Textdatenbanken

Sommersemester 2018

4. Vorlesung

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Angaben zu Wörtern: Sachgebiet

Tabelle **`subject_area`** mit Spalten **`w_id`** und **`subject_area`**

Index auf **`w_id`**

(word)	w_is	subject_area
Abfangjäger	141400	Luftfahrt
Balustrade	113484	Architektur
Beinfreiheit	131029	Auto
Blinker	107209	Auto
Bodenplatte	104364	Bauwesen
wägen	102481	Technik
Windstoß	113556	Meteorologie
Wölbung	126466	Technik
Xenon	103457	Chemie
Ziegelstein	111300	Bauwesen

Sachgebietsangaben zu Wortnummern: Die Datenbank enthält nur Wortnummern und Sachgebiete, die dünn gesetzten Wörter dienen hier nur zur Information

Angaben zu Wortpaaren: Kookkurrenzen

Nachbarschaftskookkurrenzen: Tabelle **`co_n`** mit Spalten **`w1_id, w2_id`** und **`sig`**

Index auf **`w1_id, w2_id`**

(word_1)	(word_2)	w1_id	w2_id	sig
geschmolzenes	Blei	202641	13296	25
gehacktem	Blei	848257	13296	30
Mikrogramm	Blei	10975	13296	36
flüssiges	Blei	55179	13296	36
...	...	...	...	...

Nachbarschaftskookkurrenzen. Linke Nachbarn zum Wort Nummer 13296 *Blei*. Dabei enthält die Datenbank nur die Wortnummern, die dünn gesetzten Wörter dienen hier nur zur Information

Angaben zu Sätzen: Tabellen

Sätze:

- Sentences (Sätze)

Quellen:

- Sources (Quellen)
- inv_so (Verweise von Satznummern auf Quellen)

Inverse Liste:

- inv_v (Verweise von Satznummern auf Wörter)

Angaben zu Sätzen: Tabelle sentences

Tabelle **`sentences`** mit Spalten **`s_id`** und **`sentence`**
Index auf **`s_id`**

s_id	sentence
3	Leder: Vielleicht ringt Normann nur um Anerkennung.
6	Das können die Fachleute beraten.
10	Rose sei Realist, einer, der erst denkt und dann handelt.
12	Ich liege im Bett, im Krankenhaus, träume tief und dunkel.
17	Das von Seoul finanzierte Projekt ist in der Anfangsphase.
18	Auf dem Bus steht in kursiven Lettern: Die Wölfe kommen.
29	Jeder westliche Soldat wird im Irak als Besatzer angesehen.
31	Damit ist sie aber hoffnungslos überfordert.
32	Vogts Konzept trug schnell Früchte.
36	Erich Mielke ist an allem Schuld.
38	Wir haben den Willen zur aktiven Zusammenarbeit.
49	Die beiden Insassen blieben jedoch unverletzt.
59	"Das hätte uns gut getan."
60	Demnächst will er auf die Auer Dult.

Angaben zu Sätzen: Quellen

Tabelle **`sources`** mit Spalten **`so_id, source`** und **`date`**

Index auf **`so_id`**

so_id	source	date
1	Berliner Zeitung vom 30.11.2001	2001-11-30
2	Süddeutsche Zeitung vom 15.03.2002	2002-03-15
3	Süddeutsche Zeitung vom 27.09.2001	2001-09-27

Tabelle **`inv_s`** mit Spalten **`so_id`** und **`s_id`**

Index auf **`s_id`**

so_id	s_id
118823	1
118823	2
1527	3

Angaben zu Sätzen: Inverse Liste

Tabelle **`inv_w`** mit Spalten **`w_id, s_id`** und **`pos`** (wievieltes Wort im Satz)

Index auf **`s_id`** und **`w_id.`**

w_id	s_id	pos
1	92	26
1	104	21
1	527	15
1	647	8
1	728	18

Qualitätssicherung: Sprache

Nur Sätze der gewünschten Sprache berücksichtigen:

Sprachidentifikation auf Satzebene: Bestimme

- die wahrscheinlichste Sprache L_1 (korrekt mit Wahrscheinlichkeit p_1)
- die zweit-wahrscheinlichste Sprache L_2 (korrekt mit Wahrscheinlichkeit p_2)

Sprache L_1 wird akzeptiert, wenn

- p_1 ausreichend groß und
- p_1/p_2 ausreichend groß

Qualitätssicherung: Musterbasiert

Sätze werden entfernt (oder gar nicht erst aufgenommen), wenn sie einem der folgenden Muster entsprechen:

- Länger als 256 Zeichen
- Satzanfang und -ende kontrollieren: `regexp '^[A-ZÄÖÜ"'].*[.:!?"']$'`
- **Sätze mit gesperrtem Text löschen:** `like '% _ _ _ _ _ _ %'`
- **Sätze mit zwei aufeinanderfolgenden Leerzeichen löschen**

  ```
  Kommunale Abfallwirtschaft rechnet mit 500 DM jährlich      aha München
  (Eigener Bericht).
  ```

