Introduction to Information Retrieval and Text Mining Lecture 02: Term Vocabularies and Postings Lists

Roman Klinger

Institute for Natural Language Processing, University of Stuttgart

2021-10-26

Overview

Outline

Inverted index

For each term t, we store a list of all documents that contain t.

BRUTUS	$\longrightarrow$	1	2	4	11	31	45	173	174
CAESAR	$\longrightarrow$	1	2	4	5	6	16	57	132	...
CALPURNIA	$\longrightarrow$	2	31	54	101				

$\vdots$

dictionary **postings**

Intersecting two postings lists

BRUTUS $\longrightarrow$ 1 $\rightarrow$ 2 $\rightarrow$ 4 $\rightarrow$ 11 $\rightarrow$ 31 $\rightarrow$ 45 $\rightarrow$ 173 $\rightarrow$ 174

CALPURNIA $\longrightarrow$ 2 $\rightarrow$ 31 $\rightarrow$ 54 $\rightarrow$ 101

Intersection $\Longrightarrow$ 2 $\rightarrow$ 31

Query processing: Exercise

FRANCE $\longrightarrow$ 1 → 2 → 3 → 4 → 5 → 7 → 8 → 9 → 11 → 12 → 13 → 14 → 15

PARIS $\longrightarrow$ 2 → 6 → 10 → 12 → 14

LEAR $\longrightarrow$ 12 → 15

Compute hit list for ((paris AND NOT france) OR lear)

6 . .

Constructing the inverted index: Sort postings

term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

$\Longrightarrow$

term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

ambitious → 2

be → 2

brutus → 1 → 2

capitol → 1

⋮

Westlaw: Example queries

Information need: Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company

Query: "trade secret" /s disclos! /s prevent /s employe!

Information need: Requirements for disabled people to be able to access a workplace

Query: disab! /p access! /s work-site work-place (employment /3 place)

Information need: Cases about a host's responsibility for drunk guests

Query: host! /p (responsib! liab!) /p (intoxicat! drunk!) /p guest

Does Google use the Boolean model?

- On Google, the default interpretation of a query [w_1 w_2 ... w_n] is w_1 AND w_2 AND ... AND w_n
- Cases where you get hits that do not contain one of the w_i:
  - anchor text
  - page contains variant of w_i (morphology, spelling correction, synonym)
  - long queries (n large)
  - boolean expression generates very few hits
- Simple Boolean vs. Ranking of result set
  - Simple Boolean retrieval returns matching documents in no particular order.
  - Google (and most well designed Boolean engines) rank the result set – they rank good hits (according to some estimator of relevance) higher than bad hits.

Take-away

- Understanding of the basic unit of classical information retrieval systems: words and documents: What is a document, what is a term?
- Tokenization: how to get from raw text to words (or tokens)
- More complex indexes: skip pointers

Outline

Remarks/Formalities

Recordings and Communication

- Teaching in this hybrid setup is new to me. It might be easy to miss something. Please give me feedback if something could be improved.
- If you have questions, please write them in the forum. Then, others can also benefit from answers. If you send mails, I might answer them, but probably I will only answer them in the next lecture/video. We prefer questions in the forum.

Attendance

- Not mandatory. Not in lecture, not in exercise discussions.

Remarks/Formalities

Exercise Q&A Sessions

- Organization details will be announced when the sheet is published.
- Solutions will be presented, questions can be discussed.
- No recording.
- Attendance not mandatory.
- Will only work with your active participation.
- We will also answer questions there that you asked in the forum.

