Factors Indicating Novelty of Tweets

Wenjie Zhong
VU University
Amsterdam, The Netherlands
wzg600@student.vu.nl

Abstract

With the advent of the internet and social networks, the production side of information has shifted to a larger amount of content makers. Especially in the social network of tweets, the ever-growing amount of noise and duplication has called the need for curation. Current approaches for filtering duplicate data and noise are focussed on either text analysis, relevance, machine learning or a combination of these three methods. This paper proposes an approach that uses the combined traits of relevance, similarity between tweets, sentiment scores and factors gathered from the Twitter platform. Similarity distance between tweets can aid in filtering repetitive old information, seed words can be used to filter non-relevant tweets and statistical factors like popularity of tweets can be used in training the predictive model. Training and testing data for machine learning is gathered from crowdsourcing experiments. The CrowdTruth platform and methodology is used for filtering and improving the quality of the annotation data. The collected features as input produced promising results, several features showed a significant correlation with novelty. The collected features and methods were effective in predicting novelty and removing noise or duplicate tweets.

Categories and Subject Descriptors

D.2.8 [**Novelty Detection**]: Crowdsourcing

General Terms

Machine learning, Information Retrieval

Keywords

CrowdTruth, Novelty, Crowdsource

1 Introduction

One of the main obstacles of the web is the large amount of unstructured data, which causes an explosion of information. People can upload anything they want, resulting in an impenetrable pile of information, consisting of duplicate and unreliable articles, blogs and other web media [49]. In 2008, the amount of monthly active Twitter users was between 6 and 7 million. [44], this number has grown to 302 million active users. About 500 million tweets are published every day on the network. An average of 350.000 tweets are sent every minute [47]. On the side of online news articles the size of information data is somewhat smaller, but similarly unwieldy. A commercial news aggregator Google News has more than 4000 news sources, Yahoo News aggregates from more than 5000 content publishers. Several other news engines like Newsbot, Findory and NewsInEssence do the same task. All of them acknowledging the need for content curation [16]. To better visualise the size of the internet, by the time the reader has come to this sentence, 1 million gigabyte of data has been transferred over the internet by the most popular websites and online services [1].

Data science has become an important research area for these vast amounts of unstructured data. In the past decades, several approaches in this field are being developed. Some methods are used to extract meaningful entities from raw text, including annotation enrichment, and relation propagation between the discovered entities [33]. Semantic web technologies provide mechanisms to store the extracted data. The semantic database offers the possibility to query the data in a human-friendly way, which may raise the accuracy of the results [5]. A highly integrated project by IBM of several artificial intelligence techniques is named Watson. Watson uses a combination of natural language processing, machine learning, entity type extraction and coercion to analyse both unstructured and structured data [24]. This system could be used to answer Jeopardy questions or solve medical diagnostics for doctors[23].

The aim of this paper is to calculate novelty of tweets. We also examined if features like links and pictures have an influence on novelty. We accomplished this in the following way. First, we collected tweets and its meta-data (containing the features). Afterwards, the gold standard needed for machine learning is retrieved via crowdsourcing. With the gold standard, a model is created with supervised learning. This model can distinguish between old and novel content. The remaining part of this chapter focuses on the context of the research, the problem statement and the research questions. Chapter 2 describes the related work, diving into the definition of novelty in relation with ongoing events. Subsequently, chapter 3 describes the event space and chapter 4 explains the techniques and methods proposed for the experiment. In chapter 5 the crowdsourcing task and it accompanying metrics are described. Chapter 6 details the set-up of the experiment, the results and discussion. After that, the paper is concluded in chapter 7, answering the research questions based on the results of the experiment. Finally, chapter 8 presents possible future work.

1.1 Context

Activism and specifically the whaling topic are the domain of this research. Activism encompasses the use of confrontational action, such as protests, in support of a cause. Activism is important, because it plays a key role in shaping the future. Whaling is in this scenario a fitting example of activism, since it is relatively well covered in both academic research and general news networks [41, 39, 36]. In social science, activists are a huge research topic, due to the fact that activism can lead to new political activities and significant changes. For example, in 2011 the activist group Sea Shepherd Conservation Society obstructed Japanese whaling ships. As a result of the global attention and pressure, the Japanese government was forced to call back their whaling fleet and lower their whaling quota [21]. This example stresses the importance of consequences of activism [35]. It is therefore important to improve the detection of such events, either by curation of content or noise

filtering. Analysing news sources and opinions in tweets, can provide new important information about activist events.

1.2 Problem Statement

Currently, there is an overwhelming abundance of information available on the web [49, 44, 47, 1]. The over-abundance of that information caused many users to have issues managing and absorbing these information resources [16]. Many attempts in the past tried to overcome these issues with algorithms to filter the noise and retrieve both the relevant, reliable and novel stories [25, 12, 50].

Present Twitter features do not encompass a good method to find novel information on one specific event. For instance, if you search on an event term on Twitter you will only get unfiltered tweets, most of the times completely irrelevant for the event. This problem is introduced by the advent of web democracy, the notion that everybody on the internet has the ability to publish content [19]. The information overload is a problem for the general audience. It would be helpful for them if the social network could present fewer and better tweets. For another type of audience, namely social scientists, the abundance of information on the new web is a social goldmine. For example, whispers that occur between guests of an event would be lost. Now these discussions are shared on Twitter and possibly identifiable with a hashtag [46].

However, there is another issue coming with web democracy, the bias introduced by the stream of new authors can give a wrong perspective on some study cases. One team of scientists suggests that large-scale studies of human behaviour should be held to higher methodological standards [42]. For example, researchers should consider possible biases in Twitter pollings. People who elect to polls could have a bias for either side of an argument [13], even large groups of people on the web. Novelty detection will not help in negating bias, but it can help with detecting new biases.

To summarise, the present web introduces new kinds of information overload, caused by the democratisation of content authorship. Everyone on the internet with a social network account duplicates or creates new information pieces or biases. This poses a problem for members of the general audience or professionals. Finding and absorbing unique and relevant information becomes harder and avoiding biases in research projects are a problem.

Nowadays, some users consume news stories through a news-feed, like RSS. This method is useful because a user can only receive new information from the specified sources. This prevents the user from being overloaded with too much information and only new information articles are pushed. However as specified in the previous sections, when the information overload is too big, manual curation becomes harder. Some social networks like Twitter provide a news stream of tweets, some of them containing interesting and novel news articles, pictures or videos. However duplication of certain tweets are rampant, not relevant or old. An automated curation system that can decide when a new message is novel, could be a solution to the information overload problem.

1.3 Research Questions

This research aims at identifying novelty measures for tweets. Thus, the main research questions are as follows:

- Can we measure the novelty of tweets in the domain of activist events?
  - What factors indicate the novelty of tweets?
  - What is the level of influence of different factors for determining novelty?
  - What is an accurate and web scalable approach to determine novelty of tweets?

2 Related Work

2.1 Mapping Activist Events

MONA project (Mapping Online Networks of Activism)

As discussed in 1.1, the use case in this research concerns the whaling event. Activist events such as the whaling event are highly important, because they have a significant role in shaping social perspectives. Scientists studying these phenomena and activists want to know how activists shape these social perspectives. One project called MONA [40], helps by visualising important movements and events by analysing and presenting them over large amounts of data. People from both social and computer sciences worked on developing the MONA project. Computer techniques like named entity disambiguation and date normalisation are used for finding important activity patterns. A pipeline of these tools can help detect certain events and its actors or places.

2.2 Existing approaches to Novelty Detection

2.2.1 Similarity distance measure

A relatively early research paper concerning novelty detection experiments with several similarity distance measures [3]. If a new document is highly dissimilar because of new words, this document can be considered as new. The authors rank a dozen of different measures in four different conditions:

- Performance of novelty measures on known relevant sentences in the training set.
- Performance of novelty measures on best relevance results in the training set.
- Performance of novelty measures on known relevant sentences in the testing set.
- Performance of novelty measures on best relevance results in the testing set.

These four conditions exist because the authors use two different retrieval methods for training and testing data: TFIDF and a two-stage language modelling method [52]. Some of the novelty measures are cosine distance, variation of the word count method and multiple language models. Two distance measures are discussed in detail here, one that uses language models and the other one that counts new words.

$$N_{ds}(S_i|S_{1,\ldots,}S_{i-1}) = \frac{min}{1 \leq j \leq i-1} KL(\Theta_{S_i}||\Theta_{S_j}) \tag{1}$$

$$p(w|\Theta_{S_i}) = \frac{len(_{S_i})}{len(_{S_i}) + \mu} p(w|\Theta_{ML_{S_i}}) + \frac{\mu}{len(_{S_i}) + \mu} p(w|\Theta_{ML_{S_{1,\ldots,}S_n}}) \tag{2}$$

In Dirichlet Smoothing (equation 1), the novelty score (N_{ds}) is calculated between sentence i (S_i) and its most comparable sentence analysed earlier (S_j). Kullback-Leibler(KL) determines the distance between the two language models(Θ_{S_i} or Θ_{S_j}), one based on S_i and the other S_j. Both models are given by equation 2). Chance of word w in sentence Si $p(w|\Theta_{S_i})$ determines what

maximum likelihood model to use. If the sentence length is small, len(S_i), than the focus is on the model that accounts all sentences ($ML_{S_{1,...,S_n}}$). If the sentence is big, the focus will be on the model that only accounts the sentence Si (ML_{S_i}). The advantage of this algorithm is the dynamic smoothing dependent on sentence length, this method ensures that each sentence influences the outcome relatively equally. To summarise, Dirichlet Smoothing checks if a new sentence has enough novel words, to distinguish itself with the most comparable earlier sentence. Additionally, using different maximum likelihood model ensures that each sentence has equal importance. The other distinct similarity measure simply counts the amount of new words a second document has. More new words means a more novel second document.