- **Sätze mit allzu vielen Kommas löschen:**
  `like '%,%,%,%,%,%,%,%,%,%,%'`

  ```
  Krupp AG Hoesch-Krupp, Gea Vorzugsaktien, Gehe, Gerresheimer, Glunz
  Vorzugsaktien, Grohe Vorzugsaktien, Hornbach Vorzugsaktien, Kampa Haus, Moksel,
  O. Reichelt, SAP Vorzugsaktien, Spar Handel Vorzugsaktien, Strabag, Villeroy &
  Boch sowie Weru.
  ```

Mehr Muster

- **Sätze mit allzu vielen Punkten löschen:** `like '%.%.%.%.%.%'`
  Reiseberichte - Linksammlung: 1001 Reiseberichte in deutscher Sprache
  ....
- **Sätze mit mehr als 50 Leerzeichen löschen (45 sind noch sinnvoll möglich!):**
  `length(sentence)-length(replace(sentence," ",""))>50`
  FC Köln - 29 796 Karlsruher SC 25 000 29 060 Eintracht Frankfurt 28 300 28 487 VfL Bochum - 24 274 Bayer Leverkusen 16 800 22 875 MSV Duisburg - 20 447 Bayer Uerdingen 31 415 17 349 SC Freiburg - 17 029 Dynamo Dresden - 16 588 Insgesamt 292 345
- **Sätze mit anteilig zu vielen Leerzeichen löschen:**
  `length(replace(sentence," ",""))/length(sentence)<.7`
  Elferwette: 2 1 2 1 1 1 2 1 1 0 2.
- **Sätze mit Links und "¦" löschen:**
  `(sentence like "%|%" or sentence like "%[%[%")`
- **Sätze mit Sonderzeichen löschen:**
  `(sentence like "% >> %" or sentence like "%++%" or sentence like "%*%" or sentence like "%~%")`

Noch mehr Muster

- **Sätze mit mehreren Ausrufe- oder Fragezeichen löschen:**
  ```
  regexp "[!?].*[!?]"
  DU SOLLST AN MILZBRAND KREPIEREN!!!!!!!
  Wenn nicht dieses Tape dann gar keinz!!!! |
  ```
- **Sätze mit vielen gleichen Zeichen löschen:**
  ```
  (sentence like "%)%)%)%)%" or sentence like "%/%/%/%"
  or sentence like "%&%&%&%")
  ```
- **Sätze mit vielen Großbuchstaben oder vielen Ziffern hintereinander löschen:**
  ```
  sentence regexp "[[:upper:] .,/-]{20}" or sentence
  regexp "[[:digit:] .,/-]{16}"
  DANKE AN EUCH ALLE, DASS IHR SO NETTE SACHEN HIER REINSCHREIBT.
  Kathrin Sadowski Borsteler Bogen 27 22453 Hamburg Telefon: 040 / 553 87 58
  Telefax: 040 / 553 88 86
  ```
- **Sätze mit Leerzeichen vor Satzzeichen löschen**
- **Sätze mit Initialen am Ende löschen:** `regexp " [A-Za-z].$"`

Letzte Muster

- **Kurze Sätze mit vielen Ziffern vor Punkt löschen:** `sentence regexp "[[:digit:].,/-]{6}$" and length(sentence)<45`

  ```
  " 29. Kapitel Ev.Joh.4,27.

  Meine Maße sind 91-60-91.
  ```

- **Extrem kurze Sätze mit Punkt löschen:** `length(sentence)<13 and sentence regexp "\.$"`
- **Schachpartien löschen (mit mindestens zwei typischen "Schachwörtern"):** `regexp '[KDSTL][a-h][1-8].*[KDSTL][a-h][1-8]'`

Achtung: Erst die Wirkung der Statements prüfen (z.B. Sätze markieren). Speziell, wenn die Statements verändert wurden!

Erst dann löschen!

Nicht alle Muster für jede Sprache

- Manche Alphabete haben keine Groß-/Kleinschreibung (Arabisch, Hebräisch, Devanagari, …)
- Andere Satzendezeichen (in vielen Sprachen)
- Mehr (Vietnamesisch) oder weniger (Thai, Japanisch) Leerzeichen.
- Setzung von Leerzeichen um Satzzeichen nach anderen Regeln.

Kompromiss: Internationalisierte Version

- Vereinfachter regulärer Ausdruck mit Zeichenklassen für Sätze
- Viele zulässige Satzendezeichen
- Weniger Ausschlussregeln

Ranking für Sätze

In welcher Reihenfolge sollen Sätze abgespeichert werden, falls später in der entsprechenden Reihenfolge zugegriffen werden kann?