Groups and On-Campus Attendance

- There were 9 people last Thursday.
- There are 14 people here today.
- I will wait for another week and then decide if and how to combine groups.
- **Please move yourself on CAMPUS from the groups "Ungerade/Gerade" to "Standardgruppe" if you do not participate on campus, that will free space for those who might want to come.**

Statistics ($\sum$=181)

Informatik (LHG)	53	29.1%
Computational Linguistic (LHG)	41	22.5%
Softwaretechnik (LHG)	24	13.2%
Informatik (LHG)	14	7.7%
Maschinelle Sprachverarbeitung (LHG)	13	7.1%
Data Science (LHG)	12	6.6%
Computer Science (LHG)	11	6.0%
Medieninformatik (LHG)	4	2.2%
Mathematik (LHG)	2	1.1%
Digital Humanities (LHG)	2	1.1%
Geschichte (LHG); Informatik	1	0.5%
Mathematik (LHG)	1	0.5%
Wirtschaftsinformatik (LHG)	1	0.5%
Simulation Technology (LHG)	1	0.5%
Softwaretechnik (LHG)	1	0.5%
Deutschkurs/TestDaF (VorStud)	1	0.5%

Ungerade 1: 50; Ungrade 2: 42;
Gerade 1: 35; Gerade 2: 41 (as of 10:22 today))

Questions?

Any other organizational questions?

Outline

Documents

- Last lecture: Simple Boolean retrieval system
- Our assumptions were:
  - We know what a document is.
  - We can "machine-read" each document.
- This can be complex in reality. Examples? Why?

spoken language

signed languages

handwritten texts

scanned/paper documents

lay doc.

filter-out non-documents

Parsing a document

- We need to deal with format and language of each document.
  - What (proprietary) format is it in? pdf, word, excel, html etc.
  - What language is it in?
  - What character set is in use?
- Classification problems: Studied later
- One (alternative) approach: use heuristics
- General approach:
  Converter for each format to a generic representation.

Format/Language: Complications

- A single index usually contains terms of several languages.
- Document may contain multiple languages/formats.
  - French email with Spanish PDF attachment
  - Abstracts in different languages in one PDF
- What is the document unit for indexing?
  - A file?
  - An email?
  - An email with 5 attachments?
  - A group of files (ppt or LaTeX in HTML)?
  - What about XML?
  - Part of a file?

  ⇒ Answering the question "what is a document?" is not trivial and requires some design decisions.

Issues with PDF – Example

OCR/scan

RESPONSE GENERATION IN QUESTION - ANSWERING SYSTEMS

Ralph Grishman
New York University

hyphen

columns separate

equations

1. INTRODUCTION

As part of our long-term research into techniques for information retrieval from natural language data bases, we have developed over the past few years a natural language interface for data base retrieval [1,2]. In developing this system, we have sought general, conceptually simple, linguistically-based solutions to problems of semantic representation and interpretation. One component of the system, which we have recently redesigned and are now implementing in its revised form, involves the generation of responses. This paper will briefly describe our approach, and how this approach simplifies some of the problems of response generation.

Our system processes a query in four stages: syntactic analysis, semantic analysis, simplification, and retrieval (see Figure 1). The syntactic analysis, which is performed by the Linguistic String Parser, constructs a parse tree and then applies a series of transformations which decompose the sentence into a operator-operand-adjunct tree. The semantic analysis first translates this tree into a formula of the predicate calculus with set-formers and quantification over sets. This is followed by anaphora resolution (replacement of pronouns with their antecedents) and predicate expansion

the predicate satisfied by the set, add a universal quantifier over the extension of the set, and convert the resulting formula into an English sentence. For our example, this would mean

print-English-equivalent-of'$(\forall x \in S_1)$

passed (x, French exam)'

where $S_1 = \{s \in \text{set-of-students} \mid \text{passed}(s,\text{French exam})\}$

and

print-English-equivalent-of'$(\forall x \in S_2)$

failed (x, French exam)'