2.2.2 Novelty detection using Local Context Analysis (LCA)

The experiment from Fernandez and Losada (2007) consists of two tasks [22], selecting relevant sentences and novel sentences. The method for selecting novel sentences is the same as previously discussed paper (new word count). The novel part of this paper is their use of LCA for relevancy measure. A group of researchers stated that novelty should not be reliant on novelty measure alone but should also be based on a set of seen sentences with common meanings [54]. This method ensures that novelty score is not tainted by a past set of sentences that are completely irrelevant. For example, if one uses documents from disparate events and topics. The chance of retrieving new documents in a chronological time line is big. However, if one only retrieves documents from one specific event, the chance on finding new words will become smaller. The new words that this method finds belong to relevant and more novel documents.

The last paper [48] discussed in this paper uses distance measures described earlier (word count, TF-IDF and language models). The interesting part is their utilisation of entities. The authors come to the conclusion that using vectors with only named entities perform better than using vectors with all words. The core reason of this method, is the fact that an entity has more value than normal words. Sentences contain a lot of noisy words that do not point to any distinct topics. If one only looks at the entities, you can gain more novelty information without the disturbance of non-words.

2.2.3 Hybrid-human machine

Besides detecting new words, objects or entities in text documents, a combination of crowdsourcing and machine learning is needed for making a predictive model. Currently, crowdsourcing is used in several tasks: relevance judgements[4], improving search systems (with tweets)[20, 9] or digital humanities[37]. The combined usage of curation with human intelligence and scalability of machine learning is a promising approach for making predictive models, this approach is also called hybrid-human machine information systems [17].

2.2.4 Utilised methods

The previously discussed papers all describe interesting methods and insightful results. The first section discussed similarity measures to calculate how much novelty a new document brings. Cosine distance is a highly successful method to determine new event detection [30]. However, existing literature has shown that it effectiveness decreases with short documents [43]. Using Levenshtein distance is more effective for short texts. This method is better suited for tweets, because it not only measures the distance between documents but also the distance between individual words.

Besides similarity, researchers also use LCA to ensure that novelty measures are not tainted by unrelated sentences [22]. This advantaged can be implemented with the use of relevancy of events, identified by seed words from domain experts. Shared resources like Wikipedia can also help in achieving a better relevancy. Another important information piece is the added value of an entity. Including entities as a separate feature can hopefully improve the novelty scores [48]. Lastly, tweets have extra features besides entity that can bring more predictive value [44]. For example, the popularity of a tweet can point to novelty, because of its uniqueness. Credible authors with a lot of followers and a known profile picture, probably have lower chance of posting noise and spam. Another separate feature is sentiment score, a sub-event with a new sentiment can point to new changes. With these variables, a formula can be produced with machine learning, that can assign different weights to the related variable equation 3. The novelty could be calculated by summing all the products of variables and their respective weights. If the score is higher or lower than the threshold, the tweet is either novel (+1) or not-novel (-1).

$$f(x) = sgn(w \times x_* + b) \tag{3}$$

2.3 Novelty

2.3.1 Novelty Definition

In the previous chapter, an automated system for detecting novelty was discussed. For such a system to work, it has to know what being novel entails. Some prior research works use TF x IDF or cosine similarity to define how much two documents differ [2, 53, 3]. A general paper about novelty detection in texts discusses several techniques [48]. The basic principles of these techniques state that documents are different when the amount of utilised words are different. This difference could be expressed in new words, so the general techniques transform texts in a bag of words. When a new bag of words contains a proportional new set of words, then it is deemed as novel. The proportion could be calculated with methods like TF*IDF. Using these techniques to analyse novelty in small documents creates subpar results, because different words can point to the same information. Novelty can also mean more than just new words or subtopics, an added sentiment dimension can point to new and important information. Some existing works have been done to solve this issue [51, 43].

In the domain of novelty detection in twitter feeds, the aforementioned techniques are partially irrelevant, considering that tweets have added dimensions that can aid novelty detection not present in flat short texts. Tweets contain information about the original author, the person the tweet was directed to, hashtags or topics and other multitudes of statistical information. This information can be harnessed to aid in novelty detection.

2.3.2 Novelty dimensions

In the domain of social networks, some existing research work has been done concerning news and relevancy detection in tweets [7, 11, 16, 38, 44]. In the early days of the Twitter network research was focussed on why people use Twitter and unique properties of Twitter. One research particularly focussed on analysing tweets to target the topological and geographical properties of tweets [28].

For news and relevancy detection in tweets, one exemplary approach takes three dimension of information in consideration, content sources, user interests and social voting [11]. These dimensions are deemed most efficient in selecting the highest scoring tweets (table 1). The score of the tweet is determined by the user, the score is given after the user is presented with a collection of tweets gathered by the algorithm. This approach is useful when both relevancy and novelty are needed. Furthermore,

Table 1. The different dimension of a tweet, each containing its category of variables describing the tweet.

Dimension	Description
Content	The body of a tweet, words, hashtags, mentions, sentiment and urls.
Author	The user who posted the tweet, followers count, verified status by Twitter and Wikipedia and client used.
Interactions	The properties that can be mutated by other users as in retweet count, favorite count and reply status.

Table 2. Novelty factors and their dimensions

Factor	Description	Dimension
Content of tweet	entities, similarity score, hashtags and mentions	Content
Sentiment	difference in sentiment	Content
Author	New author	Author
Geography	A tweet could have a new location	Content
Source	tweet has new or other sources	Content
Social Status of tweet	Has the tweet been retweeted, liked or is it part of an conversation	Interaction

the relevancy is dictated by the interest model from the users. For the original application of news story detection, this method is deemed sufficient. For novelty detection the sources should go further than only URLs, all words and entities should be taken into consideration. For an event-based novelty detection model, the user interest dimension can be removed. Instead of the user providing words of interest, seed words given by domain experts can dictate the relevancy of tweets. Social voting could be broadened to statistical properties of tweets, like followers count, favourite count and retweet count.

2.3.3 Novelty Factors

When looking for novelty in new tweets, one has to look for multiple reasons why a tweet is novel. Although novelty at face value can mean different words, but different words can have the same meaning. Vice versa, the same words can have a different meaning. For example, the sequence of words can result to different moods or an added question mark can bring novelty to the same sentence. When you look at two identical tweets by different authors, a new credible author can bring novelty to a tweet. For example, an authoritative figure that republishes old content, can give new found importance to a tweet. So the novelty of a tweet can come from new hashtags, mentions, source, different words (similarity distance) or a sentiment difference. Table 2 contains these factors and their corresponding dimensions from a tweet.

3 Event space and Data

As stated in the context chapter, the event space is created around the activist event whaling. We asked domain experts from the social science department to provide a collection of seed words to create a general topic space about the whaling event. The seed words are used to mine tweets. The data mining process and data set characteristics are described in chapter 4. All the tweets are related to the event of whaling. The tweets were gathered using the Twitter streaming API in the spring of 2015. Between the end of March and the beginning of April a gap exists in the dataset, because of problems with the API.

Table 3. Seed words gathered from social scientists describing the whaling event

Events	Location	Actors/organizations	Other
commercial whaling	Japan	Japan Whaling Association	harpoon cannon
whaling	shops	International Whaling Commission (IWC)	harpoon
hunting	restaurants	Institute of Cetacean Research	markets
moratorium	North Pacific Ocean	pro- and anti-whaling countries and organizations	whale meat
quota	Southern Ocean	Nations	
	Antarctica	Scientists	
	factory ship	environmental organizations	
	factory ship Nisshin Maru	United Nations International Maritime Organization	
	security patrol vessels	Japan Fisheries Agency	
		Antarctic Treaty System	
		Anti-whaling governments	
		Anti-whaling groups	
		Greenpeace	
		Japanese government	
		World Wildlife Fund	
		Ocean Alliance	
		Sea Shepherd Conservation	
		Society	
		NGOs	

3.1 Seed words for selecting relevant Tweets

The seed words in table 3 are provided by domain experts, related with their research on activist events [14, 15]. Each seed word is categorised by the domain experts as event, location, actor/organisation or words without a particular category.

3.2 Crowdsourcing Relevance

Even though the tweets are gathered based on seed words, there is still a level of difference between relevance among the tweets. This graduation of difference is expressed in the relevance span score. The whaling tweets are also used in the crowdsource task [27], whereby the workers have to rate how much a tweet is related to the whaling event. It would be interesting to see what kind of aid this score can provide for detecting novelty. The snippets and the 'relevance span scores' are used for the current experiment. The scores here are reduced to mean scores of unique tweet snippets. To generate relevance score for new tweets, the presence of the snippets in tweets are collected.

Thereafter, the mean of the scores of known snippets are calculated.

4 Methodology

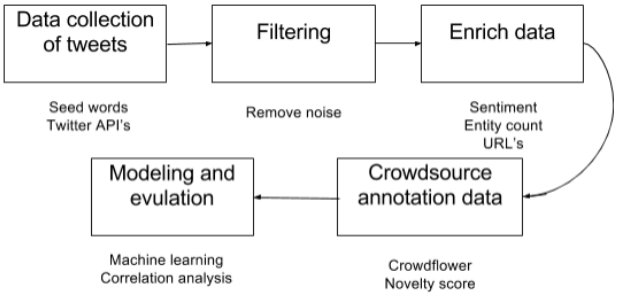

Figure 1. Workflow overview

This research aims at measuring the novelty of tweets. Figure 1 is a general view of the process of achieving this goal. First, tweets are collected via the Twitter streaming API based on the seed words (table 3). The whaling topic could be tracked by presenting the API a comma separated list of seed words. Secondly, the tweets are filtered, for example, by removing non-relevant features. Next, the metadata is extended by calculating new features, such as sentiment scores and entity count from the whaling Wikipedia page. After that, the feature set is complete and ready to attribute selection and feature weight determination. The weights of the model are determined with a machine learning approach using a crowdsourced training set. Next, the novelty scores are calculated using the features and weights from the previous step. Finally, an evaluation is performed using a set of tweets that was manually checked for novelty.

4.1 Data collection

The tweets on whaling were mined in a timespan of roughly three months using the Twitter streaming API. The search keywords consists of most of the words in the seed list (table 3). Only words that are too common like 'Japan', 'markets' and 'hunting' are avoided. The streaming API is used since the general Twitter API only provides large searches for tweets not older than a couple of weeks. All available tweet fields, as described by the Twitter API Developer documentation, are collected.

