transferring speech onto text form, utterances can be indexed and searched, monitored and interlinked, exactly like written sources, with minimal delay. These applications have considerate impact on people's daily lives. With regard to media providers, manual summarization, transcription and annotation, as is performed in different contexts including political debates, lawsuits and journalistic interviews, can be facilitated. For the media consumers, information can be presented in a tailored, enriched, summarized, or on-demand fashion, and all spoken media such as radio and television material can be searched. However, automatic speech recognition is still under development, and it is not unusual that only half of the words actually spoken are correctly perceived. Therefore, information retrieval systems based on the speech recognizer output will likely be less successful than based on regular text.

The research at hand is performed in the context of interlinking an archive of speech-recognized news broadcasts and a newspaper archive. This functionality is to be embedded in a web-based audiovisual search system that was co-developed by TNO in a project entitled "Multimedia extraction services (MES)". The system records streams of TV broadcast and processes the audio signal by a general-purpose speech recognizer. By indexing the recognized speech or, if available, the closed captions, users can search TV materials by their content. It is currently explored how to automatically embed links to written media sources, such that users are only one mouse-click away from finding more information about a topic or seeing how it was covered in the newspapers.

For reasons of generality, it was chosen to impose on this project the restriction that no data other than the speech-recognized text itself could be used to find related written sources. In the domain at hand, it is common for closed captioning to be available through Teletext services. However, our aim is to develop a linking method that works even in absence of closed captioning or any other metadata. Therefore, the techniques experimented with do not rely on any external information sources. However, this research does use the closed captioning corresponding to the ASR text as reference materials in order to inspect whether a technique was successful for ASR text in particular or for regular text as well.

This study treats the task of finding interesting documents as a similarity search problem. That is, we assume that users are interested in reading documents that have at least

some topical overlap with the source item, and the higher the topical overlap with the source item, the more relevant a document is perceived to be. In information retrieval, a common method to assess topical overlap is by documents' term similarities. However, since the wrongly recognized terms are generally not indicative -often even counterindicative- of the document's semantics, we cannot assume documents similar in terms of lexical commonalities to also be similar in terms of content, as was confirmed in preliminary experiments with the data at hand.

The aim of this study is to find in what form the speech-recognized text can be represented in order to retrieve its topically most similar documents. We hypothesize that we need to take into account the coherence between the terms in the documents in order to detect such outliers. Therefore, we aim to transform the representation of the input document and/or the target collection into a set of meaningful, i.e., semantically coherent, term features. We do so by including background knowledge in two forms: an LDA- trained topic model of a background news corpus; and the link structure of the general-purpose online encyclopedia Wikipedia. The two techniques can be paraphrased as follows:

- exploiting the link structure of Wikipedia to detect those terms in the input document that are semantically coherent, and to assign to them more weight in search;
- applying Latent Dirichlet Allocation for the same purpose, i.e., to counterbalance the effect of the speech recognizer by weighting terms according to their semantic coherence in the document;
- applying Latent Dirichlet Allocation to convert the document to a semantically higher level of abstraction.

Our research questions assess the use of LDA and Wikipedia in finding a document representation that is an effective basis for similarity measurement. We formulate our main research questions as follows:

1. *Can Wikipedia entities be used to overcome the difficulties of retrieval with ASR?*
2. *Can LDA be used to overcome the difficulties of retrieval with ASR?*

Finally, we outline a survey to assess the validity of our evaluation approach and inter-rater agreement in the assessment. The survey addresses the following matters:

1. *Does our evaluation method capture how general users perceive similarity and link-worthiness?*
2. *Do users' assessment based on our evaluation protocol correspond to rater's assessment?*

The structure of this document is as follows. In chapter 2, we position our work in the context of information retrieval, focusing especially on document representation and the previous use in this domain of semantic techniques. Then we present the proposed Wikipedia- and LDA-based methods in chapter 3. Next in chapter 4 we describe the experimental setup we use to test and validate our methods in the actual study as well as the survey, followed by chapter 5 which describes the results and chapter 6 which discusses them. Finally, in chapter 7 we sum up the document and in chapter 8 we identify future work.

Chapter 2

Background

The study at hand fits into several areas of research such as information retrieval, machine learning, and web science. First, by the nature of the data used, it relates to the persistent problem of extracting semantics from text that lacks syntactic segmentation and that contains errors. Second, in its aim it connects to content linking and similarity search. The methods that are explored touch upon several well-established fields in information retrieval, such as probabilistic models as well as query reformulation and semantic enrichment, and in machine learning, such as Latent Dirichlet Allocation. Finally, this study connects to web science by using linked Wikipedia instances as a web of data.

In this chapter, we discuss the related work that has been done on the topics relating to our research questions, as defined in Chapter 1. The structure of this chapter is from general to specific. We start with outlining the general frameworks this study adheres to: content linking methods in general, the information retrieval approach of finding related documents, the vector space model of representing documents, and the laboratory model of evaluating search outcomes. Then, we outline the basic ingredients of search, from which we move on to more advanced document representation methods, including term reweighting, term selection and document transformation. This section describes the techniques that are used in these fields by others and by us, including the way in which Wikipedia and LDA have been employed in this domain and will be employed in this study.

2.1 Approaches

Studies that aimed to find links *for speech-recognized documents* are scarcely present in the literature. The most well-known effort in this domain is likely the TREC Spoken Document Retrieval (SDR) competitions, which have been held since the late 1990s to bridge the gap between ASR and regular text in search tasks, albeit while taking the opposite approach of searching in speech-recognized collections based on regular text queries. However, rather than documents, the queries that have been used in SDR are sentences or phrases [26], hindering direct comparison with the current study. Because of this research gap, we broaden the discussion to the interlinking task in general.

The task of interlinking document sets has been approached in several ways, that in general split up into content- or similarity-based techniques on the one hand and user- or recommendation-based techniques on the other hand. The first group of techniques addresses linkworthiness as a static property between documents that can be evaluated by the textual content only, in isolation of the envisaged end users; as such, this approach fits into the *laboratory model* of information retrieval [32]). The second group of techniques, on the other hand, considers the human as the only valid model. It uses explicit or implicit forms of user-item preference data, e.g., obtained by large-scale click-through data, to find appropriate linking patterns or make individual predictions (e.g., [27] [30] [45]). Since

no user preference data are available, the approach in the current study falls within the content-based group. The remainder of this chapter will focus on that type of techniques.

First, it should be pointed out that manual techniques are still being used in the task of finding articles based on textual content. Editors search for linking materials altogether or label the documents based on a thematic vocabulary, reducing the linking task to a trivial tag-matching task. Studies like the current one aim exactly to automate such labour-intensive tasks.

Machine-learning techniques such as clustering and classification have proven to be of help in finding related materials for a given input document. For an overview, we refer to the survey edited by Berry [8]. Clustering is the task of grouping together similar documents without preimposing any structure other than the desired number of clusters [44]. In classification, the partition of the corpus is preset, i.e., documents are assigned to one of a predefined set of categories. These categories can be defined on the basis of topics (e.g., [14]), but also of relevance judgment, i.e., "relevant" or "non relevant" to a given query (e.g., [20]). The strength of machine learning techniques is that they can take as input a wide variety of data (features) and are able to select among these the most powerful ones. However, both clustering and classifying have certain caveats. Classification requires generating a representative training set, which is time-consuming. Also, it is not always possible to define the classes beforehand. For clustering, difficulties include choosing an appropriate number of clusters, and evaluating the outcome, not only for practical reasons such as corpus size, but also since there are generally several valid partitioning possibilities.

A common method for finding similar materials for an input item, which is used in the current research, is the information retrieval approach of *similarity search*. As a directional method which does not structure the document collection, nor involves training any model, similarity search does not display the difficulties that clustering and classification suffer from. In similarity search, a ranking algorithm is used to sort documents by their term-based similarity to the input item. Likewise, the top-x of this ranked list is intended to represent the set of x documents most similar to the input document. In the most common search applications, the input takes the form of a sparse keyword-based information request from a user, i.e., a *query*; in similarity search, it consists of a complete text document.

In similarity search, the *vector space model* [46] forms the basis of document representation and distance measurement. The input document (query) as well all the documents in the search collection are represented as a vector form, such that their distance can be quantified in the same way as multidimensional vector distances in space. In an information retrieval system, the distance between a query and a document decides upon the position of the document in the returned result list, the one with the lowest distance score featured at the first position. In the vector representation, every element corresponds to a term in the collection, and its value (also called "weight") expresses its relative importance in the document (see section 2.2). The order in which terms occur in a document is often neglected, an approach called *bag-of-words*. Examples of studies that used vector space search for interlinking archives include Bron et al., [12], Ceylan et al. [16] and Ikeda et al. [29].

As a follow-up of the vector space model, *probabilistic models* are the current state-of-the art in information retrieval. These are based on *language models*, which express the likelihood of finding specific words in a document. With regard to relevance, the idea behind probabilistic modeling is that *the more likely it is that the query could have been generated from a document's language model, the more relevant to the query this document is*. Term probabilities can be modelled as either independent or dependent of their preceding term(s). The approach where a term w has a fixed probability $P(w|d)$ of appearing in a document d, regardless of its position, is equivalent to the *bag-of-words* approach in vector representation and is referred to as the *unigram language model*. The likelihood of finding a query Q under the unigram language model D of a given document d is the combined independent probability for each of the query terms: $P(Q|D) = \prod_{q \in Q} P(q|D)$ [53].

Consequently, the search task is transformed into a document modelling task of estimating term likelihoods. Such document models are generally based on statistical properties of the search collection only (e.g., [48, 55], but can also include background knowledge to make the probabilities more realistic (e.g., [40]).

Although common in regular query-based search, probabilistic models have rarely been used in document-to-document matching tasks [10]. Bogers and van den Bosch [10] in 2007 were the first to compare the basic tf-idf method with a probabilistic model for exactly this purpose. Similarly to the current study, they used the whole documents as their basic input materials and experimented with reduced versions. Their findings indicated, first, that the probabilistic model outperformed the tf-idf method. The current study, which, unprecedentedly, uses speech-recognized input documents for similarity search, does use tf-idf as a "baseline" for future work. Second, Bogers and van den Bosch found that when using tf-idf, results benefitted from reducing the input document to a lower number of words. This variation is adopted in this study. The current study uses another type of probabilistic model for document matching as well, which is topic-based rather than term-based. We defer the discussion of this model, called Latent Dirichlet Allocation, to the equally named section in 2.3, and elaborate on the vector space model first.

2.2 Basics of search

The most common system of term weighting in the vector space model is *tf-idf*. Tf- idf, short for *term frequency - inverse document frequency*, is based on the assumptions that the more frequently occurring terms in a document are indicative of the document's semantics, and that this holds even more for terms with a low frequency in the collection, which are more specific in nature, than for common words [1]. Therefore, a term's tf-idf weight is obtained by trading off its frequency (*tf*) and its frequency in the collection (*df*) such that $(tf-idf)_t = tf_t \times idf_t$. As the term frequency is a simple count of occurrences of the term in the given document, most approaches add a factor to their weighting scheme to normalize term frequency by the document length. Also, terms are generally reduced to their stems before they go into the vectorization process, such that, for example, *conform*, *conformed*, and *conforming* are all mapped onto the same vector dimension, which represents the stem *conform*. Due to bag-of-words, the position of a term in the vector is random, and in a set of vectors any two columns can be switched. Finally, the *idf*, which expresses the *rareness* of a term, can be modelled either as the number of times the term occurs in the collection, which is usually called "collection frequency", or as the number of documents it appears in, i.e., the "document frequency". The latter is the most common approach.

The last step of search, after converting all documents to *tf-idf* vectors, is to calculate their similarity. Comparison takes into account only those vector dimensions that correspond to a term in the query. The most common distance calculation in the vector space model is *cosine similarity* [7]. As the name indicates, it is the complement of the cosine *distance*, which is $1 - cosinesimilarity$. The similarity between two vectors is taken to be the cosine of the angle between them. Mathematically, the cosine of the angle between two vectors is their inner product, normalized by their magnitudes. Let A and B be two vectors, then their cosine similarity $sim(A,B)$ is $\frac{A \times B}{||A|| \times ||B||}$, where the length of a vector is defined as the square root of the sum of its n squared elements, i.e., $||X|| = \sqrt{\sum_{i=1}^{n} X_i^2}$ [7]. After calculating the cosine similarity between the input document and every document in the collection, a search system returns a list of documents in decreasing order of similarity.

However, in many situations, the unaltered input document does not provide a good basis for similarity search, both on the term- and on the vector- level. A classic example for term-related problems is the sparse query "python", fired by an internet surfer who is interested in learning about the Python programming language: the system will inevitably

return documents about the snake carrying the identical name, too. Whereas sparsity is an issue in short queries, long queries frequently suffer from redundancy, i.e., the presence of extraneous terms that cause topic drift [34] [12]. For the study at hand, non-relevant terms are generated not only by the inherently lengthy and unfiltered nature of text documents, but by the ASR errors even more. A process called *query reformulation* aims to overcome this problem and make a query representative of the underlying information need. With respect to the vector representation, tf-idf provides a fixed weighting scheme that ignores inter-term relationships, such as their relative importance to the information need. For example, in query "British airlines reservations", the last term is clearly the focus [4]. *Query (re)weighting* is the technique of adjusting the weights of the query terms in order to *boost* key terms. The next section describes query reformulation in more detail.

2.3 Advanced search: Term selection, reweighting and document transformation

Query reformulation is a highly heuristic process [5]) for which general guidelines are nonexistent. Consequently, it has been addressed in variety of ways. We report some findings here.

Earlier work has demonstrated that query reduction can improve search results in long or document-like queries. Starting from textual queries of up to thirty terms, Kumaran and Carvalho [34] trained a classifier to indicate the most informative part of the query, which was then used as a substitute for the original query. Their classifier used a variety of features that had been reported in other studies to correlate with query success, such as its constituent number of terms and the so-called *mutual information* between every two query terms, a statistical measure that is based, among others, on the number of times the terms co-occur within a certain window in the corpus [34]. With this approach the authors report a significant increase in search performance. The queries in the current study are longer than theirs, but are also preprocessed using relationships between query terms. However, in the term selection process we do not use any information based on the search corpus. Ceylan et al. [16] used a grammar-based heuristic to extract eventful sentences from textual source articles, and maintained the predicate as query materials. They defined key phrases to contain an action verb in past tense and a named entity. This approach is difficult to apply to our materials, which are not separated into sentences. Finally, Bron et al. [12] attempted to link items from a Dutch newspaper archive to items in a video catalog that were represented by sparse metadata such as a description, summary, and keywords. To reduce their textually rich queries, they first ranked the terms by their tf-idf value and selected to the top-k percentage; from that selection, they maintained only the named entities as query materials. They evaluated their technique twice: once on a task of finding same-event documents, and once for retrieving similar-event documents. Selecting the top-60 % of terms led to a significant improvement over the baseline in same-event linking, but not in similar-event linking. This study is interesting for the one at hand, as the source collection is of the same nature as ours. Our goal is to retrieve same-event documents as well as similar-event documents, with the former ranked higher than the latter.

The studies reported thus far in this section relied on within-corpus information only. However, external information sources, such as vocabularies, taxonomies and encyclopedias, are extensively used for modelling concept dependencies in query representation tasks, both for selection and weighting. A widely used tool for this purpose is Wordnet (e.g., [13, 43, 35]), a large and freely available lexical database of English. The likely closest Dutch counterpart for Wordnet is Cornetto [52]. However, Cornetto does not contain any named entities, nor is regularly updated to account for new phenomena. For these reasons, in the currents study it was decided to use the Dutch version of general-purpose online

encyclopedia Wikipedia. The next section goes into detail about the use of Wikipedia in query reformulation.

2.3.1 The use of Wikipedia in query reformulation

Wikipedia has been attributed some qualities that are believed to make it appropriate for knowledge modelling [39]. As it is widely used and freely accessible and extendable, Wikipedia content is generally of sufficient quality and is constantly updated to include new phenomena of interest. Second, its content scope is not restricted to any domain or category. Wikipedia's large variety of topical categories make it a genuinely general-purpose information source. Third, Wikipedia has a uniform, prescribed structure, which allows for targeted information extraction. For example, each page covers exactly one topic; and the first sentence on the page is the concept's definition. Also, every phrase that denotes a significant topic is embedded in a link to that topic, such that a topic page's incoming links provides a valuable set of alternative formulations of the topic.

Using Wikipedia as a reference requires *entity matching* or *entity linking*, i.e., connecting terms in the source material with their correct counterparts in Wikipedia, a task that we will also refer to as *wikification*. The major challenge of entity linking is to resolve *ambiguities*, which are mainly caused by polysemy (e.g., "python"). The counterpart of polysemy, which is synonymy or the diversity of names used to refer to one single entity (e.g., "president of the United States", "president Obama", "Barack") is dealt with by the structure of Wikipedia itself. The encyclopedia accounts for this problem by the text-embedded labels (i.e., anchor texts) that are used to refer to a certain topic, and which presumably reflect the denotative diversity for the topic. The problem of polysemy, however, cannot be solved without taking into account the semantic *relatedness* of a term with its surrounding concepts.

For disambiguating in wikification, we cite the approaches of Cucerzan [19] and Milne and Witten [41]. Cucerzan addressed disambiguation by comparing a vectorial representation of the original document with a Wikipedia-composed vector for each of the possible senses of the ambiguous terms in the document. The Wikipedia vector for each target concept consisted of several information types, such as additional (parenthesized) expressions in its title, and the labels that occur in the article's first paragraph. Milne and Witten's disambiguation approach, which is used in the current study, emulated Cucerzan's by its learning ability and computational lightness. They did not use any descriptive text from Wikipedia other than the (sparse) anchor texts, and they used machine learning rather than a deterministic approach. Specifically, they trained a classifier to detect the most likely sense for an ambiguous concept. Thereby the Wikipedia corpus was used as a training and test set, where every anchor term provided a positive training example through its actual link target, and -usually- several negative examples through its alternative destinations. The features included, among others, the Wikipedia-based prior probability of the sense for a given phrase, and the semantic relatedness between each of the senses and the context of unambiguous terms. The approach by Milne and Witten has proven its potential. Using a training set of 500 content articles with a minimum of 50 links each and a test set of 100 pages, the authors reported a disambiguation precision score of over 98 %. The current study uses Milne and Witten's disambiguation method, and applies their relatedness measure for term weighting as well. Therefore, this measure is described in more detail in the following.

Milne et al [41] proposed that the mere structure of Wikipedia suffices to assess the degree of relatedness between concepts. They based their theory of relatedness on the existing theory of the *Google distance*, which was developed for similar purposes but was based on returned web pages for keyword search in Google. The intuition behind the Google distance was that for any two terms, the number of web pages that mentioned them both, relative to the number of pages that mentioned only one of them, expressed

the strength of the semantic relationship between those terms [17]. In line with this idea, Milne et al. hypothesized that the semantic distance of two pages (hence concepts) in Wikipedia could be inferred from the commonalities and differences in their incoming and / or outgoing links, with overlap in links as counterevidence, and distinction as evidence for conceptual distance [41]. Hence, they defined the semantic distance (*sd*) of two articles *a* and *b* as

$$sd(a,b) = \frac{\log \max |A|, |B| - \log |A \cap B|}{\log |W| - \log(\min(|A|, |B|)} \tag{2.1}$$

, where W is entire Wikipedia, and A and B are the sets of articles that link to or are linked to by a and b respectively. The measure can be applied based on incoming or outgoing links, but Milne et al. found that the inlink-based formulation was most successful [54]. Since this value ranges between 0 and 1, they defined the semantic *relatedness* of a and b to be $1 - sd(a,b)$. Evaluating the obtained relatedness measures on a standard set of manually generated *word similarity scores* against preceding text-based and computationally heavier techniques of Strube and Ponzetto [49] and Gabrilovich and Markovitch [25], it within reasonable distance underperformed against these [54]. However, the current study puts this semantic similarity score to practice in a wider context than just lexical proximity, in assessing conceptual relationships *of any kind*.

Wikipedia has been used in query reformulation in different ways, of which we give two examples. Bendersky et al. [4] used Wikipedia in query term weighting. They used a weighted linear combination of heterogeneous features to assign weights to query terms, including the number of times the term occurred within the query, and the frequencies by which it was used as a Wikipedia title and as part of a Wikipedia title. As their results were above baseline, their study suggests that Wikipedia statistics can be useful in weighting schemes, and we adopt this method. Building onto their method, we use Wikipedia data not only to interpret individual terms, but also to assess inter-term relationships. Gabrilovich and Markovitch [25], in the study that was cited before, did not alter the terms or their weights but the vectorization method itself, by directly integrating relatedness information in the vector. They represented each document by a vector of length *the number of topics in Wikipedia*, where every entry was modelled to expressed the degree of presence of that topic in the query. Thereby, the weight for each dimension was the summed combination of each term's tf-idf weight in the query and its association with the corresponding topic. The association between a word and a topic, in turn, was taken to be this term's weight in the tf-idf vector of the topic page text. The current study also uses terms' topic adherence in weighting, but does not alter the document representation itself.

Put in a broader context, the use of Wikipedia fits into research practices of extracting semantics from the web. First, through the use of link-based statistics. For a given web document, the overlap in its incoming and outgoing links with another web document has been shown to be indicative of their similarity (e.g., see [38], [37]). The current study uses this principle, which originates from the Google distance, to assess semantic relatedness between Wikipedia pages. Second, on the term rather than the document level, through its application as a *web of data*. The way in which entities in Wikipedia link towards their unique identifier is reminiscent of the Semantic Web [6], a framework that promotes the use of machine-readable semantic markup, envisaging to integrate each web document's content into a unified web-based ontology. For content to be truly semantic, not only the entities should be rooted in a controlled vocabulary, but also their semantic relationships in the text. Since Wikipedia cannot directly account for this, it has been tried to extract formal relationships from its content in a project called DBPedia [2]. DBPedia has been used in interlinking materials in the news domain [33], but lacks an operational Dutch counterpart.

The approaches outlined in this section form a step beyond a purely term-based approach towards an approach that takes into account concepts. *Topic-based modelling*, of which Latent Dirichlet Allocation (LDA) is an example, denotes an approach that in turn

surpasses the concept-based level by taking into account semantically motivated groups of concepts, i.e., *topics*. We first provide a general introduction to LDA before outlining its possibilities in query reformulation and document transformation.

2.3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a text modeling approach that aims to represent documents in a lower dimensionality while maintaining their semantic characteristics. The new dimensions are topically motivated, where topics are modelled as sets of prevalent words. In practice, the parameters of the model are first estimated on a training corpus. The training phase explicates to the user the likelihood for each term to occur under each topic. Then, the model can be used to represent new documents in terms of their topical distribution.

The key assumption behind LDA is that a document is a mixture over topics, which in turn are sets of words. Hence, as a *generative probabilistic model*, LDA assumes that the observed textual data have been generated by a two-step sampling process over topics and words [53]. Note the difference with the unigram language model, where each word in a document is hypothesized to be drawn independently from a single distribution. The distribution that supports the LDA model is the Dirichlet, which is the multivariate generalization of the beta distribution [50]. By fitting the parameters of the Dirichlet to maximize the likelihood of the observed data, a model is obtained that explains the observations as accurately as possible.

Formally, the hypothesized generation process is as follows:

1. For each of the K topics z, pick a multinomial distribution ϕ_z from a Dirichlet distribution with β as its parameter;
2. For each document d, pick a multinomial distribution θ_d from a Dirichlet distribution with parameter α;
3. For each token position w in d, pick a topic $z \in \{1,..,K\}$ from θ_d;
4. Pick a word form for w from ϕ_z.

Based on an input corpus and a defined number of topics K, the LDA algorithm provided by Blei et al., which is used in the current study, estimates the parameters of ϕ_z, which describe the assignment of the words to topics, and α, one of the two *hyperparameters* of the model, which relates to the distribution of topics over word positions. The other hyperparameter is β, which describes the distribution of words over topics. The lower the value of α and β, the sparser θ and ϕ, which are drawn from the Dirichlet that corresponds to these hyperparameters [28]. In practice, a low α will result in a lower variability of topics within a document, and a low β leads to lower term variability within a topic. As a rule of thumb it has been proposed to take $\alpha = \frac{50}{K}$ and $\beta = 0.01$ [28]. However, the implementation by Blei et al. does not use any fixed user-defined hyperparameters: although α is user-initialized, both parameters are estimated on the training corpus itself. For the details of parameter estimation we refer to Blei et al. [9].

With the resulting LDA model, as described by ϕ and α, the likelihood for each term to occur under each topic can be inspected. This gives an impression of the quality of the model and of the suitability of the chosen number of topics. Then, *inference* can be performed on a collection of unseen documents. In this step, based on the learned word-topic associations, the algorithm outputs the likelihood for each topic to occur in each of the documents. The next section describes the possibilities of LDA in information retrieval.

2.3.3 The use of LDA in query reformulation and document transformation

The use of LDA in information retrieval fits into the aim to refine probabilistic retrieval by surpassing the lexical level to a more semantically motivated level. The probability of

a term to occur in a document, $P(w|D)$, trivially follows from the estimated $p(w|z)$ and the inferred $p(z|d)$ by the LDA phases of training and inference, respectively. Azzopardi et al. [3] used a linear combination of LDA-based $P(w|D)$ and some smoothing a-priori likelihood for each term to obtain its likelihood estimate. When applied to a standard set tests, they found improved performance over regular language models.

Another way to use LDA in document modeling in search, and which is adopted in the present study, has been referred to as *topic querying* [28]. It models the query and the collection documents as vectors of topic distributions $p(z|d)$ for all topics z, and uses standard vector distances to rank the documents by their similarity to the query. Wei and Croft [53] adopted this method and reported decreased retrieval performance in their preliminary experiments. However, Heinrich [28] points out that although topic querying might yield less precise results, the abilities of the topic model to map closely in vector space documents which are topically related but lexically distinct can be expected to improved recall. Moreover, Wei and Croft applied this technique to regular documents, therefore the benefit of their method in speech-recognized materials cannot be excluded. We do not know of any studies that used topic querying based on speech-recognized documents.

Finally, as another application, LDA has been used to remove semantic outliers in speech-recognized queries. The main idea here is that an outlier is a term whose topic distribution is different from that of the other terms in the query. Senay and Linares [47] computed a small topic space (of approximately 5 topics) for each term *separately* based on a targeted part of text in which the term was extensively covered. Then, they defined the consistency of a term in the query to be the sum of its consistency in each of the topic classes. The latter, in turn, was modelled as a linear combination of the number of query terms with an above-threshold probability in the given topic cluster, and the probability of the term to occur in the cluster. The current study takes a similar approach of using the document-dependent term prior $p_d(t) = \sum_z p(t|z)p(z|d)$ to detect outlying query terms. The difference with the approach of Senay and Linares is that our technique is simpler and computationally lighter, as the model is trained on all concepts simultaneously, hence on the same corpus, and the probability calculation is based on each term's individual probability only.

2.4 Evaluation

A retrieval system is commonly evaluated by its ability to find and appropriately rank documents that are relevant to the queries it is presented with. Relevance is an ill-defined notion that in a real-world setting depends on a subjective information need, as well as on the user's background knowledge and on the effects of redundancy and saturation by the documents as they are presented to the user [56]. Therefore, in information retrieval (IR) studies it is not uncommon to make the following two assumptions [56]. First, the relevance of a document is independent of other documents, including those already retrieved; this is known as the *independent relevance assumption*. Second, as was already mentioned, the relevance of a document is assumed to be user- and situation-independent; it is a semantically motivated relationship between two documents that can be objectively assessed. This is known as the *topical relevance assumption*. The degree of relevance of a returned document for a given query depends on its similarity in terms of the number of topical aspects in the query that are covered by the document. As a result of these assumptions, evaluating an IR technique requires only a set of test documents, a set of test queries, and a protocol of relevance assessments [32]. This principe has been called the *laboratory model* [32] and forms the basis of all theoretical IR evaluation tasks, including the one at hand.

Chapter 3

Method

In this chapter, we propose three different techniques for query rewriting or document representation using Wikipedia and LDA:

- Our Wikipedia-based method of query rewriting, i.e., term selection and weighting
- Our LDA-based method of query rewriting, i.e., term selection and weighting
- Our LDA-based method of vector representation

For every approach we explain the steps involved and we propose some variations. As a notational remark, note that this report uses the terms *input document* and *query* as synonyms. After all, the approach taken, i.e., tf-idf cosine similarity search, is the default functionality of search engines, where the input consists of user-defined information requests or *queries*.

Due to the use of speech-recognized text, the traditional term-based method of representing a document is likely not entirely suitable. Conventionally, each term in the collection is taken as a vector component, and its value in a document vector is taken as a statistic of its frequency in the document and its rareness in the collection. However, because of the many wrongly recognized terms, which are not indicative –often even counterindicative– of the document's semantics, we cannot assume documents similar in terms of lexical commonalities to also be similar in terms of content. Preliminary experiments indeed confirmed that matching on the raw speech recognizer output was not satisfactory. When evaluating the top-10 ranked documents on predefined relevance criteria for a set of test queries, it was found that on average about half of the results were not relevant.

Investigating the data set used in this study, it was observed that especially named entities were frequently missed (e.g., Marx -> mais) or erroneously perceived (e.g., mais -> Marx) by the ASR component, causing mismatches between the corpus documents and the input document in the corresponding term weights. As named entities are generally low-frequent terms, these mismatches heavily influence the similarity scores of the corresponding documents. Also,it was observed that quite frequently, the (wrongly) recognized terms were not the type of semantically rich terms that are expected to occur in the news domain, as demonstrated, for example, by the excerpt *dan kunnen we maar ja je in een uh nee ik je bent er erg veel ouders vaarwel*.

Consequently, an approach was taken that aimed to reduce the effect of wrongly recognized terms on document comparison. Thereby, two main assumptions were made. First, we assumed that a correctly recognized term could be recognized in two ways:

- by its *frequency* of occurring in the document. The more often a term appears, the more likely it is that it was correctly transcribed;

- by its *semantic coherence* with the other terms in the document. The more a term is related to the remainder of the input document, the more likely it is that it was correctly transcribed.

Second, we hypothesized that the key meaning of a news document can be captured by those terms with a certain semantic load or meaningfulness, such as the words that occur in encyclopedias or other news corpora. As a result of these assumptions, we developed the following general approach:

- reduce the query to a set of semantically rich terms by means of an external corpus;
- weight each resulting term according to its frequency in the input document and/or its semantic coherence with the other document terms.

Latent Dirichlet Allocation (LDA) and Wikipedia were used in this study exactly for the purpose of detecting meaningful terms for term selection and assessing semantic coherence for term weighting.

Wikipedia and LDA each provide a vocabulary and method for maintaining meaningful terms or units. Wikipedia was used as a reduction tool by maintaining from the original query only those terms or phrases that are connected to a Wikipedia topic. That is, the full set of terms that carry a link to a Wikipedia page, i.e. the *anchor vocabulary*, was regarded as a general source of semantically rich terms. LDA was applied to generate search input materials in two different ways, one of which operated on the vectorization level and the other on the term level. First, by generating a low-dimensional topic vector space and using this as a basis of comparison rather than the high-dimensional tf-idf standard. In this method, the document was reduced to meaningful *units* rather than terms. Second, by maintaining from the original query only those terms that occurred in the vocabulary that the algorithm was trained on. By training on a corpus similar to the link document pool, this vocabulary was intended to reflect the most important terms from the news domain.

As outlined in chapter 2, the link statistics in Wikipedia allow to quantify the *semantic coherence* of a concept within another set of concepts. Milne et al. [54] provided the link-based measure to capture relationships between pairs of Wikipedia pages that is used in this study. Given the prescribed one-to-one mapping of concepts and pages in Wikipedia, this measure was intended to represent the relationship between any two *concepts* present in Wikipedia alike. It trades off the number of shared and unique incoming links (i.e., user-generated references) of a set of Wikipedia pages to express their conceptual proximity. By aggregating a term's similarity score with each other term in the query, a coherence score is generated for a term in the query.

Latent Dirichlet Allocation can be applied as a tool to assess semantic coherence as well. Semantic coherence of a term in a document is then modelled as the *likelihood* for this term to have been generated from the topical mixture identified in the *query* document in which it occurs. This likelihood is derived from two types of probabilities that are outputted by the LDA algorithm: the likelihood of the term to occur under each of the topics, and the likelihood of each of these topics to be present in the document. Likewise, a term is assigned a high probability when it is a prevalent term in some topic (or several topics) that is (are) recognized with high probability in the document.

A general overview of the processes involved in query reformulation using Wikipedia is displayed in Figure 3.1. The grey-colored ellipses denote processes for which several alternatives are proposed, as treated in the upcoming sections, and the thickly bordered rectangles indicate that the obtained information source is part of the end product needed, i.e., of the reformulated query or the new document representation. The workflow consists of the following components:

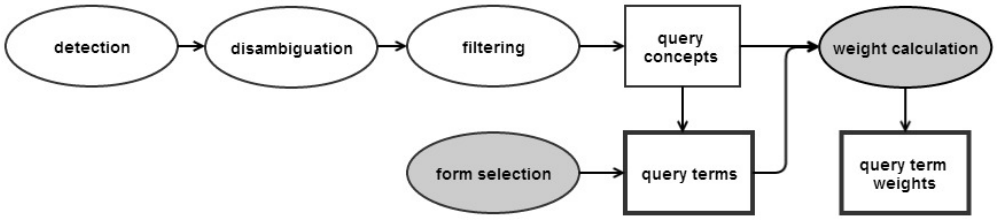

Figure 3.1: General processes in the Wikipedia-based methods

- detection: finding all possible matches between the query and the concepts from the Wikipedia vocabulary
- disambiguation: deciding on the most likely match between query segments and concepts
- filtering: removing matches based on certain criteria
- form selection: choosing the lexical form of the concepts as the terms or phrases of the new query
- weight calculation: deciding on the weight to assign to each of the query terms or phrases.

The following section describes how we implemented each of these components and combined them into our different Wikipedia-based algorithms.

For LDA, a general overview of the different tasks involved is given in Figure 3.2. As these processes apply to both the query reformulation and the vector representation application of LDA, the outcome of the document representation process is not further specified. The following main processes are involved:

- training: with the use of a background news corpus, generating a topic model that describes the likelihood distributions for each term to occur under each topic.
- inference: using the topic model to discover the topic distribution of previously unseen documents from the search task (the search corpus and/or the query documents). The result of this step is a document model of topic distributions for each document separately, which, for the sake of applicability to the two different methods, we conceptually merge with the topic models in a unified ”topic and document model”.
- document representation: with the help of the collected model(s), create a new topic-based document representation.

Section 3.2 describes how these components were implemented and combined into our two LDA-based algorithms.

3.1 Our Wikipedia query reformulation method and variations

Before the steps formulated in Figure 3.1 are explained, recall that Wikipedia is structured such that *topics* correspond to *pages* in a one-to-one fashion, where -trivially- the title denotes the topic name. Furthermore, the texts that are used to embed a link to a certain topic, i.e., the anchor texts or so-called *labels* is what this study takes as the Wikipedia *vocabulary*. Generally, for a given topic, its labels include the page title (i.e., the topic name) itself,

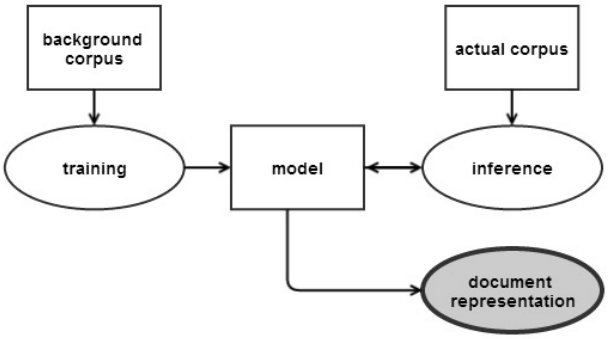

Figure 3.2: General processes in the LDA-based methods

grammatical and orthographical variations, and alternative formulations. For example, for topic *boete [fine]*, its label set includes the terms *geldboete [fine], boetvaardigheid [repentance], bekeuring [ticket], boetedoening [atonement]*. Likewise, a topic's label set can be regarded as a list of synonyms or closely related terms. Note also that labels can consist of more than one term. Therefore, term *sequences* of n ($n \geq 1$) terms, i.e., *phrases* or also *n-grams*, of the original query are mapped onto the Wikipedia vocabulary rather than just terms. Consequently, the notions *term*, *n-gram*, and *phrase* are used interchangeably in the remainder of this report.

The approach of using Wikipedia's link structure as a vocabulary was developed by researchers at the University of Waikato, New Zealand, and is supported by their open-source software package WikiMiner [1] [42] , which provides object-oriented access to topics and labels. Every step formulated in Figure 3.1 is supported by functionalities in the WikiMiner software. The models used for disambiguation and similarity measurement, which form the basis of our approach, were adopted with no modifications from Milne and Witten [41] and Witten and Milne [54] and are built-in in Wikiminer. The only requirement is to download a Wikipedia dump of choice and to follow the included instructions found on the website [2] to train the disambiguation algorithm on the downloaded Wikipedia data.

3.1.1 Detection

The first step in the process matches the input query onto Wikipedia. That is, all possible mappings are generated of query terms or phrases onto the Wikipedia label vocabulary. To match a query to Wikipedia, every possible sequence of n terms ($n \geq 1$) is compared to the total label set, ignoring differences in case, such that the entire set of lexical correspondences is generated.

The resulting set of candidate matches is likely to contain mutually exclusive solutions that need to be chosen between. For example, a query phrase such as "Third World War" can be projected onto several combinations of Wikipedia label sets, such as "World" and "War", "Third World" and "War", and "Third World War". The choices made on this level affect the resulting query in two ways. First, in the choice of terms that will make up the query. In the given example, it is likely that all three terms *Third*, *World*, and *War* are projected onto the vocabulary, regardless of the segmentation; however, if the term "war"

[1] http://sourceforge.net/projects/wikipedia-miner/
[2] http://sourceforge.net/projects/wikipedia-miner/

was lacking in Wikipedia and the choice were between "Third World" and "World War", then, respectively, either the term "third" or 'war" would be eliminated from the final query. Second, in the weight assigned to each term, as the weighting procedure takes into account information related to the term's corresponding topic, such as its semantic coherence with the other query topics. This requires disambiguated phrases. For example, to judge the relatedness between "plane" and "KLM" in a document, it should be decided whether the former refers to a means of transportation or a mathematical concept [54]. For these reasons, a disambiguation step is required to estimate the best match of a phrase and a (set of) topic(s) by taking into account the meaning of the other document terms.

3.1.2 Disambiguation

The context for disambiguation is obtained by the unambiguous terms, i.e., those n-grams which match with a label that refers to one topic only and that cannot relate to any other topic by adding one of its neighboring terms. With the help of these reference materials, the ambiguous terms are mapped onto the most appropriate topics.

The disambiguation stage is performed by a machine-learned classifier that is trained and tested on the downloaded version of the downloaded (Dutch) Wikipedia corpus itself [41]. Given a phrase and a set of candidate concepts, the classifier outputs the concept with highest probability, optimized on the complete set of query phrases. That is, the algorithm disambiguates all phrases in a document *simultaneously*. The key to the technique is that every anchor term in a Wikipedia corpus provides a positive training example by means of its actual link target, and -usually- several negative examples by its alternative destinations.

Although the disambiguation algorithm is a black box, we do know the features the model is trained on to output the likelihood of a sense (i.e., topic) for a given phrase and set of unambiguous context terms. They are as follows [41]:

- sense commonness, i.e., the prior probability of the topic as a link destination for the phrase, as defined by the number of times the phrase is used in Wikipedia as a link to the topic, divided by the total number of times the phrase is used in Wikipedia as an anchor text.
- sense relatedness, i.e., the average semantic coherence between the topic and each of the (unambiguous) context terms, as defined by the inlink-based formula that will be explained in section 3.1.5.
- context term quality: the contribution of each *context term* to the relatedness measure is weighted by the supposed quality or meaningfulness of the term in the context, which is taken to be its prior probability of being used as a link in Wikipedia. This is simply the number of times the context term is used as an anchor text over the total number of times it is mentioned in Wikipedia.

3.1.3 Filtering

The detection and disambiguation steps assign topics to the query in a deterministic fashion: any phrase that matches a (i.e., at least one) label from the Wikipedia vocabulary is designated to be a phrase in the new query. There is no mechanism, other than mere presence in the Wikipedia anchor set, to ensure that the found concept is not marginal, nonsensical, or deviant from the other concepts. The weighting scheme that we propose is intended to account for this by assigning a lower weight to exactly such concepts. However, additionally, we propose to filter out topics based on the relatedness measure obtained in the disambiguation stage. We use a heuristic that trades off a term's frequency and its average semantic relatedness to all other recognized topics in the query document (i.e., after

disambiguation). That is, concepts with a semantic coherence score below a certain threshold are filtered out, with a lower threshold for terms that occur repeatedly than for terms that occur only once. Hereby, we use a more generous definition of frequency: not only as a count of occurrences of the term (phrase) in the query, but aggregated (multiplied) with the count of occurrences of the term (phrase) in the merged *set of unique labels* of all the topics recognized in the query. This process results in the final set of concepts that make up the new query.

3.1.4 Form selection

As the lexical form of new query terms, the default method that we propose is to maintain the original query phrases, i.e.,every label in the query document that is detected (and disambiguated) to belong to a certain topic in Wikipedia. However, since the Wikipedia vocabulary for each of the concepts contains a variety of related terms, it can be easily experimented with alternative lexical bases. We propose four alternatives. First, the following two:

1. Substituting each of the query phrases with the name (i.e., title) of its corresponding topic (page).
2. Enriching each of the query phrases with the most frequently used other labels of its corresponding topic.

Each of these techniques can be hypothesized to improve search results based on the following assumptions, each corresponding to its equally numbered counterpart above:

1. Topic names are a clear-cut expression of what the term refers to.
2. Alternative labels provide synonymic and near-equal formulations that can be matched upon.

Second, although rather as an experimental setting than a separate method, we propose to merge the newly obtained query terms to the original ones, in an uncontrolled and a selective way respectively, as follows:

1. By appending the original query phrases to the new ones rather than substituting the former for the latter, while assigning to the original terms a significantly lower term weight.
2. By adding to the new query phrases only those phrases from the original query that occur more than once in the label set of the detected topics in the query, again while assigning to the original terms a lower weight.

Adding the original query terms with a lower weight ensures that meaningful terms that do not appear in Wikipedia can still be matched upon, while still maintaining the boosting effect of the Wikipedia algorithm on terms that are likely important to the document semantics. Methodologically, these variations are added to provide more insight into the relative quality of the new terms as compared to the original terms.

3.1.5 Weight calculation

When the phrases for the final query are selected, they are assigned a weight that is intended to reflect their semantic reliability. It is aimed to weight the terms such that ASR errors are assigned a low weight, reducing their influence on document matching. Based on the stated assumptions on how to recognize wrongly interpreted terms, we propose two variants on regular tf-idf weighting: one which stresses frequency information, and one which, additionally, includes the proposed coherence measure. The next section explains these frequency- and coherence- enhanced versions of tf-idf and their combination. After, we explain how we incorporated these weighting schemes into different algorithms that are experimentally validated.

Coherence-based weighting

Our coherence-based weighting assigns to each query phrase (i.e., to each label) a regular tf-idf weight *multiplied by the average of the relatedness of its topic to every other topic in the query*. Thus, this type of weighting, which takes place after disambiguation, is based on the *final* set of topics detected in the query. By averaging the relatedness between a found topic and each of the other detected concepts, a *semantic coherence* score is obtained that is intended to express the extent to which the concept *fits in*. For every two topics a and b, their semantic relatedness (sr) is defined as [54]

$$sr(a,b) = \begin{cases} 0 & \text{if } A \cap B = \emptyset \\ 1 - \frac{\log(\max|A|,|B|) - \log|A \cap B|}{\log|W| - \log(\min(|A|,|B|)} & \text{else} \end{cases} \tag{3.1}$$

, where W is entire Wikipedia, and A and B are the sets of articles that link to a and b, respectively. Given the number of articles that refer to only one of the two articles a and b, respectively, the formula takes the highest of these two as the numerator of the fraction, and normalizes this by the highest of the two numbers of articles that refer to a nor b, respectively. The more articles that refer to A(B) refer to B(A) as well, the smaller the fractional part of the formula, hence the larger the semantic relatedness. However, the formula is overruled in case the in-link intersection of a and b is empty; in that case, the distance measure becomes 1, such that relatedness is zero. It should be noted that the WikiMiner software supports a machine-learned model of similarity measurement as well, which takes several other link-based measures as input. For simplicity, we use the incoming links only.

As an example of measuring relatedness between concept pairs, consider the terms *boete [fine]*, *indianen [indians]* and *rechter [judge]*, which all occur in the ASR transcription of *query 2*, which is about the fine imposed on students whose studies exceed the nominal duration. In this case, *indianen* is erroneously perceived, and indeed is experienced as semantically distant from the other two concepts. Taking the concept *boete* as a starting point to assess its semantic proximity to *indianen* and *rechter*, we first need to gather some in-link statistics. The Wikipedia page on the topic *indianen* has 1035 incoming links, *boete* has 144, and *rechter* has 849. The number of shared incoming links of the topic *boete* with *indianen* is 1 (the page *eigendom [property]*), and with *rechter* it is 9. Furthermore, the Dutch Wikipedia dump used in this study contained around 1.3 million pages. Then, the semantic relatedness of *boete* with *indianen* is $1 - \frac{\log 1035 - \log 1}{\log 1300000 - \log 144}$ or 0.24; and that of *boete* with *rechter* is $1 - \frac{\log 849 - \log 9}{\log 1300000 - \log 144}$ or 0.50, which is twice as high. Thus, the semantic relatedness measure displays that *boete* is more closely related to *rechter* than to *indianen*. In the application at hand a term's coherence score is composed of its average relatedness to every other Wikipedia-matched query term; thus, the higher its relatedness to these terms individually, the higher its resulting coherence score.

It is important to realize that an incoming link can be motivated by any sort of semantic relationship. As we do not take into account the anchor text that carries the link, but just the *page* that contains it, an incoming link for a given page simply denotes a mention of the page topic in a different page. However, as Wikipedia prescribes to link to any other topic that helps the reader understand the text on a given topic [24], the links that depart from a page are intended to represent semantically rich concepts relevant to the page topic. As an example, table 3.1 lists the page title of a random set of 20 pages that link to the pages *boete*, *indianen*, and *rechter*, respectively.