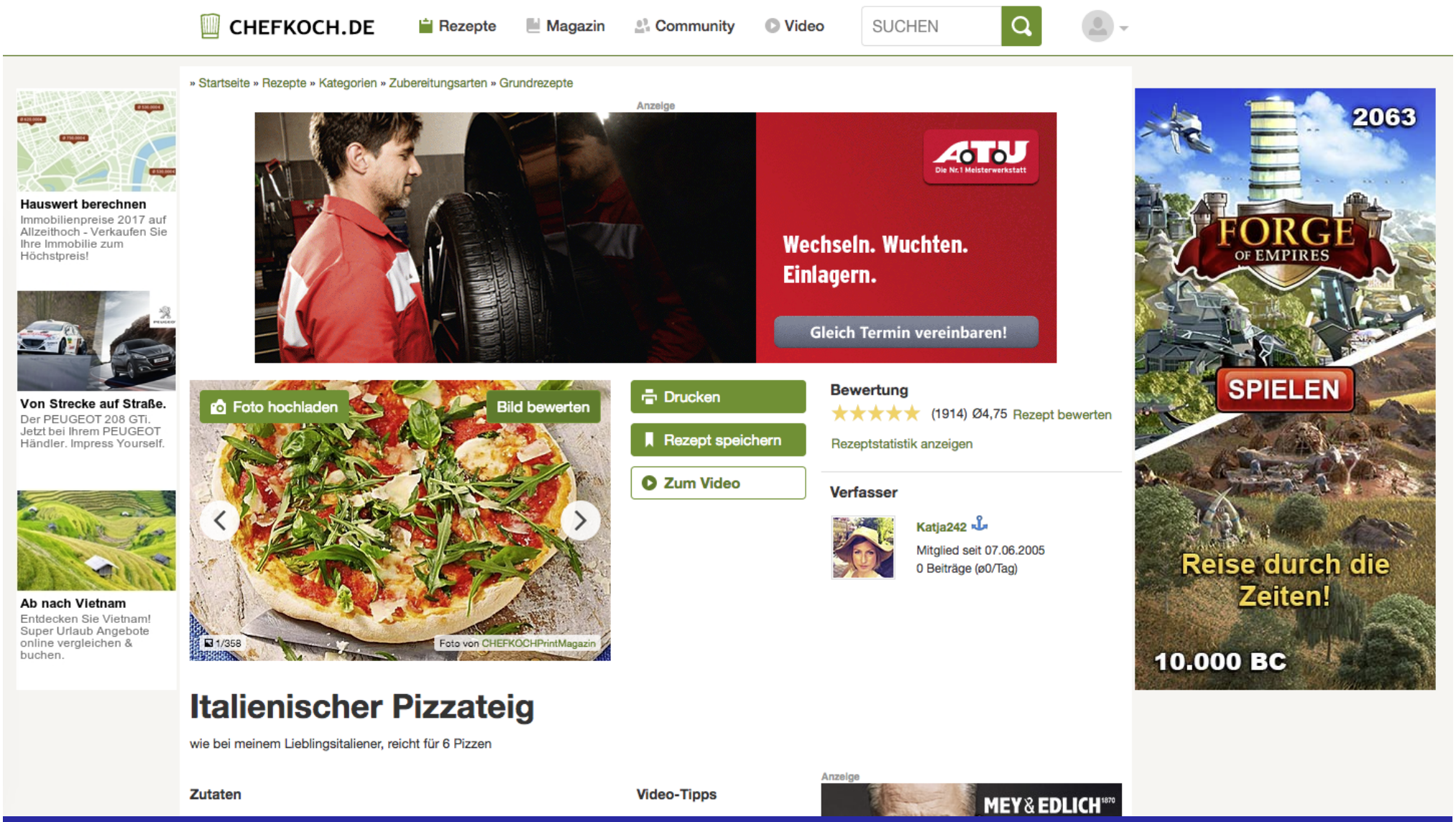
where $S_2 = \{s \in \text{set-of-students} \mid \text{failed}(s,\text{French exam})\}$