The retrieved data sets from the Twitter API contains almost all of the useful features, with one exception. Similarity distance is the measure of difference between tweets. This is calculated with the Levenshtein distance [31]. This measure is better suited for tweets, because it not only measures the distance between documents but also the distance between individual words. As in the amount of changes of letters are needed to reach the same words.

A week after streaming and storing the tweets, the retweets and favourites are recollected using the normal Twitter API. The average life-cycle of a tweet, including retweeting and favouring it, is a couple of hours [26], so to be safe the tweets are updated at least one week later. First, we collect all the tweet identifiers from our current dataset. Next, we use these identifiers to recollect the tweets, using the statuses_lookup function of the Twitter API. Collecting the retweets and favourites at this stage in the process helps us to reduce the corpus using the activity filter in the next paragraph, which means less computations later in the process.

Table 4. Utilised filters for tweets

Filter	Description	Rationale
chat and retweets filter	This filter removes tweets that start with: RT @, MT @, using regular expressions.	We only want to work with the source tweets, since retweets are ambiguous and chatter is normally not meant as a news message for the crowd, but to talk to someone specific.
Activity filter	This filter removes the tweets without at least one retweet or favourite (like), based on the retweet_count and favourite_count fields of the tweet.	By filtering tweets with at least one retweet or favourite, we indicate the tweet as active. Hence, it is assumed that at least one person read the tweet, which makes it probable to be a credible tweet.
Spam filter	A random excerpt of the tweet database contains many quickly recurring messages, these can be identified as spam.	The distance metric Levenshtein is used to detect repetitive messages that occur in a short time span. For example, if one phrase repeats itself 20 times within a couple of minutes, it is detected as spam.
English language filter	This filter removes all tweets that are written in non-English, based on the lang field of the tweet.	The tools and plugins we use for this experiment are based on the English language.

4.2 Filtering

The raw tweet corpus gathered by the Twitter streaming API may only consist of tweets related to the whaling domain, but it contained a vast amount of unrelated tweets. To clean noise like repetition, non-English tweets and chatter, several custom filters are used. Table 4 describes the various types of custom filters and the rationale behind the choice.

4.3 Data enrichment

The dataset is enriched with new possible novelty features, using natural language processing tools. Table 5 describes these extracted features and the plugins and tools used in the extraction process.

4.4 Feature Extraction

The next step is to extract the features related to novelty. By default, the Twitter API returns advanced result sets with a vast amount of features. Only features related to novelty are extracted, where the selection is based on the novelty factors in table 2. Consequently, the extracted features are categorised in the following dimensions: interaction, content and author.

4.5 Model

Weights are assigned to important features from the lists elaborated earlier. The weights are determined using feature selection. As

Table 5. Utilised scores for tweets

Feature	Description	Tools and plug-ins
Sentiment score	This feature calculates the sentiment of all tokens in a tweet, based on the sentiwordnet corpus. The sentiment scores are added altogether, which represents the tweet sentiment.	NLTK sentiwordnet corpus
Entities count	This feature counts the number of entities from Wikipedia and other sources in a tweet	Tagme, DBpediaSpotlight
Relevance Span Score	Score [27] regarding the relevance between the tweet snippet and the event.	Crowdflower

Table 6. Extracted features in the interaction dimension of tweets

Feature	Description
favourite count	number of likes
retweet count	number of retweets

Table 7. Extracted features in the content dimension of tweets

Feature	Description
entities count	number of entities in text
mentions count	number of mentions in tweet
URL count	number of URL's in tweet
hashtag	hashtags in tweet
similarity distance	Levenshtein similarity between tweets
sentiwordnet	sentiment analysis based on the sentiwordnet corpus

Table 8. Extracted features in the author dimension of tweets

Feature	Description
friends count	number of friends
listed count	number of lists that contains the author
followers count	number of followers
has url	profile contains a URL
description length	length of profile description
created at	the date the profile was created
verified	the account is officially verified by Twitter
has defeault profile image	user uses the default profile image
has banner image	user uses custom banner image
relevance span score	strength of relation between tweet snippet and event

this is a supervised machine learning approach, the following step concerns the creation of a training set. We use Crowdflower in order to create several crowdsourcing task instances to determine the novelty of tweets. With this training set, machine learning will be possible and the learned model can be utilised for detecting novelty in new sets of tweets. The next paragraph explains the setup of the crowdsourcing tasks used to create the training sets.

5 Crowdsourcing task

The task in figure 3 is presented to the crowd. The task consists of two tweets, the job of the worker is to define the novelty factor of the tweet. In other words, the worker has to decide if one tweet is more novel, less novel or equally novel in relation with the other tweet. The worker can also annotate the tweet as irrelevant to whaling. The event is described by a summary text gathered from the Wikipedia whaling page. Sentences with the seed words (table 3) given by the experts are used to retrieve the sentences in the summary. The worker can also highlight at least one word in each tweet, that contribute to the novelty of the tweet. The goal of the crowdsourcing task is to gather gold standard data needed for supervised learning. For this kind of machine learning you need the input data together with its expected output.

The tweets used in the task range between March 9th and March 14th, 2015. In the pairwise comparison, the first tweet is compared with the second the tweet. Afterwards, tweet 2 in the timeline is compared with tweet 3 and so on. It is unnecessary to repeat the comparison of an earlier tweet, because you want to only know if a new tweet is bringing more novelty.

5.1 Crowdtruth

To collect the gold standard data, the Crowdtruth methodology is used. Crowdtruth works on the premiss that disagreement does not necessarily point to bad annotation data [6]. Disagreement could perhaps point to bad task design or bad input data.

Worker vectors are created from the raw Crowdflower results. Each vector tells what the worker stated, concerning which tweet was more novel, relevant or which words were important. The tweet vectors are then in turn created from the aggregated worker vectors. In other words, the annotation data from one tweet is aggregated from all worker vectors concerning that tweet.

5.2 Tweet Vectors

The Crowdtruth methodology requires its input data in specific vectors. Each vector consists of variables, denoted as 1 ore 0. The vectors in this experiment are divided in four different categories. Aggregating everything to one vector, lowers the agreement data unfairly. This is unfair, because the chance of workers agreeing on both word highlighting and novelty is too small. The worker has to agree on two different categories. Selecting the same words in a long tweet and with the same combination of highlighted words, the worker also has to agree on the novelty of a tweet.

5.2.1 Novelty Selection Vector

The novelty vector contains information about the novelty annotation of the crowd worker. The vector has the identity of both the worker and tweet with the binary condition of the four options. Only one of the following options can be true: the tweet is more novel, equally novel, less novel and non-applicable (table 9).

5.2.2 Highlighted Words Vector

Besides the novelty factor that the worker can annotate, the crowd can also give information about the words in the tweet. As such, which words or entities specifically in the tweet was novel. In the

STEP 1: Read carefully the description of the topic *"Whaling"*

During the 20th century, Japan was heavily involved in commercial whaling. This continued until the International Whaling Commission (IWC) moratorium on commercial whaling went into effect in 1986. Sea Shepherd Conservation Society contends that Japan, as well as Iceland and Norway, is in violation of the IWC moratorium on all commercial whaling. Japanese whaling is currently conducted by the Institute of Cetacean Research, using the scientific research provision in the IWC agreement. The whale meat from these scientific whale hunts is sold in shops and restaurants. The International Court of Justice (ICJ) ruled that the Japanese whaling program in the Southern Ocean, begun in 2005 and called "JARPA II", was not for scientific purposes and ordered the cessation of JARPA II in March 2014. Japanese whaling hunts are a source of conflict between pro- and anti-whaling countries and organizations. Nations, scientists and environmental organizations opposed to whaling consider the Japanese research program to be unnecessary, and that it is a thinly disguised commercial whaling operation. Greenpeace argues that whales are endangered and must be protected. Japanese whaling activities have historically extended far outside Japanese territorial waters. Factory ships were not used by Japan until the 1930s. As whale catches diminished in coastal waters, Japan looked to Antarctica.

Figure 2. Summary of the whaling event given in the crowdsourcing task.

Figure 3. Example of an annotated novelty task.

Table 9. Vector of novelty selection

worker_ID	tweet_ID	more_novel	equally	less	NA
3587109	38654	0	1	0	0

crowdsource task this step is given as: STEP 3: Highlight words in the tweet that point to new information. Again the vector contains both the identity of the tweet and the worker. The remaining units are indexes of the words in a tweet, the values indicates if the words point to new information in relation with the whaling event.

The highlighted words vectors are also divided in two categories, one category is called the novel words vector the other one is called non-novel words vector. Besides the word indexes, the two words vectors contain an indicator named 'NONE', affirmative when the worker did not highlight any words in that tweet. Also, the novel words vector has one extra column for detecting the case when the worker chooses more novel or equally novel and does not highlight any words. In this case the worker is contradicting himself, a tweet can not contain zero novel words and be considered novel.

5.2.3 Relevance Vector

The last data vector contains the same worker identity and tweet identity (table 12). The other two values can either be 0 or 1 exclusively. The first unit of the besides the worker identity and tweet identity, indicates if the tweet is relevant to the whaling event. The last unit indicates if the tweet is irrelevant to the whaling event.