Frequency-based weighting

In the frequency-based weighting method that we propose, the term (or phrase) weight is declared to be the product of regular tf-idf (see section 4.2.1) and *the number of detected*

Indianen	Boete	Rechter
Brazilië	Algerije	Islam
Communicatie	Aswoensdag	Nederland
Cuba (land)	Paus Johannes XXIII	Nationale Ombudsman
Chili	822	Pakistan
CaliforniÃ«	Rooms-katholieke Kerk	Recht
Zondvloed	Adder	Kennis
Frans-Guyana	Vlaginstructie	Appingedam
Geschiedenis van de Verenigde Staten	Gevangenpoort (Den Haag)	Politieagent
Groenland	Monopoly	Kelten
Genocide	Eigendom	Paus
Gaia (mythologie)	Salische Wet	Boris Dittrich
Thor Heyerdahl	Tabakswet	Wet
Thomas Jefferson	Zwerfafval	Nederlandse wetgeving
Verenigde Staten	Leerplicht	Civiel recht (Nederland)
Zuid-Amerika	Nazarener	Strafrecht (Nederland)
1990	Huis van bewaring	Louis Couperus
18e eeuw	Magna Carta	Oorzakelijkheid
16e eeuw	Rechtsbijstandverzekering	Levenslange gevangenisstraf
4e eeuw v.Chr.	Opus Dei	Gratie
2005	Misdaad	Édouard Manet

Table 3.1: Examples of pages that link to the Wikipedia pages *Indianen*, *Boete* and *Rechter*, respectively.

topics in the query that have the term (or phrase) as a label. This can be regarded as an intervention in the usual WikiMiner functionality, where after disambiguation the connections between query phrases and topics are fixated. Instead, we cut these links and permutate every topic's labels for each query phrase, allowing the latter to be connected to more than one topic. Likewise, an extra boost is given to phrases that adhere to multiple topics. As an example, consider a query in which the term "tour" occurs three times. Suppose that from the (disambiguated) topics recognized in the query, it happens that two of them have "tour" in their label set; say, a topic with title "Tour de France" and a topic with title "wielrenwedstrijd [cycle race]". In that case, the phrase "tour", if included in the new query just once, is assigned a weight in the query vector that is the product of its term frequency 3, its inverse frequency in the search collection, and the number of query topics it belongs to, i.e., 2.

Combined frequency- and coherence-based weighting

When combining frequency and coherence in weighting, the weight –trivially– becomes the product of the term's tf-idf and its average similarity with each of the other detected topics as explained in 3.1.5. Consequently, if the term adheres to *multiple* topics, as explained in 3.1.5, it is included in the new query as a separate term for each of these topics, weighted by its tf-idf and coherence score. To continue with the example given in 3.1.5, the term "tour" would be added to the final query twice, once for each topic it adheres to, i.e., once with a weight $3 \times SR_{AVG}$ (*Tour de France, [other topic]*)) and once with a weight $3 \times SR_{AVG}$ *(wielrenwedstrijd, [other topic])*.

3.1.6 Algorithms

The two proposed *weighting techniques*, incorporated into two separate basic algorithms, are summarized in this section. The first one uses only frequency information in weighting;

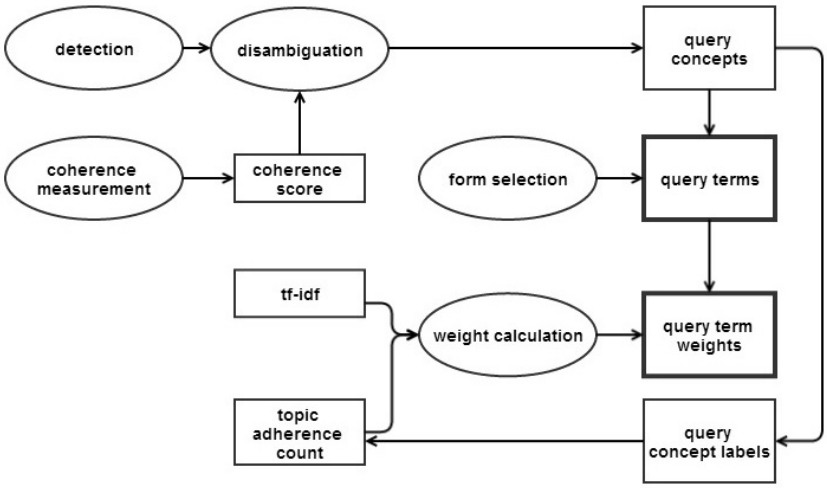

Figure 3.3: Frequency-weighted Wikipedia-aided query reformulation workflow

the other one combines frequency and similarity weighting. The latter is used as a basis for the tried term selection variations.

Algorithm 1: Frequency-weighted Wikipedia-aided query reformulation algorithm

Algorithm 1 uses frequency-based weighting only, as shown in the diagram in Figure 3.3. Detection and disambiguation are as described in sections 3.1.1 and 3.1.2, respectively. Coherence measurement is used in disambiguation only. Filtering, as described in section 3.1.3, does not take place. As form selection, the default method is adopted, i.e., the original query phrases that are found to be a label of a Wikipedia topic are taken as the terms for the new query. The weights for each of the terms are obtained by counting their matches with the set of labels of all recognized query topics, resulting in what is referred to in the figure as the *topic adherence count*, and multiplying this with tf-idf, as explained in section 3.1.5. For easy implementation, the term selection and weighting procedure is expressed algorithmically in pseudocode in algorithm 1, starting from the complete set of matched terms in the original query. Note that *tf* is implemented indirectly, not by its weight but by including the term tf times in the new query.

Algorithm 2: Frequency- and coherence-based weighting algorithm

Algorithm 2 uses frequency- and coherence-based weighting. Its components and dependencies are shown in the diagram of Figure 3.4 and the pseudocode is displayed in Algorithm 2. Detection and disambiguation are as described in sections 3.1.1 and 3.1.2, respectively. In this algorithm, the filtering stage based on the terms' coherence score is included, see section 3.1.3. For the specific coherence thresholds for filtering it is referred to chapter 4. As the form selection for the new query terms, both the basic approach and each of the four variants described in 3.1.4 are implemented, permutating this algorithm into five different versions. The weights for each of the new query terms are obtained from coherence and frequency information as outlined in section 3.1.5, where the Wikipedia-based frequency count is referred to as *topic adherence count*. With regard to the term selection variants, the implemented weighting method of the added alternative labels or

Input: The Wikipedia-matched terms (phrases) from the original query $\vec{P_{old}} \leftarrow [p_1, .., p_n]$, including duplicates for each occurrence, and their corresponding topic vector $\vec{T} \leftarrow [t_1, ..., t_n]$
Output: The terms (phrases) for the new query $\vec{P_{new}}$ and their corresponding weight vector $\vec{W}$
initialization
foreach *topic* $t \in T$ **do**
  get all unique labels $L \leftarrow \{l_1, .., l_m\}$
  foreach *label* $l \in L$, *ignoring capitalization differences* **do**
    if $l \in \vec{P_{old}}$ **then**
      get frequency of occurrence f_l of l in P
      $w_l = f_l \times idf_l$
      Add l to $\vec{P_{new}}$ and w_l to $\vec{W}$, at the same position
    end
  end
end

Algorithm 1: Frequency-weighted Wikipedia-aided query reformulation pseudo-code

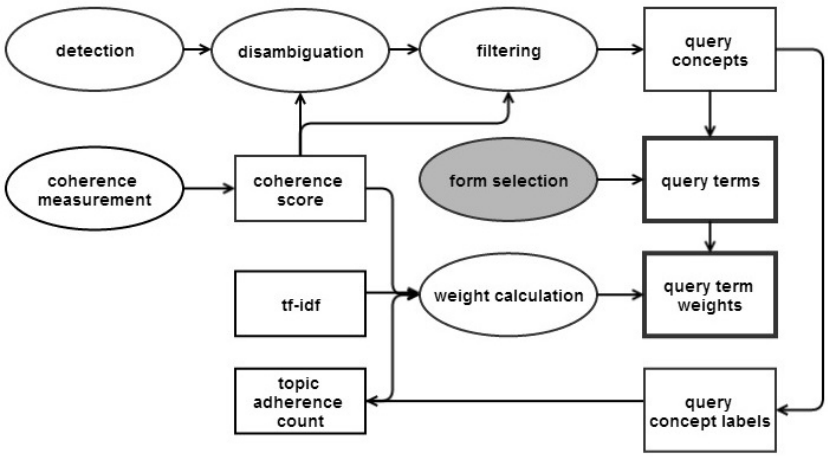

Figure 3.4: Frequency- and coherence- weighted Wikipedia-aided query reformulation workflow

original terms is deferred to the next chapter, which describes the experimental setup. The techniques of term selection and weighting are expressed in pseudocode in Algorithm 2. Again, note that *tf* is implemented indirectly, not by its weight but by including the term tf times in the new query.

Input: The Wikipedia-matched terms (phrases) from the original query $\vec{P_{old}} \leftarrow [p_1, .., p_n]$, including duplicates for each occurrence, and their corresponding topic vector $\vec{T} \leftarrow [t_1, ..., t_n]$; parameters $threshold_1$ and $threshold_2$, with $threshold_1 \geq threshold_2$
Output: The terms (phrases) for the new query $\vec{P_{new}}$ and their corresponding weight vector $\vec{W}$
initialization
foreach *topic* $t \in T$ **do**
  get all unique labels $L \leftarrow \{l_1, .., l_m\}$
  foreach *label* $l \in L$, *ignoring capitalization differences* **do**
    if $l \in \vec{P_{old}}$ **then**
      get frequency of occurrence f_l of l in $\vec{P_{old}}$
      get average of relatedness $r(t, t_i)$ for all $t_i \in \vec{T} \setminus \{t\}$, i.e., r_{AVG}
      if $r_{AVG} > threshold_1$ ***OR*** $(r_{AVG} > threshold_2$ ***AND*** $f_l > 1)$ **then**
        $w_l = f_l \times idf_l \times r_{AVG}$
        add l to P_{new} and w_l to W, at the same position
      end
    end
  end
end

Algorithm 2: Frequency- and coherence- weighted Wikipedia-aided query reformulation pseudocode

3.2 Our LDA query reformulation and document representation methods

Our second method for obtaining a document representation that accounts for ASR errors involves the use of Latent Dirichlet Allocation. Training and applying an LDA model (and some post-processing), as will be elaborated upon in the next sections, provides us with the following information:

- $p(z|d)$, i.e., the probability for each topic to occur under each document (in the test corpus).
- $p(t|z)$, i.e., the probability for each term in the (training) vocabulary to occur under each topic.

These methods apply LDA in two manners. The first one consists of regular vector space search using the $p(z|d)$ vectors. Likewise, documents are compared on the basis of 50 topically motivated dimensions instead of on several thousands dimensions of vocabulary terms. The second approach consists of constructing a *topic-based language model* and applying this to the query terms. Each query term t in a query document d is weighted according to $p(t|d)$, as obtained by $\sum_z p(t|z) \times p(z|d)$. That is, the likelihood of a term is modelled as the summed likelihood of finding the term in each of the topics that the LDA model is trained on. As such, for topics that have a low probability of being present in the query, little is added to to the query weight, as holds for terms that have a low probability of occurring under the topic.

The next sections first describe the training and inferencing stages that are inherent in both methods and then describe the two document representation algorithms. The software used in this study is the LDA implementation by David Blei, which can be downloaded from `http://www.cs.princeton.edu/~blei/lda-c/`.

3.2.1 Training

The LDA algorithm requires a collection of documents for topic estimation as well as some preknowledge about word and topic distributions, such as the number of topics in the model, which was set at 50. We propose to use a corpus of the same nature and origin as the test set, which in this situation translates to taking a set of articles from the same newspapers as the search collection. Also, the training corpus should be sufficiently large. This does not necessarily result in more clear-cut topic models, but does yield a larger vocabulary, which reduces the likelihood that any of the input documents does not contain any vocabulary term and cannot be rewritten. After training, besides the vocabulary, the algorithm outputs the (log) $p(t|z)$, i.e., the likelihood for every term t in the vocabulary to occur under each of the (50) topics z. This is referred to as the *topic model*.

3.2.2 Inferencing

The inferencing step of the LDA algorithm consists of applying the parameters estimated in the training phase onto the target corpus, which, depending on the algorithm used, consists of the queries only or the search collection as well. The result of inferencing is a vector of length [number of topics], i.e., 50, expressing the so-called variational Dirichlet parameters for each document. Since these parameters are proportional to the likelihood of occurrence in the document for the corresponding topic, the document vectors can be easily transformed into *document models*, i.e., 50-dimensional vectors of topic likelihood $p(z|d)$.

3.2.3 Document representation

Our LDA query reformulation method

The proposed query reformulation method based on LDA is schematized in Figure 3.5 and expressed in pseudocode in Algorithm 3. As a query reformulation method, the document representation is based on the components of term selection and term weighting.

Term selection The new query terms are taken from the background corpus that the algorithm is trained on. That is, all of the terms from the original query that appear in the vocabulary go into the new query, and all terms that do not are filtered out.

Term weighting Every resulting query term is weighted by the likelihood of finding this term under the topic model of the query document. We define the probability of a term t to occur in a document d to be the sum, for every of the 50 topics z, of $p(t|z) * p(z|d)$. Note that the frequency of the term in the query does not play any role here, and thus the only frequency information that is included in weighting is the inverse document frequency of the term in the search collection.

The LDA-based likelihood is derived from two models: the topic model and the document model. The topic model describes the distribution of terms over topics z, i.e., the likelihoods $p(t|z)$, and is generated in the training phase. The document model refers to the topical distribution detected in the query document, i.e., the likelihoods $p(z|d)$, which requires inferencing on the query documents.

Note that, as absolute probabilities, these values are extremely small. Instead of the raw probabilities, we therefore added a value 1 and then took the logarithm. After multiplying with the $p(z|d)$ values, the term weights still had exponents in the order of 10^{-10}; therefore every term's weight was normalized by a constant factor, keeping the relative weights unchanged.

Given the vocabulary terms and all values $p(z|d)$ and $p(t|z)$ for a query document d, the query reformulation algorithm is described in pseudocode in Algorithm 3.

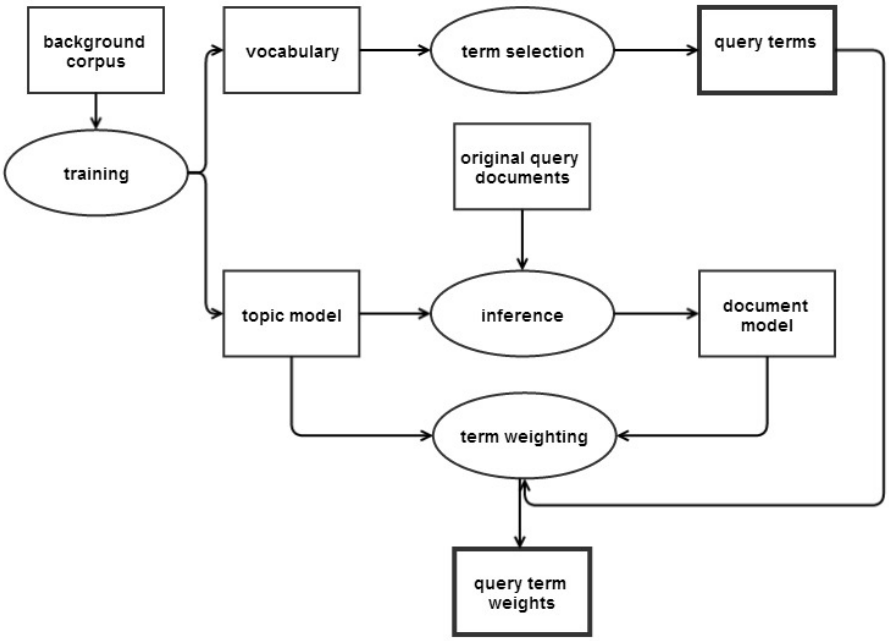

Figure 3.5: Query-based LDA algorithm

Recall that the remainder of the procedure consists of merging the weights into a vector representation, just like the target document collection, and then calculating pairwise cosine similarities between the query vector and every target document, resulting in a ranked list of most similar documents.

Input: original query document q; all values $p(t|z)$ and $p(z|q)$; constant factor *normalization*
Output: labels that make up the new query and their weight, i.e., new_query
initialization: $new\_query \leftarrow \emptyset$
get the set $T \leftarrow \{t_1, ..., t_n\}$ of all unique terms from q that occur in the model vocabulary
foreach *term* $t \in T$ **do**
  $prob \leftarrow 0$
  foreach *topic* z **do**
    $prob += p(t|z) \times p(z|q)$
  end
  $weight_t = prob \times normalization \times idf_t$;
  new_query.add$(t, weight_t)$
end

Algorithm 3: LDA-based query reformulation algorithm

Our LDA document representation method

The vector-based application of LDA is schematized in Figure 3.6. The idea behind this approach is to use the topic vectors derived from the document as comparison materials

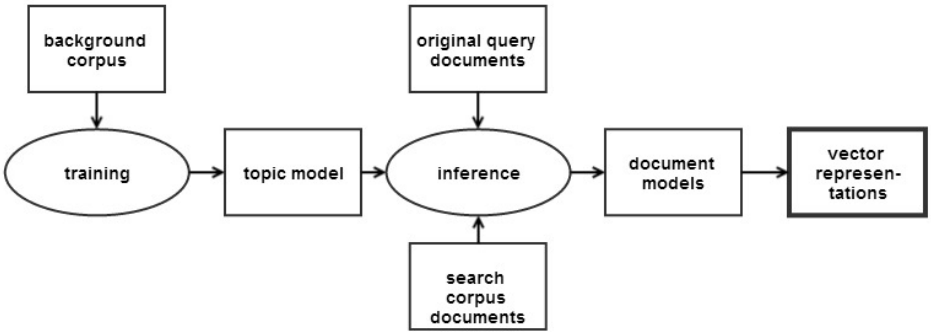

Figure 3.6: Vector-based LDA algorithm

instead of some tf-idf representation of (a selection of) the document, as in the method explained above. The topic vectors are exactly the *document models* obtained from the inferencing step. As it was chosen to work with 50 topics, the resulting vectors are of length 50. Since the input document and the target documents can only be compared when they are equally represented, this algorithm requires inferencing not only on the query documents but on the entire corpus of queries and search collection.

The remainder of the procedure is equal to the outlined query reformulation methods: pairwise cosine similarities are calculated between the query vector and every document in the collection, resulting in a ranked list of most similar documents.

Chapter 4

Experimental Setup

In order to test the benefit of the proposed document representation techniques in overcoming ASR errors in similarity search, we implemented and evaluated the methods described in Chapter 3. The goal of this experiment was to answer the research questions outlined in Chapter 1, which will be discussed as part of our results. Recall that our main research questions were as follows:

- *Can Wikipedia entities be used to overcome the difficulties of retrieval with ASR?*
- *Can LDA be used to overcome the difficulties of retrieval with ASR?*

As a model of ASR-based retrieval, the unmodified speech recognized documents were taken as baseline queries. They can be found in A.1. Recall that LDA was applied in two different ways, on the query term level as well as the vector representation level. Since our aim was to retrieve documents relevant to the input document, the research questions were operationalized into the following questions:

1. *Do the queries obtained by any of the Wikipedia conditions return a higher number of relevant documents than the unaltered input document?*
2. *Do the queries obtained by the LDA algorithm return a higher number of relevant documents than the unaltered input document?*
3. *Do the vector representations obtained by the LDA algorithm return a higher number of relevant documents than the tf-idf representation of the input document?*

To exploit the full potential of Wikipedia, we tried its anchor text structure in the tasks of term selection and weighting, respectively. With respect to the proposed term selection methods, *question 1* above was assigned the following subquestions:

- *Does appending the original query terms to the Wikipedia algorithm output lead to a higher number of relevant returned documents than substituting them for the latter? Is it better to append the full original query or only the more-frequent terms?*
- *Does using the topic's title, rather than its label that is found in the query, lead to a higher number of relevant documents in retrieval?*
- *Does enriching the label by the most frequent alternative labels of its corresponding topic lead to a higher number of relevant returned documents?*

To gain insight into the effect of coherence-based term weighting, the following question was added:

- *Does weighting the new query terms by their coherence in the query document lead to better retrieval results than weighting them by frequency information only?*

In this chapter we describe the materials used for the study as well as important settings. Also, we explain how the relevance of the returned documents was assessed. Finally, we explain the different experimental conditions used to answer the questions above.

4.1 Data

4.1.1 Target corpus

We evaluated our algorithms on a background corpus of 25100 newspaper articles. They were taken from two of the most widely circulated Dutch newspapers, *Telegraaf* and *NRC Handelsblad*, and dated between May 31 and August 31, 2012. Note that the original dataset was over 28000 documents in size, but was filtered to be interpretable by the LDA algorithm (see 4.2.2); likewise, all experiments could be performed on the same data. All text materials from the articles were used as indexing materials, including headlines and picture subscripts.

4.1.2 Source corpus

As our input documents we used 30 news items that were produced by national broadcast service NOS and broadcast on Dutch TV in July and August 2012. They were selected using a random generator for the day (1-30), month (7 or 8), and a number referring to the order of appearance in the bulletin (1-10), where a maximum number of 10 items was assumed. Generated articles were selected for the final set if they met the following requirements:

- availability in both speech-recognized and closed captioning form;
- a minimum length of 50 words, based on the ASR version of the document;
- a topic other than weather forecasts and sports game results;
- novelty, i.e., absence of another item on the exact same event in the data set.

The resulting queries can be found in Appendix A.1.

The closed captioning data for these news bulletins were gathered from the NOS editorial office through their live Teletext service, and were exactly time-stamped to be synchronized with the recognized speech. Note that closed captioning was not entirely verbatim. It was edited to include punctuation and special characters such as numbers and symbols indicating the end of a news item. Also, interjections such as “uh” were omitted, whereas metadata were added in capitals (e.g, *zucht diep [sighs deeply])*, and occasionally, rephrasing took place. However, as the closed captioning did reflect the intended meaning of the article, it was judged that they could safely be used as reference materials.

The ASR transcriptions for the queries were generated by X-MI, a 2008-founded spin-off company of research conducted by the Human Media Interaction department of Twente University. No word error rate (WER) was known for the X-MI ASR tool; however, its performance reflects the state-of-the-art in speech recognition. The query materials in Appendix A.1, where both the transcripts and the Teletext counterparts are given, give an impression of the ASR’s accuracy level. The speech recognizer phrased all information perceived, including numbers and interjections. However, it was trained to recognize a set of names (e.g., *France-Soir*). Since the ASR software did not segment its output in any way, we transferred the segmentation present in Teletext onto the speech output, using the assigned time stamps. An initial round of experiments revealed that some news items

were clustered due to poor segmenting in Teletext, which showed to have a distortive effect on search results. Therefore, whenever necessary, the Teletext input documents (and speech texts alike) were manually separated from their neighboring items. Note that our segmentation method did not violate the intended use case, in which closed captioning is not assumed available, since recognizing transitions from one news item to another is likely to be unproblematic by means of acoustic or image analysis.

4.1.3 LDA training corpus

Training the LDA model was performed on a background corpus equal in nature and similar in size to the test set (i.e., to the indexed documents): 23746 documents from *NRC Handelsblad* and *Telegraaf*. To prevent overlap with the test set and simulate a realistic setting, they were taken from the period directly preceding the test set materials and covered a period of 4 months (February up to May 2012). The recency of the training set as compared to the test set was not likely to influence the resulting thematic clustering, but did ensure that the *vocabulary* was updated with recent names and phenomena, which was considered desirable in the highly dynamic news domain.

4.1.4 LDA topic model

The topics that resulted from training the LDA model, as was visualized by printing the 20 most likely terms for each of the 50 topic clusters, are displayed in Appendix C. They appear to be fairly coherent and well-defined, and most of the topics can intuitively be labeled to discuss, for example, "food", "art and culture", "foreign politics", etc. This strengthened our confidence that the model could successfully be applied to capture document themes.

4.1.5 Wikipedia corpus

The Dutch Wikipedia corpus that was used was freely downloaded from `http://download.wikipedia.org` in June 2012. At that moment, it contained 1.288.615 articles.

4.2 Settings

For each of the techniques there are some settings that need to be specified in order for our results to be reproducible. We list the technical details about the information retrieval setting, such as the weighting scheme and text preprocessing, as well as the settings used in training and inferencing using the LDA software.

4.2.1 Information retrieval settings

Our query-based experiments were executed using open-source library Lucene version 4.2.1. The similarity measurement supported by this software is a tf-idf measure that takes into account document length and user-defined query term boosts. Conceptually, the similarity score between a query q and any indexed document d in Lucene is as follows [23]:

$$sim(q,d) = \sum_{t \in q} tf_{t \in d} \times tf_{t \in q} \times idf_t^2 \times boost_t \times lengthnorm_{t,d} \tag{4.1}$$

As the inner product of two vectors (see 2.2 on cosine similarity) the formula contains *tf* and *idf* for both query and document, which are equal; this results in a squared *idf*. A term with an intended frequency f can be incorporated in the query either once with $boost_t = f$, which has the same effect as $tf_{t \in q} = f$, or three times such that $tf_{t \in q} = 1$. The effect is equal, ignoring minor deviances due to the compression of the length norm in byte form. Note that the queries listed in Appendix A might differ in this respect. These queries are

displayed in the form in which they entered Lucene, hence their weight is actually the *boost* in the formula above. Term frequency tf_{tind} takes the form of $frequency^{\frac{1}{2}}$, i.e., this value weighs less-frequent terms relatively more heavily than high-frequent terms. A similar smoothing effect for high values is applied to the document length normalization, which is taken to be the square root of the number of tokens of the documents. Finally, inverse document frequency is calculated as $1 + \log \frac{N}{docFreq+1}$, again discounting the relative effect of very high-frequent terms.

For text preprocessing tasks such as stop word removal and stemming, the default settings from the *DutchAnalyzer* class [21] for Lucene 4.2.1 were used. It includes a very general stop word list and a small stemming exclusion list of four irregular Dutch nouns. Tokenization was done by the *standardTokenizer* class [22], which, except for a few exceptions, splits tokens on punctuation characters and hyphens. For stemming, the Porter algorithm was used [11]. Again, note that the queries in A, which are displayed in the form in which they entered Lucene, are not yet stop-filtered or stemmed and hence represent an overcomplete and overly diverse set of terms that was matched upon.

4.2.2 LDA settings

The LDA algorithm, whether in training or inferencing, took as its input word count vectors of length *vocabulary size*. For generating the training corpus (and hence the vocabulary as well), the input vectors were obtained from the training documents through the steps of tokenization, stop word removal, and frequency counting and filtering. We designed the stop word list to not only include the mainstream stop words, but also specific terms from newspaper sections that were not of interest, such as TV channels or terms such as "crosswords". Assuming that moderately sized articles provide the most reliable training materials, we decided to add a filtering step by removing documents with a word count below 100 or above 800. After stop word removal and filtering, 25000 unique terms were maintained in the vocabulary. Also, we filtered out documents that did not contain at least 2 different vocabulary terms. The resulting document set for training included nearly 24000 documents. For testing, all 60 (speech and Teletext) queries passed the vocabulary requirement and could be easily transformed into word count vectors.

With regard to the model, we imposed a number of topics $K = 50$ and, according to the well-established [28] rule of thumb $\alpha = \frac{K}{50}$, an initial value of 1 for topic distribution parameter α, which was estimated along with the topic distributions. Topics were chosen to be initialized randomly rather than to a distribution smoothed from a randomly chosen document, which is also supported by the used software. As stop conditions, variational inference in both training and testing were defined to have converged when the difference of the new and the previous score over the previous score was below 1×10^{-6}. For the training phase, variational inference continued until either convergence was reached or after 20 iterations; for testing, no maximum number of iterations was imposed on variational inference. With regard to expectation maximization, convergence was set to 1×10^{-4} for both training and testing, with a maximum of 50 respectively 100 iterations.

To transform the variational Dirichlet parameters outputted in the inferencing phase into topic likelihood vectors, we used standard normalization operations. It was experimented with the maximum norm, based on the highest element; the L2 norm, based on the absolute squared sum, and the Euclidean norm. Intuitively, it was preferred to normalize the vector of topic probabilities by the most apparent topic detected in the document rather than, for example, the sum, such that documents with a wide topic distribution would not be discounted by low probability values, and therefore the maximum normalization method was chosen. However, examining the resulting queries for different ways of normalization, this did not seem to greatly affect the proportions of the probabilities.

4.3 Evaluation

Throughout the experiments, *relevance* of a document for any of the 30 unique queries was assessed by means of a protocol defined for each of the queries. The rationale behind this approach was to avoid bias in the judgment of the rater, who was the author of this document. By defining the relevance statement *before* being presented with the resulting documents as well as *before* generating the different query variations, any significant decision process was removed from the assessment phase, where the rater could (unconsciously) favor high scores for certain experimental conditions. Since no estimation could be made in advance of the potential of the queries to generate documents of the desired thematic types, the protocols were based entirely on *what the rater (and the informally consulted peers) considered the most salient topics in the input document*.

The protocols were formulated in terms of (boolean) combinations of topics on a graded scale in such a way that

- a document was given the highest relevance score *2* if it contained all substantial topics present in the query;
- a document was given the intermediate relevance score *1* if it contained at least one substantial topic present in the query;
- a document was given the lowest relevance score *0*, i.e., was considered irrelevant, if it contained not a single substantial topic from the query.

For judging how *substantial* a topic was in the document, we took into account its frequency of occurring in the news in general. For example, for a news item about the financial position of airline company *KLM- Air France*, we judged the mere topic *financial news about a company* not discriminative enough for relevance, and required a combination of *airline industry* and *financial news* to be present in a document for getting a score of *1*. On the other hand, for an input news item about a power cut in India we did consider *India* a significant topic, as it was judged to be not frequently covered in Dutch news. A complete overview of the formulated protocols is given in Appendix B. Note that, as the same protocols were used across experimental conditions for each query, a certain degree of subjectiveness which is inherently involved in formulating relevance protocols was not likely to hinder our main goal, which was to *contrastively* assess the proposed techniques.

After assessment, search performance of any of the experimental conditions was evaluated by means of averaged *normalized discounted cumulated gain* [31]. Discounted cumulative gain is a rank-based measure designed for graded relevance. Assuming that relevant documents higher up in the result set are more valuable than lower-ranked results, DCG discounts a result's contribution to the score by some b-logarithm of its rank. Likewise, the gain (G) contributed by a document at rank r with relevance score s is $s \times^b \log r$. For ranks that are lower than the base of the logarithm, this formula would result in a boost rather than a discount; therefore, the gain for each $r < b$ is not discounted. We took the base of the logarithm to be 2, as suggested in [31]. Cumulating the values for each of the rank positions that we considered resulted in the *discounted cumulated gain vector* of length 10, where the element at position r was defined recursively as [31]

$$DCG[r] = \begin{cases} CG[i] & \text{if } i < 2, \\ DCG[i-1] + G[i] \times^2 \log r & \text{else.} \end{cases} \tag{4.2}$$

The DCG vector for a query in an experimental condition was normalized by the recall base for the query, which was taken to be the maximum possible or *ideal* DCG vector over all results for that query. That is, the ideal vector was based on the total number of

documents with relevance scores 2 and 1 and was cut off either at 10 or, if the total number of relevant documents was below 10, at the last non-zero score. For example, if for a given query across all conditions a total of eight documents were assessed as 2, one as 1 and some irrelevant number as 0, then the *ideal non-discounted gain vector* would be $\{2,2,2,2,2,2,2,2,1\}$. Then, discounting the gain values following formula (4.2) would result in the *ideal normalized discounted cumulated gain vector* {2.00,4.00,5.26,6.26,7.12,7.90,8.61,9.28,9.59}. Finally, to obtain a test statistic for a given combination of a query and an experimental condition, its corresponding DCG vector was divided , i.e., *normalized* , by the query's ideal DCG vector, element-wise, until the last position of the ideal DCG. Finally, the average of this vector represented the test statistic for the given combination of query and condition.

4.4 Validation of our evaluation method

In order to validate the assumptions behind the approach of this study, a survey was set up. It was designed to answer the following questions:

1. Given the used protocols, would other raters assess a target document's relevance to a source document the same way?
2. Does the used protocol capture perceived similarity between document pairs?
3. Does the used protocol capture perceived linkworthiness of the source document to the target document?

To address (1), a specific method was thought out. Rather than presenting respondents with a protocol and a pair of source and target documents, which would unnecessarily complicate the task, we gave them only the target document and a list of thematic tags. The tags were exactly the terms that the relevance protocols of the input document were composed of. Likewise, a translation was made of the input document to a set of thematic terms. Respondents were asked to select all terms that applied to the (target) document, which was referred to as [item A]. In case of uncertainty, respondents could indicate their hesitation and were asked to clarify their doubts in a blank field. Below, an example question from the survey is displayed, as well as a translation.

> Vink alle onderwerpen aan die van toepassing zijn op krantenartikel [item A]. Kies *geen* als geen enkel genoemnd label van toepassing is en *ik weet het niet* als je geen keuze kan maken. Licht in dat laatste geval je twijfel toe in het tekstvak. [Tick all the subjects that are applicable to newspaper article [item A]. Select *none* if none of the listed labels are appropriate and *I don't know* if you cannot make a choice. In the latter case, explain your hesitation in the text field.]
>
> - Olympische Spelen [Olympic Games]
> - sport [sports]
> - doping [doping]
> - wedstrijdfraude [game fraud]
> - *geen* [*none*]
> - *ik weet het niet (geef nadere toelichting)* [*I don't know (please explain)*]

To address the remaining questions, i.e., (2) and (3), users were presented with the source document as well. Note that all questions for a document pair appeared on the same page. After reading the input document, that was clearly labeled as [item B], respondents were asked to rate the relatedness between this new item [item B] and the target document [item A] they had already read and that was displayed at the top of the page. Relatedness was rated on a 3-point qualitative scale with labels that translate as *directly highly related*, *indirectly partly related*, and *not related* Also, there was a response option *I don't know* with a blank text field, that was requested to be filled with an explanation in case this option was chosen. Below an example of the relatedness question formulation is displayed.

In hoeverre is krantenbericht [item A] gerelateerd aan journaalbericht [item B]? Kies *ik weet het niet* als je geen keuze kan maken. Licht in dat geval je twijfel toe in het tekstvak. [To what extent is newspaper article [item A] related to news broadcast item [item B]? Select *I don't know* if you cannot make a choice. In that case, explain your hesitation in the text field.]

- Direct / sterk gerelateerd [Directly / strongly related]
- Indirect / deels gerelateerd [Indirectly / partly related]
- Niet gerelateerd [Not related]
- *Ik weet het niet (geef nadere toelichting)* [*I don't know (please explain)*]

Finally, the question that assessed the degree of linkworthiness was formulated as follows: *Imagine that you are watching an online news broadcast and that you hear [item B]. Would you consider a link from this item to the previous one, [item A], relevant? That is, would you find it useful to be able to click through from the broadcast item to the newspaper article?* Following a three-point scale, answer options for this question included *yes*, *maybe*, and *no*. Again, there was the possibility to respond with *I don't know* and to give a reason for this. See example below.

Beeld je in dat je online een journaal kijkt en het vorige item [item B] te horen krijgt. Zou je een link van dit item naar het eerstgenoemde item [item A] relevant vinden? M.a.w., zou je het waardevol vinden om van het journaalbericht door te kunnen klikken naar het krantenartikel? Kies *ik weet het niet* als je geen keuze kan maken. Licht in dat geval je twijfel toe in het tekstvak. [Imagine that you are watching an online news broadcast and that you hear [item B]. Would you consider a link from this item to the previous one, [item A], relevant? That is, would you find it useful to be able to click through from the broadcast item to the newspaper article? Select *I don't know* if you cannot make a choice. In that case, explain your hesitation in the text field.]

- Ja [Yes]
- Misschien [Maybe]
- Nee [No]
- *Ik weet het niet (geef nadere toelichting)* [*I don't know (please explain)*]

In total, 76 document pairs were included in the survey, representing 20 queries (1 up to 20) from this study. However, due to an error in the tag list, the results for one of the queries (query 14) were omitted, yielding a final experimental set of 71 document pairs. For each query, target documents were carefully selected to at least include the more difficult cases encountered during assessment. Consequently, queries for which the assessment process had been straightforward were represented to a lesser extent in the question set. Also, it was assured that the overall distribution of protocol-based relevance assessments by the rater over the documents was approximately equal. In the final set, out of 71 target documents, 26 documents had been assessed by rater as highly relevant to the corresponding query, 20 out of 71 as relevant, and 25 as non-relevant. The survey was composed such that each page assessed all three questions (1), (2) and (3) for one input and target document pair.

Respondents were collected through an online social networking site. Pages were randomized for each respondent. Every respondent was asked to fill out 10 pages of the survey, although they were free to do more or less. Participation was voluntary and was not refunded in any way. In total, 15 people filled out the survey, with 237 filled out pages in total. One participant completed the entire survey, the others on average assessed 12 pages.

4.5 Experimental conditions

To be able to answer all proposed research questions in detail, we created a set of variants of every input document. The names are composed in such a way that their components express, respectively, the source(s) the terms were selected from (the original query, Wikipedia, or the LDA vocabulary), the factor(s) included in the weighting scheme (tf, idf, the Wikipedia-based frequency count, or the Wikipedia-based coherence score) and, if applicable, the term selection variant that was applied (the topic pagename or the top-5 most frequent labels) or some other specification.

1. ***original_tfidf***: the unaltered query baseline, weighted by tf-idf;
2. ***original_tfidf_freqselective***: those terms from the original query which occurred more than once, weighted by tf-idf;
3. ***wiki_idf***: the Wikipedia-matched query phrases, weighted without any frequency information (i.e., a constant term frequency 1);
4. ***wiki_freq&tfidf***: the Wikipedia-matched query phrases, weighted by tf-idf and the Wikipedia-based frequency count;
5. ***wiki_freq&sim&tfidf***: the Wikipedia-matched query phrases, weighted by tf-idf, the Wikipedia-based frequency count, and the Wikipedia-based coherence measure;
6. ***wiki_freq&sim&tfidf_pagename***: the *topic pagenames* of the topic associated with the Wikipedia-matched query phrases, weighted by tf-idf, the Wikipedia-based frequency count, and the Wikipedia-based coherence measure;
7. ***wiki_freq&sim&tfidf_altlabels***: the *top-5 most frequently used labels* of the topic associated with the Wikipedia-matched query phrases, weighted by tf-idf, the Wikipedia-based frequency count, and the Wikipedia-based coherence measure;
8. ***original&wiki_freq&sim&tfidf***: the Wikipedia-matched query phrases, weighted by tf-idf, the Wikipedia-based frequency count, and the Wikipedia-based coherence measure; and all of the original query terms, whether already included in the new query or not, weighted by tf-idf;
9. ***original&wiki_freq&sim&tfidf_selective***: the Wikipedia-matched query phrases, weighted by tf-idf, the Wikipedia-based frequency count, and the Wikipedia-based coherence measure; and all the original query terms that occurred more than once in the original query but not yet in the new query, weighted by tf-idf;
10. ***LDA_query***: the result of our LDA-based query reformulation method;
11. ***LDA_vector***: the result of our LDA-based vector representation method.

The following sections describe every variation and its parameters (if applicable), and provide an example of the rewritten query as it went into the weighting scheme described in section 4.2.1. The examples are based on the speech version of the shortest query in the set (query 1). Recall that every condition was applied to the Teletext queries as well.

original_tfidf

This condition was formed by the unaltered speech-recognized queries as found in Appendix A.

Example. bij een schietpartij uh bij discotheken frankrijk zijn afgelopen nacht tien mensen gewond geraakt het dader was boos omdat die de discotheek vlakbij de plaats kan brengen niet in mochten daarop haalde hij jachtgeweren uit zo'n auto beschoten zo buiten bij de discotheek op binnen na de schietpartij ging de man er vandoor is waar leven aangehouden aan

original_tfidf_freqselective

This condition was formed by taking every term from the original query which occurred more than once.

Example.(Bij)3 (Schietpartij)2 (Discotheek)2 (De)4

wiki_idf

Disregarding their frequency in the original query, this condition was made out of all *unique* Wikipedia-matched terms from the original query, which in practice is equal to assigning to each of the terms a frequency of 1.

Example. Frankrijk Auto Dader Discotheek Discotheken De Man

wiki_freq&tfidf

The Wikipedia algorithm that used frequency-based weighting was fully described in Algorithm 1, and since this was one of our main conditions we listed all speech-based queries in Appendix A. To recap, the query was reduced to contain every term that occurred as a label of a Wikipedia topic recognized in the query, weighted by the number of recognized topics of which it was a label and regular tf-idf.

Example.(Frankrijk)1 (Auto)1 (Dader)1 (Discotheek)2 (Discotheken)1 (De Man)1

wiki_freq&sim&tfidf

The Wikipedia algorithm that used frequency as well as similarity weighting was fully described in Algorithm 2. We examined the query outcomes for different threshold values for frequency and coherence and judged that the best terms were maintained with a default similarity threshold of 0.20 and a value of 0.05 in case the term's frequency was above 1. To recap, the query was reduced to contain every term that occurred as a label of a Wikipedia topic recognized in the query, duplicated for each of these topics of which it was a label, on the condition that its corresponding topic was related to the other detected query topics by at least 0.20 on average, or by at least 0.05 in case the label occurred more than once. Each term was then weighted by multiplying its averaged similarity to the other query topics with its tf. Since this is one of our main conditions we listed all speech-based queries in Appendix A.

Example.(Frankrijk) 0.51 (Dader) 0.32 (Discotheek) 0.77 (Discotheken) 0.38

wiki_freq&sim&tfidf_pagename

In this variant of wiki_freq&sim&tfidf, the terms outputted by algorithm 2 were replaced by the name (i.e., page title) of their corresponding topic.

Example.(Frankrijk)0.51 (Dader)0.32 (Discotheek)0.77 (Discotheek)0.38

wiki_freq&sim&tfidf_altlabels

In this variant, for each of the terms (recall that these are Wikipedia labels) outputted by algorithm 2 we added its most-frequent alternative labels, with a maximum of 4. To keep the relative weights among the concepts equal, we divided the weight of the parent term by the number of alternative labels that were found, which overall resulted in a doubled, hence relatively unchanged, weight for each concept.

Example.(Frankrijk0.51) (Franse Frans Fransen Fransman)0.128 (Dader0.32) (Plegers)0.32 (Discotheek0.77) (Discotheken Disco Club Danszaal)0.192 (Discotheken0.38) (Discotheek Disco Club Danszaal)0.095

original&wiki_freq&sim&tfidf

The variant that combined the outcome of algorithm 2 with the original query terms was trivially obtained by appending these two in random order. However, the key question in this condition was the weight to assign to the original terms. As a simple heuristic, we decided to weight them by the *lowest weight* that the algorithm had assigned to the *new* query terms.

Example.(Frankrijk) 0.51 (Dader) 0.32 (Discotheek) 0.77 (Discotheken) 0.38 (bij) 0.32 (een) 0.32 (schietpartij) 0.32 (uh) 0.32 (bij) 0.32 (zijn) 0.32 (afgelopen) 0.32 (nacht) 0.32 (tien) 0.32 (mensen) 0.32 (gewond) 0.32 (geraakt) 0.32 (het) 0.32 (was) 0.32 (boos) 0.32 (omdat) 0.32 (die) 0.32 (de) 0.32 (vlakbij) 0.32 (de) 0.32 (plaats) 0.32 (kan) 0.32 (brengen) 0.32 (niet) 0.32 (in) 0.32 (mochten) 0.32 (daarop) 0.32 (haalde) 0.32 (hij) 0.32 (jachtgeweren) 0.32 (uit) 0.32 (zo'n) 0.32 (auto) 0.32 (beschoten) 0.32 (zo) 0.32 (buiten) 0.32 (bij) 0.32 (de) 0.32 (op) 0.32 (binnen) 0.32 (na) 0.32 (de) 0.32 (schietpartij) 0.32 (ging) 0.32 (de) 0.32 (man) 0.32 (er) 0.32 (vandoor) 0.32 (is) 0.32 (waar) 0.32 (leven) 0.32 (aangehouden) 0.32 (aan) 0.32

original&wiki_freq&sim&tfidf_selectivelyappended

This condition appended the outcome of algorithm 2 to a selection of the original query terms, specifically, to terms that occurred in the original query at least twice *and* had not been selected for the new query by the algorithm. Similarly to *original&wiki_freq&sim&tfidf* we took the lowest weight of the algorithm-generated terms as a basis for weighting the added terms, using the formula $weight_{min} \times \frac{freq}{2}$. As such, we incorporated the frequencies of the original query terms while maintaining the baseline value $weight_{min}$ that was also used in *original&wiki_freq&sim&tfidf*.

Example.(Frankrijk)0.51 (Dader)0.32 (Discotheek)0.77 (Discotheken)0.38 (Bij)0.32 (Schietpartij)0.32 (De)0.64

LDA_query

The generation process of the LDA-based query was explained in Algorithm 3, and since this was one of our main conditions we listed all speech-based queries in Appendix A. To recap, we reduced the query (q) to the terms that occurred in the LDA training set and weighted every term t by its idf as well as its likelihood of having occurred under the topic-motivated query language model, i.e., by the sum, for every of the 50 topics z, of $p(t|z) * p(z|q)$. Note that tf did not play a role here.

Example. (daarop) 0.2 (frankrijk) 4.91 (vandoor) 0.27 (haalde) 0.02 (gewond) 2.75 (buiten) 0.61 (geraakt) 0.26 (auto) 20.27 (schietpartij) 0.89 (vlakbij) 0 (plaats) 0.65 (brengen) 0.05 (beschoten) 0.02 (afgelopen) 4.12 (ging) 4.36 (mochten) 0 (boos) 0.02 (aangehouden) 9.33 (binnen) 3.18 (dader) 5.34 (discotheek) 0 (nacht) 0.51

LDA_vector

This condition was applied on the deeper level of vector representation rather than the query formulation level, and on the full collection rather than just the query. Every document (q) was reduced to its 50-dimensional vector of topic likelihoods $p(z|q)$.

Chapter 5

Results

5.1 Main study

This chapter describes the results of our main experiment, the setup of which was treated in chapter 4. Eleven experimental conditions were composed, which differed along the dimensions of interest of the used external information source, and the document representation method. As a reference, we list in tables 5.1 and 5.2 the characteristics of the different conditions. For the Wikipedia-based conditions, the two main weighting variations were expressed as two separate algorithms in Algorithm 3.1.6 and Algorithm 3.1.6. The document representation variants for these conditions affected the lexical form of the terms in the rewritten query, and were integrated only in the frequency- and similarity-based weighting algorithm. For the LDA-based conditions, the key difference was in the way LDA was applied to represent the documents: as a vector of topic probabilities, or as a query reduction and weighting mechanism on the term level.

Condition name	Term weight	Term filtering
original_tfidf	tf-idf	none
original_tfidf_freqselective	tf-idf	tf > 1
wiki_idf	idf	none
wiki_freq&tfidf	$tf-idf \times freq \times sim$	none
wiki_freq&sim&tfidf	$tf-idf \times freq \times sim$	sim > .2 OR [sim > .05 AND freq > 1]
"_pagename	$tf-idf \times freq \times sim$	sim > .2 OR [sim > .05 AND freq > 1]
"_altlabels	$tf-idf \times freq \times sim$	sim > .2 OR [sim > .05 AND freq > 1]
original&wiki_freq&sim&tfidf	$tf-idf \times freq \times sim$	sim > .2 OR [sim > .05 AND freq > 1] (Wiki) none (original)
original&wiki_freq&sim&tfidf_selective	$tf-idf \times freq \times sim$	sim > .2 OR [sim > .05 AND freq > 1] (Wiki) tf > 1 AND not in Wiki terms (original)
LDA_query	$likelihood \times idf$	none
LDA_vector	not applicable	not applicable

Table 5.1: Experimental conditions and their differences

The analysis was performed separately for speech (ASR) and Teletext conditions. Within both groups, we compared results in two manners. First, for the combined set of conditions, which indicated whether there was a reliable difference between the different conditions in the first place. Second, we performed comparisons between pairs of conditions that were of interest, such as the unaltered speech-recognized baseline condition and every other experimental condition, between the conditions with the different weighting schemes and between the term selection variants and their parent condition. This allowed us to state which query formulation method(s) performed better than the unaltered query,

Condition name	Form selection
original_tfidf	Original terms
original_tfidf_freqselective	Original terms
wiki_idf	Wiki-matched terms
wiki_freq&tfidf	Wiki-matched terms
wiki_freq&sim&tfidf	Wiki-matched terms
"_pagename	Page name of Wiki-matched terms
"_altlabels	Wiki-matched terms + altlabels
original&wiki_freq&sim&tfidf	Wiki-matched + original terms
original&wiki_freq&sim&tfidf_selective	Wiki-matched + original terms
LDA_query	Original terms
LDA_vector	Topic-based vector

Table 5.2: Experimental conditions and their differences [continued]

and which of the tried weighting method and term choice was best-performing.