which would generate a response such as

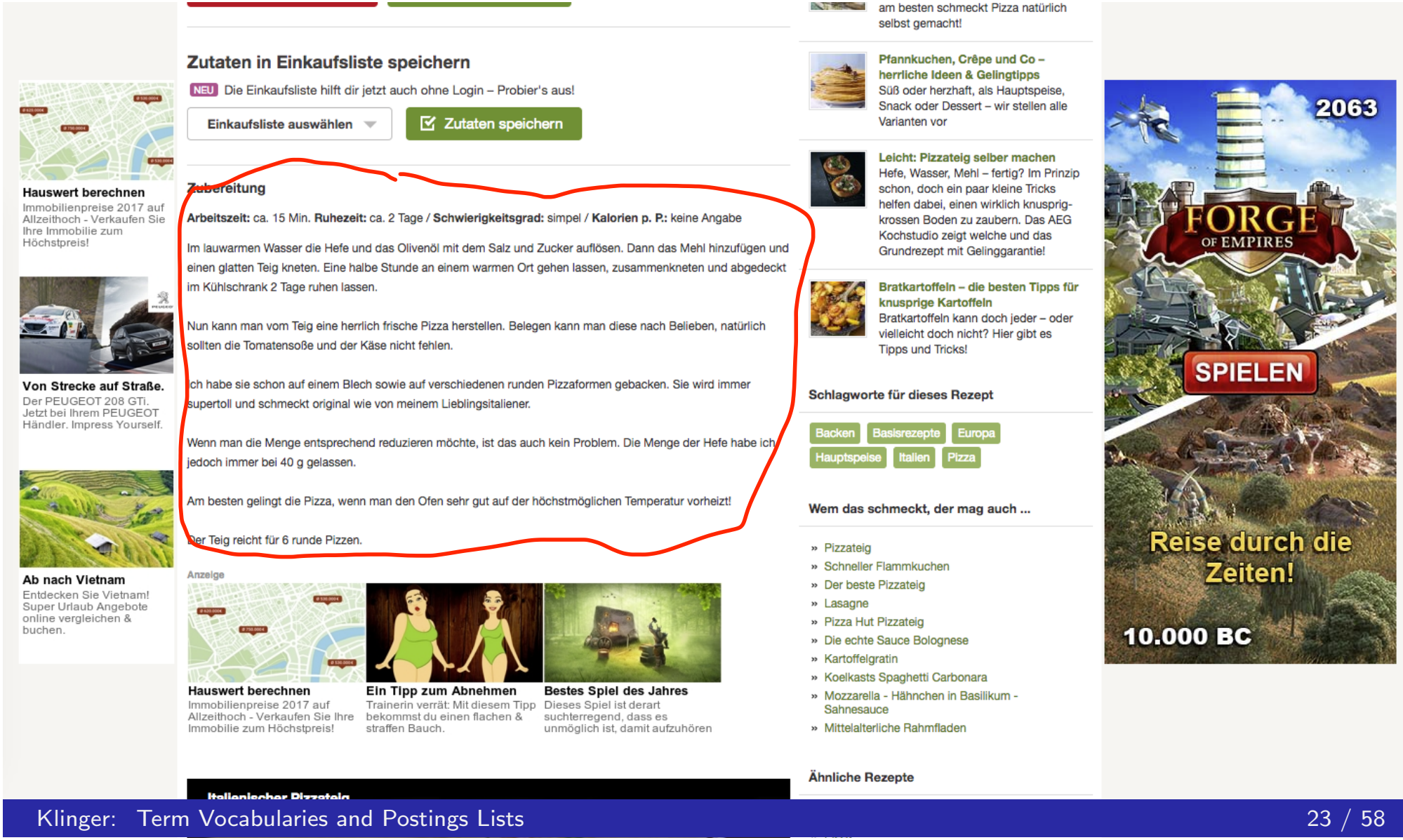
John, Paul, and Mary passed the French exam;
Sam and Judy failed it.

The same technique will handle set-formers within the scope of quantifiers, as in the sentence

Issues with HTML pages – Example (1)



Issues with HTML pages – Example (2)



Hauswert berechnen
Immobilienpreise 2017 auf Allzeithoch - Verkaufen Sie Ihre Immobilie zum Höchstpreis!

Von Strecke auf Straße.
Der PEUGEOT 208 GTi. Jetzt bei Ihrem PEUGEOT Händler. Impress Yourself.

Ab nach Vietnam
Entdecken Sie Vietnam! Super Urlaub Angebote online vergleichen & buchen.

Zutaten in Einkaufsliste speichern

NEU Die Einkaufsliste hilft dir jetzt auch ohne Login – Probier's aus!