5.3 Worker Metrics and Spam Filtering

5.3.1 Cosine Similarity and worker-worker agreement

To check the agreement factor of crowd workers, two measures of the Crowdtuth platform are utilised [58]. The cosine similarity measure and worker-worker agreement are utilised as an evaluation measure or a method to define if one specific worker diverts too

Table 10. Novel vector, contains information about highlighted words in a tweet and if the worker failed to highlight necessary data.

worker_ID	tweet_ID	word0	...	w27	NONE	check_failed
20043586	38654	1		0	0	

Table 11. Not-novel vector, contains information about highlighted words in a tweet.

worker_ID	tweet_ID	word0	...	w27	NONE
20043586	29347	0	1		1

Table 12. Vector of relevance

worker_ID	tweet_ID	relevant	irrelevant
3587109	38654	1	0

much with the rest. To further explain the specifics of the measures, the cosine similarity is calculated with dot product and magnitude, as expressed in equation 4. In the equation, V_{u,w_i} is the annotated vector of worker i for unit u. V_u is the aggregated annotation vector of the crowd for unit u (minus the annotations of worker i). Cosine similarity expresses the degree of similarity between these two vectors. For example, if the aggregated vector is 0,1,11,0 and the worker chose 0,0,1,0 -> the user has a high agreement score, regarding the cosine distance between vectors. With the cosine measures of both vectors, one defines "how close" these measures are.

$$CosineSim = \sum sim(V_{u,w_i}, V_u) = \cos(\theta) = \frac{V_{u,w_i} \cdot V_u}{|V_{u,w_i}||V_u|} \quad (4)$$

Subsequently, for the worker-worker agreement, one looks into the agreement factor between a worker and the crowd. You look how many times a worker agrees with another worker divided by the total amount of annotations of that worker. In equation 5, the worker-worker agreement is defined as the agreement factor between two workers divided by total annotations of worker i. To calculate the average worker-worker agreement for one worker. The pairwise comparisons between one worker and the crowd is aggregated and weighted by the shared units with every other worker in the crowd. The reason for the worker-worker agreement measure is that spammers generally disagree with no one because they show erratic behaviour. On the other hand disagreement does not immediately point to spam behaviour. It is possible that within a group there are multiple subgroups with its own opinion. An honest worker can agree with one specific sub-group.

$$wwa(w_i, w_j) = \frac{\sum_{u \in U_{w_i,w_j}} \sum V_{u,w_i} \cdot V_{u,w_j}}{\sum_{u \in U_{w_i,w_j}} \sum V_{u,w_i}} \quad (5)$$

The worker metrics are combined with other features to detect spammers. One such feature is named the **worker consistency score**. This score is measured for each worker as the number of times the worker did not highlight words that refer to new information even though the worker chose more novel or equal novel option, thus contradicting him or herself. A tweet can

not be novel if it only contains irrelevant or non-novel words. Another feature, **worker irrelevant behaviour** score is measured for each worker as the number of times the worker said that at least one tweet is not relevant, averaged by the total number of units the worker solved. A worker shows suspicious behaviour when he or she annotates more than 50% of the tweets as irrelevant. The next feature is called **Worker annotation frequency**, this feature indicates when a worker continuously chooses the same answer in the experiment. The remaining features concern the annotations of words, the total words annotated and the average.

The previously described features are used in different combinations to detect whether a worker is a spammer. The pseudo-code snippet in listing 1 contains all conditions. To test how these parameters perform on identifying spammers, a subset of annotations are gathered. A subset of 298 annotations are gathered and manually checked if they are from spammers. The set of parameters scored an F1 = 0.82 with an accuracy of 93%.

6 Setup of Experiment

Chapter 5 covers the task of the experiment. As for the conducted experiments on Crowdflower, only filtered data is used. The used tweets for the experiment occur in a span of 6 days. Using all the tweets of one day for one job is too much. So the tweets of a day are spread across multiple jobs. The jobs have a minimum of 24 units or a maximum of 50 units. The workers were presented with three units per page, with a pay of 3 cents per page. There are about 50 workers per unit and each unit has 15 judgements.

6.1 Overview of Results

A Crowdflower job consists around 20 tweets resulting into a range between 400 and 675 tasks for a Crowdflower experiment. Furthermore every job consists around of 60 workers and the workers take between 50 hours and 4 hours to complete a task. A Crowdflower experiment consisting of a maximum 675 tasks takes 50 hours and another experiment consisting of 495 tasks took 4 hours to finish. A total of 1331 units and 34 jobs are issued on Crowdflower. Appendix A contains an overview of the conducted experiments and the corresponding information.

6.2 Results and Discussion

All the data, functions and features used are accessible online [1] As mentioned before, two Crowdtruth measures are calculated of the workers, namely cosine similarity and worker-worker disagreement. Figure 4 shows a subset of workers, a subset of 298 workers. Based on both the cosine similarity and worker-worker disagreement measures, the overall quality of the workers on Crowdflower hovers around 0.75, and there are some bad quality workers scoring under 0,25. However the majority of the workers tend to agree with each other.

With these two measures, spammers in the crowd are removed from the data used for analysis. Based on the four conditions listed in the pseudo-code (listing 1), 29.41% workers in the crowd were identified as spammers. The subset of workers is also utilised to further test the spam filtering. For these workers, their annotations were manually checked to see if they were suspicious of giving bad annotations. Several different conditions and parameters of the spam filter were tested, but the chosen filter scored the best at an F1 measure of 0.82.

After removing the spammers, the novelty score is calculated for every tweet. The score was calculated as following, if a tweet was more novel it gains a +1. If the tweet was deemed equally

[1] https://github.com/CrowdTruth/Novelty_Detection

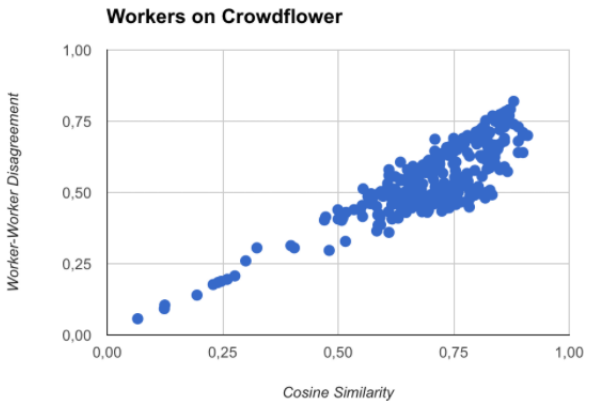

Figure 4. Quality overview on a subset of workers.

Table 13. Calculating normalised novelty score

Tweet ID	more_novel	equally_novel	less_novel	irrelevant
38977	23	96	8	3
Calculations				
Aggregated score = 23 + (96 * 0.5) - 8 = 63				
Normalised score = 63 / (23 + 96 + 8 + 3) = 0.48				

novel it gains halve a score, if the tweet is less novel it loses one point. Table 13 contains an example of one calculation of a tweet and the normalisation step. There is no correlation between the normalised score and the number of times a tweet is used (r = -0.02, p = 0.83). To show an example of the content of tweets, table 14 contains the top 5 scoring tweets and table 15 has the 5 lowest scoring tweets. As the tweets are very clearly showing, crowdworkers value important information like statistical data, whaling activity, time stamp and important actors. On the other hand, the lowest scoring tweets show unimportant data. Although they are related with whaling, they either show parts of the tweet in the non-English language or just show inflammatory comments. The tweets that scored around 0 show a nice transition from novel information, to duplicate or unimportant information.

6.3 Correlation between novelty and features

As the novelty state of tweets are known, we can use them as a dependent variable. A bivariate correlation analysis was conducted to see if there were any significant correlations between the gathered or constructed features. From the utilised features discussed in chapter 4.4, only the significant or almost significant are shown in

Table 14. Five tweets with the highest normalised novelty core

Tweet Content	Novelty Score
@AbelValdivia: Commercial hunting wiped out almost 3 million! whales last century. #whales #whaling	0.57
@IrinaGreenVoice: Humans slaughtered nearly 3 MILLION whales in the 20th century.... according to a new study	0.56
@SeaShepherd_USA: Nearly 3 million whales were killed by commercial whaling in the last century.	0.55
#Worlds #whaling #slaughter tallied. #Hunting wiped out ~3 million last century |@NatureNews	0.55
@spalumbi: New total. 3 million whales have been killed by whaling.	0.53

```
#Condition tree for spam detection
def spam_detection():
  threshold_cos = half_of_mean_std_cosine_agreement
    threshold_wa = half_of_mean_std_worker_agreement
    i = 0
    while < len(workers):
      if worker_cosine[worker] < threshold_cos and worker_agreement[worker] < threshold_wa:
          #spammer
      elif worker_cosine[worker] < threshold_cos or worker_agreement[worker] < threshold_wa:
          if worker_consistency[worker] == True or worker_irrelevant_behaviour[worker] > 0.5
    \
          or worker_annotation_frequency[worker] == True or avg_novel_words[worker] < 2:
              #spammer
          else:
              #non-spammer
      elif worker_irrelevant_behaviour[worker] > 0.5 and worker_annotation_frequency[worker]
    == True:
          #spammer
      elif worker_annotation_frequency[worker] == True and avg_novel_words[worker] < 1.2 and
    \
      total_annotation[worker] > 7:
          #spammer
      else:
          #non-spammer
      i +=
```

Listing 1. Pseudo-code snippet for spam detection, features described in chapter ??

Table 15. Five tweets with the lowest normalised novelty score

Tweet Content	Novelty Score
@jumpingGrendel the whaling is the hardest part #mobydick.	-0.10
@SP00KY: stop illegal whaling	-0.14
Quel whaling #sum #cutty.	-0.15
WHY am I just finding out about whaling omg	-0.38
Bestu vinir rokinu #whaling #whale-watching by sigurjonthr #socialreykjavik.	-0.50

figure 6. First, the most obvious notion, the time that the tweet was posted has an important role in determining novelty. If two tweets contain the same information, the tweet that was earlier is the more novel tweet, this effect is visible as a significant correlation (r = -0.170 α= 0.048). The correlation is negative because an older tweet is probably more novel.

The 'relevance span score' [27], was gathered from crowdsourcing tasks. The tasks asked workers to score how strongly a tweet was related to the whaling event. The score is calculated with the cosine similarity measure [6], between the unit case and aggregated vector. It shows the likelihood that a give snippet of text in the tweet is relevant for the whaling event. The correlation analysis indicates the 'relevance span score' as a strong significant factor in determining novelty (r = 0.297 α<0.00). One possible explanation for this phenomenon is that the workers not only scored relevancy but also the quality of the tweet snippet. So tweet content is highly important for novelty detection, a tweet that only contains rubbish information or incorrect usage of language would score lower. The next factor named distance in figure 6, indicates the similarity between tweets. Similarity is a popular method to determine if a new document contains new words. Of course if two tweets are alike, the newer tweet has a lower chance to bring novelty to the event space. This assumption is expressed by an almost significant correlation of the feature (r = 0.168 α= 0.051).