Recall that in the following, each experimental result denotes a cumulated discounted gain value that was normalized (nDCG) by the ideal cumulated gain value obtained from the cross-conditional recall base for the query document. See section 4.4 for details. Unless stated otherwise, a significance level of 0.05 was maintained. Query 16 and 24 were left out of consideration whenever necessary, as these did not give a query result in all conditions. For more information about the experimental conditions and examples, see section 4.5.

5.1.1 Speech queries

Overall performance and differences

The cross-query average nDCG for each expermental condition and their deviations are displayed in table 5.3. Values ranged between 0.19 and 0.75, with a baseline performance of 0.63.

Condition	Mean	Standard deviation
original_tfidf	0.625	0.280
original_tfidf_freqselective	0.575	0.342
wiki_idf	0.707	0.299
wiki_freq&tfidf	0.745	0.286
wiki_freq&sim&tfidf	0.691	0.348
wiki_freq&sim&tfidf_pagename	0.610	0.341
wiki_freq&sim&tfidf_altlabels	0.542	0.367
original&wiki_freq&sim&tfidf	0.667	0.268
original&wiki_freq&sim&tfidf_selective	0.725	0.308
LDA_query	0.311	0.302
LDA_vector	0.187	0.209

Table 5.3: Average nDCG and standard deviation for each speech condition

Techniques that scored higher than the unboosted tf-idf speech baseline (*original_tfidf*) in terms of average nDCG included *wiki_idf*, *wiki_freq&tfidf*, and *wiki_freq&sim&tfidf* together with its variants *original&wiki_freq&sim&tfidf* and *original&wiki_freq&sim&tfidf_selective*. The highest value was found in *wiki_freq*, shortly followed by *original&wiki_freq&sim&tfidf_selective*. Below-baseline performance was found in the condition that did not include any Wikipedia information, but only the higher-frequent terms from the original query, i.e., *original_tfidf_freqselective*. Also, two variations on *wiki_freq&sim&tfidf* underperformed

against the baseline: the condition that included the alternative labels as well as the original terms, i.e., *wiki_freq&sim&tfidf_altlabels*, and the condition where the Wikipedia-matched terms were substituted by the topic pagename, i.e., *wiki_freq&sim&tfidf_pagename*. Both LDA techniques had averages markedly lower than the baseline.

The variation within each condition, as expressed by the standard deviation of nDCG values between the queries, was higher than the unaltered baseline in every experimental condition except *original&wiki_freq&sim&tfidf* and *LDA_vector*. This indicates that the variations in nDCG for the separate queries was higher in the Wikipedia-based conditions than in the baseline condition, suggesting that the effect of these conditions was less *stable* or *reliable*. This was especially the case in the condition *original_tfidf_freqselective*, which selected from the original query only the more-frequent term. This technique yielded very low scores whenever the resulting query was very sparse. However, the best-performing condition, *wiki_freq&tfidf*, did not deviate substantially from the baseline in terms of standard deviation.

A one-way repeated measures ANOVA was used to test for overall differences in results. The assumption of equal variances of pairwise condition differences was not met, as indicated by Mauchly's test ($\chi^2(54) = 115.76, p = 0.00$); therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.52$). Also, for the majority of conditions, the data were not normally distributed. However, this has been reported to have only a small effect on the chance of so-called *type 1 errors* [36], where an effect is claimed that is not actually present; therefore, we accepted this violated assumption. The test showed that nDCG values differed significantly across conditions ($F(5.26, 141.88) = 18.00, p = .00$), allowing us to perform post-hoc tests.

Differences between conditions

For a more detailed insight into the performance of each of the different algorithms, we performed pairwise comparisons of the means. Since the data were not normally distributed for the majority of the conditions, we chose the Wilcoxon signed rank test.

Wikipedia conditions First, to examine whether any of the conditions produced significant performance gain over the unmanipulated input documents, we compared each of them to the unaltered speech baseline condition.

The condition that did not include any Wikipedia information and was based solely on the more-frequent terms from the original query, i.e., *original_tfidf_freqselective*, although not significantly ($Z = -0.31, p = 0.76$), scored below baseline level. Thus, using frequency information alone did not suffice to reduce the query such that it returned more relevant documents.

The condition that was based on the Wikipedia-matched terms, but without including any frequency information, i.e., *wiki_idf*, was significantly better than the unaltered baseline at the 0.1 level ($Z = -1.66, p = 0.10$). Thus, merely reducing the original query terms to the unique set of terms that occurred in the Wikipedia anchor text vocabulary increased performance.

Looking at the difference with the baseline of the weighting variants that were tried on basic tf-idf, it was found that *wiki_freq&sim&tfidf* did not significantly differ from the speech baseline ($Z = -1.33, p = 0.18$). Therefore, we cannot reliably state that the Wikipedia algorithm with the most elaborate weighting scheme was any better than the unaltered speech-recognized query in combination with tf-idf weighting. On the other hand, *wiki_freq&tfidf* did reliably *outperform* the baseline ($Z = -2.38, p = 0.02$), indicating that our Wikipedia rewriting method combined with frequency and tf-idf weighting indeed increased performance in the context of speech-recognized text.

Contrasting the four term variants of *wiki_freq&sim&tfidf* to the baseline condition, the findings were as follows. First, no significant difference could be demonstrated for the conditions based on the pagename (*wiki_freq&sim&tfidf_pagename*, $Z = -0.01, p = 0.99$) and the different labels associated with the recognized topics in the query (*wiki_freq&sim&tfidf_altlabels*, $Z = -0.51, p = 0.61$). However, both conditions scored below baseline level, indicating that any benefit obtained from our Wikipedia-based method, as displayed by the above-baseline mean of *wiki_freq&sim&tfidf*, was lost when substituting the topic-matched query terms for their corresponding topic name or when adding their alternative labels. Second, it was found that the two conditions where materials from the original query were attached to the rewritten query both scored above baseline level. The condition *original&wiki_freq&sim&tfidf_selective*, which maintained solely the *more-frequent* original terms that were not yet in the query, significantly outperformed the baseline ($Z = -2.30, p = 0.02$) at the 0.05 level; for *original&wiki_freq&sim&tfidf*, the difference was significant only at the 0.1 level ($Z = -1.64, p = 0.10$). This finding suggests that keeping the more-frequent original terms was a valuable improvement of the proposed Wikipedia algorithm *wiki_freq&sim&tfidf*, which boosted performance above baseline level.

Second, we compared the three different weighting schemes among each other: that without frequency information (only idf); with frequency information in the form of tf-idf merged with our Wikipedia-based frequency count; and tf-idf, frequency and similarity together. To assess the added value of including any frequency information in the weighting procedure, we took *wiki_idf* as a baseline. Adding frequency as well as coherence information, as displayed by *wiki_freq&sim&tfidf*, did not significantly alter performance ($Z = -0.71, p = 0.48$). However, adding frequency information to the weight, i.e., *wiki_freq&tfidf*, improved performance, and this effect was significant at the 0.1 level ($Z = -1.66, p = 0.10$). These findings suggest that frequency information was a useful addition to the tf-idf term weight, but coherence was not. The difference between *wiki_freq&tfidf* and *wiki_freq&sim&tfidf* itself, with the former displaying the highest mean, was significant at the 0.1 level ($Z = -1.73, p = 0.08$). Thus, it can be stated with fair confidence that our Wikipedia algorithm that used only frequency information was more successful than the technique that took into account similarity scores as well. Differently put, including in the weighting scheme the proposed coherence measure, i.e., the averaged inlink similarity of the term‘s corresponding topic with the other detected query topics (see 3.1.5), harmed performance.

Finally, to assess the benefit of the different form selections tried, we contrasted each of the variations on *wiki_freq&sim&tfidf* with their base form. The pagename and alternative label conditions, which were already mentioned to display under-baseline means, trivially scored lower than their base form as well. For both *wiki_freq&sim&tfidf_altlabels* ($Z = -2.14, p = 0.03$) and *wiki_freq&sim&tfidf_pagename* ($Z = -2.32, p = 0.02$), the underperformance against *wiki_freq&sim&tfidf* was significant. Thus, both including the alternative labels and substituting the query terms for their corresponding page title harmed performance. Second, appending to the rewritten query (all or a selection of the) terms from the original query, as was displayed by conditions *original&wiki_freq&sim&tfidf* ($Z = -1.19, p = 0.23$) and *original&wiki_freq&sim&tfidf_selective* ($Z = -0.71, p = 0.48$) was found to not statistically affect performance. However, the latter did display a higher mean than its parent condition. Although not statistically significant ($Z = -0.71, p = 0.48$), this suggests that allowing the more-frequent terms from the original query into the rewritten query can improve results. The difference between the selective and the non-selective method of adding the original query terms was significant at the 0.1 level ($Z = -1.90, p = 0.06$), which suggests that keeping from the original query only the more-frequent terms works gives better results than adding them all.

LDA- based conditions Both LDA techniques differed significantly from the tf-idf speech baseline. That is, both *LDA_query* ($Z = -3.91, p = 0.00$) and *LDA_vector* ($Z = -4.42, p =$

.00) lead to a *decrease* in performance as compared to taking the unaltered speech-recognized documents as input materials. Of the two conditions, the query-based one displayed the highest mean. The difference between *LDA_query* and *LDA_vector* itself was significant at the 0.1 level ($Z = -1.92, p = 0.06$), suggesting that our method of using LDA for weighting query terms was more successful in finding related documents than our method based on the topic probability vector.

5.1.2 Teletext queries

Overall performance and differences

The average nDCG across queries for each experimental condition and their deviations are displayed in table 5.4. The baseline performance for Teletext was 0.82 and average nDCG scores ranged from 0.21 to 0.82.

Looking at the overall means of the Wikipedia-based experimental conditions, two observations are striking. First, as could be expected, that they are higher than the means in the speech counterparts. This demonstrates that the similarity search task was less problematic on regular text than on speech-recognized text. Second, that the effect of the Wikipedia-based experimental conditions on the results was clearly smaller than in the speech findings. The results of the query rewriting methods were all slightly below or around baseline-level. This suggests that our Wikipedia methods did not greatly affect performance in the case of regular text.

Observing the individual means, the highest-scoring Wikipedia-based conditions were *wiki_freq&sim_plusfreqselective* and *wiki_freq*, similarly to the findings in the speech queries; however, the scores differed from the baseline by mere fractions of a percentage. Also, similarly to the speech results, both LDA techniques displayed lower means than both the baseline and the Wikipedia-based techniques.

A one-way repeated measures ANOVA was used to test whether the differences within the Teletext results were statistically reliable and post-hoc tests were justified. The assumption of sphericity was not met, as indicated by Mauchly's test ($\chi^2(54) = 303.27, p = 0.00$), i.e., the variances of the differences between conditions could not be assumed equal; therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.41$). Also, the data within each condition was not always normally distributed. However, as this has been claimed to have only a minor effect of the chance of making type 1 errors [36], we accepted this violated assumption. The test showed that nDCG values differed significantly across conditions ($F(4.07, 109.91) = 41.17, p = 0.00$). Therefore, a series of pairwise Wilcoxon signed rank tests was carried out to locate the differences between conditions.

Differences between conditions

A series of Wilcoxon signed rank tests showed that *none* of the Wikipedia-based conditions significantly differed from the baseline condition of speech-recognized input documents and tf-idf weighting. This indicates that for regular input documents, the tried methods of term selection and weighting did not have any effect on retrieval.

However, the condition *original_tfidf_freqselective* did significantly differ from the baseline ($Z = -2.81, p = 0.01$) as a *less successful* condition. This shows that retrieval did not benefit from naively reducing the Teletext documents to their more-frequent terms and boosting the tf-idf weights by each term's frequency of occurrence.

Of the four variations on *wiki_freq&sim&tfidf* that were tried, *wiki_freq&sim&tfidf_altlabels* ($Z = -0.37, p = 0.72$) and *original&wiki_freq&sim&tfidf* ($Z = -1.11, p = 0.27$), although not significantly, scored (slightly) *above* their parent condition. This is in contrast with the

Condition	Mean	Standard deviation
original_tfidf	0.818	0.239
original_tfidf_freqselective	0.732	0.329
wiki_idf	0.811	0.210
wiki_freq&tfidf	0.819	0.220
wiki_freq&sim&tfidf	0.796	0.242
wiki_freq&sim&tfidf_pagename	0.720	0.267
wiki_freq&sim&tfidf_altlabels	0.807	0.206
original&wiki_freq&sim&tfidf	0.814	0.227
original&wiki_freq&sim&tfidf_selective	0.817	0.234
LDA_query	0.453	0.309
LDA_vector	0.205	0.165

Table 5.4: Average nDCG and standard deviation for each Teletext condition

findings for the speech queries. Finally, inclusion of (a selection of) the original query terms in conditions *original&wiki_freq&sim&tfidf* and *original&wiki_freq&sim&tfidf_selective* also improved over the mean found in *wiki_freq&sim&tfidf*, but this effect was not statistically reliable ($Z = -1.11, p = 0.27$, and $Z = -1.27, p = 0.21$, respectively). Again, this is not similar to the findings in the speech queries. These differences suggest that the success of the Teletext and speech queries was affected by the term variations in different ways.

Similarly to the findings for the speech queries, the LDA conditions *LDA_vector* ($Z = -4.70, p = .00$) and *LDA_query* ($Z = -4.36, p = .00$) both performed significantly below baseline level. Within LDA, the query condition led to significantly higher scores than the vector condition *LDA_vector* ($Z = -3.48, p = 0.00$).

5.2 Validation study

This section describes the results of the survey that was set up to test the validity of our evaluation method. The latter relied upon two main assumptions: first, that the formulated protocols allowed for little interpretation differences ; and second, that if a document satisfied the protocol formulated for a given query document for one of the three relevance levels, it would be *to the same extent* perceived as thematically *related* to the query document and *suitable* as a link target.

Therefore, the analysis consisted of the following tests:

- pairwise correlation tests between the factors protocol, relatedness and linkworthiness, assessing the general relationship between the factors;
- pairwise inter-rater agreement tests between rater's assessment on the one hand and respondents' judgments of (either) relatedness and linkworthiness on the other, assessing the degree of substitutability of the concepts of similarity and linkworthiness by the protocol;
- inter-rater agreement test between rater's and respondents' relevance assessment, measuring the inter-rater agreement and the extent to which the protocols were sensitive to interpretation;
- disagreement analysis for differences between rater's relevance assessment on the one hand and (either one of) respondents' relevance assessment, similarity judgment and linkworthiness judgment on the other hand. This addressed whether the protocols could be stated to be conservative or rather tolerant in assessing relatedness, similarity and linkworthiness.

A Spearman rank correlation test between respondents' pairs of judgments for every two out of the factors assessment, relatedness and linkworthiness on the queries showed that there were highly significant ($p < 0.01$) correlations between every combination of these three factors. The highest correlation ($\rho = 0.77$) was found between perceived relatedness and linkworthiness, the lowest (0.63) between the protocol and linkworthiness; these values fall within ranges that have been reported as *moderate* and *high* correlations [15].

With regard to the inter-rater agreement tests, because of the difficult interpretability of Cohen's kappa [18] for weighted values, we added a transformed version of the dataset that disregarded the graded nature of the assessment results. That is, we transformed all results to a two-level scale, mapping values 1 and 2 onto a value 1. Thus, a value *0* indicated *not relevant similar linkworthy*, and a value *1* indicated *(at least somewhat) relevant similar linkworthy*.

Results of the inter-rater agreement test between rater's relevance assessment and each individual respondent's relevance assessment are displayed in table 5.5. Values with a p value above 0.05 were left out, as they indicate a high chance of having resulted from chance [51]. Note that κ values always lie between 0 and 1. Besides the absolute agreement on the rescaled 2-categorical data, the table also displays the *weighted kappa* of the original three-scale data as a reference in order to judge the degree of information loss from rescaling. Weighted agreement adds less to the score as the judged categories in a data pair are further apart. We chose to weight agreement linearly to their categorical distance. For example, a disagreement between assessment scores 0 and 2 was given twice the weight of a disagreement between 1 and 2. Unweighted κ values varied between 0.53 and 1, or, in terms of a commonly cited scale [51], between qualitative levels *moderate* and *perfect*. The average level of agreement on the protocol-based assessment across participants was 0.75 or *substantial agreement*. The weighted kappas on average did not differ greatly from the unweighted counterparts, which suggests that rescaling did not to a large extent modify the data.

Respondent ID	nr of responses	two-scaled		three-scaled	
		κ	p	κ	p
1	1	-	-	-	-
2	12	0.86	0	0.83	0
3	4	-	-	1	0.05
4	11	0.53	0.03	1	0
5	3	1	0.01	-	-
6	9	1	0	1	0
7	24	0.75	0	0.91	0
8	71	0.64	0	0.77	0
9	14	0.78	0	0.71	0.01
10	25	0.69	0	0.84	0
11	5	-	-	-	-
12	17	0.91	0	1	0
13	16	0.6	0	0.88	0
14	13	0.6	0	0.53	0.05
15	12	0.63	0	0.67	0.01
AVERAGE		0.75		0.85	

Table 5.5: Interrater agreement between assessment by rater and assessment by respondents

Results for inter-rater agreement between rater's set of relevance assessments and each individual respondent's set of judgments for relatedness and linkworthiness are shown in table 5.6. Both the unweighted kappas on the 2-categorical data and the linearly weighted

kappas on the original 3-categorical data are displayed. Values with a p above 0.05 were left out. Looking at the rescaled data, for relatedness, κ values varied between 0.41 and 0.85, spanning the ranges *moderate agreement, substantial agreement* and *almost perfect agreement*, with an average of 0.65 or *substantial agreement*. For linkworthiness, the κ range was almost equal, between 0.39 and 0.85, however with the average just below the lower bound of *substantial agreement*. With regard to the original three-scale data, for both relatedness and linkwortiness, average weighted kappa differed from its unweighted counterpart by just a few percentages and in a non-systematic way. This illustrates that rescaling did not greatly affect the data.

Respondent ID	nr of responses	Relatedness two-scaled		Linkworthiness two-scaled		Relatedness three-scaled		Linkworthiness three-scaled	
		κ	p	κ	p	κ	p	κ	p
1	1	-	-	-	-	-	-		
2	12	-	-	0.5	0.05	0.44	0.01	0.53	0.01
3	4	-	-	-	-	-	-	-	-
4	11	-	-	-	-	-	-	-	-
5	3	-	-	-	-	-	-	-	-
6	9	0.77	0.02	0.77	0.02	0.61	0.01	0.46	0.02
7	24	0.47	0.02	0.43	0.02	0.61	0	0.63	0
8	71	0.72	0	0.39	0	0.78	0	0.49	0
9	14	0.85	0	0.85	0	0.78	0	0.78	0
10	25	0.41	0.04	0.58	0	0.58	0	0.75	0
11	5	-	-	-	-	0.55	0.03	0.71	0.03
12	17	-	-	0.66	0	0.59	0	0.74	0
13	16	-	-	-	-	0.56	0	-	-
14	13	-	-	0.53	0.05	0.57	0.01	0.55	0.01
15	12	0.67	0.02	0.67	0.02	0.58	0.01	0.73	0
AVERAGE		0.65		0.60		0.60		0.64	

Table 5.6: Interrater agreement between assessment by rater and relatedness linkworthiness by respondents

Finally, based on the binary assessment data, we analyzed which patterns of disagreement were more prevalent. Tables 5.7, 5.8 and 5.9 display the response frequencies and percentages for relevance assessment, relatedness, and linkworthiness, respectively, each broken down by type of assessor (rater or respondent) and assigned score (0 or 1).

		Rater	
		0	1
	0	85	17
Respondents	1	4	131
TOTAL		89	148
	0	96%	89%
Respondents	1	4%	11%

Table 5.7: Agreement and disagreement frequencies in protocol-based assessment

With regard to the protocol, it was more likely that users' responses mapped onto relevance assessment *not relevant (0)* while rater's assessment was *relevant (1)* than the opposite: 4% (4 out of 89) of responses were not in agreement with rater when the latter responded by 0, as compared to 11% (17 out of 148) when rater's assessment was 1. This

suggests that, in the cases where rater and a respondent disagreed on relevance assessment, rater had usually assigned a higher relevance score than the respondent.

		Rater	
		0	1
	0	58	22
Respondents	1	31	126
TOTAL		89	148
	0	65%	85%
Respondents	1	35%	15%

Table 5.8: Agreement and disagreement frequencies in relatedness assessment

		Rater	
		0	1
	0	54	22
Respondents	1	35	126
TOTAL		89	148
	0	61%	85%
Respondents	1	39%	15%

Table 5.9: Agreement and disagreement frequencies in linkworthiness assessment

For both perceived *relatedness* and perceived *linkworthiness*, disagreement rates were higher in the cases where rater judgment was 0 than when it was 1, with a difference of 35% as opposed to 15% for relatedness, and 39% as compared to 15% for linkworthiness. This indicates that respondents more frequently judged a document to be similar to the input document or worthy of being linked to by the input document when rater did not, than the other way around.

Chapter 6

Discussion

6.1 Validation study

The survey that assessed the validity of our evaluation approach indicated that *agreement* between rater's and users' document assessment in terms of *relevant* or *not relevant* to a given input document was 0.75, which is regarded as *substantial*. This suggests that the formulated protocols were not overly sensitive to interpretation differences, and that we succeeded in developing an objective evaluation strategy.

We found respondents' protocol-based assessment and similarity and linkworthiness judgement on individual questions to strongly *correlate*. These findings suggest that the notions of similarity and linkworthiness are highly linked among themselves as well as to the defined protocols, in the sense that high perceived values and low perceived values tend to co-occur in each of the three variables.

However, although closely linked, these variables were not found to be mutually *substitutable*, as inter-rater agreement between rater's assessment on the one hand and respondents' perceived document relatedness on the other hand was merely 0.65. The same finding held for respondents' perceived linkworthiness of the target document, which had an agreement score of 0.60 with rater's protocol-based assessment. This suggests that it is preferable to assess the correspondence between users' perception of these factors before putting the protocol to practice, rather than to assume them equal.

With regard to the protocol-based assessment, cases of disagreement were more prevalent when rater judged a document to be relevant than when rater assessed a document to be irrelevant. This suggests that rater handled the protocol in a liberal rather than a conservative fashion, and as such, was biased towards judgments of relevance more than irrelevance. However, the disagreement frequencies were low, as supported also by the high level of inter-rater agreement on this variable, and therefore we can assume the impact of a possible bias to be sufficiently small.

For similarity and linkworthiness judgments, respondents disagreed more when rater had *dismissed* a document as being relevant to the input document than when rater had credited the document. That is, they judged the documents to be similar to the input document or worthy of being linked to more frequently than rater did based on his protocol. These findings indicate that the protocols were too narrowly formulated to capture perceived relatedness and appropriateness as linking material. Put in a broader perspective, these findings suggest that an article can be perceived as related to the input document as well as an interesting link target even if it does not show a large degree of direct topical overlap with the input document.

6.2 Main study

The experimental results show that, in line with the goal, our Wikipedia technique improved similarity search results for speech-recognized materials; even more, it did so to a level that approximated the results of the corresponding flawless Teletext counterparts. This suggests that the Wikipedia anchor vocabulary can successfully be used to select meaningful terms from imperfect ASR transcriptions. Also, term selection based on Wikipedia proved to yield better and more stable results than based on frequency information inherent in the query, indicating the additional value of including an external information source. The Teletext documents remained relatively indifferent under the various query manipulations. Therefore, the Wikipedia technique showed to *effectively* as well as *specifically* address imperfect transcriptions.

The lack of benefit of our technique in Teletext seems to illustrate that query reduction is beneficial in flawed text only. Regular text contains valuable matching materials beyond its most informative nouns, whether in verbs, trivial nouns, adverbs, adjectives or numbers. For example, when reducing *query 1* to the terms "Frankrijk", "auto", "jachtgeweer" and "discotheek", the details contained in the words "nacht", "tien", and "gewond", are lost. Support for this hypothesis was found in the difference in performance of the unboosted Wikipedia condition and the unaltered baseline: reducing the queries to the terms that occurred in the Wikipedia anchor vocabulary *increased* performance for the speech queries, but not for the Teletext queries. This effect is also displayed in the - albeit not statistically significant - ordering in the performance of three Teletext conditions: the baseline condition, the Wikipedia condition with similarity measure, and the appended combination of these two. The more the original query was replaced by Wikipedia terms, the less successful it was. The observations in the speech condition, where the ordering was exactly the opposite, can be explained using the same line of reasoning: while regular text does not benefit from term reduction, noisy text does. The remainder of this discussion primarily focuses on the outcomes for the speech documents, which is our primary interest.

With respect to the different weighting schemes that we tried, with *idf* weighting as a baseline, we found that adding to the weighting scheme frequency information in the form of $tf-idf \times frequency$ led to better search results than mere idf weighting. However, augmenting the $tf-idf \times frequency$ weighting scheme with semantic coherence information, i.e., $tf-idf \times frequency \times coherence$, harmed retrieval. We conclude from this three things.

First, that the used semantic coherence measure of a term in the total query did not adequately predict its importance. Recall that semantic coherence of a term in the query was modelled as its average inlink similarity measure, based on the Google distance, with each of the detected concepts in the query. Taking into account the domain of the study, we explain this finding as being due to *cross-theme* measures that are inevitably reflected in this measure, besides the *within-theme* relationships that are of interest. News items are frequently composed of several thematic dimensions that refer to, for example, the location of the event, the actor(s), the action, and other circumstantial aspects. Within a thematic dimension, conceptual connections are not unlikely to be reflected in a general-purpose encyclopedia; for example, the concept *politician* is likely to be related to concepts such as *campaigns*, *laws* and *elections*. However, in the case of multiple themes, the coherence score of a term was affected by the (coincidental) factor of how many of the Wikipedia-matched query terms belonged to the same theme as compared to the other themes. As an example, suppose that an input text about the fine for study delays has as its Wikipedia-connected terms *school, study, student, fine, euro*. Then, because of the accidental majority of study-related terms over fine-related terms, the two last terms will likely be given a lower weight than the three first, even though they constitute a coherent thematic dimension. Thus, we feel that a more suited coherence measure would credit the presence of *any* term in the query, whether ambiguous or not, that is semantically close to a given term. This

explains the increase in performance after including the missing more-frequent terms from the original query in this weighting scheme, which corrected exactly for the problem that important terms had been down-weighted or filtered out due to low coherence values.

Second, that frequency weighting helped reduce the effect of the ASR errors even more than did the Wikipedia-based term selection alone. In the frequency-based condition, there was no filtering stage involved, therefore wrongly recognized terms still appeared in the final query. However, since a specific error generally occurred in a query only once, erroneous terms did have the lowest possible score in the query, diminishing their effect in document matching. This effect manifested itself not only in the frequency-based *weighting* variant, which was better than the mere idf weighting variant, but also in the *term selection* variant where the *more-frequent* unused terms from the original query were appended to the Wikipedia-based query, which was more successful than its parent condition. The high-frequent terms apparently provided better matching results than the unfiltered set of terms, supporting the hypothesized strength of frequency information in ASR similarity search. Based on these findings, we expect that a frequency-based algorithm combined with the overlooked more-frequent terms, a condition that was lacking in the current study, will be most powerful.

As an intermezzo, an important question to address is whether the specific frequency count used in this study was an influential factor in the success. Rather than the raw count of occurrences of the term in the query, we factorized this by the number of recognized topics in Wikipedia of which it was a label. However, inspecting five random queries, we noticed that a query label *rarely* belonged to two different topics found in the query: only 4 out of 228 terms in total had been added because of this rule. Thus, this barely made any difference. Therefore, the number of times a term appears in the query seems to have contributed most to the frequency feature.

Third, and most importantly, following from the finding that Wikipedia information did not notably contribute to frequency weighting, it was inferred that the beneficial effect of applying Wikipedia was due to the term selection rather than the weighting process. The Wikipedia terms were found to be better indexing materials than the error-prone speech terms, which suggests that Wikipedia labels form a representative repository of semantically meaningful terms. However, due to the harmful effect of the used coherence measure in the weighting method, as opposed to the beneficial effect of the use of query-inherent frequency information, Wikipedia could not be used to improve weighting.

As for the lexical variants we tried, let's first sum up the results of the conditions that merged the Wikipedia-matched terms with (some of) the original query terms, which were already pointed at in different parts of this discussion. It was found, first, that adding the original query terms in a non-selected fashion, in spite of the lower weight assigned to these terms, harmed performance. This once more indicated that the speech-recognized terms were inherently poor indexing materials, and suggests that elimination was more effective than weight reduction to counter the effect of semantic distracters. However, adding only those overlooked terms that *repeatedly* occurred in the original query did increase the scores. This suggests that frequency, which is generally known as a reliable indicator of the semantic quality of a term in a document, is also a reliable indicator of whether a term was correctly interpreted by a speech recognizer.

With regard to the form variations that included the topic name and the top-5 most used topic labels, these manipulations both led to decreased performance. Note that these variants were -rather unfortunately- combined with the *expected* most successful weighting scheme of frequency and similarity (hereafter called the *parent condition*), rather than the *actual* most successful weighting scheme of frequency only. However, since the difference in performance between the two weighting algorithms was relatively small as compared to that between the parent condition and each of its two term variation conditions, these fin-

dings can with some confidence be generalized beyond the weighting method. Examining why these term variations failed, we point at three causes.

First, as a feature inherent in Wikipedia, the added terms were not always common and synonymic in general use to the original. For the alternative label condition, we observed that the most-frequent alternative label was generally a valid substitute for the concept, e.g., "staat [state]" for "land [country]", "regering [government]" for "kabinet [ministry]", and "ontploffing [blast]" for "explosie [explosion]". However, further down in the top-5 the alternative labels included forms such as "H20" for water, "Aram-damascus" for "Damascus", and "klokuren [clock hours]" for "uren [hours]". For the pagename condition, similarly, some pagenames had a title that is not commonly used, such as "Verenigd Koninkrijk [United Kingdom]", which is often referred to as "Engeland [England]" or "Groot-Brittannië [Great Britain]".

Since these uncommon terms must have had a low frequency in the corpus, they likely affected scoring more than was intended by the weight assigned to them. In the alternative label condition, it was aimed to keep the overall weight of a certain concept constant by equally dividing the weight of the covering concept among the added alternative labels. However, this intuition was probably incorrect, since it implicitly assumed document frequency for each label to be equal. Consequently, the intended effect of the term weights, i.e., to correct for recognition errors by assigning low weights and to boost correct terms by higher weights, was distorted.

Second, as a factor beyond Wikipedia, these term variations exposed the effect of wrong disambiguations. For example, in *query 8*, which was about the residents of an asbestos-affected neighborhood, the term "bewoners [residents]" was mapped onto the concept "Big Brother (televisieprogramma) [Big Brother (television show]" rather than onto the intended concept of *house*, i.e., "huis (woning) [house (residence)]". Clearly, substituting the original term for the term "Big Brother" or any of its alternative labels leads away from the document's semantics.

Third, and similarly, whenever there was no exact match between a term and the set of available topics, this was made visible by the page name or the alternative labels. For example, the term "kant [side]" in the context of "de kant van de regering [the side of the government]", by lack of the more abstract notion of political affiliation, was projected onto the topic of the famous philosopher "Immanuel Kant". Similarly, the label "defensie" in query 4 was intended to refer to the Syrian defense ministry but could only be projected onto one of the following concepts: *Defensie (landsverdediging) [Defense(land protection)], Defensie van België [Defense of Belgium], Defensie van Nederland, Defensie van Noorwegen, Defensie van Oostenrijk, Ministerie van Defensie (Verenigd Koninkrijk) [Defense Ministry (UK)], or* United States Department of Defense (DOD). The effect of such unfit best matches is equal to the wrongly disambiguated terms: whereas the term "kant" in itself has some chance of finding its counterpart in another article on the Syrian conflict, the term "Immanuel" certainly does not. Hence, the label extension or substitution distracted from the actual meaning. However, the difference with the previous point addressed is that this flaw could not be ascribed to the disambiguation procedure, but rather to the term selection technique, which did not take into account the context and hence was *deterministic*.

The reported findings led us to question the added value of using Wikipedia for the task at hand, rather than some unstructured document collection. After all, our main ways of deployment of Wikipedia structure - term rewriting and coherence-based weighting - were not helpful to the search task, and this was at least partly due to fundamental flaws in disambiguation and to the level of detail of topic coverage. Also, the Wikipedia-inspired frequency measure used barely differed from a regular occurrence count. Therefore, the only Wikipedia-aided task that was found to be valuable in query reformulation was the *selection* of terms that reflected the document semantics. Wikipedia's role in this was

twofold: to provide the vocabulary, and to optimally segment the matched sequences into conceptual units. However, using an unstructured background, co-occurrence statistics might be able to underpin these functionalities equally well as Wikipedia did.

Based on the present study, the LDA- based techniques seemed incapable as a substitute for the original input in finding related documents. However, this impression might be due to the evaluation method used. Within the set of returned documents for a given query, which were examined and assessed one by one by the rater, it was generally possible to describe the common denominator in a small number of topical aspects, each of them with a clear, albeit sometimes abstract, link to the original document. For example, for *query 10* about the power cut in India, results were either about bad conditions in underdeveloped countries, e.g., in factories in Bangladesh and Ukraine, or about energy- and power-related issues. For *query 11*, which reported on house-searches of drug criminals in the Noord-Brabant province and the many stolen goods that were found, results dealt with either or a combination of the theme components of theft, crime and drugs. Most documents failed to display the prescribed mixture of topics necessary to be considered relevant to the query, and therefore results were considerably under baseline level. However, unlike in the baseline speech condition, none of the returned documents were found to be clear semantic outliers that had been accidentally matched on some term(s); instead, they seemed to *make sense*. Also, it cannot be excluded that a higher level of abstraction from the original document might result in interesting linking material to a user. After all, the survey results suggested that actual users have a more liberal opinion about documents' potential as a link destination than was reflected by the protocols. Therefore, we are not certain that a user would prefer as links the documents generated by the baseline query over those generated by the LDA vectors or the LDA-based rewritten query.

Chapter 7

Conclusion

In this chapter we summarize our findings related to the formulated research questions.

This study aimed to alter speech-recognized news texts for finding related documents to link to. Two external sources were explored. First, we used Wikipedia as a vocabulary to select terms from, and Wikipedia link statistics as an indicator of semantic relatedness to weight every term. Second, we used a background corpus of news documents to create a topic model using Latent Dirichlet Allocation. This topic model was applied directly, as an alternative vector representation, as well as indirectly, to weight terms selected from the background vocabulary to become the new input document.

We succeeded in improving similarity search results with this approach. In fact, performance approached that of the Teletext input documents, which represent flawless transcriptions. However, weighting the terms by means of the semantic coherence score did not contribute to performance. Also, we found that overruling the Wikipedia algorithm by adding all unused more-frequent terms from the original query, whether matched to a Wikipedia topic or not, boosted performance. From these findings, we infer three things. First, that Wikipedia page labels form a representative repository of semantically meaningful terms. Second, that reducing a speech-recognized document to its semantically meaningful terms, while maintaining frequency information, leads to better similarity search results than its unaltered form. Third, that frequency is an important cue in finding reliable, meaningful terms.

We also tried alternative Wikipedia-based term representations, but without any success. These variants showed off the effects of flaws in the disambiguation procedure and mismatches between the intended and the best-matching topic in the encyclopedia. Combined with the lack of benefit of the semantic coherence measure, which is also inherently *wikipeadic*, this raises the question of the added value of Wikipedia, or a structured corpus in general, for the task at hand. Topic detection and segmentation might be performed just as well by using, for example, a non-structured news vocabulary.

Finally, we deployed Latent Dirichlet Allocation to weight the query terms by their topically motivated likelihood of being generated by the query‘s document model, and second, to translate the document to the strongly reduced space of topic likelihoods. Both techniques harmed performance. However, in a survey we assessed the correspondence between our evaluation criteria and real users‘ perception of similarity and link-suitability. The results indicated that users more often considered a document similar to the input document and suitable as a link destination than our protocols expressed. Also, qualitatively, the returned document sets for each of the two LDA applications seemed to be coherent, with some, albeit sometimes abstract, link to the input document. Therefore, this study cannot exclude the potential of LDA in the task of linking speech-recognized documents.

Chapter 8

Future Work

In this chapter, we briefly describe the shortcomings of our methodology that we discovered after running the experiments and propose some different approaches for future studies that wish to build onto our experiments. Second, we reflect on what we feel are the next steps to take to solving the given task.

With respect to the query reformulation methods, we suggest four aspects of amelioration, which all concern the Wikipedia-based techniques. First, when experimenting with term variants, to take the frequency-augmented tf-idf weighting as a basis rather than frequency-and-similarity-augmented tf-idf, as the former was not only more parsimonious, but also more successful. Second, as we found its effect to be negligible, to leave out the redistributing feature in calculating term frequencies and, thus, to keep the correspondence between topics and query phrases one-to-one. Third, to include a condition that combines frequency-augmented weighting (i.e., based on term occurrence counts and tf-idf) with as term selection not only the Wikipedia-matched terms, but the more frequent original terms as well. Based on the findings in the current study, we expect this to be a high-performing condition. Finally, we recommend to experiment with different applications of the relatedness measure, using heuristic rules to relate the number of semantically close terms in the query for a given term to a suitable weight. Alternatively, we propose to optimize the similarity measure. The Wikiminer software supports training a classifier for this purpose, which uses not only in-link but also out-link statistics and an alternative tf-idf based measure [54]. This allows the researcher to state with more certainty whether the Wikipedia-based relatedness measure, and Wikipedia in general, has any particular beneficial effect in this problem setting.

With regard to the evaluation method, we suggest to use a more user-oriented approach, as our protocol revealed a certain discrepancy between our evaluation and the user perspective. To mimic a realistic setting, users can be presented with the video news fragment itself and two alternative lists of proposed related documents, each taken as a result of a different experimental condition, and asked to state their preference. Differently put, we propose to shift the research towards the field of research on *recommendation* rather than pure information retrieval. Alternatively, we suggest, ideally, to repeatedly experimentally validate any relevance protocols until a sufficient level of agreement is reached; or at least to thoroughly negotiate them among several people before putting them to practice.

Anticipating future work on the given use case, we feel that our study has pointed at two major knowledge needs. First, we need to know users' generalized preferences in the envisaged application. With what purpose do they use the MES search platform? Do they prefer links to directly related or more distantly related items, or a combination? Do they usually click through because of a specific information need or do they like to be *surprised* by the suggested reads? These questions need answers in order to judge the benefits of our proposed techniques, particularly the ones based on LDA, which seem to increase the

surprise factor in link suggestion. Second, we need to investigate the added value of a (partly) structured information source such as Wikipedia for filtering ASR queries over an unstructured document corpus. Are corpus-based methods of term filtering, e.g., based on term co-occurrence statistics, readily available? Are such methods sufficiently able to detect semantic outliers? In case the answer to these questions is positive, then the computational cost of including a more structured source cannot be justified.

References

[1]

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer Berlin Heidelberg, 2007.

[3] L. Azzopardi, M. Girolami, and C. J. Van Rijsbergen. Topic based language models for ad hoc information retrieval. In *IEEE International Joint Conference on Neural Networks*, volume 4, pages 3281–3286, 2004.

[4] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 31–40. ACM, 2010.

[5] Michael Bendersky, Donald Metzler, and W. Bruce Croft. Effective query formulation with multiple information sources. *WSDM '12*, 2012.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[7] Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM review*, 41(2):335–362, 1999.

[8] M.W. Berry, editor. *Survey of Text Mining I: Clustering, Classification, and Retrieval*. Springer, 2004.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[10] T. Bogers and A. Van den Bosch. Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 141–144. ACM, 2007.

[11] Richard Boulton. Dutch stemming algorithm. `http://snowball.tartarus.org/algorithms/dutch/stemmer.html`.

[12] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *TPDL 2011: International Conference on Theory and Practice of Digital Libraries 2011*, 2011.

[13] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[14] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 78–87. ACM, 2004.

[15] K.G. Calkins. Applied statistics lesson 5: Correlation coefficients. http://www.andrews.edu/~calkins/math/edrm611/edrm05.htm.

[16] H. Ceylan, I. Arapakis, P. Donmez, and M. Lalmas. Automatically embedding newsworthy links to articles. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1502–1506. ACM, 2012.

[17] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[18] J. Cohen. A coefficient of agreement for nominal scales. educational and psychological measurement. *Educational and psychological measurement*, 20(1):37–46, 1960.

[19] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, volume 7, pages 708–716, 2007.

[20] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 181–190. ACM, 2010.

[21] Apache Software Foundation. Class dutchanalyzer. http://lucene.apache.org/core/4_2_1/analyzers-common/org/apache/lucene/analysis/nl/DutchAnalyzer.html.

[22] Apache Software Foundation. Class standardtokenizer. https://lucene.apache.org/core/4_2_1/analyzers-common/org/apache/lucene/analysis/standard/StandardTokenizer.html.

[23] Apache Software Foundation. Class tfidfsimilarity. http://lucene.apache.org/core/4_2_1/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html.

[24] The Wikipedia Foundation. Wikipedia: Manual of stylelinking. http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking.

[25] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, January 2007.

[26] J. S. Garofolo, C. G. Auzanne, and E. M. Voorhees. The trec spoken document retrieval track: A success story. *NIST SPECIAL PUBLICATION SP*, 246:107–130, 2000.

[27] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479. ACM, 2004.

[28] G. Heinrich. Parameter estimation for text analysis, 2005. http://www.arbylon.net/publications/text-est.pdf.

[29] D. Ikeda, T. Fujiki, and M. Okumura. Automatically linking news articles to blog entries. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 78–82. AAAI, 2006.

[30] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7, 2007.

[31] J. Kekäläinen and K. Järvelin. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[32] J. Kekäläinen and K. Järvelin. Evaluating information retrieval systems under the challenges of interaction and multidimensional relevance. *Proceedings of the 4th CoLIS Conference*, 2005.

[33] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, and R. Lee. Media meets semantic web: How the bbc uses dbpedia and linked data to make connections. In *The semantic web: research and applications*, pages 723–737. Springer Berlin Heidelberg, 2009.

[34] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 564–571. ACM, 2009.

[35] C. Leacock, G. A. Miller, and M. Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

[36] Lund Research Ltd. One-way anova.

[37] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on World Wide Web*, pages 107–116. ACM, 2005.

[38] F. Menczer. Combining link and content analysis to estimate semantic similarity. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters*, pages 452–453. ACM, 2004.

[39] R. Mihalcea. Using wikipedia for automatic word sense disambiguation. In *HLT-NAACL*, pages 196–203, April 2007.

[40] D. R. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221. ACM, 1999.

[41] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[42] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 2012.

[43] S. Patwardhan and T. Pedersen. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, 2006.

[44] K. Premalatha and A. M. Natarajan. A literature review on document clustering. *Information Technology Journal*, 9(5):993–1002, 2010.

[45] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 239–248. ACM, 2005.

[46] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[47] G. Senay and G. Linares. Confidence measure for speech indexing based on latent dirichlet allocation. *Interspeech*, September.

[48] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM, 1999.

[49] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 1419–1424. AAAI, July 2006.

[50] G. G. Tiao and I. Cuttman. The inverted dirichlet distribution with applications. *Journal of the American Statistical Association*, 60(311):793–805, September 1965.

[51] A. J. Viera and J. M. Garrett. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[52] P. Vossen, K. Hofmann, M. de Rijke, E. T. K. Sang, and K. Deschacht. The cornetto database: Architecture and user-scenarios. In *Proceedings DIR*, pages 89–96. 2007.

[53] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[54] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30. AAAI Press, Chicago, USA, 2008.

[55] H. Zaragoza, D. Hiemstra, and M. Tipping. Bayesian extension to the language model for ad hoc information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–9. ACM, 2003.

[56] C Zhai. *Risk minimization and language modeling in text retrieval*. PhD thesis, University of Massachusetts, Amherst, 2002.