Einkaufsliste auswählen | Zutaten speichern

Zubereitung

Arbeitszeit: ca. 15 Min. **Ruhezeit:** ca. 2 Tage / **Schwierigkeitsgrad:** simpel / **Kalorien p. P.:** keine Angabe

Im lauwarmen Wasser die Hefe und das Olivenöl mit dem Salz und Zucker auflösen. Dann das Mehl hinzufügen und einen glatten Teig kneten. Eine halbe Stunde an einem warmen Ort gehen lassen, zusammenkneten und abgedeckt im Kühlschrank 2 Tage ruhen lassen.

Nun kann man vom Teig eine herrlich frische Pizza herstellen. Belegen kann man diese nach Belieben, natürlich sollten die Tomatensoße und der Käse nicht fehlen.

Ich habe sie schon auf einem Blech sowie auf verschiedenen runden Pizzaformen gebacken. Sie wird immer supertoll und schmeckt original wie von meinem Lieblingsitaliener.

Wenn man die Menge entsprechend reduzieren möchte, ist das auch kein Problem. Die Menge der Hefe habe ich jedoch immer bei 40 g gelassen.

Am besten gelingt die Pizza, wenn man den Ofen sehr gut auf der höchstmöglichen Temperatur vorheizt!

Der Teig reicht für 6 runde Pizzen.

Anzeige

Hauswert berechnen
Immobilienpreise 2017 auf Allzeithoch - Verkaufen Sie Ihre Immobilie zum Höchstpreis!

Ein Tipp zum Abnehmen
Trainerin verrät: Mit diesem Tipp bekommst du einen flachen & straffen Bauch.

Bestes Spiel des Jahres
Dieses Spiel ist derart suchterregend, dass es unmöglich ist, damit aufzuhören

Italienischer Pizzateig

am besten schmeckt Pizza natürlich selbst gemacht!

Pfannkuchen, Crêpe und Co – herrliche Ideen & Gelingtipps
Süß oder herzhaft, als Hauptspeise, Snack oder Dessert – wir stellen alle Varianten vor

Leicht: Pizzateig selber machen
Hefe, Wasser, Mehl – fertig? Im Prinzip schon, doch ein paar kleine Tricks helfen dabei, einen wirklich knusprig-krossen Boden zu zaubern. Das AEG Kochstudio zeigt welche und das Grundrezept mit Gelinggarantie!

Bratkartoffeln – die besten Tipps für knusprige Kartoffeln
Bratkartoffeln kann doch jeder – oder vielleicht doch nicht? Hier gibt es Tipps und Tricks!

Schlagworte für dieses Rezept

Backen | Basisrezepte | Europa | Hauptspeise | Italien | Pizza

Wem das schmeckt, der mag auch ...

- Pizzateig
- Schneller Flammkuchen
- Der beste Pizzateig
- Lasagne
- Pizza Hut Pizzateig
- Die echte Sauce Bolognese
- Kartoffelgratin
- Koelkasts Spaghetti Carbonara
- Mozzarella - Hähnchen in Basilikum - Sahnesauce
- Mittelalterliche Rahmfladen

Ähnliche Rezepte

Issues with Proprietary formats – Text Document Example

MS Word

Example document 1

This is an example document.

Roman Klinger 19.10.2017 14:25
Gelöscht: el

Apple TextEdit via RTF

Example document 1

This is an exampleel document.

RTF (saved via TextEdit)

```
{\rtf1\ansi\ansicpg1252\cocoartf1504\cocoasubrtf830
{\fonttbl\f0\fnil\fcharset0 Calibri;\f1\froman\fcharset0 TimesNewRomanPSMT;\f2\f
}
{\colortbl;\red255\green255\blue255;\red52\green90\blue138;}
{\*\expandedcolortbl;;\csgenericrgb\c20392\c35294\c54118;}
{\info
{\author Roman Klinger}
{\*\company University of Stuttgart}}\paperw11900\paperh16840\margl1417\margr141
\deftab708
\pard\pardeftab708\ri0\sb480\partightenfactor0

\f0\b\fs32 \cf2 Example document 1\
\pard\pardeftab708\ri0\partightenfactor0

\f1\b0\fs24 \cf0 \

\f2 This is an exampleel document.
\f1 \
}
```