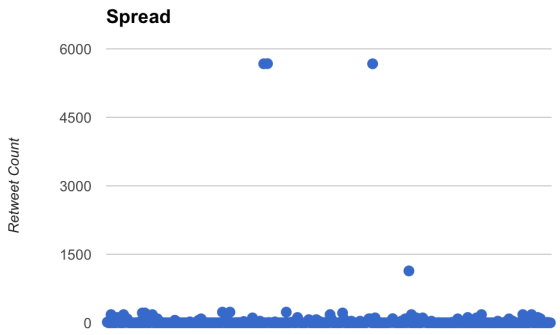

Figure 5. Spread of the retweet count of tweets

The next two features are 'user.url' and 'user.description_length' both are futures telling something about the author, specifically the credibility or expertness of the author. The 'user.url' feature tells if the tweet author has his or her own website shown on the profile page, the latter feature states the length of the user biography the author places on Twitter. Both of these features show significant (r = 0.195 α= 0.023) or near significant correlation with novelty (r = 0.153 α= 0.074). The final feature named retweet count is proven to be a somewhat ambiguous result. Although the correlation is significant, the relation is negative of a kind. One would think that more retweets, thus a higher popularity should point to an important and novel tweet. Alas, the results show a small counter result (r = -0.175 α<0.042). This could be caused by the lack of diversity and quantity in data. Figure 5 shows that most of the retweet counts hovers between 0 and 100, with some extreme exceptions. Although the outliers skewer the results of retweets negatively, the tweets themselves provide valuable information in other factors (user expertness and tweet quality). Gathering more instances with

Table 16. Overview of all testing results of the model

Feature excluded	Accuracy	Recall
Without weights		
All features Included	47.37%	11.11%
Relevance Span Score	59.21%	100.00%
Similarity Distance	46.05%	13.33%
Tweet Date	59.21%	100.00%
User.description_length	59.21%	100.00%
Duplicative tweets test-set		
All features Included	71.43%	58.82%
Relevance Span Score	33.77%	0.00%
Similarity Distance	78.57%	71.57%
Tweet Date	33.77%	0.00%
User.description_length	33.77%	0.00%
Duplicative tweets test-set, sampled		
All features Included	71.54%	60.24%
Relevance Span Score	32.52%	0.00%
Similarity Distance	78.86%	72.29%
Tweet Date	32.52%	0.00%
User.description_length	32.52%	0.00%
Unique Tweets test-set		
All features Included	67.11%	73.33%
Relevance Span Score	60.53%	48.89%
Similarity Distance	59.21%	100.00%
Tweet Date	46.05%	8.89%
User.description_length	64.47%	68.89%
Unique Tweets test-set, sampled		
All features Included	66.67%	74.29%
Relevance Span Score	60.00%	47.06%
Similarity Distance	56.67%	100.00%
Tweet Date	48.33%	11.43%
User.description_length	61.67%	69.44%

high retweet counts is necessary to come to more conclusive results.

6.4 Modelling

Now with the knowledge of the importance of the factors, the features themselves can be utilised to predict novelty. The support vector machines (SVM) algorithm is used. SVM is proven to be effective in text analysis tasks [32, 34, 45]. The variables used are a mix of continuous values and categorical values. The training data consists of data annotated and gathered from the Crowdflower experiments, also tweets that were deemed duplicate and unfit for the crowdsourcing experiments were also added. A cost matrix was also utilised, predicting a false negative is punished three times as more than predicting a false positive. Training without weights, produced non-sensible results (table 16). This seems to reflect the data more, because in reality there are far more not novel tweets in circulation.

A total of 355 tweets were used for training. It would be preferable to have more data in hand to divide the data in bigger parts for testing. According to Beleites et al. [8], you need a minimum of 5 times more data as features for reliable results. The data at hand roughly conforms to these rules. However, additional training data would be more reliable, as evident in the ascending learning curve (figure 10). A test set of 82 tweets were gathered by manually checking for novelty. Testing with a tweet set with duplicative tweets, thus with more noise, resulted into pronounced results. Duplicative tweets have a lot of tweets with low distance similarity, thus this feature was overwhelming the model, as visible in figure 7. The model should not be used with duplicative tweets. Low similarity tweets should be omitted in application of the model, as explained in 6.4.1. Testing the same set with sampling, resulted in near-exact same results (table 16).

Another test was conducted because the 'distance similarity' feature was proven to be too powerful in predicting novelty. Two tweets that are very alike can be easily detected as duplicate tweets by simply checking if they contain the same words. The Levenshteins distance used for this feature, has proven itself very effectively with short length texts [31]. Duplicate and thus not-novel tweets were easily identified by the model. With more novel tweets to work with, the model could predict more true positive results. The accuracy resulted from this test is 67.11% and the recall rate is 73.33%. Various other combinations of omitted variables were used (9), to make sure the model does not hinges too much on one feature. The figure contains results with the complete test dataset. Testing with sampling did not incur any meaningful changes (figure 8). Omitting similarity distance from the model, caused the model to fail at predicting true negatives. The model always predicted true, resulting into a 100% recall rate. It should be noted that the model could be improved with more training data. This is visible in the learning curve in figure 10, because the curve has not reached a plateau yet.

6.4.1 Scalability

To use this approach on the web and at a greater scale, some additional explanation is needed. The first step of detecting novel and relevant tweets, is to use the seed words. So the current approach continuously fetches tweets about whaling, with the aid of seed words. This step creates a huge set of tweets relevant to the event. To further decrease the amount of tweets, filtering is used for spam, exact duplicative tweets or near-exact duplicates. Similarity distance as explained in chapter 4.3 is the utilised method for this. Finally, the SVM model acquired by machine learning is used to check if a tweet contain the necessary characteristics of novel tweets. The collected features of tweets are then calculated with the equation described in chapter 2.2.4. This approach is usable on a greater scale, because all steps can be automated and as a bonus are computationally light. On a personal computer, checking the presence of seed words over thousand tweets was done instantaneously. Similarity distance calculation took significantly longer, an hour or more. However, processing time can be accelerated by moving to more powerful servers. Finally, the computationally expensive task of machine learning can be avoided, by using the completed SVM model. Detecting novel tweets can be done by implementing the model, with the collected tweet features and its weights in the model.

6.5 Tweet Comparison Examples

For clarification purposes, 3 pairs of different tweets are compared. Each pair is either novel, not-novel or ambiguous on novelty status. With these tweets the appurtenant features are included in the tables 17, 18 and 19. Firstly, the tweets are selected based on their novelty level and ambiguity. The novel examples are clearly bringing in new content. The not-novel tweets are clearly duplicative in kind or irrelevant, thus bringing very little new information to the whaling event. The ambiguous tweets are harder to distinguish and are low in novelty score accordingly.

The novelty score is calculated by aggregating the novelty annotations and normalising the score (table 13). An example of an novelty annotation vector is shown in table 9. Also, for the 6 example tweets discussed earlier, the total annotation vector for novelty is shown in table 20. The first number in the time-stamp indicates a date in March. The more important features are explained in section 6.3 about features and their correlation with novelty score. Some remaining variables need to be explained further. All features preceded by the 'user' notation indicates user properties. Listed count

Correlations

		TweetDate	TweetEventScore	Distance	retweet_count	user.url	user.description_length	novelty
TweetDate	Pearson Correlation	1	-,062	-,514**	,016	-,017	-,023	-,170*
	Sig. (2-tailed)		,474	,000	,852	,846	,789	,048
	N	136	136	136	136	136	136	136
TweetEventScore	Pearson Correlation	-,062	1	,092	,058	-,006	,141	,297**
	Sig. (2-tailed)	,474		,285	,501	,945	,103	,000
	N	136	136	136	136	136	136	136
Distance	Pearson Correlation	-,514**	,092	1	,010	,106	,085	,168
	Sig. (2-tailed)	,000	,285		,906	,218	,323	,051
	N	136	136	136	136	136	136	136
retweet_count	Pearson Correlation	,016	,058	,010	1	-,057	-,055	-,175*
	Sig. (2-tailed)	,852	,501	,906		,512	,526	,042
	N	136	136	136	136	136	136	136
user.url	Pearson Correlation	-,017	-,006	,106	-,057	1	,182*	,153
	Sig. (2-tailed)	,846	,945	,218	,512		,034	,074
	N	136	136	136	136	136	136	136
user.description_length	Pearson Correlation	-,023	,141	,085	-,055	,182*	1	,195*
	Sig. (2-tailed)	,789	,103	,323	,526	,034		,023
	N	136	136	136	136	136	136	136
novelty	Pearson Correlation	-,170*	,297**	,168	-,175*	,153	,195*	1
	Sig. (2-tailed)	,048	,000	,051	,042	,074	,023	
	N	136	136	136	136	136	136	136

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Figure 6. Significant or near significant correlations between features and novelty score.