Appendix A

Queries

A.1 *original_tfidf*

A.1.1 Speech

1. bij een schietpartij uh bij discotheken frankrijk zijn afgelopen nacht tien mensen gewond geraakt het dader was boos omdat die de discotheek vlakbij de plaats kan brengen niet in mochten daarop haalde hij jachtgeweren uit zo'n auto beschoten zo buiten bij de discotheek op binnen na de schietpartij ging de man er vandoor is waar leven aangehouden aan

2. ze brengen weer mensen te veel slecht nieuws dan kunnen we maar ja je in een uh nee ik je bent er erg veel ouders vaarwel bang voor maar nu heeft een het is een zegen dat aangegeven dat overheid mag een boete opleggen als ze te lang door over je studie duizenden euro's bovenop het collegegeld van verstaan studentenorganisaties vingen en een pot bij de politiek en dus probeerden ze het bij de rechter studenten moeten voortaan op strikte te krijgen jaar spelen als een meer dan een jaar vertraging oplopen betalen ze moeten ook studenten die actief zijn voor een studentenorganisatie moeten bitter het ziet er al heel wat jan maar op het moment dat je inzetten en je weet doet nu even een fulltime zit jaren ik ja en je zet je een jaar voltooid in voor de studentenbelangen dat beloond wordt zegt maar met een boete van drieduizend één zestig het bestuurder is niet uh enige ook studenten die ziek zijn geweest op een verkeerde studie hebben gekozen krijgen moeten daarom stapte ze naar de rechter wij delen niet het oordeel van die zo kan de thans bestuderen de student aanspraak kan maken op het gedurende een onbeperkt aantal coulissen jaren volgen van onderwijs tegen een vooraf vastgesteld werd in de collegegeld dat afgezien van en inflatiecorrectie blijft gelden tot al het diploma is behaald met zijn je heel erg teleurgesteld dat waren dat we dat we de spelregels tijdens het spelen van het spel van het mag worden in het volle staat een uitzondering voor een kleine groep alleen tien altijd studenten die voor een tweede waarin twee werd het hele zijn begonnen aan een lange studie over geen te ja ik ben ook lid wil het is niet in te een studie wijsbegeerte die zes jaar duurt en dan zit ik hier met indianen ziet samenraapsel termen moeten gaan betalen en dat is nu stapt een heel hebben haar alle andere studenten die te lang over hun studie doen heb ik geen reden om blij te zijn zij kunnen rekening van meer dan drie duizend euro verwachten ja ik ben een zelf ook gaan betalen en dat die boete komt er dus toch maar de vraag wel voor hoelang want een eerste partijen schrijven in hun verkiezingsprogramma's dat ze die in lang studeer boete volgend jaar alweer de terug naar bij ons gaat het om miljarden

3. in die een aanzoek zijn het zware tijden want vind maar ze werken de jongeren van vooral turkse marokkaanse of surinaamse afkomst blijkt dat het helemaal moeilijk het is een cijfers ruim twintig procent van de jongeren had begin vorig jaar geen baan het opgelopen tot bijna dertig procent dus in op de drie als verschil van de autochtone jongeren zich bijna tien procent zonder werk hoe heb je een goed opleiding het is echt heel moeilijk om iets te maar evenals laten onvoldoende aansnijdt als je met een uh onze mensen toch weer op van junior zelfs hier is een van de tientallen afwijzingen die wärtsilä afgelopen maanden heeft ontvangen orgie aan zo'n man ons haar opleiding arbeidsrecht af en nu zoekt ze een baan als advocaat ze is vier keer uitgenodigd voor een gesprek maar tot nu toe niet aangenomen dit idee dat het met je achtergrond smaak heeft toen ik aan mijn op hol de taal niet maar naarmate de tijd voor het begin ik wel steeds meer op de vraag hoe het kan zijn dat beiden in heel makkelijk aan van een ander als er een uh krijgt ze vaak te horen dat ze niet bij het controlepost werkloosheid is onder nederlanders met een turkse marokkaanse of surinaamse achtergrond altijd zo hoog geweest dan bij autochtonen maar recente stijging is opvallend zegt vooral een allochtone jongeren als eerste erbuiten bood vallen omdat ze el vakken dan nieuw bij aan hebben als eerste rode loper uit handen ze heel vaak tijdelijke banen en wat was er ook meespeelt is dat heel veel uh jongeren dan was er geen doorleren dat een crisis was en die komen nu ook allemaal op de markt en werkgevers en de wat te kiezen nu er zo voor werkzoekenden zijn en dan kiezen ze toch voor autochtone er werd het idee dat we werden geven ze vooral als er zelf al gauw werknemers in dienst en daar helemaal ik dat er een niet zo ook uit maar me dat wel want zijn dat klanten overal nemen ze hij liet een allochtone werknemers willen zien dat ze daar voor bepaalde functies die van autochtone werk is werkgevers dit ook zo zien weten we niet in willen vandaag niet reageren kunnen we na twee naast rust en china heeft de heel zwaar bijkwamen opgepakt maar blijft solliciteren en ook ooit haar eigen advocatenkantoor te hebben natuurlijk op de vondst kabinet hier zou moeten helpen het ministerie van sociale zaken zegt we doen niet seaport voor bepaalde groepen alle jongeren moeten vooral een richting kiezen waarin ook banen zijn als je denkt dat een werkgever discrimineert moet je dat melden soms uh

4. het is een aanslag in het hart van het syrische bewind stond het oog van president als dat tijdens een vergadering van het syrische veiligheidskabinet ontploften om die aan zeker twee sleutelfiguren van het regime het leven heeft gekost minister van defensie en de onderminister van defensie die laatste is een zwager van present als valt een van de meest gevreesde figuren vond het syrische regime het is een klap in het gezicht van de president een bom zou er is meer in het kunnen hebben op het syrische regime het irissen vol dan zestien maanden van oorlogsgeweld zwarte rook boven damascus vlak na de eerste berichten van de aanslag op het gebouw van de veiligheidsdienst als een vergadering van veiligheidsfunctionarissen vooral ministers en topmilitairen een lijfwacht van een van die functionarissen zou zichzelf hebben opgeblazen een van de plek zelf zijn er niet die werd om in leuk afgegrendeld wel kwamen deze beelden van de straat in de buurt stukje bij beetje begon de syrische staats-tv na de aanslag berichten van andere bronnen de rest als marilyn is oranje was het niet te leveren wordt uit alle torma gezworen wordt dat hij nadien nog wat op attent waarin ze ja abera daar zeer uh in delft waar de yen is daarbij van macbeth aandeel komen hier op afgenomen uit beide weer zou willen terzijde zijn je bereikbaar wel abdallah was rapporteerde diva volgens sommigen is het een lijfwacht van deze minister van defensie die zich heeft opgeblazen defensieminister aïsja werd binnen een paar uur een opvolger aangewezen als je bereikt maar als dat zo kaart maribor twee rudi van asociaal kaart is deze man de zwager van president assad foto werd genomen bij de begrafenis van de oude president als dat in twee duizend shaukat was onderminister van defensie een sleutelfiguur in het veiligheidsapparaat het zeer gevreesd ja dat voor het nieuws van de aanslag maakte correspondent sander verloor deze beelden in damascus maar nu vier dagen achter elkaar gevechten zijn het sideris relatief rustige uit ook is te zien hoeveel mensen met bezittingen aan de wanden zijn mogelijk op zoek naar veilige plekken in de hoofdstad de uitgever er met het vallen van hoorde in dallas uh streuer zal het moeten de enorme slag voor we voor de syrische president zijn nou ja dat is het zeker al was 't maar moreel kijkt ze doen er alles aan om te voorkomen dat het zo lijkt er en met een nieuwe minister van defensie aanstellen en er vooral in de elkaar in willen die ze uitstralen dat eenheid is dat dit het syrische volk alleen maar sterker maakt maar gisteren leek het nog heel ver weg voor de mensen in dan was dus uh de buitenwijken waar zwaar gevolg de wet en dan nu zoiets in het zet een van de mast is ja dat is natuurlijk een zware sla er is het nou zeker dat het om

een zelfmoordaanslag gehad want dat doet toch heel denken aan irak en afghanistan ik vind het best zou er nog altijd heel vreemd ik heb er nog geen verklaring gehoord van iemand op wat ik van daar zou ik probeerde in de buurt van die aanslag te komen dat het natuurlijk niet dat snap ik ook wel maar dicht bij en vrij dicht bij dat gebouw bij de wet staande gehouden door een aantal militairen die is er is maar heel vriendelijk bedoeld keer nou dat hebben gedaan en zijn dat eist gaven ingereden en uh kommil lijken op tientallen meters misschien van dat gebouw waar mensen gewoon doorging met waar ze mee bezig waren auto's die gewoon reed winkels die gewoon open waren 't was compleet surrealistisch en daar gaat het dan allemaal gewoon door en lijkt het wel maar er toch zagen net in je reportage beelden van uh de enorme vluchtelingenstromen de hoorn van de maskers moeten daar dan ja dit die kijken waar nu zwaar gevochten wordt er zelf nu banen praten zijn horen kan de bron opvingen en uh de er zal was jaren daar inzien witte taferelen maar ook in het zet een van de macht is heeft dit toch echt wel wat gedaan om in is nerveus dat merk je gewoon continu die in devon en dan bellen met mensen maar ben je wat doe je hoe maak je dat woord hij wat dat betreft uh banen mensen zitten natuurlijk een van de masters weinig als je er is een explosie woorden heb is ja dat is een even die buitenwijken waar al dagen zo zwaar gevochten voor hoe je nu een explosie dat hier met naar buiten lopen want dat kan dus ook een explosie heel dicht bij zijn daar waar het vierendertig gesneuvelde militairen die hier in dit ziekenhuis worden opgehaald door mensen die niets dan is er een zouden stappen zoals dat van manschappen een middag hele middag na zaterdag met een berg met de beren kwam aan een keer weer een het is dat we het zo waren er maar dat zou willen dat er een in de loop is een heel is een hoe dan de kant van de regering is op alle kanten op positie dagelijks komen mensen familieleden op palen droomt er de kans dat het uh ophoudt cyanide de elke dag op weer kleiner en dat zie je ook achter dit gebouw daar is er zwarte rook ruimte zien we weten dat daar grote stromen vluchtelingen uit die wijk waar vandaan kwamen daar weer burgerzin het lijkt neemt alleen maar toe elke dag is ja dit ziekenhuis smit meer dan duizend bedden is vol schotwonden explosieven onder uh weer aan wat je vroeger al een naspelen en de maar nu weer op andere ooit aan die supermensen zappen maar later aan de zoenen maar zet roma geluiden aan 'm aan naser op alle tal van het zijn de wedstrijd maleiers nog echt een kennis aan nadat aan het eind van het nou daar is de meeste militairen zijn al dood als ze hier aankomen ja het leek zich aan de kant van de regering en natuurlijk het leed aan de kant van de oppositie met die zelfs het taferelen waarbij opgemerkt moet worden dat daar de medische zorg stond geheel ontbreekt het zal nog even over de oppositie er eerder al het over de lichte wapens die door positie vooral heeft de kerk zijn ze toch in staat om zo'n zware aanslag te plegen ja als ze d'r nou gedaan hebben waar ze en waar we 't net over had dan blijkt dat uh beschikken over organisatie ook over communicatie apparaten uur en ja dit had gisteren niemand voorspelt en die weten is wat er morgen gaat gebeuren dus het is echt een enorme opsteker voor de oppositie en de eentje merkte ook meteen in uh gebieden maar de oppositie heel sterk is in ons in dit jaar daar konden we in de vreugde niro en zelfs alle van hoorn in damascus

5. je merkt in de winkel in is die nog niet zo heel veel van maar we moeten er rekening mee houden dat de prijzen van onze eetlust er gaan stijgen en hierdoor komt het overstromingen als deze op zijn en op de één plek in de weer otten die hevige droogte daardoor dreigen om een maand veel oogsten te mislukken mogelijk is harst aan grondstoffen en nu al tot hoge prijs zoals prijs van marx en drie maanden tijd meer dan dertig procent gestegen en die van tarwe zelfs meer al veertig procent dat maakt veel producten de komende tijd een de tuurlijk tarik komt onder meer hier terecht en meelfabriek in uithuizermeeden laat ze malen tweehonderd ton bloem en meel per dag uit de grondstof waar het bedrijf werd om steeds meer voor moet betalen maar in de prijsschommelingen zien van afgelopen twaalf maanden ging het van honderd zeventig wonen twee in het even te komen en dat zegt de fabrieksdirecteur kan er niet helemaal voor z'n eigen rekening nemen achter zijn van onze was de laatste is de grondstof bij de eerste en de molen concentratie op de dus wij komen niet omheen reprazent ontbreken de prijs van graan wordt sterk beïnvloed door de wereldwijde weersomstandigheden het midden van de verenigde staten daar wordt vooral maïs verbouwd tussen alle tijd droog en heet meer dan veertig graden akkers droge uit ook in rusland en kazachstan hebben boeren last van de droogte in india valt juist te veel regen en verdrinken de gewassen ook dichter bij frankrijk en duitsland een voorbeeld rijk de oogsten mislukken door te veel regen en dat denkt allemaal door in de uithuizen bij de ware directeuren boek ook de prijs niet alleen ziet stijgen maar ook sterk bewegen aan de wordt ook in parijs op de beurs genoteerd het ziet en de schommelingen heel groot geworden we hebben op dat moment schommelingen op de dag dat we vroeg op een jaar india hebben gezien in het maar eens de uiteindelijk gaat het deel van de fabriek in aardig pakken daarbij worden ook de hogere prijzen door geven en de onze hé rijdt en als je bruin vanmorgen afgerekend voor één euro veertien wordt duurder dus als ik hoeveel en maar neer precies dat moet nog blijken waarschijnlijk al snel en die hogere broodprijs is op ons in het westen het nog niet zo heel veel uitmaken echt een koppel een is het voor het veel armere afrika moet aangeven en ze nu al veel groter over een inkomen uit aan voedsel en toch nog duurder wordt houden ze helemaal niet zo in over voor andere

6. de rabobank is een kopje hard geraakt en dat heeft alles te maken met het intro zou banken schandaal rond de zogeheten libor rente daarmee is gesjoemeld rode britse bank barclays en mogelijk heeft ook de taal osborn zich niet aan de regels gehouden reden voor een en onze bank om samen met de autoriteit financiële maar het uh de toezag ouder van de banken om ons om te doen naar de rabobank en de aanleiding is dus dat schandaal onder libor rente en een van de belangrijkste ruimtes ter wereld elke ochtend bellen vijftien banken met elkaar om door te geven tegen welke rente zijn geld kunnen lenen op de bekendgemaakt de lipo rente die liep horen en te is van belang voor de consument die bepaalt ook de hoogte van diepe teken en spaarrente is de britse bank barclays heeft dat toegegeven dat ze gezoem om te hebben met de rentepercentages om financieel sterker over te komen vragers nu of ook de rabobank fraude heeft gepleegd door de rente te beïnvloeden ga maar ik uh meer elk durven gert-jan elk andere hoofdkantoor van rabobank in een utrecht ja maar wordt de rabobank navo verdacht garcia ja wat precies heeft geleid tot het ontslag van vier mei medewerkers bij de rabobank dat weten we niet wat zich precies achter maar je op het hoofdkantoor en op het kantoor in londen heeft als de is ook niet maar wat we wel eten is een vergelijkbare zaken in londen al blijkt dat eigenlijk twee belangrijke redenen zijn om te sjoemelen met die tarieven eerste is om jezelf mooier voor te stellen dat je en het sommige banken hebben moeite om geld aan te trekken en moet er dan een hoogte lief betalen en ze geven eigenlijk aan lage tarieven op onze mooier voor te stellen omdat het niet fijn vinden om voor de buitenwereld aan te geven dat het moeilijk is voor een om geld laten we als we daar wat lijkt dat men er waarschijnlijk ook maar rabobank ondanks de crisis toch wel in staat was geld aan te trekken de afgelopen jaren en de tweede reden is ordinair geld verdienen doet er even aan passen zodat alle handelaren meer verdienen en dat lijkt in dit geval bij de rabobank meer waarschijnlijk de reden ja het enige deel klanten ook daarvan de dupe van hoor nou die tarieven die worden gebruikt bij het vaststellen van heel veel liever voor kredieten en voor hypotheken en in die zin hebben niet alleen daar ook klanten maar alle klanten van alle banken en mogelijk mee te maken of ze gedupeerd zijn dat is heel lastig denk ik te bewijzen voor de toezichthouder de nederlandsche bank eigenlijk ook emma niet zo relevant wat ik kijk naar de integriteit van de bank en als zou blijken dat medewerkers van de rabobank hebben gesjoemeld om ervoor te zorgen dat de bank meer gaat verdienen dan lijkt integriteit in de data te ik kan de nederlandse banken moeten opleggen dat kan flink oplopen heeft het voor wat een lijkt engeland laten zien dat ali te boeten op tot drie honderd zeventig miljoen euro gert-jan takje alle wekenlang was tot op fors onder

7. er vals kalenders van acht nieuwe slechte cijfers gekomen over de eerste halfjaar was het nettoverlies ruim één komma twee miljard euro dat aantal of ruzie eerste licht steeg blijft het luchtvaart verrast houden van de bezwaren wereld even zoals een goed een grote rol onder moeilijke omstandigheden moet kunnen vliegen zo moet de luchtvaartmaatschappij een turbulente tijden zich als eenheid dan de tienerpubliek je daarbij snelle onwezenlijke werd dus dat is dat zij rené midden in dit jaar werd steeds prins het is dat ja als michel is vind ik zes jaar een portie komen d'r in dongen kende ik niet kan en toch maakt iedereen de vergelijking tussen de beide partners ja dat is niet uh te bekennen vraag wie doet het beter dan wij doen het te weten dan en frans de splitsen van de twee kan ik unive van maar ook alemán had nu in negentien resultaat wat vrij normaal is voor deze twee kwartaal het toch is het vooral er france-soir bezuinigd moet worden ruim vijfduizend banen gaan verloren nog niet alle franse vakbonden op dit moment met de bezuinigingsplannen instemmen iedereen donker opbouwde was de tijd dat voorzieningen en dat het werd tijd dat het met je praten dat deed ze op het altijd hierop open dat op de grond of zo uh diervoer ouwe dringen steeds uw haar man verwacht niet dat het nieuwe cijfers ook gevolgen hebben voor de werkgelegenheid in ons land wij hebben bedenkingen historie dat we sinds twee duizend acht bij de kale en al onze programma's met succes we kunnen doorlopen zonder dat we iemand gedwongen moeten ontslaan dat houdt nu nog steeds vol en niet alles is levert die het aantal passagiers dat omdat dat koos voor en van staal en stegen is maar daarmee is positiever wat ik er inzien zakt het tweede kwartaal aanzienlijk beter hebben gepresteerd aan de kwartaal voor ja daar ziet zeker lichtpuntjes op het dus als de tweede half jaar te vechten sterven is het te de steeg het aandeel ervan staal-en vandaag met maar liefst negentien procent

8. uh om utrecht wist begonnen met de sanering van de als westwijk in kanaleneiland en kan een deel van de geëvacueerde bewoners nog deze week nou het opruimen van asbest nog op alleen al vroeg zij moet over enkele dagen klaar rechtse burgemeester omschreef alle bewoners beloofd dat ze daar als beste een tiende van hun eigen huis mogen inzien zo kunnen de bewoners zelf controleren ook hoeven een van de schadelijke stoffen is marcel een flat en anand de straat kunnen voorlopig nog niet naar huis voor die flat wordt nog gewerkt aan een saneringsplan aan

9. ze ioc-baas volgelingen twee grote bedreigingen voor de olympische spelen door op in het zogeheten matchstick zien het manipuleren van wedstrijduitslagen voor de illegale gokmarkt ook de olympische spelen ontkomen daar niet aan denkt hoger vandaar dat het ioc en engeland intensief samenwerken met uh gokindustrie om schandalen te voorkomen ja ik kan er wordt dan alleen maar winst is bijna het zitten britse in hun bloed god coorte net zo blij als een kopje thee kunt hier werkelijk overal op over het uiteraard ook op de uitslagen van de olympische spelen daar kun je me voor inzet dat de kans dat nederland meer dan zestig oude medailles wint deze spelen naam had dat was probeer maar steeds vaker blijkt het folkore stuk minder onschuldig afgelopen maanden waren er verschillende god schandalen en waarmaakt marat voetballers cricket spelers roeiers ze bleken te zijn opgekocht volgens diocees dat misschien nog wel een groter probleem dan de open en dan naar je ook maar dit ik zeg ik zeg maar is hij is de difranco kennen mij niemand peper over en die ing zelf af als vis heeft vijf aandelen worden in dokkum daarbij horen zeelieden op met sieraden mee zijn het zit ik weer pijn speeksel wel goran pakken staal en de je operator zuiderpark graal bevelen in en daarom houden ze hier op het hoofdkantoor van het blokes op verzoek van het ioc in de gaten of er verdachte dingen gebeuren twee maar ja je eraan moeten worden de wilde mij nog maar wij vinden waarop te missen en wat er met een wel met hem die moet plaatje bekendste en dan lijkt mij wel indymedia is tegen dat het voor die wordt verder wilde gaan failliet dus monet eenmaal in nou we dat zeker niet business dan al tussen een soort visuals langs de bal jelle er zij aan buju het met de alija juist zeer aan een ding kermde in wijken weten die uh kraaien simon is ons ze wisten ze het en dan nog even terug naar m'n eigen weddenschap het aantal keer goud voor nederland meer bij het zwemmen naar mensen toe endstra dat komt de liefhebber er kan maar tot maar mag ik heb het uit we letteren te koop maar door 't spant strekt wervelend misschien zijn er normaal komen er toch voor een riekend naar dat willen we hebben ik tot levert dus lang niet zoveel op als die andere wetenschappen die tot afgelopen vrijdag om ons inzetten op de kans dat een niveau zou verschijnen bij de openingsceremonie kreeg er duizend terug altijd dat had gedaan

10. grootste stroomstoring aller tijden zes honderd miljoen mensen in india zaten urenlang zonder elektriciteit door een heel europese unie en turkije in één klap de stroom uitvalt optreden van binnengekregen twintig deelstaten in het noorden en oosten van het land last van de stroom storing viel voor de helft van alle indiërs de stroom uit het als voor de tweede achtereenvolgende dag gisteren zoals ook al een groot deel van noord-india zonder expliciet uit de gloednieuwe metropool moet doen is het ook weer niet is in die ja school op moderne tijden vandaag even niet aan maar die was rense westra maar dat mag maar waren met doris day maar dat zou wat ze aan die uh nou voor daar werd ik weet zeker achttien was door zijn waar alleen met de auto start nou wordt-ie legt bij de kuyper die sterk is het zou niet zo raar is dat ook een paar daar in een in die jaren beleeft op het moment de heetste zomer van de afgelopen eeuw een soort hoe die in het energieverbruik ooit tara elders en erik teleur dat was dus de hangar uitstak al achter aan de leidde jarenlang uh alcom horen wat de bal wordt er dan maar de hitte is niet de enige reden voor deze in mensen storing in de stroomvoorziening ze vinden de deelstaten trouwens jullie dus door afnemen van het centraal uh nou de het is een ook dick dees had weten een buitengewone en beredeneren maar horen waarin dat bouwen maar binnen is dit ze overtrek is er bij storing elkaar als dominostenen om weer in de eerste van een maar belangrijkste oorzaak van het alles is gek genoeg de enorme economische groei van de afgelopen jaren daardoor neemt de vraag naar is die in india heel snel toe en de overheid houdt het niet bij je komt op zich genoeg geld in de bouw van nieuwe centrales maar als slecht bestuur corruptie ook levert dat allemaal veel te weinig op met de helft van india in het donker van bach werd dat wel even heel waarin ik daar een

11. de niet schuur van nederland zou werd noord-brabant en de knipoog al om de genoemd de provincie heeft omdat het uh eigen belgië aan zegt heel veel last van het criminele twee jaar geleden werd gekozen voor een nieuw aanpak justitie en politie probeerde criminelen zo hard mogelijk te raken vooral in het wordt een wilbur vanochtend rechercheurs wachten tot ze deze woning van de vermeende hennep crimineel binnenstebuiten kunnen keren speciaal het is mee om daar geld of drugs te speuren we zagen daar bezig met de actie dak van het afpakt in dat zit het met name richt op de voor zitten en de teelt en handel als een aantal verdachten dat het arrestatieteam aangehouden andere arrestanten gewoon met reguliere politiemensen thaise op pakt aan op dit moment zijn er al die alle woningen huiszoeking bezig met name omdat crimineel verdiende geld te vinden daarvoor zou ten landmacht aanwezig die beschikt over apparatuur 'm even morgen ruimte te kunnen zoeken geen overbodige luxe want het criminele geld blijkt goed verstoppen voorbeeld om er voor jaren tussen de begraven tussen de wortels van bal we ronden tegen ons uh zelfs een onder een klaar is de opvolgen en dus overal trof wel foster meer dan acht honderd duizend euro is er gevonden ook is er beslag gelegd op bankrekeningen scooters ze juist het begin hadden moeten driehonderd literair omfietsen in deze is aangetroffen uh in uh villapark of we daar doen de munitie uh een alle andere vermogensbestanddelen solse auto is sieraden logies en bij de ste actie van vandaag zijn twaalf criminelen aangehouden de meesten komen uit dezelfde familie heleen willen het vijftien woningen doorzocht binnenstebuiten gekeerd enorm machtsvertoon wat zitten geachte ja vroeger was het doel van justitie om criminelen zo lang mogelijk achter de tralies te krijgen en tegen eigen woorden gaat het erom omzet leven zelden zuur mogelijk te maken en hoe doe je dat dan nou gewoon door alles af te pakken vaak werd ook een gevangenisstraf gezondheidsrisico wie zeer ontmantelen en ik heb het bedrijfsrisico maar uh dit bedrijf citroen door als je daar komt wat pijn doet het afpakken van datgene wat ik nu al verdiend hebben dat er mee op de juiste wijze de hij dit rainey en dat betekent dus dat er uh verhalen gaan van grote criminelen die met daaraan een hele hoge te moeten kijken moeilijkste maltin overture sportwagen wordt afgevoerd het politie instituut in dit met veel machtsvertoon allerlei ene kant aan de muur te laten zien ja wat pakken zei echt wel halen en o'malley criminelen te laten zien dat misdaad echt niet aan

12. in rotterdam vannacht waren feest ontaardde in een relletjes onder invloed van drank en drugs richtte een groep boze feestgangers spoor van vernielingen aan dat oorlog was het al tot een confrontatie met agenten gekomen maar die konden situatie nauwelijks aan als een dolle moeten ze tekeer zijn gegaan in een mum van tijd zijn team winkelpanden beschadig al toen de ruiten ingegooid en is een nicolay geen rol daar maakte ze vanmorgen de schade op exclusieve zonder brillenverkoper zie je en ook twee keer de doelwit van plunderaars zit je toch een ravage en joris geplunderd je leest geplunderd seeding eersel stadhuisplein zal om even vlak nadat ze het is hier gebeurt en dan schrik je noemt wil je hele ligt want de die heb ik uh resten liggen daar en in staat kunnen plukken en meer dan ook wat je te breken als de raad gaan het begon allemaal rond kwart voor drie vannacht toen een groep van vijftien man uit dit café aan het stadhuisplein werd gezegd dat er gevochten om de politie lieten de arm die geeft waartoe tweehonderd vijftig mensen bij waren stilleggen op het plein van het vervolgens tot een confrontatie de politie moest assistentie vragen aan de andere korpsen omdat de aanwezige agenten de situatie niet aankonden en zo een confrontatie met de politie door uh glaswerk en stenen naar en het gooien het mits heeft het op moment besloten om deze mensen uit eten gaan draaien maar ze luisterden eigenlijk nergens meer naar ze waaraan jaap op onze zal dalen onder invloed van op een drugs dat cynisme onder de indruk waren van honden die werden ingezet tussen pritstift besloot om diverse charles uit te voeren en we daarmee uh zei wel dat de groep uiteengejaagd uiteindelijk heeft de politie na de rellen in de vernielingen zes mannen van rond de twintig opgepakt de schade loopt in de tienduizenden euro's de privé deed dan daar kan gewapende overval heb en daar een eind te worden ja nou ja het kan uh scherven opruimen en uh en doorgaan

13. vu in de keniaanse havenstad mombasa te hebben aangewakkerd tot voor kort leefden moslims en christenen daalt het vreedzaam naast kan stad die ook populair bij toeristen waren er deze week hele plunderingen staken moslimjongeren kerken in brand de rellen in kenia hebben alles te maken met het buurland somalië militairen uit kenia en ethiopië strijden er samen bij de afrikaanse unie tegen de radicale moslimbeweging als japan wereldwijd zijn er grote zorgen over deze groep die samenwerkt met al-kaida en somalische jongeren werft voor terreurdaden ook in de keniaanse havenstad mombasa zijn er banden tussen radicale keniaanse moslims en als japan de overheid zou dit jaar daarom al vijf keniaanse moslims zonder proces hebben laten vermoorden de laatste moord afgelopen maandag op ambon vogel uh moslimgeestelijken sloeg de vlam in de pan vandaag bij het vrijdaggebed werkte opnieuw veel moest verwacht de nederlandse ambassade in kenia niet mensen op om weg te blijven uit om wat ze kosten renske is broere ging te ook de vaste klanten van slagerij hassan staan voor gesloten deuren veel winkels in mombasa blijven dicht in de wijken waarin de rellen plaatsvonden zijn nu vooral speciale eenheden van de politie te zien wat doe je samen mogelijk relschoppers strak in de gaten en beschermen ook de gebouwen die eerder deze week werden aangevallen zoals deze wet plaats van het leger des heils afgelopen maandag ging het hier helemaal niets je moet mensen in de blanken als deze steden het is een beetje ramen waren op de worsteling zo wordt er nooit de commissie wel is er onder andere the christian zijn er dus een beetje uit is aan het doen we het moet wel een charles de hel soldaten zijn zondag hier opnieuw samen te komen zoals de moslims dat deze vrijdag in de roo door een horror hoe er drie maanden van deze moskee zelden zijn preek de moord op cabooter ook over oordelen het geweld dat daarop volgde heeft voor hem niets te maken met religieuze spannen enorme is daar maar ook het we hebben daar namelijk waarnaar worden verwelkomd worden je daar dan worden we weten ook dat het in het witte ooit werpen maar ook vandaag moeten moslimleiders en mombasa optreden om de jongeren in bedwang te houden de politie maakte ook duidelijk geen enkel geweld te dulden wederzijdse argwaan is groot de erkent geen geschiedenis van willekeurig geweld de beste wetten zeggen ook dat de jongeren in de wereld is eerder dat van al-qaeda en worden de overheid hij had de combinatie van een lichtje heb word ik maak deze situatie die zonder heeft voor zijn het blij is deze vrijdag bij dreigen maar het verleden leert hoe makkelijk het ook in de toekomst weer fout kan speciale politie-eenheden die afgelopen dagen waarom was er zijn gebracht zullen daar waarschijnlijk voorlopig blij

14. sns reaal dreigt in grote financiële problemen te komen de belangrijkste oorzaak is de crisis in de vastgoedmarkt en zes real moet nu mogelijk bedrijfsonderdelen gaan verkopen en dat zal niet makkelijk zijn nog meer de in de financiële crisis en ga naar everything er die werd hoofdkantoor staat van est en chio in utrecht zijn even moeten rekening houden zich nou zorgen maakt nee helemaal niet het is geen geval zoals bijvoorbeeld bij deze week weer ze bang maakt nogal wat een bescheiden winst en boven een heeft de overheid deze bank aangemerkt als een systeem bank dat zeggen dat het nederlands gaat vindt dat deze bank absoluut niet om mag alleen overeind gehouden zal worden desnoods met staatssteun nou is het natuurlijk niet zo dat het idee dat zou willen wat het is juist door die staatssteun uit twee duizend acht tijdens de bankencrisis dat deze banken problemen is gekomen want toen kreeg sns reaal zeven vijftig miljoen euro ook dat geld moeten ze volgend jaar eind van want jaren terug betraden inclusief ontwennen gaat het om een bedrag van acht honderd vijftig miljoen euro en het geld dat is niet ondanks die bescheiden winst ja zeg om als ultieme zorgeloosheid en als hij later ook in het verleden wil praten dan dezelfde dit streven een van de onderdelen van de sns bank de banken die als enige nog maar heel weinig staatssteun heeft terugbetaald sns kan ruim achthonderd miljoen euro niet vrijmaken er moesten extra kapitaal reserves orde opgebouwd volgens nieuwe europese regels tientallen miljoenen euro's liggen daardoor vast ook moet sns rekening houden met schadeclaims van houders van boeken paulussen die kunnen de sns veel geld gaan kosten en dan is er nog het verlies van en er twee miljard en de harder dan in hetzelfde spoor met mensen om het wat je wilt worden van het aan mijn stand staan voor het gewone vastgoedprojecten op papier zagen zo mooi uit maar ze maakte de financiële verwachtingen niet waar zo is sns eigenaar van de wolf langs de uit twee bij utrecht het laat staan imperiale en zanger van staat en marina port in florida uh stond er te een om te groeien die voor de sns niet goed uit het hij heeft ook als die problemen al aanpakken uit onderzoeken dus of ze bedrijfsonderdelen kunnen verkopen dat zou dan gaan de verzekeringstak van de dochters van en stress reaal om reaal en om zwitserleven en zijn er voor de zakenbank in de armen genomen ja de vraag is natuurlijk of dat midden in de financiële crisis genoeg geld ontbreekt om ook in staatssteun echt te kunnen afbetalen en ik begrijp ook dat ministerie van financiën bovenop zit ja de minister wel waar helemaal niets over zeggen maar je kan je voorstellen dat ze inderdaad daarbovenop zitten bij de weer kwam ik als je daar en dat komt het uh koken omdat alle banken hierbij betrokken zijn we hebben in nederland het depot stoppen franse stelsel het teken dat ik eerst de honderd duizend euro aan spaargeld gegarandeerd wordt door alle bach in nederland en stel dat het mis gaat en dat van ach het niemand nog maar stel dat het mis gaat dan zouden alle banken eiland ook opdraaien voor de verliezen iedereen zit hier kwam bovenop eva dakje

15. goed luisteren of ze duurder somaliërs in de polsen opgestapt uit het als wilson pocent vugts zo'n terug hard aan om eens eerder in een tentenkamp zaten ze kunnen niet terug naar somalië terecht gezegd schoonmakers willen zich hier vrij kunnen bewegen vanmiddag staat ze voor de deur bij het idee een lintje in den bosch zijn eerste wat ze teruggaan naar het vaderland zodra daar vaak is

16. en dat bosbranden in het zuiden van spanje een surroi gisteravond uit in de bergen en drukte vandaag op naar de bal plaats waarbij je veel zeker een dode zo'n vijf jaar in de ze hebben hun huis moeten verlaten worden brandweerlieden uh strijden is uur van de zijn vermoedelijk aangestoken

17. kanalen natuurlijk alles el meisje maar sinds kort hebben we ook de zelf broertjes twee dertien het vijftien jaar oud zijn dyslectisch maar ze zijn ook hoogbegaafd en daar een eerste blijkbaar geen school die ze wil hebben om die reden wilde ze morgen aan een grote zeilreis beginnen en hij is om naar school toch al kinderen zijn weer plichten gepaard volgens de familie klaassen is er geen enkele school die en erick-e en hugo met een combinatie van dyslexie en hoogbegaafdheid les willen geven thuisonderwijs zou misschien een oplossing zijn maar dat is in nederland in principe verboden vandaar het reisplan want op dat verbod wordt een uitzondering gemaakt voor kinderen die tijdelijk in het buitenland verblijven maar de kinderbescherming daagde de ouders in een spoedprocedure voor de haarlemse rechter er zouden namelijk wel degelijk schoolde zijn waar de jongens terecht kunnen maar de ouders vinden die goed genoeg hij wilde graag een spoedprocedure en te opraken namelijk een uh tussenkomst van door jeugdzorg die wil gaan bemiddelen ze dat het een school om te zorgen dat de salon in nederland was volgens de me wilden ook precies nou hiervan geen en instanties die gaan zeggen wat wel en niet goed is voor mijn kind als helemaal niet duidelijk is dat ik niet al doen wat goed is voor mijn kind in de rechtbank koos de kant van de ouders van het spel van god te zijn is afgerond teken dus dat er geen toezichthoudend voogd voor die jongens komt dat denk dat ook dat je dan alleen cd's kunnen beginnen als het dan ineens een heel erg blij met binnenlandse acties voortzetten en zegt er blij mee het telefoontje vinger staat dat het was wel een spanningen en een keer jaar was dat jaar met uh dat ik ooit in dat jaar nog wel ja het werk van de rechter is overigens een voorlopige woonhuis later dit jaar geeft hij een definitief oordeel die jongens willen morgen uitvaren zodra ze in belgische territoriale wateren komen zullen ze de schoolboeken openslaan een paar dagen had we hebben andere boten of je de twee klaassens wil met het avontuur ook aandacht vragen voor de term tien daarvan andere kinderen die thuis zitten omdat er geen gepast onderwijs voor ze is ja en dan

18. jarenlang zaten ze die onder de grond en zijn ze nooit een dag echt een sekte in de russische stad kan zo'n zeven om het vijftig kilometer van moskou in de ban van een drieëntachtig jaar gelijk zeventig mensen om wie veel kinderen zaten met ze alleen kleine rokjes ondergrond de kinderen zelf er wordt naar jaren onder de grond komen ze eindelijk boven kinderen een tot zeventien jaar oud die nooit naar school gingen nooit een dokter zagen en een grote islamitische voorman moesten volgen ze worden naar weeshuizen gebracht in redelijke gezondheid het is een van de heet dit voor heel het land dit was een voorman waar ze rachman satarov de vonken van een trolleybussen zag hij als een boodschap van god dat hij de nieuwe islamitische profeten is uh waren er onder de grond gezet hij zijn eigen instemming samenleving op en als de rus sische politiewerk in valt accepteert hij hun gezag niet aan een ja maar dat de ruwe en dat er wel wat staat daarop sekteleden voegen vandaag die zelf de hij onder dit gebouwtje zitten de acht verdiepingen van kleine om verwarm de kamertjes waarin de sekte komt het gebouw is volgens de op drie tijd die legaal en wordt dus gestraft de sekte heeft aangekondigd optredens dus toch te blijven zelfs op hoofdpunten op

19. ik gelopen jaar zo'n veertig duizend eilanders besmet waren met de cup-koorts spa die de bacterie die verspreid wordt via geiten maar nu blijken veel meer mensen besmet te zijn raakt maar liefst honderd duizend in uw onderzoek in de kunt kort spaulding valletje uhm bossen de huizen een bols wijst uit dat de bacterie vooral in de labohm ze heeft haar schouw minstens één op de tien bezoeken stalletje voor ons ziekenhuis het is met zijn geraakt met de cup-koorts pakte in sindsdien in eigen internet en nieuwe als ze maar alle verschijnselen waren maar zingen en daarna d'raan denken aan het strand denkt u dat u het gehalte nee we hebben internetstek heeft gehad zijn belevenissen in de leeuw te al een jaar te kunnen rekenen die brute onderzocht zijn baia mare gezinnen dat het in een dag kan ik het misschien wel of niet in ieder een die met de bacterie in aanraking komt merkt dat en wat ook niet ziek enkele duizenden krijgen dan af uh lopen jaren lichte of flink in de enkele honderden houden en chronische vermoeidheid verschijnsel aan wal en tot nu toe werd gedacht dat de vijfentwintig mensen aan zijn overleden dat aantal ligt ook hoger denken zie je de boel ze dat verwacht ik eigenlijk wel ik denk niet dat een factor tien hoger zal zijn naam is in welk vak de twee hoger steken in twee duizend negen twee duizend tien een wet nog niet te bij iedereen die het uh een complicatie had van uh ziet indien mogelijk bewust ook was de kortst daar ook echt nagezocht ook al lopen er veel meer mensen rond die besmet zijn met de bouw deri het advies is die zich niet ziek voelt gaat niet naar de huisarts ze zou dat zeker niet naar de huisarts gaan als je toch even zeker wil weten ja weet dat het zijn wel een ontzettende belasting van de gezondheidszorg als al die honderdduizend mensen zich nu gaan melden de korte epidemieën is dus veel groter geweest dan tot nu toe weigerde dat is verontrustend geruststellend is echter dat bijna geen nieuwe cup-koorts patiënten meer bijkomen en al die honderd duitse die besmet zijn geraakt met de cup-koorts bacterie die hebben antistoffen en die kunnen niet meer ziet worden het rijksinstituut voor volksgezondheid en mooier gaat mee in de race betaalt voor het onderzoek eenzelfde aantal bess met dingen naar boven bij dus van vijftig huizen naar honderd duizend

20. het thijs op de noordpool smelt veel sneller dan tot voor kort werd gedacht van dag voor dag macht en het laagterecord uit twee duizend zeven gebroken het is zo doorzet zal de noren keizer over dertig tot veertig jaar helemaal ijsvrij is en zo eerder gingen klimaatwetenschappers er vanuit dat pas eind deze eeuw nog zeeijs is bevroren zeewater dat drijft op de oceaan de ook een zomers meldt een enorme hoeveelheid van wat aan is op zich niets aan de hand het heeft geen effect op de stijging van de zes wiegel en vanaf eind september als de zon ondergaat op de noordpool groeit er ijskap de raad sinds negentien negenenzeventig wordt door zal de lieten gemeten hoeveel vierkante kilometer van het cda is er drijft op de noordpool die metingen laten vooral de afgelopen tien jaar zien dat er in de zomer steeds minder eisen overblijft het vorige laagterecord stond uit twee duizend zeven maar dit jaar ligt er dus nog minder hij is door veruit in de laatste wedstrijden actueel verdwenen is er in het uh juist minder wurmen beslaat ongeveer een op een apotheker nederland tot is wat stoffelen satelliet ex per van het knmi komt dit door het ooit als het werk en hoe minder ijs op de noordpool hoe sneller de aarde zal maar na het cda is de is elftal in mijn en het ndt kaatsten nog wel eens om niet te ze 't spiegel wessels peekel ja als de eis gesmolten is dan hebben de slaat en water voordat uh een neemt alles om een maand is er al ik het ook dat er weer wat trouw daardoor veranderen betekent ook dat uh de wanneer de oceaan lokaal aan pas dan kunnen worden en het echt allerlei gevallen waarvoor we iedereen alles wat leeft op een ook al en ver daarbuiten na een eis draaien noordpool heeft ook voordelen er ontstaan nieuwe kortere vaarroutes tussen de continenten en een ontroert in noordpool is een potentiële

gaan mijn voor boringen naar gas en olie verschillende landen hebben al claims gelegd op het gebied begin vorig eeuw met de noord ook gezien als de laatste niet ontdekt de plek op het noordelijk al vond uittikken missen dat over zo'n dertig jaar kun je in de zomer dus gewoon in een bootje naar de noordpool en van vakantie misschien krijgen dat

21. de computer bestaat je dat ze geïnstalleerd op je uhm ploeteren zonder dat je daar erg in hebt en dat met koen documenten van jou aan haal gaat het is een computervirus werd heeft toegeslagen geen een groot aantal overheidsinstellingen universiteiten en bedrijven opvallend veel gemeenten zijn het slachtoffer geworden op dit moment zijn al enkele duizenden toeters besmet geraakt en uh grote vraag is wat dat virus nog meer dat aids waarschijnlijk van een groep prusiner krijg maar ergens in het oosten van het boek gingen die er aan geld en stijlen van mensen bankrekening en ze hadden al netwerk uitgehold en op het doordacht en hebben ze nu een nieuwe ja virussen in gehangen en dit virus heeft een nieuwe nier om zich te verspreiden en dat doet het zo grof dat een netwerk waar daarna en onderuit ga ook een uh criminelen met dit wat hier is dan zijn van banken besteedt ja als ze door ss'er nestelen zich in de brouwer in het verkeer tussen de browser en de bank en uh ja proberen dan jouw banktransacties te manipuleren het over alle bankrekeningen we in te vullen het van ja je dat goed kunnen doel is dus bankrekeningen leggen alle mart virus wordt verspreid niet dat allemaal gebruik op de computer en daar hebben ook gemeenten er last van zoals de gemeente weer vandaag waren daar de hele dag op bezig om alles weer aan de praat ja normaal kan natuurlijk ook niet erg maar qureia zoals die nu zien we een van de is eerst de vissers op het günther nog nooit oude tijden herleven op het gemeentehuis van weert de typemachine stonden nog op zonde en het oude ms-dos wordt weer opgestart het zijn de beeldschermen die uh sinds negentien vier en een denk je gebruikt worden om uh naar personen met digitaal te registreren ikedia hè kan ik niet zien maar het was de onians kan het niet zien en lacht nog er is een oude laptop in de kast na de computerschermen blijven op zwart oneigenlijk met een uh storen op een incident willen werkplekken en een medewerker die stoornis valt er ogenschijnlijk even later een storm werd er nog zwaarder en zodoende van de ene na de ander werkplek en toen elk een loop van woensdag moira zit ministeries kwamen achter dat er iets anders aan de hand was dat uh zomer virus in het uh gewerkt aan het herstel van onze computersysteem ter ondersteuning van de eik die dei afdeling is hulp van buitenaf in uw denk toch dat we als overheid die toch nog meer aandacht aan het kan besteden aan de beveiling van onze digitale snel in weert worden vandaag geen rijbewijs verstrekt noot paspoorten wel het is lastig maar sommigen zien er de romantiek van alle die de halen dat liggende zei het is het is de hand werkt met hier zijn we hier te wonen in het jaar als dit anders is het is altijd met één ding op de maar van het virus zijn zin die het nog niet verlost zo hebben luongo maar jeroen hoe ernstig is dan ook nou ja je ziet maar weer 'ns een keer dat het ernstig is omdat onze kwetsbaarheid aantoont en het is ook een schommel om reden we weten niet hoe lang dat virus precies besmet en die computers besmet heeft het is een dag als waren tegen de lamp gelopen toen het is met die office bestanden aan de haal ging maar al veel langer misschien zelfs al maanden kan dat programma op die computers gezeten hebben en in die tijd heeft het gedaan maar voort gemaakt is tamelijk rekeningnummers wachtwoorden verzamelen en daar decentrale computeren woekeraar die nu sturen we weten dus dat is waar de criminelen het daar ook al in handen hebben en dat is best wordt de en heel opvallend dit virus slaat toe vooral in nederland ja we zagen het volk zou ik zie je wat voor ons uit op een rijtje zet en je ziet inderdaad als je naar dat graffiti kijkt dat nederland heeft ongeveer drieduizend besmette computers op dit moment een werd daarna als komen andere landen denemarken en ook uit stand nou waarom dat heeft te maken met waarden voor gemaakt is tachtig zijn van de nederlanders doet aan internetbankieren als het doek alle van het europees gemiddelde dus als je als computercrimineel iets met interpreteren doen dan doe je net in nederland en dat al zo veel gemeenten tegen de lamp lopen dat is ook opvallend een beetje vertekend beeld omdat we eigenlijk melding van computerproblemen bij de overheid eerst maar rekenen op dat er ook een en al bedrijven geraakt en daar zou het ook gewoon gebeurd kunnen zijn militairen die hadden dat liever niet aan de grote klok maar daarna is het om in zijn en ook van nee nog volgen om iets mijn collega geslaagde ver te verbergen ging naar een meldpunt het nationaal ze verstuurt its centrum en gaat hij sprak daar met mensen die ons aan uitlegde dat dit eigenlijk pas het topje van ijsberg ja ik zorg wel dat nog niet uh noem je weg is en met name en de karakters die klappen doordat uh zien is dat het tennis met de computer als echt actief wordt nadat die opnieuw op start dit en dat gebeurt is typisch de ochtend zou mensen weer aan het werk gaan we 'm en computer die vandaag besmet zijn wij mensen die op vakantie gaan en die kan ook drie weken terug door de opnieuw opgestart en dan zie je het effect weer elke en dat is dus op het werk maar moet je nou ook zorgen maken over je computer thuis ja het is goed nieuws en slecht nieuws het virus is nu bekend is als je virusscanner bui werkt wordt er afgehaald het slechte nieuws is dit is een kat-en-muisspel was gaat het wel eerst aan te praten zei niets aan de criminele alweer bezig om in je moet versies te maken en te verspreiden dus als je vandaag live-band het maar de vraag of dat moe ook nog zo is jeroen naar zich

22. een overval op een woning een voort uit een dorpje bij barneveld nietsvermoedend staten bewoners daar thuis plotseling stond er een mannen in de huiskamer en de werd geweld gebruikt zoals heel vaak bij overvallen op woningen beginnen dood een van de overvallers overleeft niet gisteravond rond tien uur ding is zeker twee en ze deze woning binnen in het huiswerk af om en ze de bewoners aanwezig een jong stel bij de ene keer zag hij die zijn niet van plan hun bezittingen mee te geven er ontstaat een worsteling waarbij ook wordt geschoten de twee bewoners worden gewond naar het ziekenhuis gebracht maar die heeft er geschoten dat is een belangrijke vraag en het politieonderzoek wat de bobo maar ze zijn niet de enige die gewond zijn geraakt en op een straat wordt die avond een man gevonden zwaargewond uiteindelijk overleeft hij de avond niet uh weet inmiddels dat het gaat om onze een twintigjarige man uit ede en hij is zeer waarschijnlijk een van de verdachte die betrokken geweest bij de woning al van de woning achter ons gaat er vanuit het even bewoners deze man geschoten heeft en in de cyprus alles nog mogelijk uh worden uitgebreid met de slachtoffers in gesprek om het ook uitgebreid in rome super-sub over de aantal van op woningen daalt als waren ze nog altijd zo'n zeven honderd vijftig per jaar gemiddeld meer entreepartij acht op opvalt bij veertig procent om de overvaller uit iemand gewond soms gaat er zelfs uh iemand dood zoals riant woord haar zo en dat is meer dan bij overvallen op bedrijven zien dat dit zo dan vallen vaak gewelddadige verloop en omdat de bewoners voor zich thuis in eigen huis wordt plotseling overvallen raak je maar die en ontstaat er uh ook van de overvaller uiten op geweld of de worsteling gisteravond heeft voorkomen dat er wat buiten is gemaakt is niet bekend de politie heeft vooralsnog niemand voor overval programma's die het de politie heeft ook nog niet bekendgemaakt hoeveel mensen door de gezocht voor je over van de bewoners liggen in het ziekenhuis