Outline

Outline

Definitions

- Token: Character sequence in document
  - closely related to Word
- Type: Equivalence class of tokens
  - related to Term: (normalized) type
    - as it occurs e.g. in the IR system's dictionary
- How many tokens? Types? Terms?
  Example: I like the coffee, coffees, and the shop.

coffee

11 tokens
9 types
8 terms

Normalization

- Need to "normalize" words in indexed text as well as query terms into the same form.
- Example: We want to match *U.S.A.* and *USA*
- Two different approaches:
  - Implicitly define equivalence classes of terms. (what does implicit mean here?)
  - Asymmetric expansion
    - window → window, windows
    - windows → Windows, windows
    - Windows (no expansion)
    - More powerful, but less efficient
- Why don't you want to put *window*, *Window*, *windows*, and *Windows* in the same equivalence class?

Normalization: Other languages

- Normalization and language detection interact.
- *PETER WILL NICHT MIT.* $\rightarrow$ MIT $=$ mit
- *He got his PhD from MIT.* $\rightarrow$ MIT $\neq$ mit

Recall: Inverted index construction

- Input:

  Friends, Romans, countrymen. | So let it be with Caesar | ...

- Output:

  friend | roman | countryman | so | ...

- Each token is a candidate for a postings entry.
- Which tokens to use in the index? Which could be ignored?

Exercises

14 tokens
12 types
11 terms

In June, the dog likes to chase the cat in the barn.

How many tokens? How many word types? How many terms?

Tokenize:

Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.

Tokenization problems: One word or two? (or several)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company
- cheap San Francisco-Los Angeles fares
- York University vs. New York University

Numbers

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333
- Older IR systems may not index numbers ...
- ... but generally it's a useful feature.
- What does Google do?

Chinese: No whitespace

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Ambiguous segmentation in Chinese

和尚

The two characters can be treated as one word meaning ‘monk’ or as a sequence of two words meaning ‘and’ and ‘still’.

Other cases of "no whitespace"

- Compounds in Dutch, German, Swedish
- Computerlinguistik → Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter
  - Which results would you like to get if this was your query??
- → leben + versicherung + gesellschaft + angestellter
- Inuit: tusaatsiarunnanngittualuujunga (I can't hear very well.)
- Many other languages with segmentation difficulties: Finnish, Urdu, ...

Japanese

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私 の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡 単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

4 different “alphabets”:

- Chinese characters
- hiragana syllabary for inflectional endings and function words
- katakana syllabary for transcription of foreign words and other uses
- Latin

Arabic script: Bidirectionality

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← → ← → ← START

'Algeria achieved its independence in 1962 after 132 years of French occupation.'

Bidirectionality is not a problem if text is coded in Unicode.

Accents and diacritics

- Accents:
  résumé vs. resume (simple omission of accent)
- Umlauts:
  Universität vs. Universitaet
  (substitution with special letter sequence "ae")
- Most important criterion:
  How are users likely to write their queries for these words?
- Even in languages that standardly have accents, users often do not type them. (Polish?)

Outline

Case folding

- Reduce all letters to lower case
- Even though case can be semantically meaningful
  - capitalized words in mid-sentence
  - MIT vs. mit
  - Fed vs. fed vs. FeD vs. FED
  - ...
- It's often best to lowercase everything since users will use lowercase regardless of correct capitalization.
- Counter example:
  Human Gene name: CES4A, rat gene name: Ces4a

Stop words

- stop words = extremely common words which would appear to be of little value in helping select documents matching a user need
- Examples: *a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with*
- Stop word elimination used to be standard in older IR systems.
- But you need stop words for phrase queries, e.g. *"King of Denmark"*
- Most web search engines index stop words.