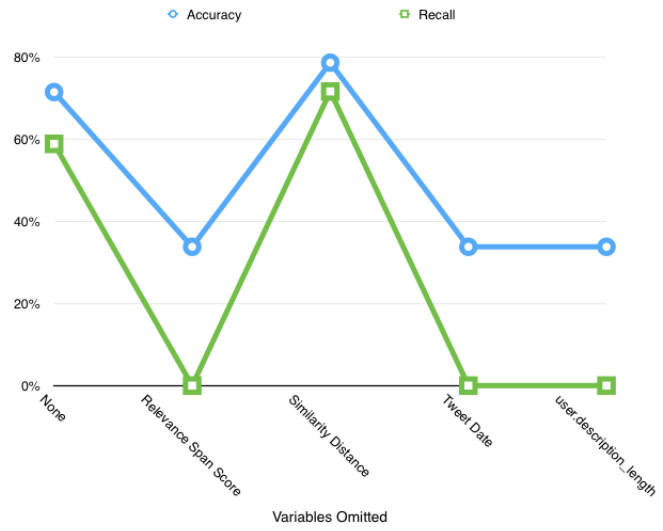

Figure 7. Accuracy and recall rate with various features omitted, with a duplicative tweets testing set

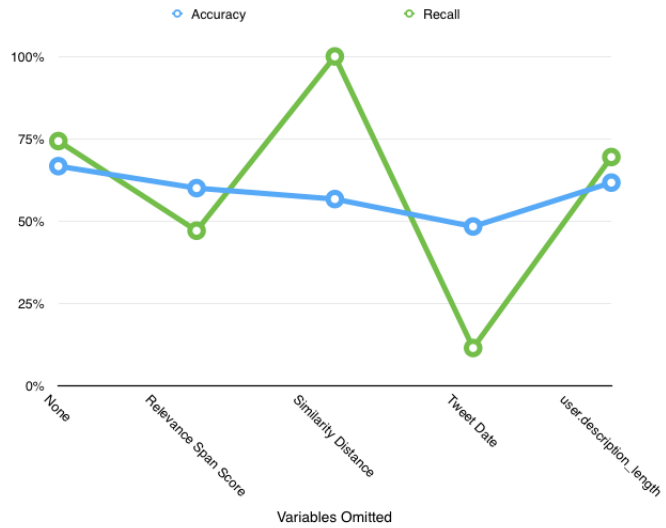

Figure 8. Accuracy and recall rate with various features omitted, with sampling enabled

states how many times an author is present in a twitter list of authors created by users. Description length indicates the length of the user biography. Entity count and sentiment score are explained in table 5.

The novelty of some tweets are easy to define based on their content. Novel tweets contain specific information about events or new data. For example, the number of hunted whales at 3 million is novel or adding new actors as Norway can help. Not-novel tweets do not contain new information relevant to the whaling event. Ambiguous tweets are harder to distinguish, although they contain relevant information about whaling. They are nonetheless scored lowly by annotators. It can be speculated that whaling museums are relevant, but are not really a news event, it is more commercial in nature. The second tweet in table 19 contains only the new actor Iceland, but the rest of the tweet is duplicative. If you analyse the tweet according to the model, features that give information about the user can be helpful. A more serious author has probably more important information to deliver. The model also uses distance similarity and relevance span score. The former indicates the amount of new words of the tweet, the latter how much the tweet is related to the whaling event.

7 Conclusion

With the rising abundance of social content and duplicating of old news, the need for curation is becoming stronger. For example, the Twitter network uses a team of credible sources and human curators to present important tweets and news articles to their users. With these curated tweets also called 'Twitter Moments', duplicate

Table 17. Comparison of novel tweets

Tweet ID	tweet/38739	tweet/38705
Tweet Content	@FinsandFluke: Norway #whale meat dumped in Japan after #pesticide finding #whaling #whales #IWC	@AbelValdivia: Commercial hunting wiped out almost 3 million! whales last century. #whales #whaling
Tweet Date	12 10:34:05	11 21:05:22
Relevance Span Score	1	1
Distance	22	23
user.followers_count	537	2332
user.favourites_count	13736	54
retweet_count	20	6
favorite_count	0	0
user.friends_count	9	2000
user.listed_count	0	48
user.url	1	1
user.description_length	25	157
user.created_at	02/19/15	02/08/10
user.profile_image_url	1	1
user.profile_banner_url	0	1
entities_count	4	6
sentiment_score	0	0
Novelty score	0.55	0.57

Table 18. Comparison of not-novel tweets

Tweet ID	tweet/38786	tweet/38626
Tweet Content	@Diversion50: PRANK: Loosen the legs on a housemates chair. When he goes to sit down disembowel him with a whaling harpoon at point b	WHY am I just finding out about whaling omg
Tweet Date	13 18:01:06	11 3:12:34
Relevance Span Score	0.20	0.70
Distance	18	23
user.followers_count	723	20
user.favourites_count	425	0
retweet_count	0	0
favorite_count	0	0
user.friends_count	100	439
user.listed_count	1	1
user.url	0	0
user.description_length	0	10
user.created_at	08/12/13	09/10/11
user.profile_image_url	1	1
user.profile_banner_url	1	1
entities_count	13	5
sentiment_score	0	0
Novelty score	-0.65	-0.38

Table 19. Comparison of ambiguous novel tweets

Tweet ID	tweet/38665	tweet/38932
Tweet Content	The New Bedford Whaling Museum & Buzzards Bay Coalition present a 3-part lecture series on Buzzards Bay's health.	National Anthems of Great Current Whaling Nations - Iceland - #Tweet4Taiji #thecove #seashepherd
Tweet Date	18 12:31:14	09 14:39:08
Relevance Span Score	0.95	0.83
Distance	11	24
user.followers_count	1992	25
user.favourites_count	190	2
retweet_count	0	0
favorite_count	0	0
user.friends_count	245	1
user.listed_count	92	3
user.url	1	0
user.description_length	160	60
user.created_at	03/19/10	02/05/12
user.profile_image_url	1	1
user.profile_banner_url	1	0
entities_count	13	10
sentiment_score	0	0
Novelty score	0.00	0.02

Table 20. Total novelty annotations by the crowd, irrelevant annotation excluded

tweetID	more novel	equally novel	less novel
Novel tweets			
tweet/38705	154	128	29
tweet/38739	143	90	47
Not-novel tweets			
tweet/38626	26	43	72
tweet/38786	16	18	86
Ambiguous tweets			
tweet/38665	7	8	11
tweet/38690	5	8	10

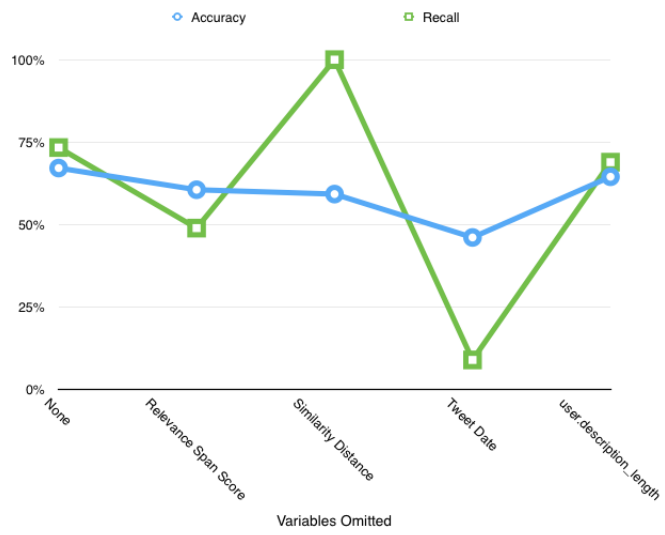

Figure 9. Accuracy and recall rate with various features omitted

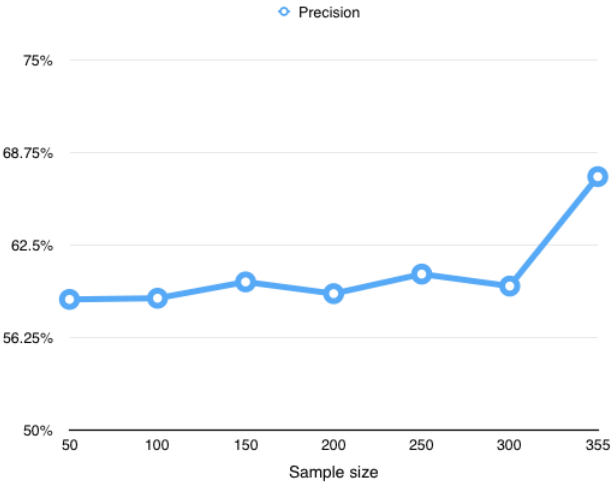

Figure 10. Precision scores and their different sample sizes

tweets and other noisy information are removed, thereby presenting users a better experience. However, this method requires a large amount of resources both in human power and credible resources. If you do not have the full cooperation of important news agencies, one can not reproduce the same results. It would be better to automatically produce these results with the aid of the crowd and machine learning algorithms.

The aim of the current research was to explore the possibility of measuring novelty in tweets. And if it is measurable, what factors indicate the novelty of tweets, or the level of influence of different features. What is an accurate and scalable approach for detecting novelty. To answer these questions, data was collected from the Twitter network. Data irrelevant to the whaling event was removed. Possible influential factors for novelty were added to the data. The crowd was also used for annotation of the data. The annotated tweets were used as training data for machine learning. The intelligence of the crowd paired with the strength of numbers of computers, models can be created for several tasks. In this case, gathering training data with the crowd to observe certain patterns in novelty detection, proved to be helpful in determining which features were important. Features from the Twitter API and other features in the author, tweet content, tweet interaction and sentiment dimension were collected for analysis. Besides these features, tweets were gathered in middle of 2015. Those tweets were later utilised for the crowdsourcing tasks on the Crowdflower platform. To narrow the scope, only six days of tweets were gathered regarding the whaling event. The tweets were deemed relevant with the whaling event, if they contained enough words in the seed words collection. The seed words were given from the social science department for activist events, specifically whaling activist events.

After the process of collecting data, filtering data and annotating data. Spam detection was used to improve the quality of the annotated data from the crowd. Two measures from the CrowdTruth platform were utilised, namely the 'worker - cosine disagreement score and the worker - worker disagreement score[6]. These two measures were used to detect spammers and low quality workers. Several other parameters were used, such as the average amount of highlighted words in tweet, answering consistency and annotation frequency. These combinations of parameters were very effective in filtering out low quality annotation data, with F1 = 0.82 and an accuracy of 93%.