23. en dan om een eerdere redenen een bijzonder verhaal uit china bijzonder omdat het de laatste bijdrage is van onze correspondent wouter zwart vanuit china na zeven halfjaar held die china voor de verenigde staten het zou het maar een bijzondere aldus het proces tegen een carla in verhaal van een van de machtigste politici van china haar boek wordt verdacht van moord maar volgens veel mensen speelt op de achtergrond nv groter politiek steekspel vol mies die intrige is hoe dan ook geen van de grootste rechtszaken in china van afgelopen dertig je de onder weersomstandigheden die bijna symbolisch leken voor de politiek is storm ging het proces vast hard achter gesloten deuren uit het zicht van burgers en journalisten speelt zich een d'raan af en niemand die precies weet hoe het zit de hoofrolspeler boek kayla je vrouw van of je daar tot voor kort baas van het station jean en kandidaat voor china's hoogste partijfuncties begin dit jaar vlucht boze rechterhand politiebaas wonen niet in de arena niet kans consulaat en verklaart daar dat de machtige van die bal verantwoordelijk is voor de dood van een britse zaken nieuw heywood was jarenlang een vriend en zakenpartner van de familie wel maar werd november vorig jaar vergiftigd aangetroffen in een hotel aan de later mede dus door boeroes verklaring wordt boek highline beschuldigd van moord ja percentage aan een enorme ook haar naar het zien van vooral mannen nu een balkon meer nu er onderdoor enigszins in die paar vermogende hongarije en weet neer omdat is het aan appelman mooi dat deze loyce zich enorm kans had willen ontdoen gondel aanvaren hier carla in lijken zoals schuldig bevonden voor de rechtszaak is begonnen een politiek proces zeggen deskundig enige op want ook guzman hoe stielike wordt ergens gevangen gehouden en is ontdaan van al zijn ik zie die officieel alleen verdacht wordt van de overtreden valpartij regels zijn er wel degelijk verhalen over corruptie machtsmisbruik en zelfs het afluisteren van partijleiders vraag is echter hoe vol bol met als een vrouw vervolgens om inziens werden peru het in leven ook haar waren al problemen door badminton oldham kereltje sri hoewel het zoen voor wordt kan ze bellen die heerenveen had waarna alles waar je ook die kwamen nooit zo dus op wat start zo ze wil geen beroep is en zo om nu wordt vindt dat de vreemde ook al actief in de stad bij louter in peking water waar dan is dit proces zo belangrijk dan dat er politieke implicatie is heel erg groot kun zij werd komende herfst krijgt china een nieuw groot communistisch leiderschap uh ja en wat die het nieuwe leiden dat het niet kunnen voor ruiken is dat onder hen een soort wenen hangt van foute zaken nou meer bosie like als zo'n anz om het politiek talent die kans maakt om daarin te komen en hoe zijn in tolwegen die zwijn dus ja hij moet het veld ruimen zonder dat die andere nieuwe leiders er straks last van en gaan krijgen aan te kunnen te niet doen nu een hoe groot politiek proces onder hebben dan gaan bouwen dat zouden de dader nog veel meer op en er vestigen dus wat er waarschijnlijk en ik moet wel zeggen de waarschijnlijk zal gebeuren dat hij alleen partijen die zich die ervan worden gestraft en dan zie je als soort achteruitgang de politiek en het publieke leven zelf al aan en dat is zelf pas zes zeven elf jaar leven en voor ons begon in china was dat al deze meer dan ook nog een gewerkt dat deels was het misschien onmogelijk geweest omdat daar al rond van het proces bijvoorbeeld daar in één dag de schil snel en zeker op een hele gevoelige daar gaat het dus zeker niet nieuw china dat gebeurt hier vrij vaak en maar dat daarbovenop nu achteraf heel ook over wordt bericht in de media dat is op zich al heel bijzonder hij het verleden zou uh zouden houdt niet hij d'rvan kozen hebben om dit in de doofpot te stoppen onder tapijt schuiven wat niet weet wat niet hier nou dit is twee duizend twaalf jaar zijn sociale media even wordt het is een internationale gemeenschap aan het worden dus ja wat kun je daar niet meer maken dat

is echt van het verhandeld werden er zijn geen na maar ook dat de chinees india achterlaat nu andere chinezen zijn dan die van zeven elf alleen dan misschien wel een klein beetje natuurlijk alles hier veranderd is het wordt steeds groter wordt steeds moderne dus natuurlijk de andere de mensen ook heel vaak in positieve zin parsons misschien ook niet heel erg in positieve zin ik bedoel maar te zeggen dat voorbeeld uh zeven jaar geleden zag je een hele mooie per als die er een soort jan prettiger soort naïviteit en een soort naïviteit waarmee men echt dol op de kans op ontwikkelingen men wilde ontwikkelen bereiken worden mee wilde leren nou nu zijn er zes zeven jaar later we merken toch tot tien daar ivic tijd die zo prettig was een beetje plaats begint te maken voor soort ja uh bezien eerder soort besef het besef bijvoorbeeld als hun land en nu eigenlijk komt er als we nu heel groot en belangrijk zijn in de wereld maar ook het besef bijvoorbeeld dat je met die welvaart die ze nu hebben die rijkdom dat je niet alleen hele mooie spulletjes kan kopen wat je ook macht en invloed kan komen er is dus ook heel veel corruptie dan dat eind dit individualisme dat toe nee wat er ook zoveel hebben in het westen en die groeiende en opgroeide gat tussen arm hij kijkt dat is toch wel iets wat we de komende jaren zeer ik er in de gaten blijven houden elke wouter dankjewel hij goeie reis straks bij de grote overste bandje

24. uh hebben we afdaling kan van de incontinentiemateriaal heeft in opdracht van apothekers patiënten benaderd fabrikant wilden aanvragen hoeveel materiaal een patiënt node minister schippers vindt het niet kun nee dat apothekers patiënt gegevens verstrekken aan bedrijven fabrikant tegen haar wilde weten wat voor incontinentiemateriaal nodig heeft om te spelen te voorkomen minister schippers is het ermee eens maar vindt dat apothekers aan de privacywetten moeten houden onderzocht wordt nu of dna dat gedaan heeft

25. het sterftecijfer in nederland is voor de eerste sterk gestegen en de eerste zes maanden van dit jaar overleden ruim twee en zeventig duizend mensen zijn er vier duizend meer dan in dezelfde periode vorig stierven vooral meer mensen op hoge leeftijd mannen ouder dan tachtig vrouwen boven de negentig tca rode weer in februari speelde een belangrijke rol zegt het centraal bureau voor de statistiek maal

26. er zal even wennen sp-leider roemer staat ineens in de belangstelling omdat je serieus de premiers kandidaat is het risico op uit te en dus is groot maar daar maakt een achterban zich helemaal geen door nee mm-hu gegaan gedurende een jaar dan ook zelfs emile roemer zijn stropdas af doet dan is het officieel worden de campagne is begonnen romme wil regeren maar dat moet dan waar ook heel anders dan bij het vorige kabinet een rijk land als nederland die zegt dat wij kidjan te hebben om het wat eerlijker te delen met is oranje dat wat ons allemaal is dat we daar een beetje saai hierop zijn dan de afgelopen tien jaar kan naar romer trekt de aandacht als mogelijke te meer ook de buitenlandse pers bewaarders weten welk beleid die wil gaan voeren da's niet moeilijk vindt sp-leider puinruimen zal het voor maar ik wil daar niet van mij en ik wil ook een beroep een beroep op de samenleving omdat niet verwijten dat maar ik vraag welk ernst aan heel veel mensen een als de kans om het waar te maken dat wij de komende jaren nederland menselijkerwijze chalet kan maar de andere kant staat gehad nu is het ons beeld de vraag is nul koppert en daarvoor met de sp wil samenwerken zeker nadat er omar nogal enthousiast aangaf dat die zich niks aantrekt van eventuele europese begroting moet dus uitspraak over maar ik that barry vond u dat een goeie uitspaart maar zitten ook mensen en oude als wat dacht is dit jaar de uitspraak op zichzelf van een nou niet zo heel slee en zo wil ik er wel even heb ik een keer fout maakt hebben we te maken een fout van vijftig miljard ja ik wil zijn allemaal mensen ja en nee ik zeg ik z'n engelse ja nee regeren lijkt dichterbij dan ooit dus de sp er vergeven en roemer graag als straks maar wij ook de verkiezingen wint eenmaal langer

27. ze juliana station onzichtbaar te ontwrichten fobici ik maak de middag was heel lange tijd in het openbaar te zien en te hoe het een korte speech tot zo'n aanhang en de verzamelde wereldpers dat deed die heilig vanaf het balkon van de ambassade van ecuador in londen als je het op de stoep wat gedaan zoals eerder aangekondigd dan was ie direct opgepakt met enige bravoure maakte het fenomeen zijn is vanmiddag zijn rentree op het wereldtoneel verblijf van zweden andreu anders zou ja en voor en twee die de politie was in de poging zijn uitlevering aan zweden onmogelijk nog recent is bang dat zweden gebleken zijn die publieksactiviteiten doorstuurt naar het virus ja het wennen aan de aan de lijnen een ja en nee ik ga maar na intuïtie had ik enkele keren in het zijn tussen advocatenteam wordt aangevoerd door de voormalige spaanse onderzoeksrechter baltasar krijgt hier in het staan buffart trad dit is een mooie toen maria ook uit de realen aart-jan in londen werd werk draait al moet ik er is een jaar in de zeilen aan een half dozijn gehouden door een meneer binnen is tekenend na de op nu komt ook zo is het er nu weer betrekkelijk rustig maar het toespraakje van een soort kost me scherp markt lijn aanloop naar het nog heel veel meer er wordt om als uh

28. en dan een hobby die niet meteen zou verwachten in de stad het houden van bijen en toch is het een hobby die juist daar steeds populairder wordt in nederland zijn er ongeveer acht duizend bijenhouders en vooral in de steden worden dat er is met want een zalm arm plaats in de utrechtse binnenstad hier wonen tienduizenden daar ja op het ook van jan-kees en rode ik daar het hartstikke leuk het was voor huisdieren gehad maar er zal weinig er was een gat in bijenvolk is van lykken maar zaten geen plaats voor mochten hier neerzetten het aantal neemt snel toe ik vind dit echt wil leiden tussen omstandigheden zijn er ook gunstig dan wanneer er veel verschillende bloemen in omgeven nee uh oh oscar begon drie jaar geleden bij te houden hier bovenop het ook in amsterdam-west chen waar je volk in ontwikkeling volgt het is fascinerend en hun ja een mooie bestuiving strook machtig uhm omdat we daar een daarvan mee te krijgen om dat te kunnen volgen eigenlijk en uh de olievlekken wil ik had al dertig jaar bij en leidt nu jongen in zo in het daar midden in amsterdam-noord de verenig in groeide in een paar jaar tijd van acht naar veertig leek is een tendens elementen kwam maar voor jonge mensen die het zeer interessant vinden om er maar bij je beesten zijn niet alleen vragen maar ook doen het zo ver het goeie nieuws want het slechte nieuws is dat zou ik die groepen wij aan een eigen volk en de wereld neemt alleen maar af en plaatst de bij je zijn van belang voor bestuiven van planten en daarom meer belangrijke ons voedsel als olga zullen denken moet dat nou per se in de stad als om aan het mis zoals vorig jaar bij ons kan uh ja d'r is een uh een uh het werk geweest in het ik heb uh wat gooien we ze maar je kon er op het dak zijn de naardense belabberde was weliswaar wordt naar een oplopen ze eerder het komt wel is voor de probeerden te voorkomen maar goed we uh hield zich niet aan de regels waar je en zo is nu eenmaal de natuur op de bij werken aanstekelijk op de mannen daarom willen ze er volgend jaar nog een kans naaien ja zeggen dan

29. uh nuchtere zeiden al dat het aantal te komen dat er eigenlijk het keuzes moeten worden gemaakt voor de gestegen kosten in de gezondheidszorg dit medicijn vergoeden we wel en dat medicijn die het college voor zorgverzekeringen komt nu met zo'n advies stoppen met medicijnen voor twee zeldzame ziekte is ziekte van pompe en om fabbri en dit is het letterlijk advies aan meningsverschil was ook rond handhaven van de vergoeding heeft dus waren kan ik me soms niet ethisch worden beschouwd om heel dure medicijnen te betalen voor relatief kleine groep patiënten in dus die onze werkt zorg zonder soort rinke van den brink wernink is de nou een omslagpunt bereikt het wel op het is voor het eerst op een werken medicijn is tegen bepaalde ziekten en dat er niet meer van goed gaat worden als het cvz zijn zin krijgt om wat er toen dürers het gaat dan met meer ziekte gebeuren maar het deskundigen m'n hart zorgde hebben gesproken voor wachten tot ons vertelt visser door kon de werkelijkheid wordt toch al een aantal anderen en je dure geneesmiddelen op dezelfde manier woorden solo gewoon bevorderen vormde tegen vergelijkbare ziekte waarbij mensen niet hebben maar ook hele dure kankerbehandelingen eenmaal bijvoorbeeld soms ook heel dure medische gevraagd daar zullen patiënten dit natuurlijk verschrikkelijk vinden maar wat een acties van de dief in de bedoeling is om als je een te dure medicijnen nodig helpt om houden het op de het gaat dus uh in dit geval nu om een medicijnen voor twee zeldzame ziekten zie van pompen en fabriek dat zijn erfelijke aandoen de ziekte van pompe een van die zijn erfelijke in heel zeldzaam aandoen de patiënten maken een bepaald en zien niet aan in nederland beide negen baby's aan pompen in medisch zijn wil het college voor verzekeringen wel blijven vergoeden of het helpt anders ligt het volgen het cvz springt weet dat zijn ruim tachtig jongeren en volwassenen de verslapping van hun spieren raken ze invalide in de medicijnkosten per persoon per jaar tussen de vier een zeven honderd duizend euro daar hebben ze dat wij zeggen ze het cvz vindt het effect onvoldoende en wil stoppen met vergoed de verhouding tussen kosten en baten is om acceptabel bij de fabriek je treft het advies ruimste stug patiënt in de en zien therapie werkt maar kost zo'n twee honderd duizend euro per patiënt per jaar de tweede set binnen ook deze vergoeding schappen het doorslaggevende argument is de buitengewoon ongeluk struck kosten effectiviteit en ik wat zij naar de gevolgen voor de patiënten van deze ziekte als ze deze medicijnen die we krijgen procent we met de ziekte van fabry die gaan we om tien tot twintig jaar korter leven daar hebben ze leven je moeders toe kon dat onze meer last voor ziekten kunnen waar de ziekte van pompe worden mensen sneller invaliden verslechteren leven en daar is dit dus maar een voorlopig advies zijn concert adviezen van het van het college zorgverzekeringen staat er op kans dat dit advies orde aangepast aan ook hier wel topper in de verlosser termen komen alle partijen die alleen te maken waarbij ook arbeidsrecht is het roer over te praten het cvz heeft in dat concept tot vis alarm en te voeren tegen het voor goede waarom worden eigen zet die gewichtheffen hogere vroeg komen soms moet stoppen te stoppen met de vergoeden dus is onwaarschijnlijk dat daar wat verstandiger was ja ik je en als er wordt besloten om in een daar die meer zeiden tegen pompen en verdriet te stoppen dan denk 't is kundige dat dit dus maar een voorbode is van nog veel die je kunt niet het had alles wat die zonder worden allemaal mogen waar is er een positief kan uitwerken dat kan hier uh gratis in het in de basis heeft adresje nooit en moeten leiden worden getrokken moeten keuzes worden gemaakt over welke medicijnen zodat in de toekomst nog meer ga de beleid aangaande wordt gedomineerd door duren is er weer een ook het dure geneesmiddelen als eerste in het vizier zei een soort voorbeelden een dot com dure geneesmiddelen zijn die voor

hele zeldzamer ziekten worden de toegepast ook de dure geneesmiddelen hij bepaalde vormen van uh kan maar hoe zat dat met deze cursus af naar kom je vaak die uh nou ik hou ik hoop dat na de advies honderd die nu gaande is dat daarna en niet alleen een advies van het cdja van op het college een zorg zijn maar dat had ook een commissie wordt tussen geschakeld dat is nu dat kan de praktijk zijn dat is natuurlijk zijn en dat in die commissie serieus wordt gekeken niet alleen naar die financiële aspecten maar ook naar de rechtvaardigheid en naar de ethische en dat moet prevaleren we in het algemeen geld niet prevaleren en zo erg vindt zijn is nu is de schippers die hierover een definitief dus daar werd en dat zal dan in de herfst zijn op onze zou dus het recht voor meer informatie over dit advies en of ze het er

30. zuid in het waterschap in zuid-limburg onderzoekt of er mate regen nodig zijn om nieuwe overstromingen te voorkomen overstroming van de rivier de groep heeft flinke schade aangericht in slenaken op het was een eerste inventarisatie de kelders van enkele horecazaken zijn veranderd in een puinhoop afgelopen nacht rond middernacht werd waken voor als ze op een hoogwatergolf stonden voor hoe rustiger lieve in als je te riviertje de gulp zo ziet kan wanden dan kun je bijna niet voorstellen dat gisteravond hier voor zoveel overlast zorgden door noodweer in belgië zestig negen en negentig een half uur kijk dat konden hier simpelweg niet aan dan buitenste wil voorts en zorgde hier heeft hij naar het voorbeeld het water drong ook naar binnen bij enkele vakantie bowling inhoudloos en waren vaak ook niet bestand tegen de kracht van zoveel waar in dan is het zondag hoe ze binnen uh nee maar komen kijken een wordt het weer eens al het is vier maanden alles weg van het stuk van 't is nu gevallen tussen uh ja het is stel dat alles nu gaat zusjes jij je al een week daar is het een stuk sleutelen gaat er wel dit stukje opbouwen via de zussen dus al of alles wel vinden in de zien ons voorbeeld het plafond van de dalende ons dat dit uh oppassen vandaag in overvloed van alle kanten dan lijkt het zo als het nodig is ja dat voor de periode in de in kilte in beeld brengen en dat ik daar waar wat kunnen doen vrouwen dat uiteraard te doen sommige ondernemers verwachten de rest van het hoogseizoen dicht te zullen zij

A.1.2 Teletext

1. Bij een schietpartij bij een discotheek in Frankrijk zijn afgelopen nacht tien mensen gewond geraakt. De dader was boos omdat hij de discotheek, vlakbij de plaats Cambrai, niet in mocht. Daarop haalde hij een jachtgeweer uit zijn auto en schoot-ie Na de schietpartij ging de man er vandoor. Hij is een paar uur later aangehouden.

2. Goedenavond. Studenten en veel ouders waren er al bang voor maar nu heeft de rechter z'n zegen er aan gegeven: De overheid mag een boete opleggen als je te lang over je studie doet. Duizenden euro's bovenop het collegegeld. Studentenorganisaties vingen eerder bot bij de politiek en dus probeerden ze het bij de rechter. Studenten moeten voortaan opschieten. Ze krijgen een jaar speling, maar als ze meer dan 1 jaar vertraging oplopen, krijgen ze een boete. Ook studenten die actief zijn voor hun studentenorganisatie moeten betalen. Het is heel erg jammer op het moment dat je inzet Ik doe nu een full-time bestuursjaar. Je zet je een jaar full-time in voor de studentenbelangen. Dat 'beloond' wordt met een boete van 3063 euro. De bestuurder is niet de enige. Ook studenten die ziek zijn geweest of een verkeerde studie hebben gekozen, krijgen een boete. Daarom stapten ze naar de rechter. Wij delen niet het oordeel van ISO dat de thans studerende student aanspraak kan maken op het gedurend een onbeperkt aantal jaren collegejaren volgen van onderwijs kan maken op het gedurend een onbeperkt aantal jaren collegejaren volgen van onderwijs tegen een vooraf vastgesteld wettelijk collegegeld dat afgezien van een inflatiecorrectie blijft gelden tot het diploma is behaald. We zijn hier heel erg teleurgesteld over. Dat de spelregels tijdens het spelen van het spel veranderd mogen worden. In het vonnis staat 1 uitzondering voor een kleine groep: hoeven geen boete te betalen. Ik ben opgelucht, eerlijk gezegd. Ik heb een studie wijsbegeerte die 6 jaar duurt. Ik zit nu in mijn 4e jaar, dus ik zou vanaf september moeten betalen. En dat is nu geschrapt. Alle andere studenten die te lang over hun studie doen hebben geen reden om blij te zijn. Zij kunnen een rekening van meer dan 3000 euro verwachten. Ja, ik moet hem zelf ook gaan betalen. Ja, die boete komt er dus toch, maar de vraag is wel: voor hoe lang. Want de meeste partijen schrijven in hun verkiezingsprogramma's dat ze die langstudeerboete volgend jaar alweer willen terugdraaien.

3. Voor iedereen die een baan zoekt, zijn het zware tijden, want vind maar eens werk. Voor jongeren van vooral Turkse, Marokkaanse of Surinaamse afkomst blijkt dat helemaal moeilijk. Dit zijn de cijfers: Een op de vijf van die jongeren had begin vorig jaar geen baan. Nu is dat opgelopen tot bijna 1 op de 3. Van de autochtone jongeren zijn er veel minder zonder werk. Van de autochtone jongeren zijn er veel minder zonder werk. Ook al heb je een goeie opleiding, het is echt heel moeilijk om iets te vinden. 'Uw gegevens laten onvoldoende aansluiting zien' 'met het door ons gewenste profiel van junior associate .' Het is een van de tientallen afwijzingen die Rachida afgelopen maanden heeft ontvangen. Vorig jaar zomer rondde ze haar opleiding arbeidsrecht af, en nu zoekt ze een baan als advocaat. Ze is 4 keer uitgenodigd voor een gesprek. Maar tot nu toe niet aangenomen. Heb je het idee dat het met je achtergrond te maken heeft? Maar naarmate de tijd vordert begin ik me steeds meer af te vragen dat het bij de een veel makkelijker gaat dan bij de andere persoon. Als reden krijgt ze vaak te horen dat ze niet bij het kantoor past. De werkloosheid is onder Nederlanders met Turkse, Marokkaanse of Surinaamse achtergrond altijd al hoger geweest dan bij autochtonen. Maar de recente stijging is opvallend, zegt Forum. Wij denken dat allochtone jongeren vaak buiten de boot vallen Het zijn vaak tijdelijke banen. En waarschijnlijk speelt mee dat veel jongeren langer zijn gaan doorleren omdat het crisis was. Die komen nu ook allemaal op de markt. En werkgevers hebben wat te kiezen nu er zoveel werkzoekenden zijn. En dan kiezen ze toch voor autochtonen. We hebben het idee dat werkgevers, vooral als ze al allochtone werknemers in dienst hebben dat het voor hen niet zoveel uitmaakt maar dat ze wel bang zijn dat hun klanten of afnemers liever geen allochtone werknemers willen zien. liever geen allochtone werknemers willen zien. En dat ze daarom voor bepaalde functies liever autochtone werknemers kiezen. Of werkgevers dit ook zo zien, weten we niet. Die wilden vandaag niet reageren. Mag ik uw legitimatie? Dank u. Rachida heeft inmiddels haar bijbaan weer opgepakt. Maar blijft solliciteren en hoopt ooit haar eigen advocatenkantoor te hebben. De vraag is natuurlijk of bijvoorbeeld het kabinet hier zou moeten helpen. Het ministerie van Sociale Zaken zegt: we doen niets aparts voor bepaalde groepen. Alle jongeren moeten vooral een richting kiezen waarin ook banen zijn. En als je denkt dat een werkgever discrimeert, moet je dat melden, zeggen ze.

4. Het is een aanslag in het hart van het Syrische bewind, haast onder het oog van president Assad. Tijdens een vergadering van het Syrische veiligheidskabinet ontplofte een bom die aan zeker twee sleutelfiguren van het regime het leven heeft gekost: de minister van Defensie en de onderminister van Defensie. Die laatste is de zwager van president Assad en een van de meest gevreesde figuren van het Syrische regime. Het is een klap in het gezicht van president Assad. Een bom zou weleens meer impact kunnen hebben op het Syrische regime en het Syrische volk, dan 16 maanden van oorlogsgeweld. Zwarte rook boven Damascus, vlak na de eerste berichten van de aanslag op het gebouw van de veiligheidsdienst. Daar was een vergadering van veiligheidsfunctionarissen. Vooral ministers en top-militairen. Beelden van de plek zelf zijn er niet. Die werd onmiddellijk afgegrendeld. Wel kwamen deze beelden van een straat in de buurt. Stukje bij beetje bij beetje begon de Syrische staats-tv na de aanslag berichten van andere bronnen te bevestigen. Dit is weer een nieuwe escalatie. Volgens sommigen is het een lijfwacht van deze minister van defensie die zich heeft opgeblazen. Voor defensie-minister Rajha werd binnen een paar uur een opvolger aangewezen. En ook Assef Shawkat kwam om. Assef Shawkat is deze man, de zwager van president Assad. De foto werd genomen bij de begrafenis van de oude-president Assad in 2000. Shawkat was onderminister van Defensie. Een sleutelfiguur in Assad's veiligheidsapparaat. Hij werd zeer gevreesd. Voor het nieuws van de aanslag maakte correspondent Sander van Hoorn deze beelden in Damascus waar nu vier dagen achter elkaar gevechten zijn. Het ziet er hier relatief rustig uit. Maar ook is te zien hoe veel mensen met bezittingen aan de wandel zijn. Mogelijk op zoek naar veiligere plekken in de hoofdstad. In Damascus, Sander van Hoorn. Dit moet een enorme slag voor Assad zijn. Dat is het zeker. Ze doen alles om te voorkomen dat het zo lijkt. Op radio en tv wordt er eenheid gesuggereerd. Gisteren bleek dit nog ver weg. En nu gebeurt dit in het centrum van Damascus. Is het zeker dat het om een zelfmoordaanslag gaat? Dat doet toch erg denken aan Irak, Afghanistan. Ik vind het nog vreemd. Ik heb nog geen verklaring gehoord. Ik probeerde in de buurt van de aanslag te komen. Dat lukte niet goed. Ik werd staande gehouden door militairen. Ik moest omkeren. Ik kwam op tientallen meters van het gebouw. Mensen gingen gewoon door met werken. Het was surrealistisch. We zagen net beelden van die vluchtelingenstromen in het noorden van Damascus. Hoe is het daar nu? In de wijken waar gevochten wordt. daar zie je die taferelen. Maar ook in het centrum van Damascus heeft het wel wat gedaan. Mensen zijn aan het bellen. Mensen waanden zich veilig. Vandaag waren er 34 gesneuvelde militairen. Aan de kant van de regering of van de oppositie dagelijks komen mensen dode familieleden ophalen. Er is een zwarte rookpluim. Daar vandaan kwamen grote stromen vluchtelingen. Dit ziekenhuis is vol. De meeste militairen zijn dood als ze hier aankomen. Het leed aan de kant van de regering en de oppositie Met dezelfde taferelen. Maar medische zorg ontbreekt. Sander, gisteren hadden we het nog over de lichte wapens die de oppositie heeft maar kennelijk zijn ze toch in staat om zo'n zware aanslag te plegen! Als ze het inderdaad gedaan hebben. Dit had gisteren niemand voorspelt. Het is een opsteker voor de oppositie. Zij kunnen hun vreugde niet op.

5. Je merkt er in de winkel nu misschien nog niet zoveel van maar we moeten er rekening mee houden dat de prijzen van ons eten sterk gaan stijgen. En hierdoor komt het: Overstromingen op de ene plek in de wereld en juist hevige droogte op de andere. Daardoor dreigen komende maand veel oogsten te mislukken. De mogelijke schaarste aan grondstoffen leidt nu al tot hoge prijzen. Zo is de prijs van mais in 3 maanden tijd meer dan 30% gestegen. En die van tarwe zelfs meer dan 40% . En dat maakt veel producten de komende tijd duurder. De dure tarwe komt onder meer hier terecht. Een meelfabriek in Uithuizermeeden. Wassen, malen. 200 ton bloem en meel per dag. Uit een grondstof waar het bedrijf per ton steeds meer voor moet betalen. Als wij de prijsschommelingen zien van de afgelopen twaalf maanden ging het van 170 euro naar 270 euro. 80% procent van de kostencalculatie is de grondstof. De rest zijn molenkosten en transport. We komen er niet omheen om die prijzen door te berekenen. De prijs van graan wordt sterk beinvloed door wereldwijde weersomstandigheden. Het midden van de Verenigde Staten. Daar wordt vooral mais verbouwd. Het is er al een tijd droog en heet, meer dan 40 graden. Akkers drogen uit. Ook in Rusland en in Kazachstan hebben boeren last van droogte. In India valt juist veel te veel regen en verdrinken de gewassen. Ook dichterbij, in Frankrijk en Duitsland bijvoorbeeld dreigt de oogst te mislukken door te veel regen. En dit dringt allemaal door in Uithuizermeeden waar directeur Buckow de prijs niet alleen ziet stijgen. Maar ook sterk bewegen. Tarwe staat ook op de beurs in Parijs genoteerd. Sindsdien zijn de schommelingen heel groot geworden. We hebben op dit moment schommelingen op een dag die we vroeger op een heel jaar hebben gezien met de prijzen. Uiteindelijk gaat het meel van de fabriek naar de bakker. En daarmee worden ook de hogere prijzen doorgegeven. En de onzekerheid. Een halfje bruin, vanmorgen afgerekend voor 1 ,14, wordt duurder. Dat is wel zeker. Hoeveel en wanneer precies dat moet nog blijken, waarschijnlijk al snel. Een hogere broodprijs zal voor ons in het Westen nog niet eens zo heel veel uitmaken. Echt een probleem is het voor het veel armere Afrika. Daar geven de mensen nu al een veel groter deel van hun inkomen uit aan voedsel. En als dat nog duurder wordt, houden ze helemaal niks meer over voor andere dingen.

6. De Rabobank is in opspraak geraakt. En dat heeft alles te maken met het internationale bankenschandaal rond de zogeheten Liborrente. Daarmee is gesjoemeld door de Britse bank Barclays. En mogelijk heeft ook de Rabobank zich NIET aan de regels mogelijk gehouden. heeft ook de Rabobank zich NIET aan de regels gehouden. Reden voor De Nederlandsche Bank om samen met de Autoriteit Financiele Markten de toezichthouder van de banken, om onderzoek te doen naar de Rabobank. En de aanleiding is dus dat schandaal om de Liborrente. Een van de belangrijkste rentes ter wereld. Banken lenen geld aan elkaar uit en vragen daarvoor een rentevergoeding. Elke ochtend bellen 15 banken met elkaar om door te geven tegen welke rente zij geld kunnen lenen. Om 11 uur wordt in Londen het gemiddelde bekendgemaakt: de Liborrente. Die Liborrente is van belang voor de consument want die bepaalt ook de hoogte van hypotheek- en spaarrentes. De Britse bank Barclays heeft al toegegeven dat ze gesjoemeld hebben met de rentepercentages om financieel sterker over te komen. De vraag is nu of ook de Rabobank fraude heeft gepleegd door de rente te beïnvloeden. Gert-Jan Dennekamp in Utrecht bij het hoofdkantoor van de Rabobank. Waarvan wordt de Rabobank verdacht? Wat precies heeft geleid tot het ontslag van vier medewerkers bij de Rabobank dat weten we niet. Maar wat we wel weten is een vergelijkbare zaak in Londen. En dan zijn er twee belangrijke redenen om te sjoemelen met die tarieven. De eerste is om jezelf mooier voor te stellen dan je bent. Sommige banken geven een lager tarief op om zich mooier voor te stellen, om geld aan te trekken. Voor Rabobank geld dat waarschijnlijk niet omdat het toch nog wel in staat was om geld aan te trekken. En dat lijkt in dit geval bij de Rabobank meer waarschijnlijk. Zijn klanten hier de dupe van geweest? Die tarieven worden gebruikt bij het vaststellen van tarieven voor kredieten en hypotheken. In die zin hebben klanten van alle banken er mogelijk mee te maken. Het is heel lastig om te bewijzen of ze gedupeerd zijn. Er wordt gekeken naar de integriteit van de bank. Als zou blijken dat medewerkers van de Rabobank hebben gesjoemeld om te zorgen dat de bank meer gaat verdienen dan lijkt de integriteit aangetast. en kan er een boete opgelegd worden. Dank je wel.

7. Air France KLM is vandaag opnieuw met slechte cijfers gekomen. Over het eerste half jaar was het nettoverlies ruim een komma twee miljard euro. Hoewel het aantal passagiers licht steeg blijft het luchtvaartconcern last houden van de zwakke wereldeconomie. zo moet een luchtvaartcombinatie in turbulente tijden zich als eenheid tonen. Toch maakt iedereen de vergelijking tussen de beide partners. Dat is de bekende vraag: Wie doet het beter? De split van de twee kan ik u niet geven. Maar ook KLM had nu een negatief resultaat wat vrij normaal is voor deze twee kwartalen. Toch is het vooral Air France waar bezuinigd moet worden. Ruim 5000 banen gaan verloren. Hoewel niet alle Franse vakbonden op dit moment met het bezuinigingsplan instemmen. Hartman verwacht niet dat de nieuwe cijfers ook gevolgen hebben voor de werkgelegenheid in ons land. We hebben een historie dat we sinds 2008 bij de KLM al onze programma's met succes hebben doorlopen zonder dat wij iemand gedwongen hebben moeten ontslaan. Dat hou ik nu nog steeds vol. En niet alles is negatief: Het aantal passagiers bijvoorbeeld dat koos voor Air France KLM steeg licht. aanzienlijk beter hebben gepresteerd dan het kwartaal vorig jaar. Daar zie ik zeker lichtpuntjes. En we hopen dus in het tweede half jaar de eerste effecten te zien van ons ombuigingsprogramma. Op de beurs steeg het aandeel Air France KLM vandaag met maar liefst 19 procent.

8. In Utrecht is begonnen met de sanering van de asbestwijk in Kanaleneiland. Waarschijnlijk kan een deel van de geevacueerde bewoners nog deze week naar huis. Het opruimen van de asbest begon in alle vroegte en moet over enkele dagen klaar zijn. Burgemeester Wolfsen heeft alle bewoners beloofd dat ze de asbestmetingen van hun eigen huis mogen inzien. Zo kunnen de bewoners zelf controleren hoeveel Van de schadelijke stof is gevonden. Bewoners van een flat in een andere straat kunnen voorlopig nog niet naar huis. voor die flat wordt nog gewerkt aan een saneringsplan.

9. Volgens IOC baas Rogge zijn er twee grote bedreigingen voor de Olympische Spelen doping en het zogeheten matchfixing. Dat is het manipuleren van wedstrijduitslagen voor de illegale gokmarkt. Ook de Olympische Spelen ontkomen er niet aan, denkt Rogge. Vandaar dat het IOC in Engeland samenwerkt met de gokindustrie om schandalen te voorkomen. Going down towards the line. Gokken zit de Britten in het bloed. Hier kun je overal op inzetten. Ook op de kans dat Nederland meer dan zes keer goud wint op de Spelen. Laat ik dat eens proberen. Hello. Maar vaak is gokken niet zo onschuldig. Thank you. Voetballer, cricketers, roeiers, ze bleken te zijn omgekocht. Volgens het IOC is het misschien een groter probleem dan doping. Daarom houden ze op het hoofdkantoor van Ladbrokes op verzoek van het IOC in de gaten of er verachte dingen gebeuren. op verzoek van het IOC in de gaten of er verdachte dingen gebeuren. En dan nog even terug naar mijn weddenschap op het aantal gouden medailles voor Nederland. Dat levert dus lang niet zoveel op als de weddenschap dat er een ufo zou verschijnen bij de opening. Als je daar een pond op inzette kreeg je duizend pond. Had ik dat maar gedaan.

10. De grootste stroomstoring aller tijden: 600 miljoen mensen in India zaten urenlang zonder elektriciteit. Alsof in de hele Europese Unie en Turkije in EEN klap de stroom uitvalt. last van de stroomstoring. Daarmee viel voor de helft van alle Indiers de stroom uit. En dat voor de tweede achtereenvolgende dag. Gisteren zat ook al een groot deel van Noord-India zonder elektriciteit. De gloednieuwe metro van miljoenenstad Delhi Het is India's hoop op modernere tijden. Maar vandaag even niet. India beleeft op het moment de heetste zomer van de afgelopen eeuw. Dat zorgt voor een piek in het energieverbruik. Verschillende deelstaten zouden stiekem te veel afnemen van het centrale net. Zo trekken ze bij een storing elkaar als dominostenen omver, is de eerste analyse. De enorme economische groei van de afgelopen tien jaar. Daardoor neemt de vraag naar energie in India heel snel toe. En de overheid houdt niet bij. Die pompt op zich genoeg geld in de bouw van nieuwe centrales maar door slecht bestuur en corruptie levert dat allemaal veel te weinig op. Met de helft van India in het donker werd dat vandaag wel even heel pijnlijk duidelijk.

11. De wietschuur van Nederland. Zo werd Noord-Brabant ook wel gekscherend genoemd. De provincie heeft, omdat het tegen Belgie aanligt, veel last van hennep-criminelen. Twee jaar geleden werd gekozen voor een nieuwe aanpak. Justitie en politie proberen de criminelen zo hard mogelijk te raken. Vooral in de portemonnee. Tilburg vanochtend. Rechercheurs wachten tot ze deze woning van een vermeende hennepcrimineel binnenste buiten kunnen keren. Een speciale hond is mee om naar geld of drugs te speuren. We zijn vandaag bezig met een actiedag van het afpakteam dat zich met name richt op de georganiseerde hennepteelt en handel. Er zijn een aantal verdachten met het arrestatieteam aangehouden. Andere arrestanten hebben we met de reguliere politiemensen thuis opgepakt. Op dit moment zijn in al die woningen huiszoekingen bezig. Met name om dat crimineel verdiende geld te vinden. Daarvoor is ook de landmacht aanwezig. Die beschikt over apparatuur om in verborgen ruimten te kunnen zoeken. Geen overbodige luxe want het criminele geld blijkt goed verstopt. Bijvoorbeeld onder een voiliere. Begraven tussen de wortels van een boom, onder tegels. Zelfs onder een kluis, in het plafond. Overal troffen we wel forse geldbedragen aan. Meer dan 800 .000 euro is er gevonden. Ook is er beslag gelegd op bankrekeningen en scooters. Zojuist heb ik het bericht gehad dat er ongeveer 300 liter amfetamine is aangetroffen. Er is een vuurwapen aangetroffen met bijbehorende munitie. En allerlei andere vermogensbestanddelen zoals auto's, sieraden, horloges en dergelijke. Tijdens de actie van vandaag zijn 12 criminelen aangehouden. De meesten komen uit dezelfde familie. Henrik-Willem. In totaal zijn er 15 woningen doorzocht, binnenstebuiten gekeerd, Enorm machtsvertoon, wat zit daarachter? Vroeger was het doel van Justitie criminelen zo lang mogelijk achter de tralies te krijgen. Tegenwoordig gaat het erom ze het leven zo zuur mogelijk te maken. Gewoon door alles af te pakken. Heel vaak werd gevangenisstraf als bedrijfrisico gezien: het Ontmantelen van de hennepkwekerij was ook een bedrijfsrisico. Maar het bedrijf ging gewoon door. Maar als je

daar komt waar het pijn doet het afpakken van datgene dat ze illegaal verdiend hebben dan ben je op de juiste wijze bezig. Het doet pijn. Dat betekent dat er verhalen gaan dat grote criminelen met tranen in hun ogen toezien hoe luxe motor of dure sportwagen spullen worden afgevoerd. Misdaad loont echt niet.

12. Onrust in Rotterdam vannacht, waar een feest ontaardde in relletjes. Onder invloed van drank en drugs richtte een groep boze feestgangers een spoor van vernielingen aan. Daarvoor was het al tot een confrontatie met agenten gekomen. Die konden de situatie nauwelijks aan. Als een dolle moeten ze tekeer zijn gegaan. In een mum van tijd zijn 10 winkelpanden beschadigd, autoruiten ingegooid en EEN winkel leeggeroofd. Daar maakten ze vanmorgen de schade op. Exclusieve zonnebrillen verkopen ze hier. En al twee keer eerder doelwit van plunderaars. Het is natuurlijk een ravage. Je hoort en leest 'geplunderd'. Je denkt eerst: Stadhuisplein, het zal wel meevallen. Maar vlak daarna besef je: Het is hier gebeurd. En dan schrik je en tril je. De resten liggen daar. Die heb ik van straat kunnen plukken. Ze hebben het opengebroken en de brillen eruit gehaald. Het begon allemaal rond 2 .45 uur vannacht, toen een groep van 15 man uit dit cafe werd gezet. Ze hadden er gevochten. De politie liet het R&B-feest, waar toen 250 mensen waren, stilleggen. Op het plein kwam het vervolgens tot een confrontatie. De politie moest assistentie vragen aan andere korpsen omdat de aanwezige agenten de situatie niet aankonden. Zij gooiden naar de politie glaswerk en stenen. De politie besloot toen om deze mensen uiteen te drijven. Maar ze luisterden nergens naar. Ze leken zodanig onder invloed dat ze niet eens van de honden nog onder de indruk waren. Dus de politie besloot charges uit te voeren. En daarmee werd de groep uiteengejaagd. Uiteindelijk heeft de politie na de rellen en de vernielingen zes mannen van rond de 20 opgepakt. De schade loopt in de tienduizenden euro's. Ik heb liever dit dan dat ik een gewapende overval heb op mijn mensen. Het is scherven opruimen en doorgaan.

13. Een moord op een moslimgeestelijke afgelopen maandag lijkt het sluimerende vuur in de Keniase havenstad Mombasa te hebben aangewakkerd. In de stad, die ook populair is bij toeristen waren er deze week rellen, plunderingen, en staken moslimjongeren kerken in brand. De rellen in Kenia hebben alles te maken met het buurland Somalië. Militairen uit Kenia en Ethiopië strijden er samen met de Afrikaanse Unie tegen de radicale moslimbeweging al-Shabaab. Wereldwijd zijn er grote zorgen over deze groep, die samenwerkt met Al Qaida en Somalische jongeren werft voor terreurdaden. Ook in de Keniaanse havenstad Mombasa zijn er banden tussen radicale Keniaanse moslims en Al Shabaab. De overheid zou dit jaar daarom al vijf Keniaanse moslims zonder proces hebben laten vermoorden. Door de laatste moord afgelopen maandag op Aboud Rogo een moslim geestelijke, sloeg de vlam in de pan. Vandaag, bij het vrijdaggebed werd er opnieuw veel onrust verwacht. De Nederlandse ambassade in Kenia riep mensen op weg te blijven uit Mombasa maar correspondent Kees Broere ging toch. Ook de vaste klanten van slagerij Hassan staan voor een gesloten deur. Veel winkels in Mombasa blijven dicht. In de wijken waar eerder rellen plaatsvonden zijn nu vooral speciale eenheden van de politie. Patrouilles houden mogelijke relschoppers strak in de gaten en beschermen de gebouwen die eerder deze week werden aangevallen. Zoals deze gebedsplaats van het Leger des Heils. Afgelopen maandag ging het hier helemaal mis. De heilsoldaten zeggen zondag hier opnieuw samen te komen. Zoals de moslims dat deze vrijdag doen. IMAM ROEPT OP TOT GEBED Het geweld dat daarop volgde, heeft voor hem niets te maken met de religieuze spanningen. Maar ook vandaag moeten moslimleiders in Mombasa optreden om de jongeren in bedwang te houden. De politie maakt duidelijk geen enkel geweld te dulden. Kenia kent geen geschiedenis van religieus geweld. De meeste mensen zeggen dat de jongeren die hier protesteren dat vooral deden uit politieke onvrede. Maar juist de combinatie van religie en politiek maakt deze situatie bijzonder explosief. Het blijft deze vrijdag bij dreigen. Maar het verleden leert hoe makkelijk het ook in de toekomst weer fout kan gaan. De speciale politie-eenheden die de afgelopen dagen naar Mombassa zijn gebracht, zullen daar voorlopig blijven.

14. De vierde bank van Nederland, SNS Reaal, staat met de rug tegen de muur. Als de bank niets onderneemt dreigt-ie in grote financiele problemen te komen. De belangrijkste oorzaak is de crisis in de vastgoedmarkt. SNS Reaal moet nu mogelijk bedrijfsonderdelen gaan verkopen. En dat zal niet makkelijk zijn, zo midden in de financiele crisis. Eva Wiessing in Utrecht bij het hoofdkantoor van SNS Reaal. Moeten rekeninghouders zich zorgen maken? Nee. Het is niet zoals bij DSB. Deze bank maakt nog altijd een bescheiden winst. En deze bank is door de overheid aangemerkt als een systeembank: hij zou overeind gehouden worden desnoods met staatssteun. Maar het is juist door die staatssteun uit 2008 tijdens de bankencrisis dat deze bank in de problemen is gekomen. Toen kreeg SNS 750 miljoen euro. Dat geld moeten ze eind volgend jaar terugbetalen, inclusief boete: een bedrag van 850 miljoen euro. En dat geld is er niet, ondanks de bescheiden winst. De bank verkoopt het Zwitserlevengevoel als ultieme zorgeloosheid. Als jij ook zo'n leven wilt praat er dan 'ns over met Zwitserleven. Zwitserleven, een van de onderdelen van de SNS Bank. De bank die als enige nog maar heel weinig staatssteun heeft terugbetaald. SNS kan die ruim 800 miljoen niet vrijmaken. Er moesten extra kapitaalreserves worden opgebouwd volgens nieuwe Europese regels. Tientallen miljoenen euro's liggen daardoor vast. Ook moet SNS rekening houden met schadeclaims van houders van woekerpolissen. Die kunnen de SNS veel geld gaan kosten. En dan is er nog het verlies van bijna twee miljard. Vastgoedprojecten. Op papier zagen ze er mooi uit, maar ze maakten de financiele verwachtingen niet waar. Zo is SNS eigenaar van The Wall langs de A2 bij Utrecht, de Plaza Imperial in Zaragosa en Maar toch, de drang om te groeien pakte voor SNS niet goed uit. Eva, hoe gaan ze problemen nou aanpakken? Ze onderzoeken of ze bedrijfsonderdelen kunnen verkopen. Bijvoorbeeld een verzekeringtak. Ze hebben daarvoor een zakenbank in de arm genomen. Maar de vraag is of het genoeg opbrengt. Ik begrijp dat het ministerie van Financien erbovenop zit? De minister wil er niks over zeggen maar ze zullen er vast wel bovenop zitten. Alle banken zijn hierbij betrokken. We hebben hier een depositogarantiestelsel: de eerste 100 .000 euro aan spaargeld wordt gegarandeerd door alle banken in Nederland. Gaat het mis dan draaien alle banken in Nederland voor de verliezen op. Dank je wel, Eva Wiessing.