More equivalence classing

- Soundex:
  phonetic equivalence, Muller = Mueller
- Thesauri/Ontologies:
  semantic equivalence or similarity, car = automobile

Lemmatization

- Reduce inflectional/variant forms to base form
- Example: *am, are, is* $\rightarrow$ *be*
- Example: *car, cars, car's, cars'* $\rightarrow$ *car*
- Example: *the boy's cars are different colors* $\rightarrow$ *the boy car be different color*
- Lemmatization implies doing "proper" reduction to dictionary headword form (the lemma).
- Inflectional morphology (*cutting* $\rightarrow$ *cut*)
  vs. derivational morphology (*destruction* $\rightarrow$ *destroy*)

Stemming

- Definition of stemming: Crude heuristic process that chops off the ends of words in the hope of achieving what "principled" lemmatization attempts to do with a lot of linguistic knowledge.
- Language dependent
- Often inflectional and derivational
- Example for derivational:
  automate, *automatic*, *automation* all reduce to *automat*

Porter algorithm

- Most common algorithm for stemming English
- Conventions + 5 phases of reductions
- Phases are applied sequentially
- Each phase consists of a set of commands.
  - Sample command:
    Delete final *ement* if what remains is longer than 1 character
    replacement → replac
    cement → cement
- Sample convention:
  Of the rules in a compound command, select the one that applies to the longest suffix.
- Implementation e.g. in `http://snowball.tartarus.org/`

Porter stemmer: A few rules

Rule			**Example**		
SSES	$\rightarrow$	SS	caresses	$\rightarrow$	caress
IES	$\rightarrow$	I	ponies	$\rightarrow$	poni
SS	$\rightarrow$	SS	caress	$\rightarrow$	caress
S	$\rightarrow$		cats	$\rightarrow$	cat

Three stemmers: A comparison

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Does stemming improve effectiveness?

- In general, stemming increases effectiveness for some queries, and decreases effectiveness for others.
- Queries where stemming is likely to help:
  tartan sweaters
  sightseeing tour san francisco
- (equivalence classes: {sweater,sweaters}, {tour,tours})
- Porter Stemmer equivalence class *oper* contains all of *operate operating operates operation operative operatives operational*.
- Queries where stemming hurts: [operational AND research], [operating AND system], [operative AND dentistry]

What does Google/Bing/DuckDuckGo do?

- Stop words
- Tokenize at which characters (hyphen? period?...)?
- Lowercasing
- Stemming
- Non-latin alphabets
- Umlauts
- Compounds
- Numbers

Outline

Recall basic intersection algorithm

BRUTUS $\longrightarrow$ $\boxed{1} \rightarrow \boxed{2} \rightarrow \boxed{4} \rightarrow \boxed{11} \rightarrow \boxed{31} \rightarrow \boxed{45} \rightarrow \boxed{173} \rightarrow \boxed{174}$

CALPURNIA $\longrightarrow$ $\boxed{2} \rightarrow \boxed{31} \rightarrow \boxed{54} \rightarrow \boxed{101}$

Intersection $\Longrightarrow$ $\boxed{2} \rightarrow \boxed{31}$

Recall basic intersection algorithm

$$\textsc{Brutus} \longrightarrow \boxed{1} \rightarrow \boxed{2} \rightarrow \boxed{4} \rightarrow \boxed{11} \rightarrow \boxed{31} \rightarrow \boxed{45} \rightarrow \boxed{173} \rightarrow \boxed{174}$$

$$\textsc{Calpurnia} \longrightarrow \boxed{2} \rightarrow \boxed{31} \rightarrow \boxed{54} \rightarrow \boxed{101}$$