Results from the correlation analysis from figure 6 shows significant correlations between certain factors and novelty. The first factor points to date and time as a significant correlation. The second factor uses the relevance span score from another crowd sourcing experiment [27]. Workers in the experiment needed to score parts of the tweet on event relevancy. A possible explanation for the relation with novelty, is that workers in the earlier experiment unexpectedly also annotated quality of the tweets. The next factor, tweet similarity, helps in detecting tweets that are highly similar and thus not novel. Furthermore, user.url and user.description length factors indicate the expertness and seriousness of the user. If the user takes the time to write a long and explicative biography, the chance is lower that the same user produces spam and noisy tweets. The last feature to be mentioned is the retweet count, this result is somewhat ambiguous. Although it is significant, it is a negative correlation, this could be caused by outlier data and the lack of diverseness of retweet count data. The large amount of tweets produced have low or 0 retweet count and small portion of tweets have an extreme high amount of retweets (>5000).

For modelling and machine learning, the support vector machine (SVM) model was used. Present literature shows that the SVM model is popular and highly effective in text analysis tasks[32, 34, 45]. Testing was done with a random set of tweets containing a lot of duplicative or near-duplicative tweets. The test shows that the model with the aid of similarity distance, is highly helpful in indicating spam data (accuracy = 75.61%), but very ineffective at recalling true positives (recall rate = 38.47%). Another test was conducted without these duplicative tweets and a cost matrix was added at the training stage. The cost matrix had a false negative and false positive ratio of 3:1. The recall rate improved to 73.33%. However, there is still enough room for improvement, the learning curve in figure 10 indicates an upwards slope in effectiveness with more data. Features concerning the relatedness with whaling is only used for data collection. The finished model only uses Twitter statistics as variables. This means that if the model is used over a longer period of time, no retraining is needed. The finished model already knows what Twitter statistics are important. However, if the situation arises wherein the existing seed words have lost its relevance completely. New seed words are probably needed, but features like similarity and author properties hold their importance.

8 Future Work

One of the limitations of the present research project is the size of the data set. It would be preferred if more training and testing data were gathered for modelling ends. However, conducting crowdsource tasks on major commercial crowdsource websites is expensive, especially with a pairwise comparison task. The pairwise comparison means that every new sentence added, the task must be repeated for every comparison with the available sentences. To collect more data in the future, more funds are needed or a more compact task design, that can produce more results per monetary unit.

Another untested issue, is to check how well the model performs on a longer time-scale. For now, the model was created and tested with limited crowd-sourced data and manually checked data. One method to grade the performance is to count novel tweets detected. The existing set of seed words may not accurately represent the new tweets about whaling. This ineffectiveness of old seed words, can be negated by checking for new entities. An important set of seed words 5 months earlier could be supplemented or exchanged for another set of words in the future. For example, a new event related with whaling, could spawn new entities from that event. Subsequently, if an entity occurs in a great amount of tweets, that entity is probably interesting as a new seed word.

Another interesting theme of exploration is to apply the same approach with another theme. As explained in chapter 1.1, whaling event had excellent characteristics for this research experiment, because of its recurrent important news events. Also multiple actors were involved with the topic, NGO's, governments, activists and multiple news agencies. When searching for new domains for the current approach, one or more of these qualities should be present.

For example, if we take another research domain of the whaling experts [14, 15], namely activism in the anti-sweatshop movement [18, 29]. We can see both domains possessing the same characteristics. Multiple actors are involved companies, governments and NGO's. The domain is also quite recurrent in the news, with the most recent popular topic about mistreating of Asian labour of technology manufacturers [10]. If the current approach is applied to this domain, we can ask experts for seed words. Or seed words can be mined from the Wikipedia page about sweatshops. Except for the 'relevant span scores', the remaining steps of the current approach can be repeated. It is to be wondered if the results are better or worse. Taken into account that the anti-sweatshop movement could be less turbulent, without numerous protest-active animal protection organisations.

9 References

[1] Internet in Real-Time: How Quickly Data is Generated: 2015.

[2] ALLAN, J., LAVRENKO, V., MALIN, D., AND SWAN, R. Detections, Bounds, and Timelines: UMass and TDT-3. *Information Retrieval* (2000), 167174.

[3] ALLAN, J., WADE, C., AND BOLIVAR, A. Retrieval and novelty detection at the sentence level. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03* (2003), 314–321.

[4] ALONSO, O., AND MIZZARO, S. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management 48*, 6 (2012), 1053–1066.

[5] ANTONIOU, G., AND VAN HARMELEN, F. *A Semantic Web Primer*, vol. 24. 2009.

[6] AROYO, L., AND WELTY, C. The Three Sides of CrowdTruth. *Human Computation 1*, 1 (2014), 31–44.

[7] BASU, S., RAYMOND, J. M., KRUPAKAR, V. P., AND JOYDEEP, G. Using Lexical Knowledge to Evaluate the Novelty of Rules Mined from Text. *Proceedings of the NAACL Workshop on WordNet and Other Lexcical Resources: Applications, Extensions and Customizations* (2001).

[8] BELEITES, C., NEUGEBAUER, U., BOCKLITZ, T., KRAFFT, C., AND POPP, J. Sample size planning for classification models. *Analytica Chimica Acta 760*, June 2012 (2013), 25–33.

[9] BOZZON, A., BRAMBILLA, M., CERI, S., MILANO, P., AND PONZIO, V. Answering Search Queries with CrowdSearcher. *Language* (2012), 1009–1018.

[10] CHAVARRIA, V., LOKER, R., AND TORRES, B. Sweatshops: How Can We Change An Industry That is So Inherently Intertwined With Todays Culture? *Hope College, Holland, Michigan* (2015).

[11] CHEN, J., NAIRN, R., NELSON, L., BERNSTEIN, M., AND CHI, E. Short and tweet: experiments on recommending content from information streams. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), 1185–1194.

[12] CHU, W., AND PARK, S.-T. Personalized recommendation on dynamic content using predictive bilinear models. *Proceedings of the 18th international conference on World wide web - WWW '09* (2009), 691.

[13] CRESPI, I. *Public Opinion, Polls, and Democracy*. 1989.

[14] DE BAKKER, F. G. A., AND HELLSTEN, I. Capturing Online Presence: Hyperlinks and Semantic Networks in Activist Group Websites on Corporate Social Responsibility. *Journal of Business Ethics 118*, 4 (2013), 807–823.

[15] DE BAKKER, F. G. A., HELLSTEN, I., AND KOK, A. M. Examining activism: Tracing networks and tactics on CSR. *Notizie di POLITEIA XXVII*, 103 (2011), 66–77.

[16] DEL CORSO, G. M., GULLÍ, A., AND ROMANI, F. Ranking a stream of news. *Proceedings of the 14th international conference on World Wide Web - WWW '05* (2005), 97–106.

[17] DEMARTINI, G. Hybrid human-machine information systems: Challenges and opportunities. *Computer Networks* (2015), –.

[18] DEN HOND, F., AND DE BARKER, F. G. A. Ideologically motivated activism: How activist groups influence corporate social change activities, 2007.

[19] DESCY, D. E. The Wiki: True Web Democracy. *TechTrends 50.1* (2006), 4–5.

[20] DIAZ-AVILES, E., SIEHNDEL, P., AND NAINI DJAFAARI, K. Exploiting Social # -Tagging Behavior in Twitter for Information Filtering and Recommendation. *Text REtrieval Conference (TREC)* (2011), 2–5.

[21] FACKLER, M. With Whaling Ships Under Attack, Japan Will Recall Fleet, 2011.

[22] FERNÁNDEZ, R. T., AND LOSADA, D. E. Novelty detection using local context analysis. *roceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.* (2007).

[23] FERRUCCI, D., LEVAS, A., BAGCHI, S., GONDEK, D., AND MUELLER, E. T. Watson: Beyond jeopardy! *Artificial Intelligence 199-200* (2013), 93–105.

[24] FERRUCCI, D. A. Introduction to "This is Watson". *IBM Journal of Research and Development 56*, 3.4 (2012), 1–1.

[25] GABRILOVICH, E., DUMAIS, S., AND HORVITZ, E. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. *Proceedings of the 13th conference on World Wide Web - WWW '04* (2004), 482–490.

[26] HIBMA, M. The Life of a Tweet: A Look at the First 24 Hours, 2015.

[27] INEL, O., AND TOMASSO, C. Salience-In-News-And-Tweets @ github.com, 2015.

[28] JAVA, A., SONG, X., FININ, T., AND TSENG, B. Why We Twitter : Understanding Microblogging. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (2007), 56–65.

[29] KNIGHT, G., AND GREENBERG, J. Promotionalism and Subpolitics: Nike and Its Labor Critics. *Management Communication Quarterly 15*, 4 (2002), 541–570.

[30] KUMARAN, G., AND ALLAN, J. Text classification and named entities for new event detection. *Proceedings of the 27th annual international ACM ...* (2004), 297–304.

[31] LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals, 1966.

[32] MA, J., AND PERKINS, S. Time-series novelty detection using one-class support vector machines. *Proceedings of the International Joint Conference on Neural Networks, 2003. 3* (2003), 1741–1745.

[33] MANNING, C. D., BAUER, J., FINKEL, J., BETHARD, S. J., SURDEANU, M., AND MCCLOSKY, D. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2014), pp. 55–60.

[34] MARKOU, M., AND SINGH, S. Novelty detection: a review-part 2:: neural network based approaches. *Signal processing*, 1991 (2003), 1–26.

[35] MOFFA, A. L. Two Competing Models of Activism, One Goal: A Case Study of Anti-Whaling Campaigns in the South Ocean. *Yale J. Int'l L. 37* (2012), 201.

[36] NURSE, A. The beginning of the end? The International Court of Justice's decision on Japanese Antarctic whaling. *Journal of Animal Welfare Law* (2014), 14–17.

[37] OOSTERMAN, J., BOZZON, A., HOUBEN, G.-J., NOTTAMKANDATH, A., DIJKSHOORN, C., AROYO, L., AND TRAUB, M. C. Crowd vs . Experts : Nichesourcing for Knowledge Intensive Tasks in Cultural Heritage. *WWW14 Companion* (2014), 567–568.

[38] OSBORNE, M., AND LAVRENKO, V. Streaming First Story Detection with application to Twitter. *Computational Linguistics*, June (2010), 181–189.