15. Een groep uitgeprocedeerde Somaliers is boos opgestapt uit het Asielzoekerscentrum in Vught. Ze moeten terug naar Ter Apel, waar ze eerder in een tentenkamp zaten. Ze kunnen niet terug naar Somalie, heeft de rechter gezegd. De Somaliers willen zich hier vrij kunnen bewegen. Vanmiddag zaten ze voor de deur bij de IND in Den Bosch. Minister Leers wil dat ze teruggaan naar hun vaderland zodra het daar veilig is.

16. Felle bosbranden in het zuiden van Spanje. Het vuur brak gisteravond uit in de bergen en rukte vandaag op naar de badplaats Marbella. Er viel zeker een dode. Zo'n 5000 mensen hebben hun huis moeten verlaten. Honderden brandweerlieden bestrijden het vuur. De branden zijn vermoedelijk aangestoken.

17. We hadden natuurlijk al het zeilmeisje, maar sinds kort hebben we ook de zeilbroertjes. De twee, 13 en 15 jaar oud, zijn dyslectisch. Maar ze zijn ook hoogbegaafd, en er is daarom blijkbaar geen school die ze wil hebben. Alle kinderen zijn leerplichtig, maar volgens de familie Claassen is er geen enkele school die Enrique en Hugo met hun combinatie van dyslexie en hoogbegaafdheid, les wil geven. Thuisonderwijs zou misschien een oplossing zijn, maar dat is in Nederland in principe verboden. Vandaar het reisplan. Want op dat verbod wordt een uitzondering gemaakt voor kinderen die tijdelijk in het buitenland verblijven. Maar de Kinderbescherming daagde de ouders in een spoedprocedure voor de Haarlemse rechter. Er zouden namelijk wel degelijk scholen zijn waar de jongens terechtkunnen maar de ouders vinden die niet goed genoeg. Wij willen met deze spoedprocedure een doorbraak namelijk een tussenkomst van Bureau Jeugdzorg, die gaat bemiddelen zodat er een school komt zodat zij gewoon naar school kunnen in september. Wat wilt u nou precies? In ieder geval geen instanties die gaan zeggen wat wel en niet goed is voor mijn kind als helemaal niet duidelijk is dat ik al doe wat goed is voor m'n kind. De rechtbank koos de kant van de ouders. Het verzoek tot voorlopige ondertoezichtstelling is afgewezen. Er volgt dus geen toezichthouding voor die jongens? Dat klopt. Dus ze kunnen aan hun zeereis gaan beginnen? Dat klopt ook. Heel erg blij. Nu kunnen we onze actie voortzetten. Het telefoontje kwam, we zaten op het terras in spanning en in EEN keer dat telefoontje Ja, gewoon geweldig. Later dit jaar geeft hij een definitief oordeel. De jongens willen morgen uitvaren. Zodra ze in Belgische territoriale wateren komen, zullen ze hun schoolboeken openslaan. Hun vader houdt vanuit een andere boot een oogje in het zeil. De Claassens willen met het avontuur ook aandacht vragen voor de ruim tienduizend andere kinderen die thuis zitten omdat er geen gepast onderwijs voor ze is.

18. Jarenlang zaten ze diep onder de grond en zagen ze nooit 't daglicht. Ze waren in de ban van hun 83-jarige leider. Zo'n 70 mensen, onder wie veel kinderen zaten met z'n allen in kleine hokjes onder de grond. De kinderen zijn nu bevrijd. Na jaren onder de grond, komen ze eindelijk boven. Kinderen, 1 tot 17 jaar oud, die nooit naar school gingen nooit een dokter zagen en hun grote islamitische voorman moesten volgen. Dit was hun voorman: Faizrakhman Sattarov. De vonken van een trolleybus zag hij als een boodschap van God dat hij de nieuwe islamitische profeet is. Onder de grond zette hij zijn eigen islamitische samenleving op. En als de Russische politie daar invalt accepteert hij hun gezag niet. Sattarovs sekteleden volgen vandaag die lijn. Onder dit gebouwtje zitten de acht verdiepingen van kleine, onverwarmde kamertjes, waarin de sekte woont. Het gebouw is volgens de autoriteiten illegaal en wordt dus gesloopt.

19. Ze dachten dat de afgelopen jaren zo'n 40 .000 Nederlanders besmet waren geraakt met de Q-koorts-bacterie de bacterie die verspreid wordt via geiten. Maar nu blijken veel meer mensen besmet: maar liefst 100 .000. En nieuw onderzoek in de Q-koorts-poli van het Jeroen Bosch ziekenhuis in Den Bosch wijst uit dat de bacterie vooral in Brabant heeft huisgehouden. Minstens 1 op de 10 bezoekers van het Jeroen Bosch-ziekenhuis moet besmet zijn geraakt met de Q-koorts-bacterie. Maar alle verschijnselen waren net als wat er naderhand in de krant stond. Denkt u dat u het gehad heeft? Nee, ik ben me niet bewust dat ik die griep gehad heb. Dat weet ik niet zeker. Hoe zou ik dat kunnen weten dan? Dan moet uw bloed onderzocht zijn. Ah, maar daar heb ik geen zin in. Want dan heb ik het misschien wel. Niet iedereen die met de bacterie in aanraking komt merkt dat en wordt ook niet ziek. Enkele duizenden krijgen de afgelopen jaren een lichte of flinke griep. Enkelen honderden houden er chronische vermoeidheids- verschijnselen aan over. Tot nu toe werd gedacht dat er 25 mensen aan zijn overleden. Dat aantal ligt ook hoger denken ze hier in Den Bosch. Dat verwacht ik eigenlijk wel. Ik verwacht niet dat een factor 10 hoger zal zijn. Maar misschien wel een factor 2 hoger. Zeker in 2009, 2010 werd nog niet bij iedereen die een complicatie had van een ziekte die mogelijk berustte op chronische Q-koorts daar ook echt op onderzocht. Ook al lopen er veel meer mensen rond die besmet zijn met de bacterie het advies is: wie zich niet ziek voelt, ga niet naar de huisarts. Ik zou dan niet naar de huisarts gaan. Maar als je het toch even zeker wilt weten? Dat is wel een ontzettende belasting van de gezondheidszorg. Als al die 100 .000 mensen zich nu gaan melden. De Q-koorts-epidemie is dus veel groter geweest dan tot nu toe werd gedacht. Dat is verontrustend. Geruststellend is echter dat er bijna geen nieuwe Q-koorts-patienten meer bij komen. En al die 100 .000 die besmet zijn geraakt met de Q-koorts-bacterie hebben anti-stoffen en kunnen niet meer ziek worden. anti-stoffen en kunnen niet meer ziek worden. Het Rijksinstituut voor de Volksgezondheid en Milieu gaat mee in de resultaten van dit onderzoek en stelt het aantal besmettingen naar boven bij: dus van 50.000 naar 100.000.

20. Het zeeijs op de Noordpool smelt veel sneller dan tot voor kort werd gedacht. Vandaag wordt naar verwachting het laagte-record uit 2007 verbroken. En als het zo doorzet, zal de Noordelijke Ijszee al over 30 tot 40 jaar helemaal ijsvrij zijn in de zomer. Eerder gingen klimaatwetenschappers ervanuit dat pas eind deze eeuw zou gebeuren. Zee-ijs is bevroren zeewater dat drijft op de oceaan. Elke zomer smelt er een enorme hoeveelheid van dat ijs. Op zich niets aan de hand. Het heeft geen effect op de stijging van de zeespiegel. En vanaf eind september als de zon ondergaat op de Noordpool groeit de ijskap weer aan. Sinds 1979 wordt door satellieten gemeten hoeveel vierkante kilometer van dat zeeijs er drijft op de Noordpool. Die metingen laten vooral de afgelopen tien jaar zien dat er in de zomer steeds minder ijs overblijft. Het vorige laagterecord stamt uit 2007. Maar dit jaar ligt er dus nog minder ijs dan toen. De hoeveelheid die de laatste decennia gemiddeld verdwenen is het slaat ongeveer een oppervlakte van twee keer Nederlands. Volgens Ad Stoffelen satelliet-expert van het KNMI komt dit door het broeikaseffect. En hoe minder ijs op de Noordpool, hoe sneller de aarde zal opwarmen. Het zij eisen is over het algemeen wit en praat veel zonlicht terug. Het is een soort spiegel. Als het ijs gesmolten is, is het bluswater. En water neemt alle zonnewarmte in zich op. En ver daarbuiten. Maar een ijsvrije noordpool heeft ook voordelen. Er onstaan nieuwe kortere vaarroutes tussen de continenten. En een ontdooide Noordpool is een potentiele goudmijn voor boringen naar gas en olie. Verschillende landen hebben al claims gelegd op het gebied. Begin vorige eeuw werd de Noordpool gezien als de laatste niet ontdekte plek op het noordelijk halfrond. Een poolexpeditie was een levensgevaarlijke missie. Maar over zo'n 30 jaar kun je in de zomer dus gewoon in een bootje naar de Noordpool.

21. Een computerbestandje dat zich installeert op je computer, zonder dat je daar erg in hebt. En dat met documenten van jou aan de haal gaat. 't Is een computervirus en het heeft toegeslagen in een groot aantal overheidsinstellingen universiteiten en bedrijven. Opvallend veel gemeenten zijn het slachtoffer geworden. Op dit moment zijn al enkele duizenden computers besmet geraakt. De grote vraag is wat dit virus nog meer gaat doen. Ik verwacht wel dat het nog niet weg is. Dit virus heeft een nieuwe manier om zich te verspreiden maar dat doet het zo grof, dat netwerken uiteindelijk helemaal onderuit gaat. Ze manipuleren banktransacties. Het doel is dus bankrekeningen leeghalen. Maar het virus wordt verspreid door iets dat we allemaal gebruiken op de computer: Tekstverwerkingsbestanden. En daarom hebben gemeenten er ook last, zoals de gemeente Weert. Vandaag waren ze daar de hele dag bezig om alles weer aan de praat te krijgen. Normaal gaat het natuurlijk op de computer. Maar zoals u het nu ziet, doen we het op de typemachine. Oude tijden herleven op het gemeentehuis van Weert. De typemachines stonden nog op zolder en het oude MS-Dos wordt weer opgestart. Dit zijn de beeldschermen die sinds 1994 gebruikt worden om personen digitaal te registreren in het GBA. Dat kan ik niet zien, mevrouw. Wij hebben een computerstoring. Er lag nog ergens een oude laptop in de kast. Maar de computerschermen blijven op zwart. Het begon met een storing op een incidentele werkplek bij een medewerker. Die storing werd verholpen, ogenschijnlijk. Maar even later was de storing weer terug, nog zwaarder en van de ene naar de andere werkplek. En in de loop van woensdagmorgen zagen we dat er iets anders aan de hand was, een virus dus. Hier wordt gewerkt aan het herstel van ons computersysteem. Terondersteuning van de eigen IT-afdeling is er hulp van buitenaf ingehuurd. Ik denk dat we als overheid toch meer aandacht moeten besteden aan de beveiliging van onze digitale snelweg. In Weert worden vandaag geen rijbewijzen verstrekt. Noodpaspoorten wel. Het is lastig maar sommigen zien er ook de romantiek van in. Nou, het digitale ligt eruit. Het is weer handwerk met papiertjes. Papieren telefoongids. Ja, iets anders vind ik altijd leuk. Maar van het virus zijn ze in Weert nog niet verlost. Jeroen Wollaars, hoe ernstig is dit? Je ziet dat het ernstig is omdat onze kwetsbaarheid aantoont. En we weten niet hoe lang dat virus die computers besmet heeft. Het staat er misschien al maanden op. En in die tijd heeft het gedaan waarvoor het gemaakt is in. die tijd heeft het gedaan waarvoor het gemaakt is. Er zijn wachtwoorden en dergelijke naar een computer in Oekrane gestuurd. Opvallend, dit virus slaat vooral toe in Nederland? Ja. Kijk maar naar dit grafiekje 80% van de Nederlanders doet aan Internet bankieren. Dat is het dubbele van het Europese gemiddelde. Het is opvallend dat zoveel gemeenten tegen de lamp lopen. Maar dat geeft een vertekend beeld. De overheid loopt eerder tegen de lamp dan bedrijven. Nee. Mijn collega verslaggever ging kijken. Ik verwacht wel dat het nog niet weg is. Met name de karakteristiek waardoor we dat zien is dat een besmette computer pas echt actief wordt nadat-ie opnieuw opgestart is. En dat gebeurt typisch 's ochtends als mensen weer aan het werk gaan. En computers die vandaag besmet zijn bij mensen die op vakantie gaan en die komen over drie weken terug, die worden opnieuw opgestart en dan zie je dat effect weer. En moet ik nou, thuis, me ook zorgen maken om mijn eigen computer? Er is goed nieuws en slecht nieuws. Dit virus is nu bekend. Dus als je virus scanner bijgewerkt wordt het ontdekt. Maar het is een kat en muisspel. Dank, Jeroen Wollaars.

22. Een overval op een woning in Voorthuizen, een dorpje bij Barneveld. Niks vermoedend zaten de bewoners thuis, plotseling stonden er een paar mannen in de huiskamer. Er werd geweld gebruikt, zoals vaak bij overvallen op woningen. Er viel een dode. Een van de overvallers overleefde het niet. Gisteravond rond 22 uur dringen zeker twee mensen deze woning binnen. In het huis zijn op dat moment de bewoners aanwezig, een jong stel, beiden in de 30 en die zijn niet van plan hun bezittingen mee te geven. Beide bewoners raken gewond. Vandaag onderzoekt de politie wat er precies is gebeurd. Wie heeft er geschoten? De politie sluit niet uit dat de bewoners een wapen in huis hadden. Later op de avond wordt verderop in de straat een man gevonden. Zwaargewond. Uiteindelijk overleeft hij de avond niet. Op basis van wat we nu weten lijkt het erop dat deze man iets met het incident te maken heeft. Het gaat om een 27-jarige man uit Ede. Het aantal overvallen op woningen daalt, al zijn het er nog altijd zo'n 750 per jaar gemiddeld meer dan 2 per dag. Wat opvalt: Bij 40 procent van de overvallen raakt iemand gewond en soms gaat er zelfs iemand dood. En dat is meer dan bij overvallen op bedrijven. We zien dat dit soort overvallen vaak gewelddadiger verlopen. Want bewoners voelen zich thuis in hun eigen huis, worden plotseling overvallen, raken in paniek en dan ontstaat er ook van de overvaller uit een hoop geweld. Of de worsteling gisteravond heeft voorkomen dat er wat buit is gemaakt, is niet bekend. De politie heeft vooralsnog niemand voor de overval gearresteerd. De politie heeft ook nog niet bekendgemaakt hoeveel mensen worden gezocht voor de overval. De bewoners liggen nog in het ziekenhuis.

23. En dan, om meerdere redenen, een bijzonder verhaal uit China. Bijzonder, omdat het de laatste bijdrage is van onze correspondent Wouter Zwart vanuit China. Na 7 ,5 jaar verruilt-ie China voor de Verenigde Staten. Ik praat zo met hem. Dat bijzondere verhaal dus: Het proces tegen Gu Kailai, de vrouw van een van de machtigste politici van China. Mevrouw Gu wordt verdacht van moord. Maar volgens veel mensen speelt op de achtergrond een veel groter politiek steekspel Het is, hoe dan ook, een van grootste rechtszaken in China van de afgelopen 30 jaar. Onder weersomstandigheden die bijna symbolisch leken voor de politieke storm ging het proces van start. Achter gesloten deuren. Je zou deze rechtbank kunnen zien als het topje van een ijsberg vol intrige. Want vooral onder het oppervlak, buiten het zicht van burgers en journalisten speelt zich een drama af. De hoofdrolspeler: Gu Kailai, vrouw van Bo Xilai tot voor kort baas van de stad Chongching en kandidaat voor China's hoogste partijfuncties. Begin dit jaar vlucht Bo's rechterhand politiebaas Wang Lijun naar een Amerikaans consulaat en verklaart dat de machtige familie Bo verantwoordelijk is voor de dood van de Britse zakenman Neil Heywood: Die was jarenlang een vriend en zakenpartner van de familie Bo maar werd november vorig jaar vergiftigd aangetroffen in een hotel. wordt Gu Kailai beschuldigd van de moord. Gu Kailai lijkt al schuldig voor de rechtszaak is begonnen. Want ook Gu's man, Bo Xilai wordt ergens gevangengehouden en is ontdaan van zijn functies. Hoewel hij officieel alleen verdacht wordt van overtreding van partijregels zijn er wel degelijk verhalen over corruptie, machtsmisbruik en zelfs het afluisteren van partijleiders. De vraag is echter of ook Bo net als zijn vrouw vervolgd zal worden. Wouter, waarom is dit proces zo belangrijk? Omdat de politieke implicaties erg groot kunnen zijn. Komende herfst krijgt China. Nieuwe leiders. Deze man had kans daarbij te zitten. Daarom willen ze geen politiek proces rond hem. Dus waarschijnlijk zou hij alleen partij

disciplinair worden gestraft. Zo'n proces, op deze manier, was dat 7 ,5 jaar geleden toen jij voor ons begon in China nou ook mogelijk geweest? Deels wel. Het afronden van het proces in een dag is niet nieuw China. Dat gebeurt hier vrij vaak. Maar dat hier achteraf veel over wordt bericht in de media. Is wel heel bijzonder. In het verleden zou het in de doofstom worden gestopt. Maar dit is 2012. Het is een internationale gemeenschap aan het worden Met sociale media. Maar dat betekent dus dat de Chinezen die jij nu verlaat echt andere Chinezen zijn dan toen jij kwam? Wel een beetje. Alles verandert hier. Er wordt steeds groter en moderner. De veranderingen zijn meestal positief. Maar niet altijd. Zeven jaar geleden zag je een mooie passie jaar geleden zag je een mooie passie en een soort naviteit. Men wilde zich ontwikkelen en leren. Die naviteit begint plaats te maken voor een soort bezinning. En besef. En het besef dat je met die welvaart niet alleen mooie spullen kunt kopen Maar dat je ook macht en invloed kan kopen. Er is veel corruptie nou. Individualisme neemt toe. En er is een groeiend gat tussen arm en rijk. Dank, Wouter Zwart.

24. Een fabrikant van incontinentie-materiaal heeft in opdracht van apothekers patienten benaderd. Minister Schippers vindt het niet kunnen dat apothekers patientgegevens verstrekken aan bedrijven. Fabrikant Tena wilde weten wie wat voor incontinentie-materiaal nodig heeft om verspilling te voorkomen. Minister Schippers is het daarmee eens, maar vindt dat apothekers zich aan de privacy-wetten moeten houden. Onderzocht wordt nu, of Tena dat heeft gedaan.

25. Het sterftecijfer in Nederland is voor het eerst sterk gestegen. In de eerste zes maanden van dit jaar overleden ruim 72 .000 mensen. Dat zijn er 4000 meer dan in dezelfde periode vorig jaar. Er stierven vooral meer mensen op hoge leeftijd: Mannen ouder dan 80 jaar en vrouwen boven de 90 jaar. Het koude weer in februari speelde hierbij een rol, zegt het Centraal Bureau voor de Statistiek.

26. Warm was het ook in Arnhem, in het Openluchtmuseum. Daar startte de SP z'n verkiezingscampagne. omdat-ie een serieuze premierskandidaat is. Het risico op uitglijders is groot. Maar daar maakt zijn achterban zich geen zorgen om. 3, 2, 1, 0! Hier is het lekker in de schaduw. Ja, dat klopt. Als zelfs Emile Roemer zijn stropdas afdoet, dan is het officieel warm. De campagne is begonnen, Roemer wil regeren, maar dat moet dan wel heel anders dan bij vorige kabinetten. Een rijk land als Nederland dat zegt dat wij geen geld hebben om het wat eerlijker te verdelen en te zorgen wat van ons allemaal is, dat we daar wat zuiniger op zijn dan we de afgelopen 10 jaar gedaan hebben Kom op. Roemer trekt de aandacht, als mogelijke premier. Ook de buitenlandse pers wil wel eens weten welk beleid hij wil gaan voeren. Dat is niet moeilijk, vindt de SP leider: Puinruimen zal het worden. Ik loop daar niet voor weg. En ik doe ook een beroep op de samenleving om daar niet voor weg te lopen. Maar ik vraag wel een gunst aan heel veel mensen. Geef ons een kans om het waar te maken dat wij de komende jaren Nederland menselijker en socialer gaan maken. De anderen hebben hun kans gehad. Nu is het onze beurt. De vraag is wel welke partij met de SP wil samenwerken. Zeker nadat Roemer nogal enthousiast aangaf dat-ie zich niks aantrekt van eventuele Europese begrotingsboetes. De uitspraak 'Over my dead body', vond u dat een goede uitspraak? Was u het er mee eens? Wat daar achter zit wel. De uitspraak op zichzelf vond ik niet zo slim. Dat is ook leren! Het is helemaal niet erg als je een keer een fout maakt. Rutte maakte ook een fout van 50 miljard. We zijn allemaal maar mensen. Ik zeg niks meer op z'n Engels! Meeregeren lijkt dichterbij dan ooit. Dus de SP'ers vergeven het Roemer graag. Als-ie straks maar wel de verkiezingen wint.

27. Maandenlang was Julian Assange onzichtbaar, de oprichter van WikiLeaks. Maar vanmiddag was-ie na lange tijd weer in het openbaar te zien en te horen. Met 'n korte speech tot z'n aanhang en de verzamelde wereldpers. Dat deed-ie, veilig vanaf het balkon van de ambassade van Ecuador in Londen. Als-ie het op de stoep had gedaan, zoals eerder was aangekondigd, dan was hij direct opgepakt. Met enige bravoure maakte het fenomeen Assange vanmiddag zijn rentree op het wereldtoneel na een verblijf van twee maanden in de ambassade. Hij vroeg en kreeg hier politiek asiel, in een poging zijn uitlevering aan Zweden onmogelijk te maken. Assange is bang dat Zweden hem wegens zijn WikiLeaks-activiteiten doorstuurt naar de VS. door de voormalige Spaanse onderzoeksrechter Balthazar Garzon. Na de opwinding rond de ambassade is het er nu weer betrekkelijk rustig. Maar het toespraakje van Assange was waarschijnlijk nog maar een kleine aanloop naar nog heel veel meer diplomatieke onrust.

28. Een hobby die je niet meteen zou verwachten in de stad: Het houden van bijen. En toch is het een hobby die steeds populairder wordt. In Nederland zijn er ongeveer 8 .000 bijenhouders. Vooral in steden worden dat er steeds meer. Zomaar een plaats in de Utrechtse binnenstad. Hier wonen tienduizenden bijen op het dak van Jan Kees en Roderick. Ik vind het hartstikke leuk. Ik heb nog nooit zoveel huisdieren gehad waar ik zo weinig last van had. Het bijenvolk is van Lieke, maar zij had er geen plaats voor en mocht het hier neerzetten. Het aantal neemt snel toe. Dit is het echt heel leuk om te zien. De omstandigheden zijn dan ook gunstig: Warm weer en veel verschillende bloemen in de omgeving. Wil jij helpen? Oscar begon drie jaar geleden bijen te houden. Hier, boven op het dak in Amsterdam-West. Als je zo'n bijenvolk in ontwikkeling volgt, dat is fascinerend. En hun rol in bestuiving is ook machtig om daar van mee te krijgen, om dat te kunnen volgen. En de honing is lekker. Dirk houdt al 30 jaar bijen en leidt nu jonge imkers op, in een tuin, midden in Amsterdam-Noord. Dit is een tendens. Veel jonge mensen vinden het zeer interessant om toch met bijen bezig te zijn. Niet alleen vragen, maar ook doen. Tot zover het goeie nieuws, want het slechte nieuws is: Het gaat nog steeds niet goed met de bijen. Het aantal bijenvolken in de wereld neemt nog steeds af in plaats van toe. Bijen zijn van belang voor het bestuiven van planten en daardoor weer belangrijk voor ons voedsel. Maar sommigen zullen denken: moet dat nou per se in de stad? Soms gaat het mis, zoals vorig jaar bij Oscar. Ze hebben een maand geleden 2 korven op het dak gezet. Ze hadden beloofd dat we geen last zouden krijgen, maar je ziet het. Dat komt wel eens voor. Dat proberen we natuurlijk te voorkomen. Maar de bijen hielden zich niet aan de regels. En zo is nu eenmaal de natuur. De bijen werken aanstekelijk op de mannen. Daarom willen ze er volgend jaar nog een kast bij.

29. Deskundigen zeiden al dat het eraan zat te komen: Dat er pijnlijke keuzes moeten worden gemaakt door de gestegen kosten in de gezondheidszorg: Dit medicijn vergoeden we wel, en dat medicijn niet. Het College voor Zorgverzekeringen komt nu met zo'n advies: Stoppen met medicijnen voor twee zeldzame ziektes: de ziekte van Pompe en van Fabry. Dit is het letterlijke advies aan minister Schippers van Gezondheid: In de studio onze redacteur gezondheidszorg, Rinke van den Brink. Is dit het gevreesde omslagpunt? Daar lijkt het wel op Het is voor het eerst dat er wel een medicijn is dat werkt, maar waarvan gezegd wordt: We geven het niet. Gaat dit dan met meer ziekten gebeuren? deskundige en artsen verwachten dat wel Door de gestegen kosten van de zorg kan dit met meer medicijnen gebeuren zoals dure kankerbehandelingen en reuma. Patienten zullen dit natuurlijk verschrikkelijk vinden. Maar wat vinden artsen er van? Artsen vinden het verschrikkelijk dat dit gebeurd: Je mag een werkend geneesmiddel niet onthouden aan mensen die er baat bij hebben. Het gaat dus om de medicijnen voor twee zeldzame ziektes: van Pompe en Fabry. Dat zijn erfelijke aandoeningen. De ziekte van Pompe en Fabry zijn erfelijke en zeldzame aandoeningen. De patienten maken een bepaald enzym niet aan. In Nederland leiden 9 baby's aan pompe. Hun medicijn wil het College voor Zorgverzekeringen wel blijven vergoeden, want 't helpt. Anders ligt het volgens het CVZ bij 'n tweede groep. Dat zijn ruim 80 jongeren en volwassenen. Door verslapping van hun spieren raken ze invalide. Hun medicijn kost per persoon per jaar tussen de 400 .000 en 700 .000 per euro. Daar hebben ze baat bij, zeggen ze. Het CVZ vindt het effect onvoldoende en wil stoppen met vergoeden. 'De verhouding tussen kosten en baten is onacceptabel .' Bij fabry treft het advies ruim 60 patienten. Hun enzymtherapie werkt maar kost zo'n 200 .000 euro per patient per jaar. Het CVZ wil ook deze vergoeding schrappen. Dat betekent voor mensen met Fabry: de10 tot 20 jaar korter leven. Van Pompe: slechter leven, eerder invalide. Het is nog maar een voorlopig advies van het College voor Zorgverzekeringen. Bestaat er nog een kans dat dit advies wordt aangepast? Op papier wel. Maar alle voor en tegenargumenten zijn al op een rij gezet. Dus de kans dat er nog iets aan gaat veranderen, is denk ik heel klein. Dank je wel, Rinke. Als wordt besloten om die medicijnen tegen Pompe en Fabry inderdaad te stoppen dan denken deskundigen dat dit een voorbode is van meer. Je kunt niet zeggen: Alles wat mogelijkerwijs positief kan uitwerken kan gratis in het basispakket. Dat red je nooit. Er moeten keuzes worden gemaakt. Over welke medicijnen zal 't in de toekomst nog meer gaan? De beleidsagenda wordt gedomineerd door dure zaken. Dus dure geneesmiddelen zullen als eerste in het vizier zijn. Wilt u voorbeelden noemen? Dure geneesmiddelen die voor heel zeldzame ziektes worden toegepast. Maar ook dure geneesmiddelen bij bepaalde vormen van kanker. Hoe zal het met deze keuzes aflopen, Pompe en Fabry? Ik hoop dat na de adviesronde die nu gaande is niet alleen een advies van het CVZ komt maar dat er ook een commissie wordt ingeschakeld. Dat is nu de praktijk. naar de financiele aspecten, maar ook naar de ethische kant. En wat moet prevaleren? In het algemeen moet geld niet prevaleren in de zorg, vind ik. Minister Schippers neemt hierover een definitief besluit: Dat zal in de herfst zijn. Op onze site kunt u terecht voor meer informatie over dit advies: nos.

30. Het waterschap in Zuid-Limburg onderzoekt of er maatregelen nodig zijn om nieuwe overstromingen te voorkomen. De overstroming van de rivier de Gulp heeft flinke schade aangericht in Slenaken zo blijkt uit een eerste inventarisatie. De kelders van enkele horecazaken zijn veranderd in een puinhoop. Als je het riviertje de Gulp ziet, kun je bijna niet voorstellen dat hij gisteravond voor zo veel overlast zorgde. Door noodweer in Belgie: 60 mm regen in een half uur. Dat kon de rivier niet aan, hij trad buiten z'n oevers en zorgde hier in Slenaken voor veel schade. Het water drong ook naar binnen bij enkele vakantiewoningen. En auto's waren vaak ook niet bestand tegen de kracht

van zoveel water. En dan is het zondag. Hoe is het binnen? U mag komen kijken: EEN grote teringzooi. Je ziet het, he? Het is helemaal Hier is de diepvries. Alles was nieuwe voorraad. Schnitzels, stokbrood, konijnenbouten. Kijk, dit rek is stukgeslagen door de kracht van het water. Ik zit hier al een half jaar. Het hoogseizoen We hadden een volle productie. En alles gaat nu de container in. Hulp was er vandaag in overvloed, van alle kanten. En dat blijft zo, als dat nodig is. We gaan dat de komende periode in beeld brengen. En waar we wat kunnen doen, gaan we ook iets doen. Sommige ondernemers verwachten de rest van het hoogseizoen dicht te zijn.

A.2 *wiki_freq&tfidf*

1. (Frankrijk)1.0 (Auto)1.0 (Dader)1.0 (Discotheek)2.0 (Discotheken)1.0 (De Man)1.0

2. (Boete)4.0 (Euro)1.0 (Euro'S)1.0 (Overheid)1.0 (Staat)1.0 (Onderwijs)1.0 (Politiek)1.0 (Staat)1.0 (Rechter)2.0 (Lid)1.0 (Indianen)1.0 (Diploma)1.0 (Ouders)1.0 (Spel)1.0 (Spelen)2.0 (Jaar)5.0 (Jaren)2.0 (Student)1.0 (Studenten)5.0 (Verkiezingsprogramma'S)1.0 (Delen)1.0 (Vaarwel)1.0 (Zegen)1.0 (Bestuurder)1.0 (Vertraging)1.0 (Fulltime)1.0 (Duizend)1.0 (Termen)1.0 (Lid)1.0 (Zet)1.0 (Spelregels)1.0 (Coulissen)1.0 (Actief)1.0

3. (Marokkaanse)2.0 (Turkse)2.0 (Werkgever)1.0 (Werkgevers)2.0 (Arbeidsrecht)1.0 (Werkloosheid)1.0 (Nederlanders)1.0 (Opleiding)2.0 (Solliciteren)1.0 (Taal)1.0 (Procent)3.0 (Jaar)1.0 (Tijd)1.0 (Tijden)1.0 (Ministerie)1.0 (Advocaat)1.0 (Advocatenkantoor)1.0 (Ander)1.0 (Helpen)1.0 (Horen)1.0 (Idee)2.0 (Dertig)1.0 (Maanden)1.0 (Moet)1.0 (Gauw)1.0 (Weten)1.0 (Zich)1.0 (Gesprek)1.0 (Liet)1.0 (Zou)1.0 (De Heel)1.0 (Vind)1.0 (Surinaamse)2.0 (Uitgenodigd)1.0 (Richting)1.0 (Rode Loper)1.0

4. (Irak)1.0 (Zelfmoordaanslag)1.0 (Afghanistan)1.0 (Oppositie)4.0 (President)5.0 (Hoofdstad)1.0 (Minister Van Defensie)3.0 (Defensie)5.0 (Defensieminister)1.0 (Onderminister Van Defensie)2.0 (Minister)3.0 (Ministers)1.0 (Regering)2.0 (Delft)1.0 (Jaar)1.0 (Jaren)1.0 (Auto'S)1.0 (Damascus)2.0 (Aandeel)1.0 (Explosie)3.0 (Opgeblazen)2.0 (Begrafenis)1.0 (Best)1.0 (Aanslag)6.0 (Uur)2.0 (Dood)1.0 (Wet)2.0 (Communicatie)1.0 (Praten)1.0 (Staat)1.0 (Regime)3.0 (Foto)1.0 (Organisatie)1.0 (Ziekenhuis)2.0 (Merk)1.0 (Denken)1.0 (Lijfwacht)2.0 (Defensie)5.0 (Zorg)1.0 (Medische)1.0 (Sla)1.0 (Kennis)1.0 (Maribor)1.0 (Kant)3.0 (Hoorn)2.0 (Niets)1.0 (Yen)1.0 (Cyanide)1.0 (Tal)1.0 (Correspondent)1.0 (Devon)1.0 (Gaven)1.0 (Horen)1.0 (Gebouw)4.0 (Bellen)1.0 (Maanden)1.0 (Relatief)1.0 (Praten)1.0 (Moet)1.0 (De Maskers)1.0 (Buurt)2.0 (Wijk)1.0 (Gezicht)1.0 (Mast)1.0 (Ruimte)1.0 (Weten)2.0 (Kennis)1.0 (Gevolg)1.0 (Vluchtelingen)1.0 (Zagen)1.0 (Attent)1.0 (Torma)1.0 (Het Zijn)1.0 (Rest)1.0 (Irissen)1.0 (Zich)2.0 (Duizend)2.0 (Het Nieuws)1.0 (Reportage)1.0 (Zoenen)1.0 (Wedstrijd)1.0 (Maleiers)1.0 (De Loop)1.0 (Zou)6.0 (Zet)3.0 (De Bron)1.0 (Vind)1.0

5. (Kazachstan)1.0 (India)2.0 (Rusland)1.0 (Droogte)2.0 (Droog)1.0 (Tarwe)1.0 (Graan)1.0 (Afrika)1.0 (Euro)1.0 (Uithuizermeeden)1.0 (Marx)1.0 (Grondstof)2.0 (Grondstoffen)1.0 (Duitsland)1.0 (De)31.0 (Parijs)1.0 (Boeren)1.0 (Uithuizen)1.0 (Tijd)3.0 (Westen)1.0 (Voedsel)1.0 (Fabriek)1.0 (Maand)1.0 (Maanden)2.0 (Inkomen)1.0 (Verenigde Staten)1.0 (De Verenigde Staten)1.0 (Jaar)1.0 (Bedrijf)1.0 (Procent)2.0 (Molen)1.0 (Dichter)1.0 (Meel)1.0 (Bloem)1.0 (Bruin)1.0 (Schommelingen)2.0 (Dertig)1.0 (Maanden)2.0 (Moet)3.0 (Het Voor)1.0 (Meelfabriek)1.0 (Malen)1.0 (Droog)1.0 (Vroeg)1.0 (Koppel)1.0 (Moment)1.0 (Graden)1.0 (Akkers)1.0 (Frankrijk)1.0 (Tarik)1.0

6. (De Nederlandsche Bank)1.0 (Rente)5.0 (Londen)2.0 (Euro)1.0 (Rabobank)9.0 (Consument)1.0 (Geld)5.0 (Miljoen)1.0 (Hypotheken)1.0 (Staat)1.0 (Britse)2.0 (Engeland)1.0 (Londen)2.0 (Fraude)1.0 (Kantoor)1.0 (Hoofdkantoor)2.0 (Utrecht)1.0 (Handelaren)1.0 (Taal)1.0 (Mei)1.0 (Engeland)1.0 (Britse)2.0 (Horen)1.0 (Bellen)1.0 (Eten)1.0 (Toezichthouder)1.0 (Moet)1.0 (Ochtend)1.0 (Financieel)1.0 (Weten)1.0 (Kijk)1.0 (Kopje)1.0 (Het Voor)1.0 (Ontslag)1.0 (Relevant)1.0 (Bewijzen)1.0 (Ouder)1.0 (Zich)2.0 (Barclays)2.0 (Hoogte)2.0 (Buitenwereld)1.0 (Wereld)1.0 (De Banken)1.0 (Schandaal)2.0 (Zou)2.0 (Van Heel)1.0 (Hoofdkantoor)2.0 (Trekken)2.0 (Lief)1.0

7. (Miljard)1.0 (Euro)1.0 (Jaar)3.0 (Aandeel)1.0 (Luchtvaartmaatschappij)1.0 (Vakbonden)1.0 (Luchtvaart)1.0 (Luchtvaartmaatschappij)1.0 (Vliegen)1.0 (Tijd)2.0 (Tijden)1.0 (Kwartaal)3.0 (Normaal)1.0 (Staal)1.0 (Ruzie)1.0 (Praten)1.0 (Moet)3.0 (Negentien)2.0 (Sterven)1.0 (Donker)1.0 (Weten)1.0 (Zich)1.0 (Duizend)1.0 (Steeg)2.0 (Franse)1.0 (Frans)1.0 (Vind)1.0 (Kende)1.0

8. (Utrecht)1.0 (Kanaleneiland)1.0 (Burgemeester)1.0 (Stoffen)1.0 (Anand)1.0 (Flat)2.0 (Moet)1.0 (Sanering)1.0 (Opruimen)1.0 (Vroeg)1.0 (De Bewoners)1.0 (Mogen)1.0 (Deze Week)1.0 (Klaar)1.0

9. (Nederland)2.0 (Aandelen)1.0 (Winst)1.0 (Staal)1.0 (Bloed)1.0 (Sieraden)1.0 (Britse)1.0 (Engeland)1.0 (Engeland)1.0 (Britse)1.0 (Dokkum)1.0 (Thee)1.0 (Kraaien)1.0 (Cricket)1.0 (Vrijdag)1.0 (Normaal)1.0 (Verdachte)1.0 (Graal)1.0 (Indymedia)1.0 (Horen)1.0 (Speeksel)1.0 (Maanden)1.0 (Moet)1.0 (Wijken)1.0 (Weten)1.0 (Kopje)1.0 (Het Voor)1.0 (Spant)1.0 (Duizend)1.0 (Zou)1.0 (Hoofdkantoor)1.0

10. (Turkije)1.0 (India)3.0 (Noord-india)1.0 (Elektriciteit)1.0 (Stroom)3.0 (Stroomvoorziening)1.0 (Europese Unie)1.0 (Miljoen)1.0 (Corruptie)1.0 (Dominostenen)1.0 (Auto)1.0 (Geld)1.0 (Overheid)1.0 (Zomer)1.0 (Eeuw)1.0 (Bouw)1.0 (Bouwen)1.0 (Oosten)1.0 (Doris Day)1.0 (Horen)2.0 (Kuyper)1.0 (Bestuur)1.0 (Moet)1.0 (Gek)1.0 (Overtrek)1.0 (Donker)1.0 (Raar)1.0 (Dick Dees)1.0 (Weten)1.0 (Weet)1.0 (Oorzaak)1.0 (Jullie)1.0 (Hangar)1.0 (Tara Elders)1.0 (Zich)1.0 (Zou)2.0 (Rense Westra)1.0 (Viel)1.0

11. (Gevangenisstraf)1.0 (Politie)2.0 (Nederland)1.0 (Euro)1.0 (Woning)1.0 (Woningen)2.0 (Auto)1.0 (Thaise)1.0 (Hennep)1.0 (Scooters)1.0 (Omzet)1.0 (Justitie)2.0 (Handel)1.0 (Misdaad)1.0 (Crimineel)2.0 (Criminelen)5.0 (Criminele)2.0 (Huiszoeking)1.0 (Bedrijf)1.0 (Sieraden)1.0 (Munitie)1.0 (Geld)3.0 (Crimineel)2.0 (Criminelen)5.0 (Criminele)2.0 (Criminele)2.0 (Crimineel)2.0 (Criminelen)5.0 (Jaar)1.0 (Jaren)1.0 (Provincie)1.0 (Kant)1.0 (Doel)1.0 (Drugs)1.0 (Schuur)1.0 (Arrestatieteam)1.0 (De Niet)1.0 (Zagen)1.0 (Sportwagen)1.0 (Auto)1.0 (Zuur)1.0 (Tralies)1.0 (Duizend)1.0 (Muur)1.0 (Keren)1.0 (Wilbur)1.0 (Teelt)1.0 (Zou)2.0 (Actie)2.0 (Klaar)1.0

12. (Euro'S)1.0 (Voeren)1.0 (Plein)1.0 (Staat)1.0 (Vernielingen)2.0 (Schade)2.0 (Drugs)2.0 (Tijd)1.0 (Politie)4.0 (Agenten)2.0 (Draaien)1.0 (Ravage)1.0 (Eten)1.0 (Drank)1.0 (Onder Invloed)2.0 (Glaswerk)1.0 (Opruimen)1.0 (De Ruiten)1.0 (Rotterdam)1.0 (Stadhuisplein)2.0

13. (Terreurdaden)1.0 (Afrikaanse Unie)1.0 (Japan)2.0 (Kenia)3.0 (Keniaanse)4.0 (Staken)1.0 (Ambassade)1.0 (Christenen)1.0 (Moskee)1.0 (Moord)2.0 (Vermoorden)1.0 (Politie)2.0 (Politie-eenheden)1.0 (Moslims)4.0 (Commissie)1.0 (Brand)1.0 (Preek)1.0 (Wet)1.0 (Wetten)1.0 (Grote)1.0 (Toeristen)1.0 (Jaar)1.0 (Zondag)1.0 (Overheid)2.0 (Geweld)3.0 (Maandag)2.0 (Vrijdag)2.0 (Soldaten)1.0 (Leger)1.0 (Eenheden)1.0 (Stad)1.0 (Steden)1.0 (Havenstad)2.0 (Kerken)1.0 (Niets)2.0 (Gesloten)1.0 (Havenstad)2.0 (Winkels)1.0 (Maanden)1.0 (Moet)2.0 (Wijken)1.0 (Zeggen)1.0 (Banden)1.0 (Ambon)1.0 (Fout)1.0 (Weten)1.0 (Vlam)1.0 (Mombasa)4.0 (Het Verleden)1.0 (Broere)1.0 (Kosten)1.0 (Zou)1.0 (De Roo)1.0 (Deze Week)2.0 (Staan)1.0

14. (Nederland)2.0 (Nederlands)1.0 (Miljard)1.0 (Miljoen)3.0 (Euro)4.0 (Euro'S)1.0 (Winst)2.0 (Mis)2.0 (Overheid)1.0 (Staat)2.0 (Bach)1.0 (Duizend)2.0 (Minister)1.0 (Utrecht)2.0 (Europese)1.0 (Jaar)1.0 (Jaren)1.0 (Geld)4.0 (Kapitaal)1.0 (Kosten)1.0 (Staat)2.0 (Verlies)1.0 (Verliezen)1.0 (Eiland)1.0 (Nederlands)1.0 (Nederland)2.0 (Papier)1.0 (Absoluut)1.0 (Ministerie)1.0 (Florida)1.0 (Niets)1.0 (Idee)1.0 (Praten)1.0 (Moet)2.0 (Zeggen)2.0 (Oorzaak)1.0 (Vrijmaken)1.0 (Dochters)1.0 (Zagen)1.0 (Zich)1.0 (Zwitserleven)1.0 (Inclusief)1.0 (Het Verleden)1.0 (Kosten)1.0 (De Banken)1.0 (Zakenbank)1.0 (Zou)2.0 (Franse)1.0 (Hoofdkantoor)1.0 (De Weer)1.0 (Deze Week)1.0 (Staan)2.0 (Streven)1.0 (Overeind)1.0

15. (Wilson)1.0 (Idee)1.0 (Deur)1.0 (Zich)1.0 (Het Vaderland)1.0

16. (Spanje)1.0 (Bosbranden)1.0 (Uur)1.0 (Brandweerlieden)1.0 (Jaar)1.0 (De Bergen)1.0

17. (Nederland)2.0 (Dyslexie)1.0 (Dyslectisch)1.0 (Hoogbegaafdheid)1.0 (Hoogbegaafd)1.0 (Kind)2.0 (Kinderen)3.0 (Onderwijs)1.0 (School)4.0 (Les)1.0 (Ouders)3.0 (Hoogbegaafd)1.0 (Thuisonderwijs)1.0 (Rechter)2.0 (Kant)1.0 (Spel)1.0 (Verbod)1.0 (Verboden)1.0 (Vinger)1.0 (Aandacht)1.0 (Haarlemse)1.0 (Jaar)5.0 (Staat)1.0 (Meisje)1.0 (Spanningen)1.0 (Zeggen)1.0 (Principe)1.0 (Rechtbank)1.0 (Rechter)2.0 (Dyslectisch)1.0 (Term)1.0 (Klaassen)1.0 (Buitenland)1.0 (Wateren)1.0 (Zou)1.0 (Thuis)1.0 (Wilden)1.0 (Territoriale)1.0 (De Familie)1.0 (Woonhuis)1.0

18. (Moskou)1.0 (Russische)1.0 (Rus)1.0 (Sekte)3.0 (Staat)1.0 (Land)1.0 (Islamitische)2.0 (Stad)1.0 (Grote)1.0 (Jaar)2.0 (Jaren)1.0 (Tijd)1.0 (Samenleving)1.0 (Gezag)1.0 (Gezondheid)1.0 (God)1.0 (Kilometer)1.0 (Vonken)1.0 (Profeten)1.0 (Gebouw)1.0 (Ban)1.0 (In De Ban)1.0 (Zagen)1.0 (Ondergrond)1.0 (De Nieuwe)1.0

19. (Huisarts)2.0 (Vermoeidheid)1.0 (Gezondheidszorg)1.0 (Bacterie)4.0 (Internet)1.0 (Belasting)1.0 (Complicatie)1.0 (Antistoffen)1.0 (Onderzoek)2.0 (Onderzocht)1.0 (Ziekenhuis)1.0 (Denken)2.0 (Denk)1.0 (Rekenen)1.0 (Duitse)1.0 (De)22.0 (Wet)1.0 (Huizen)2.0 (Bossen)1.0 (Bouw)1.0 (Eilanders)1.0 (Volksgezondheid)1.0 (Weten)1.0 (Weet)1.0 (Geiten)1.0 (Baia Mare)1.0 (Het Zijn)1.0 (Verschijnsel)1.0 (Zich)2.0 (Duizend)4.0 (Steken)1.0 (Besmet)4.0 (Zou)1.0 (Boel)1.0

20. (Aarde)1.0 (Zon)1.0 (Noordpool)7.0 (Noord)1.0 (Zomer)2.0 (Zomers)1.0 (Nederland)1.0 (Jaar)4.0 (Zeewater)1.0 (Water)1.0 (Kilometer)1.0 (Weer)1.0 (Maand)1.0 (Eeuw)2.0 (Water)1.0 (Ijs)1.0 (Vakantie)1.0 (Wiegel)1.0 (Trouw)1.0 (September)1.0 (Niets)1.0 (Draaien)1.0 (Dertig)2.0 (Negentien)1.0 (Olie)1.0 (Macht)1.0 (Zich)1.0 (Duizend)2.0