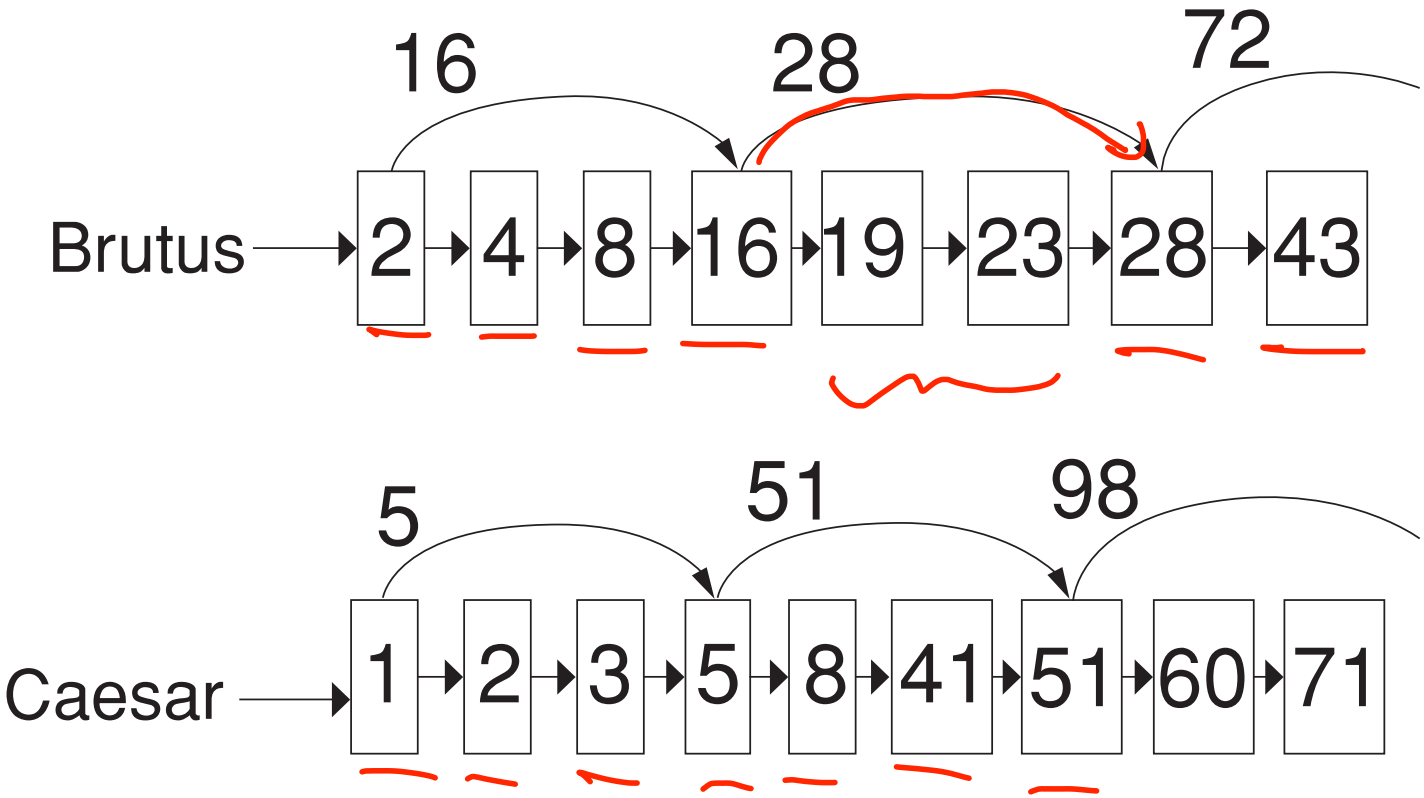
$$\text{Intersection} \Longrightarrow \boxed{2} \rightarrow \boxed{31}$$

- Linear in the length of the sum of the postings lists.
- Can we do better?

Skip pointers

- Skip pointers enable to skip postings that will not figure in the search results.
- This makes intersecting postings lists more efficient.
- Some postings lists contain several million entries – so efficiency can be an issue even if basic intersection is linear.
- Where do we put skip pointers?
- How do we make sure intersection results are correct?

Skip lists: Example



Where do we place skips?

- Tradeoff:
  number of items skipped vs. frequency skip can be taken
- More skips:
  Each skip pointer skips only a few items, but we can frequently use it.
- Fewer skips:
  Each skip pointer skips many items, but we can not use it very often.

Where do we place skips? (cont)

- Simple heuristic: for postings list of length P, use $\sqrt{P}$ evenly-spaced skip pointers.
- This ignores the distribution of query terms.
- Easy if the index is static; harder in a dynamic environment because of updates.
- How much do skip pointers help?
  - Memory/Computation trade-off

Take-away

- Understanding of the basic unit of classical information retrieval systems: words and documents: What is a document, what is a term?
- Tokenization: how to get from raw text to words (or tokens)
- More complex indexes: skip pointers