[39] PEACE, A. The whaling war: Conflicting cultural perspectives. *Anthropology Today 26*, 3 (2010), 5–9.

[40] PLOEGER, T., KRUIJT, M., AROYO, L., DE BAKKER, F., HELLSTEN, I., FOKKENS, A., HOEKSEMA, J., AND TER BRAAKE, S. Extracting activist events from news articles using existing NLP tools and services. *Detection, Representation, and Exploitation of Events in the Semantic Web 30* (2013).

[41] ROTHMAN, S. B. Unveiling the Whale: Discourses on Whales and Whaling (review). *Global Environmental Politics* (2011), 126–127.

[42] RUTHS, D., AND PFEFFER, J. Social media for large studies of behavior. *Science 346*, 6213 (2014), 1063–1064.

[43] SAHAMI, M., AND HEILMAN, T. D. A web-based kernel function for measuring the similarity of short text snippets. *Proceedings of the 15th international conference on World Wide Web WWW 06 pages* (2006), 377.

[44] SANKARANARAYANAN, J., SAMET, H., TEITLER, B. E., LIEBERMAN, M. D., AND SPERLING, J. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09* (2009), p. 42.

[45] SCHÖLKOPF, B., WILLIAMSON, R., SMOLA, A., SHAWE-TAYLOR, J., AND PLATT, J. Support Vector Method for Novelty Detection. *Advances in Neural Information Processing Systems 12* (1999), 582–588.

[46] STANKOVIC, M., ROWE, M., AND LAUBLET, P. Mapping tweets to conference talks: A goldmine for semantics. In *CEUR Workshop Proceedings* (2010), vol. 664.

[47] TWITTER INC. About Twitter, 2014.

[48] VERHEIJ, A., KLEIJN, A., FRASINCAR, F., AND HOOGENBOOM, F. A comparison study for novelty control mechanisms applied to web news stories. *IEEE Computer Society* (2012), 431–436.

[49] VOSSEN, P. H. Information Overload. 41–59.

[50] YANG, Y., ZHANG, J., CARBONELL, J., AND JIN, C. Topic-conditioned novelty detection. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02* (2002), 688–693.

[51] YIH, W.-T., AND MEEK, C. Improving similarity measures for short segments of text. *AAAI 7*, 7 (2007).

[52] ZHAI, C., AND LAFFERTY, J. Two-stage language models for information retrieval. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02* (2002), 49.

[53] ZHANG, Y., CALLAN, J., AND MINKA, T. Novelty and redundancy detection in adaptive filtering. *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2002), 81–88.

[54] ZHAO, L., ZHANG, M., AND MA, S. The nature of novelty detection. *Information Retrieval 9*, 5 (2006), 521–541.

Appendix A Overview of the conducted experiments on Crowdflower.

	Status						Input		Settings				Results							
									Job				Judgments		Cost			Time		
Day	Job	Finished	Date of Tweets	atfor	Output	Results	Batch size	Units Lang	Judg per Unit	Units per Page	Pay per Page	Unit order	Judg	%	Total	Unit	Effective Hourly Payment	Total Job Runtime	Avg sec per Unit
Day 1	762129	14/08/2015	09/03/2015	CF	f762129.csv	results762129	36	EN	15	3	$0.03	random	540	0%	$10.00	$0.28	$548,894.1	32:00:00	00:00:17
Day 1	762156	13/08/2015	09/03/2015	CF	f762156.csv	results762156	24	EN	15	3	$0.03	random	382	0%	$7.00	$0.29	$666,514.2	27:00:00	00:00:14
Day 1	762631	15/08/2015	09/03/2015	CF	f762631.csv	results762631	45	EN	15	3	$0.03	random	675	0%	$12.00	$0.27	$622,080.0	50:00:00	00:00:15
Day 1	762935	14/08/2015	09/03/2015	CF	f762935.csv	results762935	26	EN	15	3	$0.03	random	390	0%	$7.00	$0.27	$666,514.2	26:00:00	00:00:14
Day 1	763868	15/08/2015	09/03/2015	CF	f763868.csv	results763868	35	EN	15	3	$0.03	random	525	0%	$10.00	$0.29	$717,784.6	27:00:00	00:00:13
Day 1	764154	16/08/2015	09/03/2015	CF	f764154.csv	results764154	44	EN	15	3	$0.03	random	660	0%	$12.00	$0.27	$717,784.6	31:00:00	00:00:13
Day 2	762501	18/08/2015	10/03/2015	CF	f762501.csv	results762501	39	EN	15	3	$0.03	random	585	0%	$11.00	$0.28	$666,514.2	35:00:00	00:00:14
Day 2	762502	16/08/2015	10/03/2015	CF	f762502.csv	results762502	39	EN	15	3	$0.03	random	585	0%	$11.00	$0.28	$666,514.2	17:00:00	00:00:14
Day 3	764696	18/08/2015	11/03/2015	CF	f764696.csv	results764696	37	EN	15	3	$0.03	random	555	0%	$10.00	$0.27	$717,784.6	13:00:00	00:00:13
Day 3	764697	19/08/2015	11/03/2015	CF	f764697.csv	results764697	40	EN	15	3	$0.03	random	600	0%	$11.00	$0.28	$933,120.0	15:00:00	00:00:10
Day 3	764698	19/08/2015	11/03/2015	CF	f764698.csv	results764698	39	EN	15	3	$0.03	random	585	0%	$11.00	$0.28	$848,290.9	12:00:00	00:00:11
Day 3	764699	19/08/2015	11/03/2015	CF	f764699.csv	results764699	40	EN	15	3	$0.03	random	600	0%	$11.00	$0.28	$848,290.9	14:00:00	00:00:11
Day 3	764939	18/08/2015	11/03/2015	CF	f764939.csv	results764939	46	EN	15	3	$0.03	random	690	0%	$13.00	$0.28	$666,514.2	11:00:00	00:00:14
Day 3	767768	20/08/2015	11/03/2015	CF	f767768.csv	results767768	43	EN	15	3	$0.03	random	645	0%	$12.00	$0.28	$666,514.2	22:00:00	00:00:14
Day 3	767986	20/08/2015	11/03/2015	CF	f767986.csv	results767986	43	EN	15	3	$0.03	random	645	0%	$12.00	$0.28	$933,120.0	14:00:00	00:00:10
Day 3	768027	21/08/2015	11/03/2015	CF	f768027.csv	results768027	45	EN	15	3	$0.03	random	675	0%	$12.00	$0.27	$848,290.9	15:00:00	00:00:11
Day 4	768238	21/08/2015	12/03/2015	CF	f768238.csv	results768238	35	EN	15	3	$0.03	random	525	0%	$10.00	$0.29	$848,290.9	6:00:00	00:00:11
Day 4	768363	21/08/2015	12/03/2015	CF	f768363.csv	results768363	49	EN	15	3	$0.03	random	735	0%	$14.00	$0.29	$848,290.9	15:00:00	00:00:11
Day 4	768557	21/08/2015	12/03/2015	CF	f768557.csv	results768557	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$777,600.0	7:00:00	00:00:12
Day 4	768608	21/08/2015	12/03/2015	CF	f768608.csv	results768608	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$848,290.9	6:00:00	00:00:11
Day 4	769354	22/08/2015	12/03/2015	CF	f769354.csv	results769354	36	EN	15	3	$0.03	random	540	0%	$10.00	$0.28	$717,784.6	8:00:00	00:00:13
Day 4	769603	22/08/2015	12/03/2015	CF	f769603.csv	results769603	41	EN	15	3	$0.03	random	615	0%	$11.00	$0.27	$777,600.0	8:00:00	00:00:12
Day 4	769661	23/08/2015	12/03/2015	CF	f769661.csv	results769661	44	EN	15	3	$0.03	random	660	0%	$12.00	$0.27	$777,600.0	10:00:00	00:00:12
Day 4	769785	23/08/2015	12/03/2015	CF	f769785.csv	results769785	41	EN	15	3	$0.03	random	640	0%	$11.00	$0.27	$848,290.9	14:00:00	00:00:11
Day 4	769787	24/08/2015	12/03/2015	CF	f769787.csv	results769787	45	EN	15	3	$0.03	random	697	0%	$12.00	$0.27	$777,600.0	16:00:00	00:00:12
Day 4	769906	24/08/2015	12/03/2015	CF	f769906.csv	results769906	37	EN	15	3	$0.03	random	555	0%	$10.00	$0.27	$848,290.9	8:00:00	00:00:11
Day 4	769908	24/08/2015	12/03/2015	CF	f769908.csv	results769908	41	EN	15	3	$0.03	random	615	0%	$11.00	$0.27	$777,600.0	8:00:00	00:00:12
Day 5	770146	25/08/2015	13/03/2015	CF	f770146.csv	results770146	50	EN	15	3	$0.03	random	750	0%	$14.00	$0.28	$518,400.0	6:00:00	00:00:18
Day 5	770310	26/08/2015	13/03/2015	CF	f770310.csv	results770310	50	EN	15	3	$0.03	random	750	0%	$14.00	$0.28	$933,120.0	7:00:00	00:00:10
Day 5	770444	27/08/2015	13/03/2015	CF	f770444.csv	results770444	36	EN	15	3	$0.03	random	540	0%	$10.00	$0.28	$1,036,800	4:00:00	00:00:09
Day 5	770612	28/08/2015	13/03/2015	CF	f770612.csv	results770612	35	EN	15	3	$0.03	random	525	0%	$10.00	$0.29	$848,290.9	6:00:00	00:00:11
Day 5	770671	29/08/2015	13/03/2015	CF	f770671.csv	results770671	38	EN	15	3	$0.03	random	570	0%	$10.00	$0.26	$848,290.9	6:00:00	00:00:11
Day 6	770852	30/08/2015	14/03/2015	CF	f770852.csv	results770852	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$848,290.9	4:00:00	00:00:11
Day 6	770853	31/08/2015	14/03/2015	CF	f770853.csv	results770853	33	EN	15	3	$0.03	random	495	0%	$9.00	$0.27	$777,600.0	5:00:00	00:00:12