21. (Computervirus)1.0 (Virussen)1.0 (Virus)8.0 (Nederland)3.0 (Nederlanders)1.0 (Geld)1.0 (Slachtoffer)1.0 (Aids)1.0 (Rekenen)1.0 (Virus)8.0 (Virussen)1.0 (Tijd)1.0 (Tijden)1.0 (Computer)5.0 (Computers)3.0 (Computersysteem)1.0 (Overheid)2.0 (Jaar)1.0 (Verkeer)1.0 (Denemarken)1.0 (Digitale)1.0 (Digitaal)1.0 (Vakantie)1.0 (Bestanden)1.0 (Internetbankieren)1.0 (Rijbewijs)1.0 (Afdeling)1.0 (Zorg)1.0 (Herstel)1.0 (Normaal)1.0 (Laptop)1.0 (Bui)1.0 (Nier)1.0 (Graffiti)1.0 (Niets)1.0 (Doel)1.0 (Kast)1.0 (Best)1.0 (Bedrijven)2.0 (Ander)1.0 (Lamp)2.0 (Eik)1.0 (Doek)1.0 (Criminelen)2.0 (Criminele)1.0 (Maanden)1.0 (Praten)1.0 (Verzamelen)1.0 (Moet)2.0 (Ochtend)1.0 (Negentien)1.0 (Aandacht)1.0 (Weten)2.0 (Interpreteren)1.0 (Klappen)1.0 (Zagen)1.0 (Het Zijn)1.0 (Topje)1.0 (Zich)3.0 (Het Boek)1.0 (Besmet)4.0 (Jou)1.0 (Zou)3.0 (Europees)1.0 (Zet)1.0 (Karakters)1.0 (Thuis)1.0 (Herstel)1.0 (Actief)1.0 (De Haal)1.0 (Storen)1.0 (Meldpunt)1.0

22. (Dood)2.0 (Woning)4.0 (Huis)1.0 (Woningen)2.0 (Bewoners)6.0 (Thuis)2.0 (Cyprus)1.0 (Procent)1.0 (Jaar)1.0 (Verdachte)1.0 (Geweld)2.0 (Gewelddadige)1.0 (Politie)2.0 (Uur)1.0 (Bedrijven)1.0 (Woord)1.0 (Rome)1.0 (Barneveld)1.0 (Ede)1.0 (Ziekenhuis)2.0 (Tien)1.0 (Zich)1.0 (Huiswerk)1.0 (Gesprek)1.0 (Thuis)2.0

23. (Peru)1.0 (China)8.0 (Chinees)1.0 (Chinezen)1.0 (Peking)1.0 (India)1.0 (Britse)1.0 (Corruptie)2.0 (Machtsmisbruik)1.0 (Moord)2.0 (Dood)1.0 (Communistisch)1.0 (Westen)1.0 (Macht)1.0 (Politieke)1.0 (Rechtszaak)1.0 (Rechtszaken)1.0 (Proces)6.0 (Zaken)2.0 (Wenen)1.0 (Politiek)6.0 (Politieke)1.0 (Politici)1.0 (Media)2.0 (Percentage)1.0 (Individualisme)1.0 (Verenigde Staten)1.0 (De Verenigde Staten)1.0 (Britse)1.0 (Land)1.0 (Welvaart)1.0 (Tolwegen)1.0 (Dood)1.0 (Grote)1.0 (Jaar)6.0 (Jaren)1.0 (Herfst)1.0 (Beroep)1.0 (Vriend)1.0 (Politici)1.0 (Correspondent)1.0 (Rijkdom)1.0 (Tijd)1.0 (Hongarije)1.0 (Leiden)1.0 (Stad)1.0 (November)1.0 (Leken)1.0 (Zoen)1.0 (Held)1.0 (Gesloten)1.0 (Dertig)1.0 (Hotel)1.0 (Bellen)1.0 (Steekspel)1.0 (Overste)1.0 (Kopen)1.0 (Moet)2.0 (Bericht)1.0 (Zeggen)3.0 (Hen)1.0 (Dader)1.0 (Schil)1.0 (Zicht)1.0 (Leiderschap)1.0 (Journalisten)1.0 (Gondel)1.0 (Zich)4.0 (Duizend)1.0 (Implicatie)1.0 (Achter Gesloten Deuren)1.0 (Sri)1.0 (Vergiftigd)1.0 (Het Verleden)1.0 (Merken)1.0 (Zou)2.0 (Kayla)1.0 (Actief)1.0 (Partijleiders)1.0 (Gevangen)1.0 (Intrige)1.0 (De Familie)1.0 (Voor De Dood)1.0

24. -

25. (Nederland)1.0 (Zeventig)1.0 (Tachtig)1.0 (Negentig)1.0 (Meer)2.0 (Sterftecijfer)1.0 (Mannen)1.0 (Jaar)1.0 (Februari)1.0 (Ouder)1.0 (Statistiek)1.0 (Eerste)2.0 (Rode)1.0

26. (Nederland)2.0 (Land)1.0 (Rijk)1.0 (Staat)2.0 (Land)1.0 (Premiers)1.0 (Verkiezingen)1.0 (Miljard)1.0 (Europese)1.0 (Begroting)1.0 (Jaar)3.0 (Jaren)1.0 (Samenleving)1.0 (Beleid)1.0 (Kant)1.0 (Beroep)2.0 (Weten)1.0 (Engelse)1.0 (Delen)1.0 (Kabinet)1.0 (Voeren)1.0 (Uitspraak)2.0 (Stropdas)1.0 (Moet)2.0 (Slee)1.0 (Aandacht)1.0 (Risico)1.0 (Fout)2.0 (Het Voor)1.0 (Zich)2.0 (Emile Roemer)1.0 (Roemer)2.0 (Belangstelling)1.0

27. (Londen)2.0 (Zweden)3.0 (Spaanse)1.0 (Politie)1.0 (Ambassade)1.0 (Juliana)1.0 (Jaar)1.0 (Heilig)1.0 (Tijd)1.0 (Openbaar)1.0 (Ecuador)1.0 (Dozijn)1.0 (Moet)1.0 (Uitlevering)1.0 (Onderzoeksrechter)1.0 (Het Zijn)1.0 (Fenomeen)1.0 (Poging)1.0 (Keren)1.0 (Zou)1.0 (Onzichtbaar)1.0 (Speech)1.0 (Staan)1.0 (Andreu)1.0 (Bravoure)1.0

28. (Stad)2.0 (Nederland)1.0 (Nederland)1.0 (Leiden)1.0 (Utrechtse)1.0 (Stad)2.0 (Mis)1.0 (Planten)1.0 (Voedsel)1.0 (Jaar)5.0 (Tijd)1.0 (Jongen)1.0 (Jonge)1.0 (Bestuiving)1.0 (Bestuiven)1.0 (Hobby)2.0 (Stad)2.0 (Steden)1.0 (Echt)1.0 (Bijen)1.0 (Bij)6.0 (Dertig)1.0 (Moet)1.0 (Denken)1.0 (Zeggen)1.0 (Naaien)1.0 (Huisdieren)1.0 (Zich)1.0 (Duizend)1.0 (Beesten)1.0 (Zou)2.0 (Vind)1.0 (Strook)1.0

29. (Medicijnen)6.0 (Medische)1.0 (Medisch)1.0 (Gezondheidszorg)1.0 (Zorg)2.0 (Medische)1.0 (Euro)2.0 (Ethische)1.0 (Ethisch)1.0 (Nederland)1.0 (Informatie)1.0 (Geld)1.0 (Kosten)3.0 (Medicijn)3.0 (Medicijnen)6.0 (Werkelijkheid)1.0 (Ziekte)8.0 (Ziekten)4.0 (Erfelijke)2.0 (Procent)1.0 (Effectiviteit)1.0 (Rechtvaardigheid)1.0 (Kosten)3.0 (Beleid)1.0 (Fabriek)2.0 (Recht)1.0 (Verlosser)1.0 (Commissie)2.0 (Staat)1.0 (Jaar)3.0 (Medicijnen)6.0 (Medicijn)3.0 (Geneesmiddelen)4.0 (Zorgverzekeringen)2.0 (Therapie)1.0 (Arbeidsrecht)1.0 (Herfst)1.0 (Leiden)1.0 (Voeren)1.0 (Verzekeringen)1.0 (Kans)1.0 (Ongeluk)1.0 (Concept)1.0 (Relatief)1.0 (Praten)1.0 (Dief)1.0 (Moet)2.0 (Medische)1.0 (Medisch)1.0 (Verhouding)1.0 (Zeggen)1.0 (Buitengewoon)1.0 (Positief)1.0 (Moeders)1.0 (Alarm)1.0 (Het Voor)1.0 (Verdriet)1.0 (Anderen)1.0 (Persoon)1.0 (Tweede)1.0 (Ziekten)4.0 (Struck)1.0 (Duizend)2.0 (Helpt)2.0 (Termen)1.0 (Volwassenen)1.0 (Vroeg)1.0 (Het College)4.0 (Hou)1.0 (Zou)1.0 (Zet)1.0 (Mogen)1.0

30. (Rivier)1.0 (Riviertje)1.0 (Overstroming)1.0 (Overstromingen)1.0 (Uur)1.0 (Vakantie)1.0 (Middernacht)1.0 (Water)1.0 (Zondag)1.0 (Regen)1.0 (Weer)1.0 (Gulp)1.0 (Lieve)1.0 (Maanden)1.0 (De Kracht)1.0 (Kijk)1.0 (Schade)1.0 (Slenaken)1.0 (Rest)1.0 (Overlast)1.0 (De Kelders)1.0

A.3 *wiki_freq&sim&tfidf*

1. (Frankrijk)0.51 (Dader)0.32 (Discotheek)0.77 (Discotheken)0.38

2. (Boete)1.62 (Euro)0.26 (Euro'S)0.26 (Overheid)0.28 (Staat)0.28 (Onderwijs)0.27 (Politiek)0.22 (Staat)0.25 (Rechter)0.53 (Lid)0.27 (Indianen)0.24 (Ouders)0.22 (Spelen)0.38 (Jaar)0.91 (Jaren)0.36 (Student)0.2 (Studenten)1.02 (Zegen)0.22

3. (Marokkaanse)0.67 (Turkse)0.56 (Werkgever)0.28 (Werkgevers)0.56 (Arbeidsrecht)0.26 (Werkloosheid)0.27 (Nederlanders)0.23 (Opleiding)0.39 (Taal)0.2 (Procent)0.57 (Tijden)0.05 (Advocatenkantoor)0.19 (Idee)0.12 (Surinaamse)0.43

4. (Irak)0.32 (Zelfmoordaanslag)0.32 (Afghanistan)0.3 (Oppositie)1.12 (President)1.45 (Hoofdstad)0.24 (Minister Van Defensie)0.55 (Defensie)0.92 (Defensieminister)0.18 (Onderminister Van Defensie)0.37 (Minister)0.66 (Ministers)0.22 (Regering)0.45 (Jaar)0.22 (Jaren)0.22 (Damascus)0.59 (Explosie)0.71 (Opgeblazen)0.47 (Aanslag)1.15 (Uur)0.2 (Wet)0.32 (Praten)0.18 (Staat)0.23 (Regime)0.68 (Ziekenhuis)0.29 (Lijfwacht)0.27 (Defensie)0.9 (Medische)0.18 (Kant)0.21 (Hoorn)0.42 (Weten)0.11 (Kennis)0.06

5. (Kazachstan)0.21 (India)0.54 (Rusland)0.22 (Droogte)0.41 (Droog)0.2 (Tarwe)0.22 (Graan)0.23 (Euro)0.2 (Grondstof)0.43 (Grondstoffen)0.21 (De)4.41 (Tijd)0.42 (Westen)0.21 (Fabriek)0.2 (Maanden)0.24 (De Verenigde Staten)0.13 (Procent)0.21 (Bloem)0.09 (Schommelingen)0.15

6. (De Nederlandsche Bank)0.42 (Rente)2.15 (Londen)0.48 (Euro)0.36 (Rabobank)4.6 (Consument)0.3 (Geld)1.83 (Miljoen)0.25 (Hypotheken)0.4 (Staat)0.26 (Britse)0.32 (Engeland)0.16 (Londen)0.32 (Kantoor)0.29 (Hoofdkantoor)0.58 (Utrecht)0.23 (Britse)0.22 (Barclays)0.15

7. (Miljard)0.23 (Euro)0.36 (Jaar)1.03 (Aandeel)0.23 (Vakbonden)0.26 (Luchtvaartmaatschappij)0.17 (Vliegen)0.17 (Tijd)0.46 (Tijden)0.23 (Kwartaal)0.97 (Moet)0.19 (Negentien)0.12 (Frans)0.18

8. (Utrecht)0.44 (Kanaleneiland)0.4 (Flat)0.48 (De Bewoners)0.21 (Deze Week)0.22

9. (Nederland)0.71 (Engeland)0.19 (Britse)0.19 (Dokkum)0.21 (Indymedia)0.21

10. (Turkije)0.27 (India)1.04 (Noord-india)0.35 (Stroom)0.54 (Stroomvoorziening)0.18 (Europese Unie)0.25 (Bouwen)0.09 (Weet)0.12

11. (Gevangenisstraf)0.3 (Politie)0.78 (Nederland)0.25 (Euro)0.26 (Woning)0.27 (Woningen)0.54 (Auto)0.25 (Thaise)0.24 (Justitie)0.67 (Handel)0.21 (Misdaad)0.25 (Crimineel)0.49 (Criminelen)1.23 (Criminele)0.49 (Huiszoeking)0.25 (Munitie)0.24 (Geld)0.51 (Crimineel)0.46 (Criminelen)1.16 (Criminele)0.46 (Criminele)0.61 (Crimineel)0.61 (Criminelen)1.52 (Jaren)0.12 (Drugs)0.21 (Auto)0.16

12. (Vernielingen)0.48 (Schade)0.33 (Drugs)0.31 (Politie)1.29 (Agenten)0.64 (Onder Invloed)0.32 (Rotterdam)0.24 (Stadhuisplein)0.19

13. (Terreurdaden)0.33 (Afrikaanse Unie)0.3 (Japan)0.65 (Kenia)1.49 (Keniaanse)1.98 (Staken)0.25 (Ambassade)0.31 (Christenen)0.27 (Moskee)0.25 (Moord)0.57 (Vermoorden)0.28 (Politie)0.5 (Politie-eenheden)0.25 (Moslims)1.06 (Commissie)0.24 (Brand)0.24 (Wetten)0.15 (Toeristen)0.22 (Overheid)0.26 (Geweld)0.51 (Vrijdag)0.28 (Soldaten)0.21 (Leger)0.21 (Steden)0.16 (Havenstad)0.31 (Kerken)0.23 (Havenstad)0.31 (Mombasa)1.17 (Deze Week)0.16

14. (Nederland)0.56 (Nederlands)0.28 (Miljoen)0.42 (Euro)0.75 (Euro'S)0.19 (Winst)0.16 (Mis)0.33 (Staat)0.27 (Duizend)0.14 (Utrecht)0.53 (Jaren)0.18 (Geld)0.6 (Kapitaal)0.15 (Kosten)0.15 (Staat)0.28 (Verliezen)0.05 (Nederland)0.29

15. (Het Vaderland)0.52

16. (Spanje)0.45 (Bosbranden)0.27 (Brandweerlieden)0.25 (Jaar)0.24 (De Bergen)0.34

17. (Nederland)0.69 (Dyslexie)0.42 (Dyslectisch)0.42 (Hoogbegaafdheid)0.34 (Hoogbegaafd)0.34 (Kind)0.51 (Kinderen)0.76 (Onderwijs)0.25 (School)1.01 (Les)0.25 (Ouders)0.77 (Hoogbegaafd)0.23 (Rechter)0.2 (Verboden)0.07 (Aandacht)0.23 (Jaar)0.56 (Rechter)0.19 (Dyslectisch)0.29

18. (Moskou)0.25 (Russische)0.26 (Rus)0.26 (Sekte)1.06 (Staat)0.22 (Land)0.22 (Islamitische)0.51 (Stad)0.2 (Jaar)0.35 (Jaren)0.18 (Gezag)0.21 (God)0.24 (Profeten)0.22 (Ban)0.26 (In De Ban)0.26

19. (Huisarts)0.55 (Bacterie)0.73 (Onderzoek)0.19 (Onderzocht)0.09 (Denken)0.13 (Denk)0.07 (De)2.39 (Huizen)0.19 (Besmet)0.37

20. (Aarde)0.46 (Zon)0.42 (Noordpool)4.91 (Noord)0.7 (Zomer)0.91 (Zomers)0.46 (Nederland)0.24 (Jaar)1.61 (Zeewater)0.33 (Water)0.33 (Kilometer)0.24 (Weer)0.34 (Eeuw)0.15 (Water)0.26 (Ijs)0.26 (Olie)0.38 (Macht)0.34

21. (Computervirus)0.21 (Virussen)0.21 (Virus)1.7 (Nederland)0.84 (Nederlanders)0.28 (Virus)1.13 (Virussen)0.14 (Tijden)0.09 (Computer)1.11 (Computers)0.66 (Computersysteem)0.22 (Overheid)0.34 (Digitaal)0.11 (Herstel)0.11 (Bedrijven)0.18 (Criminelen)0.21 (Criminele)0.1 (Weten)0.12 (Besmet)0.29

22. (Dood)0.38 (Woning)0.61 (Huis)0.15 (Woningen)0.31 (Bewoners)0.92 (Thuis)0.31 (Cyprus)0.21 (Jaar)0.29 (Geweld)0.28 (Gewelddadige)0.14 (Politie)0.51 (Woord)0.24 (Ziekenhuis)0.22 (Tien)0.29 (Thuis)0.52

23. (Peru)0.27 (China)3.88 (Chinees)0.49 (Chinezen)0.49 (Peking)0.38 (India)0.35 (Britse)0.29 (Corruptie)0.53 (Machtsmisbruik)0.22 (Moord)0.49 (Dood)0.24 (Communistisch)0.3 (Westen)0.27 (Macht)0.23 (Politieke)0.23 (Rechtszaak)0.23 (Rechtszaken)0.23 (Proces)1.38 (Zaken)0.46 (Wenen)0.25 (Politiek)1.3 (Politieke)0.22 (Politici)0.22 (Media)0.37 (Verenigde Staten)0.21 (De Verenigde Staten)0.21 (Britse)0.21 (Land)0.21 (Welvaart)0.21 (Grote)0.2 (Jaar)1.41 (Jaren)0.23 (Hongarije)0.25 (Zeggen)0.26

24. -

25. (Nederland)0.56 (Zeventig)0.24 (Tachtig)0.25 (Negentig)0.26 (Meer)0.33 (Sterftecijfer)0.35 (Mannen)0.26 (Statistiek)0.27 (Eerste)0.37 (Rode)0.2

26. (Nederland)0.7 (Land)0.35 (Rijk)0.35 (Staat)0.42 (Land)0.21 (Premiers)0.28 (Europese)0.22 (Jaar)0.72 (Jaren)0.24 (Beroep)0.14 (Stropdas)0.21 (Emile Roemer)0.21 (Roemer)0.41

27. (Londen)0.69 (Zweden)1.27 (Spaanse)0.29 (Politie)0.22 (Ambassade)0.24 (Jaar)0.25 (Ecuador)0.25

28. (Stad)0.75 (Nederland)0.37 (Nederland)0.31 (Leiden)0.28 (Utrechtse)0.3 (Stad)0.6 (Voedsel)0.21 (Jaar)0.94 (Jonge)0.1 (Bestuiven)0.13 (Hobby)0.21 (Stad)0.29 (Steden)0.14 (Bij)0.88 (Naaien)0.22

29. (Medicijnen)2.17 (Medische)0.36 (Medisch)0.36 (Gezondheidszorg)0.32 (Zorg)0.65 (Medische)0.32 (Euro)0.57 (Ethische)0.28 (Ethisch)0.28 (Nederland)0.23 (Informatie)0.27 (Kosten)0.38 (Medicijn)1.37 (Medicijnen)2.74 (Werkelijkheid)0.27 (Ziekte)3.22 (Ziekten)1.61 (Erfelijke)0.62 (Kosten)0.55 (Fabriek)0.28 (Commissie)0.56 (Staat)0.22 (Jaar)0.33 (Medicijnen)2.06 (Medicijn)1.03 (Geneesmiddelen)1.37 (Zorgverzekeringen)0.56 (Therapie)0.3 (Leiden)0.24 (Verzekeringen)0.26

30. (Rivier)0.21 (Riviertje)0.21 (Overstromingen)0.2 (Regen)0.2 (Gulp)0.39 (Slenaken)0.38

A.4 *LDA_query*

1. daarop 0.2 frankrijk 4.91 vandoor 0.27 haalde 0.02 gewond 2.75 buiten 0.61 geraakt 0.26 auto 20.27 schietpartij 0.89 vlakbij 0 plaats 0.65 brengen 0.05 beschoten 0.02 afgelopen 4.12 ging 4.36 mochten 0 boos 0.02 aangehouden 9.33 binnen 3.18 dader 5.34 discotheek 0 nacht 0.51

2. oordeel 0.98 overheid 12.8 vooraf 0.09 ouders 42.81 spel 0.08 bitter 0 rekening 1.9 nee 16.74 aanspraak 0.01 stapte 0.02 termen 0.01 waarin 3.66 bovenop 0.08 volgend 1.02 jan 0.64 studie 9.72 bang 0.3 inzetten 0.05 verkeerde 0.32 brengen 0.95 boete 9.78 enige 1.99 zestig 0.02 politiek 0.38 behaald 0.01 reden 1.78 duizenden 0.16 gedurende 0.04 begonnen 0.14 verstaan 0.01 groep 8.17 betalen 19.71 jaren 33.38 moment 8.58 thans 0 terug 2.47 delen 0.32 partijen 0.72 ziek 0.06 hele 10.17 diploma 13.41 onderwijs 646.08 miljarden 0.21 teleurgesteld 0.02 vertraging 0.01 spelregels 0.01 jaar 382.65 spelen 7.29 beloond 0.01 bestuurder 0.35 fulltime 0.04 duizend 0.01 pot 0.03 bestuderen 0.01 hoelang 0 studenten 252.22 tijdens 16.66 alweer 0.28 weet 47.3 schrijven 0.24 blij 0.52 student 3.24 verwachten 0.09 gekozen 0.19 indianen 0 aangegeven 0.02 voltooid 0 erg 4.95 probeerden 0 onbeperkt 0 uitzondering 0.09 lange 0.65 bent 2.5 daarom 17.38 vraag 19.82 volgen 2.76 opleggen 0.18 aantal 18.9 strikte 0.01 drieduizend 0 actief 0.1 voortaan 0.22 wijsbegeerte 0 vastgesteld 0.07 volle 0.19 oplopen 0.05 afgezien 0 vaarwel 0 zegen 0.01 duurt 0.12 gelden 0.18 lid 0.25 coulissen 0

3. solliciteren 0.01 nemen 8.84 werkgever 4.24 cijfers 4.05 dertig 0.13 idee 4.01 ontvangen 0.11 vakken 0.34 advocaat 2.91 blijkt 12.49 opleiding 5.64 el 0.21 reageren 0.07 meespeelt 0 overal 0.59 waarin 3.99 naarmate 0.02 tijdelijke 0.48 dienst 1.38 richting 2.56 beiden 0.01 hoog 2.5 helemaal 8.55 vorig 77.18 recente 0.1 jongeren 22.53 onvoldoende 0.25 allochtone 0.02 procent 163.91 afkomst 0.18 aangenomen 0.07 banen 3.47 horen 0.59 marokkaanse 0.73 handen 0.3 loper 0.3 tientallen 0.07 markt 20.72 werknemers 77.08 klanten 3.94 zelfs 25.67 jaar 1456.65 surinaamse 0.01 stijging 1 denkt 2.32 natuurlijk 19.17 opvallend 0.18 afgelopen 18.35 groepen 0.38 werkloosheid 0.18 turkse 0.68 gesprek 0.17 allemaal 12.65 maanden 6.07 werden 6.33 ooit 2.18 autochtone 0.07 tijden 1.03 bepaalde 0.32 hol 0 vondst 0.02 liet 0.28 crisis 4.97 werkgevers 19.91 ministerie 3.66 verschil 2.32 zaken 4.35 china 30.52 geven 9.46 opgelopen 0.01 soms 8.29 rust 0.08 helpen 0.8 functies 0.01 achtergrond 0.05 zwaar 0.24 weten 6.74 vraag 19.87 evenals 0.01 melden 0.11 nederlanders 99.63 naast 0.35 vallen 1.4 autochtonen 0 werkzoekenden 0.02 taal 1.1 uitgenodigd 0.01 ruim 6.25 werk 34.21 gauw 0 twintig 0.45 opgepakt 1.12 junior 0.05 sociale 18.11 kabinet 7.25 moeilijk 3.06 advocatenkantoor 0 makkelijk 0.13 kiezen 2.51 smaak 1.8 begin 1.06 bood 0

4. president 57.86 leek 8.36 leed 0.04 buurt 1.3 ophoudt 0 ministers 0.05 sander 0.01 afgenomen 0 ver 0.4 zwarte 0.34 hoeveel 0.23 woord 3.47 zeer 0.53 waarin 2.61 kerk 20.42 verloor 0.19 morgen 0.3 gevechten 0.53 assad 10.38 defensie 8.46 rook 0.1 lichte 0.01 moreel 0 kanten 0.03 gebieden 0.14 nadien 0 oppositie 5.39 kwamen 3.28 syrische 25.66 roma 0.58 bom 0.18 gebeuren 0.27 gevochten 0.02 volk 0.27 geheel 0.07 kleiner 0.02 woorden 1 opgeblazen 0 begrafenis 0.01 taferelen 0 meest 0.73 horen 4.01 stappen 0.07 beschikken 0 probeerde 0.39 compleet 0.08 bezig 1.5 gesneuvelde 0 rustige 0.01 zoek 0.56 hele 13.49 sterker 0.16 lopen 1.49 wijk 1.4 terzijde 0 snap 0.09 correspondent 2.61 droomt 0 zelfs 15.74 merk 0.07 genomen 0.61 voorspelt 0 defensieminister 0 zwager 0 medische 0.02 natuurlijk 23.77 bronnen 0.1 sommigen 0.04 eerder 5.86 zie 9.68 palen 0 gisteren 97.29 diva 0 maak 0.43 berg 0.36 militairen 16.93 maanden 4.3 ooit 3.37 oranje 0.55 wapens 1.59 buiten 1.86 geluiden 0.01 begon 6.03 vreemd 0.07 opvolger 0.12 gebouw 13.53 reportage 0.03 gehad 1.85 ziekenhuis 0.92 vergadering 0.03 marilyn 0 regering 28.33 vond 16.23 praten 0.77 zwaar 0.35 weten 7.47 nerveus 0 waarbij 1.13 kennis 0.26 eenheid 0.06 nou 25.38 hoorn 0.01 vallen 0.44 opgehaald 0 aantal 2.05 smit 1.17 kant 1.08 vroeger 0.67 zoenen 0 beelden 0.45 plekken 0.09 explosie 0.04 mast 0 aangewezen 0.01 vandaan 0.35 gaven 0.04 aanstellen 0 bron 0.01 hoorde 1.46 opsteker 0 nadat 4.46 ontbreekt 0.01 gehoord 0.18 maskers 0 middag 0.04 paar 13.97 uitgever 1.13 macht 0.85 stromen 0 gezicht 2.23 loop 0.07 opgemerkt 0 wet 0.07 blijkt 1.94 yen 0.02 meteen 4.26 volgens 73.17 stond 21.1 enorme 0.42 aankomen 0.01 gedaan 2.11 eind 0.87 afghanistan 8.65 straat 1.8 kijken 6.22 beide 0.75 maakte 5.21 leveren 0.19 inzien 0 communicatie 0.02 niemand 5.07 vlak 0.19 aandeel 4.42 explosieven 0.02 dood 3.92 verklaring 0.14 regime 13.69 meeste 1 banen 0.09 jaren 13.63 sterk 0.22 binnen 5.92 vriendelijk 0.02 tientallen 0.68 merkte 0.06 sla 0 gekost 0.02 jaar 140.99 asociaal 0 zeker 5.65 winkels 0.01 misschien 13.09 ruimte 0.47 irak 3.05 zoiets 0.42 bewind 1.42 duizend 0.03 dicht 0.35 slag 0.25 allemaal 13.8 maribor 0 kijkt 1.01 beetje 12.09 beren 0 veiligheidsdienst 0.01 wedstrijd 0.44 bezittingen 0 functionarissen 0.17 hart 0.33 tijdens 10.23 berichten 0.14 zitten 13.87 aanslag 0.91 reed 0.31 onderminister 0.01 later 26.78 familieleden 0.03 organisatie 0.9 kaart 0.06 hoofdstad 4.06 doe 3.02 witte 0.46 veilige 0.01 nieuwe 14.07 klap 0.2 bereikbaar 0.02 delft 0.03 dagelijks 0.08 continu 0.01 attent 0 konden 4.53 lijken 0.11 zagen 0.67 staande 0.01 meters 0.01 bedoeld 0.09 zestien 0.04 zoals 12.2 foto 47.67 masters 0.03 manschappen 0.02 tal 0.01 gevolg 0.12 bellen 0.05 eentje 0.57 minister 7.03 oog 0.28 rudi 0 apparaten 0 present 0 uitstralen 0 dallas 0 gevreesde 0 rapporteerde 0 stukje 0.12 gehouden 0.66 figuren 0 vreugde 0.01 voorkomen 0.34 plegen 0.05 positie 0.09 daarbij 0.84 wanden 0 bedden 0 bereikt 0.06 betreft 0.15 gevreesd 0.01 denken 3.94 zaterdag 1.45 zelfmoordaanslag 0.01 zouden 6.46 rest 0.59 damascus 1.7 vluchtelingen 0.37 kwam 67.74 relatief 0.06

5. akkers 0 tweehonderd 0.03 juist 10.77 gestegen 4.48 deel 5.03 nemen 3.48 overstromingen 0.02 heet 0.67 neer 0.07 dertig 0.21 otten 0 hoeveel 0.17 meel 0 rekening 0.15 blijken 0.08 grondstoffen 0.19 pakken 0.1 prijzen 18.7 boek 60.67 helemaal 3.98 waarschijnlijk 0.47 rusland 0.65 westen 2.13 terecht 0.18 gezien 1.98 vanmorgen 0.03 ontbreken 0.01 malen 0 procent 401.77 gewassen 0.01 dichter 0.08 uitmaken 0.01 genoteerd 0.01 betalen 0.82 houden 2.51 molen 0.17 moment 4.76 stijgen 0.96 omheen 0.01 hierdoor 0.14 sterk 3.25 regen 21.51 oogsten 0.01 komende 11.32 tarwe 0 prijs 13.68 schommelingen 0 last 0.47 veertig 0.11 maand 4.56 staten 1.93 zelfs 16.19 frankrijk 1.21 weersomstandigheden 0.02 vroeg 0.13 merkt 0.02 jaar 3922.54 producten 16.12 hevige 0 denkt 1.07 afgelopen 43.99 beurs 4.04 allemaal 5.83 verdrinken 0 graan 0 maanden 12.5 rijdt 0.05 winkel 1.39 aardig 0.02 veertien 0.01 zeventig 0.15 droogte 0 twaalf 0.25 voorbeeld 1.99 dreigen 0.03 grondstof 0 parijs 0.14 hogere 0.41 verbouwd 0 rijk 0.02 wereldwijde 0.3 bruin 0.09 fabriek 4.43 droog 21.69 aangeven 0 geven 3.85 duitsland 4.84 precies 1.01 groter 0.67 duurder 0.37 afrika 2.52 kazachstan 0 geworden 1.22 ton 0.55 zoals 29.18 graden 292.36 mislukken 0 bloem 1.14 concentratie 0 uiteindelijk 1.47 ware 0.05 armere 0 india 30.82 ging 0.83 boeren 0.38 wonen 0.02 directeuren 0 daarbij 1.79 bedrijf 277.15 daardoor 1.08 voedsel 0.08 koppel 0 droge 0.48 marx 0 inkomen 0.01 honderd 0.62 afgerekend 0.02 verenigde 3.07 bewegen 0.07

6. ga 5.11 deel 3.04 toezichthouder 2.65 gepleegd 0 ouder 0.02 zodat 0.28 handelaren 0.01 fraude 0.96 consument 0.09 geleid 0.03 blijken 0.14 vergelijkbare 0.02 blijkt 6.22 intro 0 ervoor 0.11 toegegeven 0 aanleiding 0.16 daarvan 0.29 bank 694.87 boeten 0 hoor 3.85 geraakt 0.02 waarschijnlijk 0.29 eigenlijk 6.24 lenen 2.18 emma 0 buitenwereld 0.01 enige 0.55 mei 0.07 ali 0.01 londen 14.04 reden 0.55 geld 251.38 geval 2.69 denk 1.83 rente 30.06 betalen 29.73 horen 2.15 jaren 11.43 banken 396.19 dupe 0.01 ruimtes 0 sterker 0.16 jezelf 0.15 kantoor 0.05 daarmee 1.22 klanten 18.62 trekken 0.24 vijftien 0.06 schandaal 0.01 gebruikt 0.29 flink 0.47 belang 1.35 bepaalt 0.02 financieel 4.75 afgelopen 4.06 zin 0.32 relevant 0 navo 0.11 ondanks 0.2 bewijzen 0.03 verdienen 0.81 passen 0.05 wekenlang 0 ontslag 0.03 nederlandsche 1.62 lage 0.92 zeventig 0.01 medewerkers 0.13 liep 0.01 data 0.17 mooier 0.04 crisis 4.05 gedupeerd 0 belangrijkste 0.29 kopje 0.02 autoriteit 0.24 spaarrente 0 redenen 0.06 lief 0.08 utrecht 0.11 integriteit 0.02 belangrijke 0.53 engeland 0.15 zaken 2.46 geven 2.95 hoogte 0.21 hoofdkantoor 0 precies 0.96 zogeheten 0.2 garcia 0 stellen 0.87 weten 5.07 kredieten 0.06 takje 0 bellen 0.29 lastig 0.08 elk 0.52 fors 0.41 barclays 0.18 fijn 0.22 ochtend 0.01 gesjoemeld 0 durven 0.04 nou 13.1 vaststellen 0 opleggen 0.04 regels 1.18 taal 0.01 gehouden 0.47 sommige 0.55 ordinair 0 tarieven 0.08 nederlandse 10.37 britse 71.56 welke 0.82 hypotheken 2.37 moeite 0.09 bekendgemaakt 0.01 verdacht 0.02 rabobank 4.76 ter 1.14 moeilijk 0.83 honderd 0.18 diepe 0 kijk 1.49 oplopen 0.04 vinden 2.17 teken 0.01

7. duizend 0.03 werkgelegenheid 0.18 gekomen 0.19 donker 0.01 kwartaal 149.14 deed 11.12 gepresteerd 0 tijden 0.7 grond 0.19 komma 0 cijfers 21.96 verloren 0.16 bezwaren 0 moeilijke 0.12 slechte 0.16 ontslaan 0.14 negentien 0 frans 10.28 zakt 0.01 vakbonden 4.66 positiever 0.02 halfjaar 0.03 dongen 0 franse 74.5 ervan 0.4 liefst 0.42 michel 0.01 luchtvaart 0.56 portie 0 half 1.01 vliegen 0.18 licht 1.31 nieuwe 76.97 aanzienlijk 0.05 bezuinigingsplannen 0 vergelijking 0.25 beide 0.55 kende 0.07 normaal 0.14 land 30.12 dringen 0.01 koos 0.06 turbulente 0 zoals 24.96 omstandigheden 0.11 inzien 0 vijfduizend 0 praten 0.38 aandeel 67.35 bedenkingen 0 weten 5.7 vraag 10.64 gedwongen 0.04 gevolgen 0.47 banen 3.59 eenheid 0.01 lichtpuntjes 0 verrast 0.03 historie 0 houden 3.04 rol 1.22 moment 5.69 passagiers 0.52 bekennen 0.01 aantal 35.39 levert 0.65 snelle 0.05 staal 0.49 splitsen 0 doorlopen 0 daarbij 1.85 daarmee 2.53 ouwe 0 stegen 0.15 ruim 12.11 bezuinigd 0.03 voorzieningen 0.02 nettoverlies 0.19 hierop 0 verwacht 10.18 resultaat 0.99 sterven 0.01 jaar 1484.18 zeker 9.71 kale 0.01 succes 0.82 vechten 0.01 sinds 9.46 ruzie 0.04 partners 0.11 vals 0 steeg 17.36 prins 95.59 instemmen 0

8. bewoners 35.23 deel 2.71 omschreef 0 burgemeester 57.16 enkele 4 begonnen 0.49 gewerkt 0.06 sanering 0 nou 11.2 marcel 0.25 schadelijke 0 huis 5.09 utrecht 2.86 rechtse 0.44 controleren 0.02 asbest 0.95 wist 0.28 hoeven 0.06 opruimen 0 vroeg 0.17 anand 2.36 flat 0.55 straat 2.85 stoffen 0.2 voorlopig 0.07 mogen 0.73 tiende 0.01 inzien 0 beloofd 0.01

9. juist 12.69 onschuldig 0 volgelingen 0 wilde 0.64 zeg 9.78 nederland 43.73 blijkt 10.62 samenwerken 0.14 verschijnen 0.04 dingen 4.84 zeer 0.9 overal 2.05 pakken 0.46 volgens 50.06 spelers 1.39 inzet 0.07 inzetten 0.02 weddenschap 0 normaal 0.49 business 0.05 niveau 1.27 koop 1.99 openingsceremonie 0 manipuleren 0 zestig 0.04 ding 1.54 gebeuren 0.45 niemand 7.76 bleken 0.01 vrijdag 2.06 hoger 0.84 eenmaal 1.24 illegale 0 kun 12.23 zoveel 2.23 bal 0.49 medailles 0.21 zwemmen 0.43 horen 8.34 schandalen 0 houden 5.18 thee 0.32 liefhebber 0.01 terug 4.09 verdachte 5.9 probeer 1 staal 0.39 peper 2.55 plaatje 0.01 ing 3.46 strekt 0 ioc 0.35 zeker 11.15 uitslagen 0.02 eraan 0.16 denkt 2.41 spelen 136.68 afgelopen 19.52 waarop 1.17 misschien 24.4 bedreigingen 0 kunt 53.46 wetenschappen 0.03 duizend 0.07 maanden 6.19 spant 0 probleem 3.26 god 0.21 sieraden 0.05 letteren 0 zitten 27.69 dokkum 0.01 kopje 0.06 simon 0.05 pijn 0.05 engeland 0.44 intensief 0.01 blij 0.57 langs 2.48 kreeg 0.83 hoofdkantoor 0.11 kennen 0.4 groter 0.53 vaker 0.83 zogeheten 0.04 werkelijk 0.26 aandelen 7.64 verzoek 0.04 verschillende 1.38 daarom 8.06 bloed 0.09 weten 15.55 ontkomen 0.01 cricket 0 vis 1.62 nou 54.6 winst 8.43 failliet 0.34 aantal 34.77 goud 0.4 wijken 0.02 levert 0.29 voorkomen 0.19 naam 4.57 stuk 1.73 daarbij 1.66 voetballers 0.03 britse 0.98 uiteraard 0.34 bekendste 0 missen 0.02 jelle 0 olympische 125.91 opgekocht 0 vinden 7.86 vandaar 0.06 gaten 0.19 roeiers 0 bevelen 0 wisten 0.04 soort 6.52

10. allemaal 21.23 viel 0.07 ooit 2.52 unie 0.34 donker 0.07 overheid 1.64 deel 4.69 zomer 0.61 tijden 0.53 kuyper 0.03 europese 33.43 helft 8.46 centraal 0.57 oosten 7.49 paar 19.63 eeuw 0.32 turkije 8.75 leidde 0.07 belangrijkste 0.84 bach 0.21 elektriciteit 3.5 westra 0.13 weet 105.97 afnemen 0.01 legt 0.65 hitte 0 achttien 0.01 enorme 0.83 waarin 2.12 nieuwe 114.51 oorzaak 0.01 aller 7.9 bestuur 12.15 klap 0.08 bouw 2.66 storing 0.79 beleeft 0 day 0 corruptie 0 gek 3.55 land 123.08 teleur 0 enige 1.16 zoals 17.26 uitvalt 0 school 49 reden 0.48 geld 10.47 elders 0.22 weten 16.42 vraag 11.41 energieverbruik 0.15 groei 18.7 bal 0.33 jullie 18.82 erik 0.06 horen 8.39 nou 54.52 noorden 47.64 auto 4.09 zaten 0.02 metropool 0.01 jaren 44.68 urenlang 0 moment 1.23 economische 9.18 genoeg 4.7 sterk 2.84 india 58.64 levert 1.24 binnen 11.24 binnengekregen 0 last 0.91 daardoor 1.12 moderne 0.2 stroom 11.74 bouwen 1.31 twintig 0.58 deelstaten 0 grootste 18.58 buitengewone 0.01 achtereenvolgende 0 zeker 3.12 optreden 0.52 honderd 0.7 trouwens 1.14 jarenlang 0.08 dick 0.23 vinden 6.18 afgelopen 43.21 gisteren 11.95 expliciet 0 gloednieuwe 0 raar 0.34 soort 3.37

11. juist 7.56 geleden 5.03 trof 0.03 moeilijkste 0 speciaal 0.15 wijze 0.33 zei 0.52 hadden 4.08 beslag 0.4 gekeerd 0 woning 7.07 nederland 84.24 gevangenisstraf 0.57 blijkt 5.93 betekent 0.23 omzet 8.83 zeer 1.33 overal 0.93 pakken 0.36 zoeken 0.6 vanochtend 0.02 knipoog 0 morgen 0.27 dak 0.66 kijken 5.65 munitie 0 meesten 0.29 doel 1.21 name 0.12 verstoppen 0 apparatuur 0.02 erom 0.02 geld 6.16 woorden 0.36 bal 0.14 auto 20.56 probeerde 0.18 citroen 0 jaren 18.78 bezig 2.32 moment 5.54 handel 0.02 wachten 0.27 hele 10.38 foster 0 zuur 0 luxe 0.03 doorzocht 0.02 genoemd 0.07 last 0.28 literair 0.01 vijftien 0.14 politie 1126.47 zelfs 13.9 datgene 0 scooters 0 jaar 689.19 beschikt 0.01 provincie 0.07 actie 0.26 ruimte 0.2 aangehouden 9.33 schuur 0.01 duizend 0.02 criminelen 1.03 ronden 0.01 enorm 0.45 politiemensen 0.14 instituut 0.02 verdiende 0 teelt 0 woningen 0.81 opvolgen 0 tralies 0.14 sieraden 0.18 gevonden 5.53 zitten 12.42 twaalf 0.3 juiste 0.85 voorbeeld 1.17 heleen 0 begraven 0.08 machtsvertoon 0 afpakken 0 doe 4.53 verdiend 0.02 rechercheurs 0.41 daaraan 0.01 pijn 0.04 wortels 0 verdachten 7.05 dezelfde 0.96 pakt 0.02 gekozen 0.06 huiszoeking 0.01 allerlei 0.08 ontmantelen 0 aangetroffen 0.92 muur 0.73 zagen 0.12 crimineel 0.17 landmacht 0 raken 0.16 afgevoerd 0.06 richt 0.03 justitie 25.01 misdaad 0.03 aanwezig 0.25 daarvoor 0.4 drugs 0.37 verhalen 0.24 thaise 0.01 nou 22.58 keren 0.1 criminele 0.83 driehonderd 0 aantal 16.61 ene 0.88 vermeende 0.04 kant 1.04 vroeger 0.35 aanpak 0.13 bedrijf 40.16 bankrekeningen 0 gelegd 0.06 honderd 0.34 speuren 0 vinden 6.61 reguliere 0 overbodige 0 halen 0.91 zelden 0.1 arrestatieteam 0.1 begin 0.92

12. tweehonderd 0.02 mits 0.01 allemaal 8.43 gekomen 0.82 werden 37.27 waaraan 0.02 deed 16.66 lieten 0.15 ravage 0.01 korpsen 0 nadat 10.52 liggen 0.95 doelwit 0.03 begon 5.63 zei 71.9 tekeer 0 gooien 0.04 charles 0.01 gewapende 0.07 gebeurt 0.46 nauwelijks 0.86 breken 0.01 gegaan 0.97 ingezet 0.22 richtte 0.01 plukken 0 stenen 0.07 dalen 0.01 ruiten 0.01 leest 0 vragen 0.92 dolle 0 eind 0.89 opruimen 0 cynisme 0 overval 2.83 plein 0.13 confrontatie 0 eigenlijk 7.04 konden 4.92 waartoe 0 maakte 5.24 draaien 0.07 joris 0 invloed 0.08 nergens 0.24 kwart 0.02 exclusieve 0 rellen 0.05 vanmorgen 0.01 mum 0 besloot 0.57 honden 0.19 noemt 0.16 vlak 0.24 gevochten 0 resten 0.06 diverse 0.58 schade 0.98 arm 0.08 feest 0.05 rotterdam 79.48 drugs 0.36 uiteindelijk 2.77 groep 1.51 agenten 27.08 glaswerk 0 gezegd 0.87 nou 11.81 rol 1.01 mannen 12.21 moment 5.68 boze 0.05 hele 7.48 moest 49.25 doorgaan 0.03 vijftig 0.53 daarmee 0.36 vervolgens 3.55 vannacht 1.33 spoor 0.24 vijftien 0.18 politie 1126.47 situatie 0.57 ligt 3.72 twintig 0.43 opgepakt 5.35 team 0.28 scherven 0 tienduizenden 0.01 vernielingen 0.01 jaap 0.04 aanwezige 0.02 raad 59.62 besloten 0.49 zie 4.36 voeren 0.06 schrik 0.07 assistentie 0 indruk 0.25 drank 0.05 oorlog 0.09

13. commissie 2.26 werkte 0.34 relschoppers 0 overheid 1.47 gebracht 0.66 werpen 0.01 toeristen 0.01 beschermen 0.12 keniaanse 0 wet 5.4 samenwerkt 0 gebouwen 1.92 steden 0.76 enorme 0.43 waarin 2.94 verleden 0.9 radicale 0.27 helemaal 7.35 vaste 0.08 argwaan 0 waarschijnlijk 0.49 religieuze 0.19 plaats 1.38 ambassade 0.39 maakte 2.62 voorlopig 0.2 namelijk 1.29 dulden 0 jongeren 5.16 zeggen 3.62 rellen 0.06 pan 0.64 aangevallen 0.01 vlam 0.03 vogel 0 hel 0 vrijdag 2.4 kosten 0.81 zondag 0.27 worsteling 0 groep 1.63 populair 0.03 houden 2.33 zullen 4.1 afrikaanse 5.23 kenia 0.1 toekomst 1.58 bedwang 0 hele 3.16 verwelkomd 0 erkent 0.03 staan 5.82 christenen 0.1 moest 20 stad 5.79 moskee 0.16 klanten 0.67 politie 283.04 oordelen 0 jaar 376.33 winkels 0.66 christian 0.01 plaatsvonden 0 banden 0.24 havenstad 0 eerder 6.2 japan 0.25 aangewakkerd 0 afgelopen 11.51 vermoorden 0.04 sloeg 0.14 maak 0.36 dicht 0.09 geschiedenis 1.13 maanden 6.04 militairen 16.93 werden 21.3 ooit 1.82 unie 1.04 strijden 0.01 combinatie 0.05 beetje 4.17 hassan 0.01 daalt 0 geweld 8.43 deuren 0.05 moord 3.28 vu 0 charles 0.04 dreigen 0.02 renske 0 vreedzaam 0 doe 1.81 witte 0.68 maandag 0.47 blij 5.73 staken 0.07 eenheden 0.03 blanken 0 leert 0.02 strak 0.01 spannen 0 fout 0.32 zoals 15.53 willekeurig 0 daarop 0.12 daarom 5.68 weten 3.76 naast 0.73 soldaten 1.36 moslims 0.51 des 0.08 wetten 0.02 leger 37.38 duidelijk 1.87 ging 41.83 wijken 0.04 somalische 0.15 kerken 0.07 brand 10.21 horror 0 gesloten 0.16 waarnaar 0 nederlandse 5.96 situatie 0.96 enkel 0.03 verwacht 0.37 buurland 0.14 speciale 0.26 word 0.4 optreden 0.13 volgde 0.13 leefden 0 ramen 0.04 wederzijdse 0 makkelijk 0.06 zelden 0.07 gaten 0.04 proces 2.75

14. begrijp 0.07 gekomen 0.14 ga 8.39 juist 4.69 florida 0.11 koken 0.01 overheid 1.72 verkopen 1.97 mooi 4.11 ontbreekt 0 verlies 4.18 bedrag 13.29 zeg 2.54 moesten 0.72 idee 0.77 orde 0.06 rekening 5.84 nee 13.2 nederland 16.74 stond 5.56 volgens 29.05 bovenop 0.03 oorzaak 0.06 volgend 0.03 bank 695.65 eind 2.79 bang 0.27 verleden 0.38 helemaal 10.38 stress 0.06 boeken 0.22 bedrijfsonderdelen 0 opgebouwd 0.05 maakte 1.74 enige 0.72 zakenbank 0.15 systeem 0.25 zeggen 6.37 stand 0.13 niemand 2.54 geld 251.32 ultieme 0.02 kosten 18.88 gegarandeerd 0.01 geval 2.18 verzekeringstak 0.01 wilt 2.27 houden 2.92 miljoenen 1.66 bankencrisis 0.03 jaren 11.63 banken 396.31 terug 5.79 genoeg 2.04 depot 0 papier 0.03 reserves 0.1 staan 9.28 desnoods 0 tientallen 0.24 spoor 0.01 overeind 0.01 genomen 0.23 voorstellen 0.03 jaar 364.73 natuurlijk 16.19 staatssteun 0.45 sns 3.45 real 0.57 aanpakken 0 kapitaal 0.8 ach 0.4 ondanks 0.35 stel 0.44 duizend 0.01 est 0 liggen 0.68 europese 0.55 afbetalen 0 groeien 0.2 eva 0.01 tijdens 2.24 zitten 8.86 nogal 0.15 crisis 4.73 belangrijkste 0.18 later 7.28 bach 0.07 franse 0.31 ministerie 0.12 utrecht 0.34 absoluut 0.05 nieuwe 15.35 dezelfde 0.38 zanger 0.02 langs 0.39 reaal 0.28 kreeg 6.8 hoofdkantoor 0.05 eigenaar 0.36 daarbovenop 0 dochters 0.15 hetzelfde 0.49 verliezen 0.89 onderzoeken 0.01 bijvoorbeeld 5.41 spaargeld 1.54 zagen 0.17 opdraaien 0 bescheiden 0.02 vrijmaken 0 zoals 10.99 vastgoedmarkt 0 praten 0.46 wolf 0.01 streven 0.01 achthonderd 0 vraag 5.67 everything 0.01 nederlands 0.61 gewone 0.04 minister 0.45 problemen 4.84 stoppen 0.12 terugbetaald 0.02 nou 14.16 harder 0.01 port 0 winst 10.54 onderdelen 0.03 verwachtingen 0.06 regels 0.49 betrokken 0.13 aangemerkt 0 armen 0.03 mis 0.2 gehouden 0.44 dreigt 0.05 vijftig 0.12 ruim 5.51 daardoor 0.42 marina 0 hierbij 0.03 zouden 2.27 eiland 0.44 kwam 18.4 honderd 0.19 inderdaad 0.39 schadeclaims 0 inclusief 0.04 makkelijk 0.16 houders 0 teken 0.02 stelsel 0.01

15. deur 0.42 vanmiddag 0.12 tentenkamp 0.68 luisteren 0.02 bosch 72.71 vaderland 0 gezegd 1.35 zaten 3.71 opgestapt 0.02 duurder 0.18 polsen 0 lintje 0.13 terug 13.95 eerder 2.87 idee 1.02 wilson 0 schoonmakers 0.14 terecht 0.21 zodra 0.01 teruggaan 0 bewegen 0

16. strijden 0.02 gisteravond 0.89 aangestoken 0.07 vermoedelijk 0.36 drukte 0.01 bal 0.35 waarbij 0.23 verlaten 0.62 jaar 127.78 zeker 1.15 plaats 3.47 bergen 0.19 huis 3.72 brandweerlieden 0.11 dode 0.02 spanje 46.23 zuiden 0.35

17. salon 0.01 oordeel 0.63 ouders 580.67 overigens 1.14 wilde 8.57 spel 0.09 voortzetten 0 paar 10.28 ineens 0.2 nederland 21.03 kanalen 0 kinderen 5310.45 beginnen 1.55 el 0.01 vinger 0.02 meisje 28.72 volgens 68.24 daarvan 1.01 jeugdzorg 1.14 morgen 0.31 helemaal 18.06 namelijk 2.91 buitenland 0.2 haarlemse 0 koos 0.03 terecht 1.1 zeggen 8.71 reden 1.57 denk 7.73 zullen 4.5 wilden 0.46 genoeg 5.56 spanningen 0.01 telefoontje 0.02 tijdelijk 0.02 onderwijs 338.19 boten 0.4 vijftien 0.5 binnenlandse 0.01 hiervan 0.05 jaar 547.5 woonhuis 0.01 natuurlijk 38.37 oplossing 0.5 sinds 6.38 misschien 12.46 gepast 0 verbod 0.44 thuis 15.44 ooit 3.38 combinatie 0.15 verboden 0.2 enkele 1.61 gepaard 0.01 tussenkomst 0 god 0.13 zitten 8.33 later 4.97 zodra 0.07 avontuur 0.02 verblijven 0.01 plichten 0 blij 11.18 vragen 2.08 acties 0.03 geven 12.96 definitief 0.07 precies 2.6 erg 13.22 uitzondering 0.09 blijkbaar 0.04 school 351.19 graag 9.86 gemaakt 8.74 principe 0.18 hugo 0.01 nou 0.11 afgerond 0.02 kind 565.72 instanties 0.05 duidelijk 4.16 jongens 8.72 dertien 0.07 les 2.88 kant 3.32

belgische 0.11 rechtbank 18.11 aandacht 2.95 wateren 0 voorlopige 0.05 werk 25.87 broertjes 0.04 bemiddelen 0 degelijk 0.42 zouden 3.06 vinden 7.78 vandaar 0.14 term 0.03 teken 0.02 klaassen 0

18. grond 9.82 gebracht 0.17 gestraft 0.04 gelijk 0.62 heet 0.61 voegen 0.02 zeventig 0.2 god 0.36 moesten 3.07 voorman 0.02 zitten 1.85 redelijke 0.02 rus 0 verwarm 0.04 gingen 4.23 kinderen 2298.66 gezondheid 0.01 gebouw 43.57 volgens 128.13 waarin 4 optredens 0.09 nieuwe 20.99 accepteert 0 russische 11.83 aangekondigd 0.09 gezet 0.17 land 72.22 moskou 5.89 zagen 0.59 kilometer 2.94 zag 21.24 school 4.48 instemming 0 daarop 0.14 zeventien 0.01 ban 0.04 gezag 0.05 zaten 1.69 volgen 0.17 jaren 27.17 sekte 0.01 stad 9.26 vijftig 1.22 legaal 0 ruwe 0.49 zelfs 7.33 islamitische 1.31 samenleving 0.35 ondergrond 0 jaar 277.88 verdiepingen 0.04 jarenlang 0.12 dokter 0.03 boodschap 0.09

19. huizen 1.68 ieder 0.59 duitse 9.64 volksgezondheid 1.08 besmet 0.79 vijfentwintig 0 wet 1.5 nee 17.28 bijkomen 0 blijken 0.53 chronische 0.46 dingen 1.87 liefst 0.27 boel 0.07 geraakt 0.01 eigenlijk 6.73 internet 0.22 lichte 0.04 vak 0.04 eenzelfde 0 gelopen 0.01 bezoeken 0 hoger 0.54 wijst 0.11 duizenden 0.2 denk 2.44 spa 0.18 houden 2.68 belevenissen 0 jaren 11.44 geiten 0.02 korte 0.2 ziek 1.56 indien 0 lopen 1.34 gedacht 0.1 veertig 0.15 echter 2.56 vermoeidheid 0.04 flink 0.17 verspreid 0.01 merkt 0.01 jaar 143.98 zeker 4.42 overleden 0.08 rekenen 0.04 wal 0.03 schouw 0 denkt 0.9 zie 6.73 advies 0.15 gezondheidszorg 1.97 misschien 9.63 kunt 20.76 raakt 0.04 duizend 0.09 strand 0.18 verschijnsel 0.01 onderzocht 0.56 zingen 0.05 bossen 0.11 enkele 1.54 onderzoek 209.69 mooier 0.06 leeuw 0.17 factor 0.02 weet 38.7 nieuwe 32.23 gehad 0.53 bouw 0.56 betaalt 0.16 ziekenhuis 74.54 gehalte 0.01 groter 0.34 welk 0.08 geruststellend 0 huisarts 2.86 honderdduizend 0 voelt 0.75 honderden 0.14 weten 7.49 weigerde 0 melden 0.17 sindsdien 0.02 belasting 0.32 bewust 0.19 aantal 8.54 daarna 0.65 minstens 0.04 bacterie 0.73 naam 1.65 vijftig 0.22 verontrustend 0 race 4.58 steken 0.13 ligt 3.62 gezinnen 0 denken 4.15 verwacht 0.11 honderd 0.4 verschijnselen 0.01 brute 0 aanraking 0.01 pakte 0

20. juist 4.71 lieten 0.08 drijft 0 zomer 0.43 thijs 0.01 kortere 0 wedstrijden 0.72 dertig 0.15 leeft 0.01 sneller 0.21 ver 0.82 oceaan 0.82 eeuw 1.48 macht 0.08 hoeveel 0.1 gevallen 0.29 ex 0.39 nederland 39.54 betekent 0.29 bevroren 0.75 stond 7.5 enorme 0.96 eind 1.16 claims 0 helemaal 2.63 vorig 43.2 ondergaat 0 beslaat 0 voordat 0.38 daarbuiten 0 verdwenen 0.15 spiegel 0 keizer 0 noord 1.71 gezien 1.11 landen 28.28 noren 0.01 kun 1.29 gas 0.98 ijs 98.45 wanneer 0.69 ontstaan 1.05 zomers 0.09 knmi 0.04 september 0.12 water 449.05 trouw 0.02 ongeveer 2.05 gedacht 0.08 veertig 0.14 maand 1.01 olie 41.41 ontdekt 0.2 continenten 0.03 jaar 753.39 stijging 0.55 eerder 3.95 sinds 3.57 afgelopen 9.53 misschien 3.75 waarvoor 0.02 gebied 3.18 duizend 0.12 ooit 1.32 vanaf 7.62 hoeveelheid 0.06 negentien 0 actueel 0 vanuit 7.43 wessels 0.04 gingen 1.27 nieuwe 34.06 veranderen 0.07 allerlei 0.09 gemeten 0.03 eisen 0.13 vierkante 0.1 draaien 0.08 hand 0.77 noordelijk 0.01 slaat 0.03 kilometer 14.3 meldt 0.1 bootje 0.08 metingen 0.01 vond 5.71 verschillende 1.07 gesmolten 0 vorige 1.32 effect 0.65 groeit 0.47 veruit 0 doorzet 0 cda 65.87 vakantie 0.04 gebroken 0.01 zeewater 0.02 smelt 0.01 voordelen 0.01 wiegel 0 aarde 32.9 daardoor 0.95 werk 2.27 ligt 5.96 gelegd 0.02 elftal 0.04 missen 0.01 eis 0.02 raad 24.96 overblijft 0 begin 0.64 lokaal 8.18 satelliet 0.06

21. overheid 4.03 typisch 0.11 internetbankieren 0.01 besmette 0.02 toegeslagen 0 ogenschijnlijk 0 zei 73.99 hadden 45.95 versies 0.02 hebt 44.94 ver 0.66 nee 35.25 jeroen 0.38 tamelijk 0.02 gemiddelde 0.08 waarden 0.06 bank 128.27 noem 0.09 europees 0.91 paspoorten 0.02 verstrekt 0.03 geraakt 0.17 normaal 1.1 gelopen 0.03 name 0.17 ijsberg 0.05 kwamen 2.33 ding 1.14 aantoont 0 reden 0.93 kast 0.09 volk 0.14 bestanden 0.04 verlost 0 groep 9.31 incident 0.61 jou 0.71 banken 73.16 bezig 2.37 buitenaf 0.01 langer 1.85 hele 21.04 karakters 0 zonde 0.1 gemeentehuis 0 lopen 3.11 doordacht 0 doordat 1.69 gebruikt 3.18 zelfs 19.35 verspreid 0.02 digitaal 1.15 rekenen 0.07 zodoende 0.01 vissers 0 noot 0 natuurlijk 42.13 sommigen 0.54 bedrijven 14.19 opvallend 0.3 sinds 3.62 zie 15.21 zin 1.29 gezeten 0.14 krijg 3.83 militairen 0.6 maanden 5.23 thuis 2.81 vertekend 0 enkele 4.27 afdeling 0.72 onderuit 0.02 besteden 0.02 oosten 0.45 graffiti 0 opgestart 0 proberen 0.66 rijbewijs 0.14 beeld 3.49 precies 2.7 erg 7.54 nationaal 0.27 geworden 1.85 virussen 0.22 hulp 3.02 praten 1.29 effect 3.64 weten 15.42 stijlen 0 laptop 0.16 nederlanders 62.75 lastig 0.48 elk 1.33 tachtig 0.07 ochtend 0.21 nou 37.72 criminele 0.39 denemarken 1.57 computers 0.35 aantal 11.83 topje 0 ene 1.25 storen 0 lacht 1.31 wonen 0.39 ondersteuning 0.01 bui 0.28 actief 0.28 ploeteren 0 bankrekeningen 0 herstel 0.14 koen 0.05 mart 0.2 virus 1.8 jouw 1.17 rijtje 0.01 zwaarder 0.03 stonden 1.28 collega 3.28 storm 0.03 ga 29.55 zomer 0.47 nadat 5.63 slechte 0.36 computer 1.94 besmet 0.79 universiteiten 0.31 loop 0.37 nederland 301.31 boek 40.49 medewerker 0.08 gedaan 3.83 registreren 0.01 afgehaald 0 liggende 0 waarschijnlijk 1.23 eigenlijk 17.79 kwetsbaarheid 0 ministeries 0.01 doek 0.18 doel 0.59 gemeente 68.51 manipuleren 0 landen 21.41 interpreteren 0 stand 0.64 geld 61.16 klok 0.19 duizenden 0.27 denk 10.67 klappen 0.02 melding 0.06 handen 1.55 documenten 0.01 moment 9.34 terug 18.4 gehangen 0.01 ongeveer 2.2 netwerk 8.99 verspreiden 0.01 wachtwoorden 0.05 last 0.86 jaar 299.7 verkeer 0.48 gebruik 6.67 misschien 19.46 aids 0.01 besteedt 0.01 verbergen 0 allemaal 20.64 praat 0.26 criminelen 0.49 leggen 0.53 kijkt 1.56 beetje 20.12 tijden 0.41 moe 0.02 negentien 0 werkplek 0 alweer 0.16 brouwer 0.02 gebeurt 0.6 later 27.63 gingen 3.35 gebeurd 0.4 doe 20.34 verzamelen 0.02 nieuwe 119.03 bestaat 1.31 tennis 0 weken 4.31 personen 0.11 bekend 5.5 meldpunt 11.64 zagen 0.73 stoornis 0.12 hand 4.36 voort 0.03 toeters 0 slaat 0.26 zoals 23.26 beeldschermen 0 woensdag 6.74 grof 0 lamp 0.89 vullen 0.03 vraag 9.64 gemaakt 7.6 sturen 0.3 slachtoffer 14.08 gewerkt 0.18 volgen 0.32 herleven 0 haal 0.29 vakantie 2.94 ergens 1.15 sprak 0.63 gemeenten 6.26 weert 0.1 ging 108.64 zwart 198.84 drieduizend 0 daarna 7.6 geslaagde 0 digitale 11.39 aandacht 1.93 werk 43.84 centrum 3 programma 157.86 bankrekening 0.02 romantiek 0.09 ter 1.16 inderdaad 1.13 ernstig 0.8 office 0.01 halen 1.95

22. gesprek 0.09 overvallen 1.54 bewoners 3.81 overvaller 0.28 thuis 0.19 buiten 1.24 daalt 0.02 gebracht 2.25 woningen 1.21 liggen 0.46 geweld 2.55 bezittingen 0 gewelddadige 0.43 avond 0.13 gevonden 5.31 hoeveel 0.11 woning 11.54 woord 6.52 vanuit 1.14 cyprus 0 riant 0 beginnen 0.29 weet 4.91 bobo 0 slachtoffers 9.59 zeer 1.45 stond 0.92 belangrijke 0.66 verloop 0.01 gisteravond 0.46 geven 2.3 overval 2.83 ziekenhuis 3.99 straat 1.83 dorpje 0 geraakt 0.26 waarschijnlijk 0.79 opvalt 0 bekend 3.7 ede 0.01 enige 0.46 plotseling 0.04 uiten 0.01 voort 0.01 uitgebreid 0.02 soms 2.81 zoals 18.36 inmiddels 2.91 ding 0.01 overvallers 2.46 zag 0.19 rome 2.66 procent 45.45 niemand 0.43 aanwezig 0.43 gezocht 0.19 vraag 6.32 dood 5.83 gemaakt 0.95 gewond 2.75 worsteling 0.01 uiteindelijk 1.38 geschoten 0.33 waarbij 2.17 huiskamer 0.01 ontstaat 0.04 mannen 11.92 vallen 0.08 verdachte 21.57 aantal 12.1 ene 0.4 betrokken 1.3 gemiddeld 0.48 huis 51.6 binnen 6.16 voorkomen 0.24 vijftig 0.22 veertig 0.07 staten 4.92 politie 1126.48 overleeft 0 gebruikt 0.51 vooralsnog 0.04 zelfs 8.68 huiswerk 0 barneveld 0 bekendgemaakt 0.04 jaar 503.24 zeker 6.29 zwaargewond 0.23 politieonderzoek 0.05 raak 0.04 honderd 0.14 bedrijven 6.15 plan 0.14 stel 0.09

23. bijdrage 0.01 doofpot 0 dertig 0.39 publieke 0.58 waarmee 2.28 wouter 0.01 kozen 0 carla 0.09 nee 27.41 mede 0.23 zeer 1.65 bol 0.02 valpartij 0 jan 4.69 station 0.44 vorig 2.26 plaats 3.34 prettig 0.1 burgers 0.79 groeiende 0.04 westen 0.04 kwamen 1.38 zag 9.42 verantwoordelijk 0.25 gebeuren 0.62 bereiken 0.24 hoewel 2.26 kun 7.84 waarna 0.16 hen 2.61 begonnen 0.31 bal 0.28 houden 5.3 schil 0.01 eerdere 0.05 boze 0.02 rechterhand 0.01 water 22.34 partijen 2 held 0.05 hele 15.74 stad 3.33 vervolgens 1.31 correspondent 2.01 staten 25.48 zelfs 36.4 weersomstandigheden 0 bezien 0 natuurlijk 38.95 eerder 3.35 jarenlang 0.33 schuldig 0.29 zie 13.73 zicht 0.01 zin 1.18 chinezen 6.18 gevangen 0.04 werden 11.14 arena 0.03 besef 0.02 kandidaat 3.79 ontwikkelen 0.19 gestraft 0.03 positieve 0.15 begon 2.58 baas 0.18 moord 1.77 ontwikkelingen 0.03 bandje 0 ervan 0.84 weet 71.4 journalisten 1.68 china 435.94 foute 0.03 precies 3.13 groter 0.58 partijleiders 0 corruptie 0.58 erg 6.19 beroep 4.2 bijvoorbeeld 27.5 chinees 3.44 invloed 0.33 hongarije 0.06 symbolisch 0 balkon 0 percentage 0 afluisteren 0.01 elf 0.07 zakenpartner 0 problemen 1.21 bijzonder 0.73 like 0 stoppen 0.23 overste 0 nou 29.41 mooie 1.76 goeie 0.04 aldus 8.1 leken 0.02 veranderd 0.09 tapijt 0 wonen 0.26 actief 0.26 rechtszaken 0.02 gesloten 0.12 moderne 0.07 officieel 0.19 britse 39.88 verdacht 0.74 veld 0.18 media 238.66 machtige 0.1 rechtszaak 0.73 leren 0.67 merken 0.13 gaten 0.14 begin 1.52 proces 1.15 achterlaat 0 bevonden 0.02 achteraf 0.14 rijkdom 0.01 storm 0.1 vergiftigd 0 geleden 10.38 mooi 2.36 politici 1.38 heerenveen 0.62 neer 0.22 wilde 17.6 mies 0 paar 14.07 macht 1.3 halfjaar 0 leiden 0.74 achteruitgang 0 volgens 60.72 enorme 0.97 boek 108.82 vreemde 0.03 eind 0.41 verleden 2.38 machtsmisbruik 0.01 begint 1.48 waarschijnlijk 0.92 eigenlijk 21.54 wenen 0 enige 2.04 vestigen 0 beschuldigd 0.14 politiek 6.69 zeggen 15.61 dader 0.55 verhaal 3.09 niemand 4.94 enigszins 0.07 herfst 0.01 overtreden 0.04 dood 10.63 nv 0.01 verklaring 0.26 ruiken 0.01 peking 27.04 zoveel 3.79 belangrijk 15.04 leiderschap 0.01 gat 0.04 jaren 44.52 talent 1.51 louter 0.01 sri 0.32 verklaart 0.09 komende 0.24 jean 0.89 last 0.24 straks 1.67 internationale 0.87 deels 0.05 echter 3.07 november 0.25 grootste 3.48 machtigste 0 jaar 186.12 zeker 21.09 onmogelijk 0.03 vriend 0.63 afgelopen 8.85 misschien 21.4 ontdoen 0 gevoelige 0.04 hotel 7.08 duizend 0.12 ontdaan 0 dol 0.01 kijkt 1.3 beetje 13.31 vlucht 0.06 enorm 0.97 deuren 0.05 peru 0.02 bedoel 0.07 twaalf 0.14 schuiven 0.01 vanuit 3.11 voorbeeld 4.32 gebeurt 1.72 later 12 redenen 0.21 verhandeld 0.02 nieuwe 37.95 kopen 0.25 zaken 2.83 consulaat 0.02 daarbovenop 0 aangetroffen 0.13 land 12.46 lijken 0.29 bericht 0.3 prettiger 0 zoals 55.45 reis 0.09 spulletjes 0 achtergrond 0.13 ruimen 0 vraag 22.63 leiders 0.07 arm 0.12 bellen 0.05 vermogende 0 gewerkt 0.14 verhalen 1.02 guzman 0.02 deskundig 0 mannen 4.45 inziens 0 hangt 0.51 klein 0.89 ergens 0.7 india 4.31 regels 0.93 ging 46.14 zwart 60.61 gehouden 0.27 hoogste 0.08 bouwen 0.21 bijzondere 0.4 degelijk 0.4 sociale 0.69 zouden 3.55 gemeenschap 0.1 welvaart 0 daarin 0.19 politieke 7.34 verenigde 10.26 opgroeide 0 soort 8.65

24. weten 10.92 fabrikant 0.08 onderzocht 0.54 aanvragen 0.06 gedaan 1.12 kun 4.19 minister 69.1 wilde 1 houden 2.38 verstrekken 0 gegevens 1.16 bedrijven 25.26 spelen 35.38 wilden 0.05 hoeveel 0.19 dna 5.16 schippers 1.07 nee 20.11 opdracht 0.05 benaderd 0.01 voorkomen 2.85 ermee 0.08 materiaal 0.12 afdaling 0.01

25. duizend 0.09 februari 0.52 ruim 16.19 maanden 11.85 belangrijke 0.18 gestegen 4.46 dezelfde 2 periode 2.76 bureau 2.55 negentig 0.03 tachtig 0.05 stierven 0.01 ouder 0.14 statistiek 0.58 vorig 245.96 jaar 3876.14 overleden 0.05 rol 0.41 centraal 0.52 mannen 1.25 zeventig 0.11 leeftijd 0.15 speelde 0.5 sterk 3.02 nederland 97.22 maal 0

26. engelse 0.07 uitspraak 7.17 allemaal 23.33 ooit 3.14 oranje 0.27 beetje 17.63 achterban 0.27 stropdas 0 kandidaat 2.26 nadat 0.68 dichterbij 0.01 europese 3.18 zeg 6.85 ineens 0.92 pers 0.07 campagne 0.15 zitten 21.16 nogal 0.54 begroting 0.24 nee 35.73 enthousiast 0.2 omar 0.01 nederland 52.54 mogelijke 0.28 nul 0.02 gegaan 0.15 samenwerken 0.01 verkiezingen 8.24 gehad 2.29 rijk 0.05 helemaal 26.39 that 0.05 beeld 2.11 emile 0.05 beroep 4.99 welk 0.12 land 2.9 eventuele 0.02 fout 0.33 roemer 0.96 trekt 0.09 vond 0.99 buitenlandse 0.22 vorige 2.74 graag 8.8 risico 0.21 regeren 0.05 romer 0.04 weten 16.89 vraag 15.81 daarvoor 0.49 saai 0.01 eerlijker 0 eenmaal 1.07 gedurende 0.02 begonnen 0.31 premiers 0 nou 38.39 jaren 22.29 dacht 0.82 delen 0.23 goeie 0.05 ernst 0.04 beleid 0.29 langer 0.7 wennen 0.06 aangaf 0 kant 2.16 barry 0.06 komende 0.42 belangstelling 0.02 vijftig 0.1 straks 1.59 aandacht 1.42 officieel 0.14 serieus 0.6 zelfs 25.75 hierop 0 samenleving 0.37 aantrekt 0 jaar 147.31 vergeven 0 zeker 12.95 kabinet 191.8 verwijten 0.01 moeilijk 2.32 voeren 0.32 afgelopen 5.04 niks 4.29

27. ga 34.36 vanaf 9.61 deed 20.67 spaanse 9.51 enkele 12.75 maria 0.06 middag 3.2 onzichtbaar 0 nee 51.11 direct 1.22 gebleken 0.01 half 3.39 gedaan 5.26 bang 0.88 station 0.04 speech 0 verzamelde 0 juliana 1.71 aangekondigd 0.01 ambassade 0.15 maakte 6.62 fenomeen 0.01 enige 3.62 tekenend 0 zoals 13.57 zeilen 0 balkon 0.01 londen 44.69 lange 2.68 draait 1.04 rustig 1.64 heilig 0.01 rentree 0.01 aangevoerd 1.01 zweden 17.09 voormalige 1.65 lijnen 0 aanhang 0.07 keren 0.24 lijn 0.18 mooie 7.2 aanloop 0.05 uitlevering 0.01 stoep 0.06 dozijn 0 korte 0.39 wennen 0.08 gehouden 2.62 binnen 11.61 staan 32.23 openbaar 3 onderzoeksrechter 0 markt 2.53 vanmiddag 0.71 politie 767.16 werk 4.97 trad 0.01 opgepakt 3.74 poging 0.36 jaar 358.24 bravoure 0 betrekkelijk 0 scherp 0.03 virus 0.32 eerder 6.58 recent 0.03 onmogelijk 0.04 meneer 2.28 ie 0.86 verblijf 0.02 maak 1.26 soort 3.44

28. leek 14.83 oscar 0 juist 24.07 geleden 23.13 tendens 0 dertig 0.56 slechte 0.13 paar 24.32 ver 1.31 beesten 0.02 nee 35.49 nederland 54.6 leiden 1.01 zeer 1.91 steden 1.86 meteen 6.9 bovenop 0.03 volgend 0.25 daarvan 1.05 plaatst 0.04 dak 0.36 interessant 0.7 vorig 72.88 eigenlijk 24.86 hartstikke 0.03 plaats 6.24 planten 8.47 binnenstad 0.97 omstandigheden 0.16 zeggen 19.33 utrechtse 0.12 volk 0.19 chen 0.34 eenmaal 1.63 bijen 0.7 oh 0.18 gat 0.18 wanneer 3.44 zaten 2.31 houden 7.4 hield 2.74 zullen 5.2 bloemen 2.4 olga 0.02 ongeveer 2.46 stad 88.85 veertig 0.57 strook 0 belang 1.99 jaar 1363.07 eerder 4.23 groepen 0.03 duizend 0.24 groeide 0.57 begon 10.49 ontwikkeling 2.76 weliswaar 0.4 leidt 0.22 gooien 0.07 mochten 0.28 zalm 0.02 belangrijke 2.1 gehad 2.18 vragen 3.31 verwachten 0.19 se 0.09 neerzetten 0.02 probeerden 0.04 zoals 69.31 verschillende 5.2 fascinerend 0 huisdieren 0 daarom 18.05 omgeven 0 arm 0.09 hobby 0.03 volgt 0.14 nou 37.84 mooie 3.23 mannen 4.3 volgen 0.64 goeie 0.06 aantal 39.11 regels 0.21 mis 0.59 voorkomen 0.46 wonen 3.14 jongen 1.09 populairder 0 voedsel 0.07 werk 19.13 denken 14.55 tienduizenden 0.01 natuur 59.67 kwam 114.62 elementen 0.05 machtig 0 oplopen 0.01 vinden 11.58 gunstig 0.01 aanstekelijk 0

29. verhouding 0.02 onwaarschijnlijk 0 ziekte 31.73 zodat 0.31 cursus 0 zei 9.84 praktijk 0.66 vergelijkbare 0.01 hierover 0.01 termen 0.01 medisch 10.09 basis 1.23 voorlopig 0.08 vergoed 0.18 procent 77.42 gebeuren 0.22 gevraagd 0.17 zat 2.74 volwassenen 0.08 woorden 0.23 gedomineerd 0 kosten 4.35 eenmaal 0.52 geval 2.5 groep 7.23 houden 3.15 zullen 3.86 cvz 0.02 partijen 0.29 hele 8.1 helpt 0.35 schappen 0 positief 0.18 therapie 0.78 argument 0.01 medische 18.55 vormde 0 roer 0.06 natuurlijk 17.08 vormen 0.62 zie 6.17 voeren 0.08 bepaald 0.31 advies 3.4 zin 0.44 duren 0.01 kunt 4.03 voorbode 0 handhaven 0.01 vergoeden 0.13 bepaalde 0.63 moeders 1.24 informatie 0.5 anderen 0.69 medicijn 1.36 omslagpunt 0 weet 36.47 hogere 0.61 dezelfde 1.1 springt 0 acties 0.19 vertelt 4.26 zorgde 0.01 definitief 0.01 erg 3.32 bijvoorbeeld 14.5 soms 4.43 deskundigen 0.09 praten 0.63 alarm 0 effect 3.65 solo 0.02 letterlijk 0.12 tachtig 0.07 waarbij 2.58 stoppen 0.37 nou 16.42 aantal 19.69 beleid 0.13 aangepast 0.02 serieus 0.27 welke 2.28 ligt 3.49 uitwerken 0 twintig 0.4 mogen 1.3 vinden 5.62 schippers 1.12 baten 0 commissie 1.37 college 0.23 ga 6.45 gestegen 0.48 invalide 0 voorbeelden 0.03 algemeen 0.09 sneller 0.7 visser 1.5 orde 0.11 nederland 28.54 aspecten 0.01 aandoen 0 bevorderen 0.02 leiden 1.99 topper 0 pompen 0 persoon 0.21 verschrikkelijk 0.03 gekeken 0.1 verzekeringen 0 eigenlijk 8.21 beide 0.45 set 0 erfelijke 0.17 jongeren 6.11 onvoldoende 0.15 zeggen 7.16 herfst 0 geld 11.98 concert 0.3 nuchtere 0 denk 2.35 zorgverzekeringen 0 betalen 6.91 toekomst 2.39 effectiviteit 0 kom 1.94 wachten 0.29 zeldzame 0.21 binnen 5.31 dure 0.1 last 0.64 vroeg 1.43 jaar 640.1 geneesmiddelen 0.26 recht 1.14 concept 0.02 besloten 0.13 gezondheidszorg 2 duizend 0.08 verdriet 0.38 allemaal 10.2 adviezen 0.01 hou 0.13 hart 1.07 ethische 0 werkelijkheid 0.1 ziekten 0.37 manier 4.95 getrokken 0.02 beschouwd 0 fabriek 0.47 toegepast 0.04 ongeluk 0.01 raken 0.25 rechtvaardigheid 0 verstandiger 0 korter 0.02 acceptabel 0 brink 0.01 gesproken 0.08 gemaakt 1.92 treft 0.02 zeiden 0.02 medicijnen 9.45 gevolgen 0.77 ethisch 0 vis 0.33 verslechteren 0 volgen 0.12 daarna 1.43 vergoeding 0.12 stug 0 ruim 4.12 bereikt 0.17 vizier 0 dief 0 buitengewoon 0 gaande 0.01 doorslaggevende 0 spieren 0.11 honderd 0.24 bedoeling 0.15 keuzes 0.04 zeldzaam 0.01 relatief 0.36 soort 3.35

30. zussen 0.12 dicht 0.32 opbouwen 0.01 maanden 2.01 rustiger 0.01 periode 0.29 negentig 0.08 enkele 10.99 overstromingen 0.01 overvloed 0 middernacht 0 gevallen 0.18 nee 17 voorbeeld 3.82 mate 0.27 drong 0.01 voorts 0 half 0.78 nieuwe 41.38 gisteravond 0.04 verwachten 0.06 zorgde 0.03 beeld 1.33 kijken 7.5 konden 0.96 kanten 0.13 brengen 0.52 bestand 0.01 zestig 0.74 onderzoekt 0.03 noodweer 0.01 oppassen 0 rivier 0.36 zondag 19.86 schade 2.95 sleutelen 0 kun 5.63 zoveel 3.18 groep 4.53 lieve 0.09 zullen 3.01 vakantie 0.4 regen 51.88 water 92.14 stukje 0.12 binnen 12.56 zorgden 0 veranderd 0.07 voorkomen 0.75 kracht 1.75 overlast 0.71 sommige 1.26 ondernemers 0.5 nacht 17.15 stuk 2.05 wanden 0 flinke 0.06 puinhoop 0 waken 0 inventarisatie 0 uiteraard 0.15 voorstellen 0.04 dalende 0.01 rest 0.65 kijk 2.37 aangericht 0.03 zuid 432.51 plafond 0.01 vinden 5.09 afgelopen 10.29 stonden 0.03 simpelweg 0.05 zusjes 0.05 stel 0.56

Appendix B

Protocols

Relevance = 2
schietpartij **AND** discotheek **AND** Cambrai
boete **AND** studievertraging
werkloosheid **AND**allochtonen **AND** jongeren
Syrië **AND** bomaanslag **AND** president Assad
droogte **AND** voedselprijzen
(rentefraude**OR**liborrente) **AND** Rabobank
KLM/Air France **AND** financiële problemen
Asbest **AND** (Kanaleneiland**OR** Utrecht)
Olympische Spelen **AND** (doping**OR** wedstrijdfraude)
(Stroomstoring **OR** elektriciteit) **AND** India
drugs **AND** noord-brabant **AND** huiszoeking
Rotterdam **AND** (feest **OR** drank **OR**drugs) **AND** (rellen **OR** vernieling)
(geweld **OR** religieus geweld) **AND** Mombasa
sns reaal **AND** financiële problemen **AND** DSB **AND** vastgoed
asielzoekers **AND** (uitzetting**OR**asielweigering)**AND** Somalië
bosbrand **AND** Marbella
thuisonderwijs**OR** leerplicht **OR** leerstoornis **OR** hoogbegaafdheid) (leerstoornis**OR**hoogbegaafdheid)
(opsluiting**OR** bevrijding) **AND** Sattarov
Q-koorts **AND** besmettingsaantallen
Noordpool **AND** ijskap **AND** (broeikaseffect **OR**opwarming aarde)
internetbankieren **AND** virus **AND** tekstverwerker
gewelddadige overval **AND** (Voorthuizen **OR** Barneveld)
moordproces **AND** Gu Kailai
Tena **AND** klantgegevens
stijging sterftecijfers **AND** Nederland
verkiezingscampagne **AND** SP
speech **AND** Assange **AND** Ecuador **AND** VK
groeiende populariteit bijen houden in NL
vergoeding **AND** (ziekte van Pompe **OR** ziekte van Fabry)
overstroming **AND** Gulp

Query	Relevance = 1
1	(schietpartij**OR** geweld**OR** ruzie) **AND** (uitgaansgelegenheid**OR**discotheek**OR** cambrai)
2	hoger onderwijs **AND** (kosten **OR** collegegeld **OR** boete **OR** studievertraging)
3	((arbeidsmarkt **OR** werkloosheid) **AND** (allochtonen **OR** culturele discriminatie **OR** jongeren **OR** hoogopgeleiden))
4	Syrië **OR** bomaanslag**OR** president Assad
5	landbouw **OR** voedselprijzen **OR** voedselbeschikbaarheid **OR** droogte
6	rentefraude**OR** Rabobank **OR** liborrente
7	luchtvaartmaatschappij **AND** (financiële rapportage **OR** financiële problemen)
8	Asbest**OR** sociale huursector **OR** woningcorporaties**OR** Kanaleneiland
9	Olympische Spelen **OR** doping **OR** (sport **AND** wedstrijdfraude)
10	Stroomstoring**OR** India
11	(drugs **AND** huiszoeking) **OR** (drugs **AND** Noord-Brabant) **OR** (criminaliteit **AND** Noord-Brabant)
12	(feest **OR** drank **OR** drugs **OR** uitgaan**OR** Rotterdam) **AND** (rellen **OR** vernieling)
13	religieus geweld **OR** [geweld **AND** Kenia]
14	bank **AND** financiële problemen
15	asielzoekers **AND** (uitzetting **OR** asielweigering)
16	bosbrand
17	(thuisonderwijs **OR** leerplicht**OR** leerstoornis **OR** hoogbegaafdheid) **AND** reis
18	sekte
19	Q-koorts**OR** virus**OR** bacterie**OR** besmettingsaantallen**OR** giftige stoffen
20	Noordpool**OR**ijskap**OR**broeikaseffect**OR** opwarming aarde
21	cybercriminaliteit en veiligheid**OR** computervirus
22	Voorthuizen**OR** Barneveld**OR** overval
23	Bo Xilai **OR** (rechtzaak **AND** China) **OR** (corruptie **AND** China) **OR** (maatschappelijke ontwikkelingen **AND** China)
24	incontinentie **OR** Tena **OR** privacy in medische sector
25	CBS **AND** (bevolkingsstatistieken **OR** consumentenstatistieken) **OR** sterftecijfers
26	verkiezingen **OR** SP
27	WikiLeaks**OR** Julian Assange**OR** klokkenluiders
28	bijen houden**OR** imker**OR** bijzondere huisdieren**OR** natuur in de stad**OR** bedreigde dieren
29	bezuinigingen in de zorg**OR** vergoeding medicijnen **OR** zeldzame ziektes
30	waterschap**OR** preventie overstromingen **OR** overstromingen in NL **OR OR** schade door noodweer in NL

Appendix C

Topics

kwartaal bedrijf beleggers jaar winst aandelen amerikaanse aandeel beurs gisteren tnt omzet punten topman verlies vs postnl cijfers steeg aandeelhouders

wit zwart philips partij spelers blauw houten gelfand remise anand rood boris licht diagram zwarte partijen verlichting led kleur spel

media nrc krant bv rotterdam gegevens informatie handelsblad hoofdredacteur uitgever twitter kranten diensten site pers zie online journalisten wegener website

nederland nederlandse nederlanders polen oost land landen europa europese vinden poolse procent meldpunt deelnemers stelling europeanen velen japan volgens groep

britse olie londen landen shell india europa land klm volgens europese china jaar brazilie verenigd nederland koninkrijk brittannie britten nederlandse

festival muziek orkest amsterdam jaar opera theater voorstelling publiek nederlandse april tm tijdens mei maart nieuwe zondag dans nederlands voorstellingen

programma serie nederland omroep nieuwe camera televisie presentator kijkers aflevering uitzending nos waarin vpro beeld pauw avro nederlandse eo draait

oranje finale barcelona league real bayern club robben bondscoach champions madrid chelsea won rotterdam spelen wedstrijd nederlandse ploeg elftal spelers

duitse franse president hollande partij sarkozy duitsland berlijn frankrijk merkel premier verkiezingen procent nieuwe stemmen parijs fransen politieke pen politiek

muziek it do how band album zanger horizontaal made cd nummer nummers liedjes nieuwe american they zangeres you past jazz

auto werk kunst nieuwe tentoonstelling museum motor kunstenaar kunstenaars jaar expositie rijden schilderijen pk foto bmw art modellen porsche nedcar

politie jarige jaar volgens gisteren slachtoffer agenten justitie werden verdachte auto onderzoek foto aldus tijdens mannen moord rotterdam slachtoffers woensdag

onderzoek patienten ziekenhuis onderzoekers jaar ziekte arts procent patient blijkt ziekenhuizen medische artsen rotterdam volgens behandeling nieuwe medisch medicijnen zoals

prins koningin burgemeester prinses italiaanse friso laan beatrix koninklijke willem foto alexander rome occupy italie koning gisteren roma tijdens maxima

gisteren president volgens land leger iran jaar syrie israel zei regering syrische militaire ap rusland militairen amerikaanse regime poetin buitenlandse

kinderen moeder kind ouders jaar dochter foto gezin zoon meisjes meisje vertelt baby huis huwelijk jarige thuis geboren relatie moeders

bosch willem zwolle fc punten jansen amsterdam ploeg almere ii play eindhoven seizoen wedstrijd offs sparta voorsprong helmond sc nederland

kunt bent hebt breivik werk doe wilt zult anderen ga nieuwe zullen liefde partner geef maak geld leren energie gevoel

weet nou nee natuurlijk zitten helemaal koers hele ga beetje allemaal jullie zie staan misschien paar zeggen hoor zelfs terwijl

amerikaanse new york obama vs staten romney president amerikanen verenigde dollar amerika washington campagne santorum republikeinse homo zelfs zei mitt

tagesschau heute das co und lokalzeit im duitsland von aktuell duitsl aktuelle stunde aller wissen anne ein planet pasen servicezeit

bedrijven nederland energie duurzame visser bedrijf gas milieu stroom directeur jaar nieuwe nederlandse duurzaam volgens co zoals groene elektriciteit aldus

directeur raad ajax club jaar nieuwe bestuur cruijff commissarissen johan amsterdam contract vertrek commissaris voorzitter gaal advocaat jan directie leden

graden wind lucht morgen land droog regen noorden zuiden frans temperaturen weekeinde lokaal temperatuur zondag nacht vanuit overdag buien sneeuw

zout vlees water wijn smaak peper boter doe beetje minuten pan bak vis gr witte keuken recepten restaurant suiker koken

spelen olympische ronde londen jaar tijdens wk olympisch plaats race tour derde nederlandse seizoen rechtstreeks ploeg team vorig jarige etappe

nederland jaar land turkije nederlandse joan tijdens argentinie turkse foto songfestival finale mol voice zweden evenement john zweedse landen mei

noord west zuid wk meter kramer korea jaar goud oost vries kilometer sven ploeg afstanden seizoen wust allround nk heerenveen

fc psv twente ajax trainer seizoen az feyenoord wedstrijd club spelers heerenveen ploeg heracles voetbal league ado vitesse eredivisie groningen

jaar zoals jaren bijvoorbeeld zelfs juist jongeren zeker natuurlijk vraag nieuwe daarom manier nederland belangrijk weet eigenlijk allemaal misschien moment

jaar museum cultuur subsidie universiteit rotterdam amsterdam directeur onderzoek nederlandse raad fonds nederland instellingen nieuwe kunst geld advies culturele instituut

jaar procent bedrijf vorig aantal nederland nieuwe volgens markt omzet bedrijven verkoop afgelopen jaren nederlandse blijkt verkocht verkopen groei amsterdam

gemeente rotterdam stad amsterdam haag wethouder haagse jaar nieuwe schip rotterdamse boot volgens burgemeester bakker woensdag aldus haven foto gemeenteraad

facebook nieuwe internet apple ns online google klanten app mobiele amsterdam trein ipad foto digitale spoor netwerk schultz gebruikers systeem

jaar foto hotel museum jan directeur prijs amsterdam bezoekers nieuwe nederland tijdens gasten restaurant amsterdamse nederlandse voorzitter vertelde gouden hotels

regisseur vs films verhaal acteur waarin acteurs regie liefde genre rol personages drama voorstelling blijkt actrice nieuwe thriller dood soms

bank banken financiele geld vestia jaar ing rente betalen abn amro risico volgens procent kosten klanten bedrag hypotheek nederlandse sector

ging zei kwam moest hadden wilde later kreeg zag jaar zat stond deed vond wist vroeg leek bleek daarna begon

chinese kpn china mexico mexicaanse peking carlos slim bod volgens amerikaanse kony bo america video jaar spaanse chinezen klanten land

volgens justitie jaar onderzoek rechtbank uitspraak ministerie raad wet advocaat beroep hof straf openbaar boete minister commissie veiligheid opstelten amsterdam

bbc news weerbericht journaal world to nos show regionaal one night time murdoch at trip zoo tinga it whitney little

boek dood jaar oorlog schrijver boeken joodse fortuyn pim blz mei roman jaren schrijft slachtoffers schreef verhaal tijdens waarin wereldoorlog

water ijs natuur jaar aarde tocht foto meter dieren langs kilometer friesland vissen friese elfstedentocht land zee planten winter nederland

pvv cda pvda kamer vvd minister kabinet partij rutte wilders haag kamerlid groenlinks d nederland politieke leider premier gisteren politiek

europese griekenland griekse economie jaar procent spanje landen europa eu land nederland banken eurozone brussel economische groei ecb grieken spaanse

school onderwijs leerlingen studenten scholen schiphol jaar kinderen ouders opleiding scholieren leraren docenten leren examen leerling middelbare diploma klas bijsterveldt

foto jaar jan natuurlijk weet mooi terug blij ga helemaal eigenlijk vertelt winkel staan schoenen wilde trots paar smit denk

jaar werknemers fnv nieuwe gemeenten volgens werk procent werkgevers sociale betalen bonden geld cao personeel pensioen voorzitter btw kosten overheid

huis boek woord woning zoals mens werk huizen boeken staan waarin schreef misschien woorden bijvoorbeeld kunt denken nieuwe berg wonen

zuid afrika kerk brand gebouw jaar volgens bewoners afrikaanse werden rotterdam jaren meter noord nieuwe pand gebouwen foto gisteren grond