Idee: Sortiere Sätze nach „Schönheit“. Diese wird so definiert, dass negativ bewertete Strukturelemente eines Satzes diesen betrafen. Die Sätze mit der geringsten Gesamtstrafe sind dann automatisch die „schönsten“.

Einzelstrafen haben typischerweise die gleiche Größe, bei sehr großen Abweichungen werden auch größere Strafen vergeben.

Manchmal wird ein Zielwert vorgegeben, der kleiner ist als der Mittelwert. Beispielsweise bevorzugen wir Sätze kürzer als der Durchschnitt, Sätze mit weniger Satzzeichen oder Ziffern als üblich.

Idee: Adam Kilgarriff et al: GDEX: Automatically finding good dictionary examples in a corpus, Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra: 425-433

Satzstruktur: Elemente

Das Vorkommen einiger Strukturelemente soll bestraft werden:

- **Anzahl Sonderzeichen**, nicht ,.!? Strafe pro Zeichen: 5 Pkte
- **Anzahl reguläre Satzzeichen**(.!?) minus 1. Strafe pro Zeichen: 3 Pkte
- **Ziffern**: Strafe 5 für jede Ziffer
- **Aufeinanderfolgende Großbuchstaben**: Strafe 5 für jedes Paar
- **Ungerade Anzahl Anführungszeichen** Strafe: 20 für isoliertes Anführungszeichen, 60 für drei Stück.

Satzstruktur

Abweichung vom Zielwert (oder Mittelwert) um denselben Faktor nach oben oder unten wird gleich bestraft.

- **Satzlänge in Zeichen**, Ziel: 80 -110 Zeichen (Mittelwert ist 110). Strafe: Abweichung um Faktor 2 von Grenze = 10 Pkte.
- **Anzahl Top-100-Stoppwörter im Satz**: Abgeschätzt als: 40% der Wortzahl. Wortzahl=Satzlänge/7, 7=mittlere Wortlänge+1. Strafe: Faktor 2 = 10 Pkte.
- **Seltenstes Wort**: Gewünschter logarithmierter Rang für seltenstes Wort: 15. Strafe 4 für Faktor 2.
- **Mittlerer Rang**: Ziel: 8000 (Mittelwert ist 18000). Strafe: Faktor 2 = 10 Pkte.
- **Mittlere Wortlänge**: Ziel: 6 (Mittelwert ist 8). Strafe: Faktor 2 = 50 Pkte.

Sätze mit minimaler Strafe

```
use deu_newscrawl_2011;
select round(penalty,4) as p,sentence from sentences s, penalty p where s.s_id=p.s_id limit 20;
+--------+-------------------------------------------------------------------------------+
| p      | sentence                                                                      |
+--------+-------------------------------------------------------------------------------+
| 0.4103 | Gusenbauer selbst wirkte vergangene Woche verständlicherweise etwas angeschlagen.     |
| 0.4486 | Manche dieser Apps bringen ihren Benutzern sogar praktischen ökologischen Nutzen.     |
| 0.4505 | Eine kleine Berufsarmee würde endlich diesen peinlichen Zustand hinter uns lassen.    |
| 0.4514 | Diese Begeisterung wollen jetzt einige junge Extremsportler wirtschaftlich nutzen.    |
| 0.5844 | Nordrhein-Westfalen muss demnach seinen derzeitigen Bestand fast verdreifachen.       |
| 0.6767 | Bei seiner Ausarbeitung wurden unterschiedliche neuere Studien berücksichtigt.        |
| 0.6935 | Aus vielen unterschiedlichen Gründen brauchen Patienten bisweilen Zahnersatz.         |
| 0.6978 | Bahnreisende ab Erfurt Richtung Sachsen mussten erhebliche Verspätungen hinnehmen.    |
| 0.7083 | Eine entsprechende Absichtserklärung unterzeichneten Vertreter beider Staaten.        |
| 0.7601 | Regelmäßig wurden kulturelle Veranstaltungen unterschiedlichen Genres ausgerichtet.   |
| 0.7617 | Mögen alle europäischen Länder thailändische Touristen gleichermaßen behandeln.       |
| 0.7818 | Zumindest tagsüber sollen beispielsweise Berufstätige ungehindert fahren dürfen.      |
| 0.7967 | Unentschieden hieß es auch zum Schluss der Begegnung zwischen Dynamo und Moskwa.      |
| 0.7969 | Denn bei einem Konjunktureinbruch geht die Teuerung in der Regel schnell zurück.      |
| 0.7986 | Auch in der Dauerausstellung lässt sich in vielen Bereichen selbst Hand anlegen.      |
| 0.8072 | Die Rumänin widmete sich fortan voll und ganz ihrer Karriere an der Universität.      |
| 0.8084 | Die Nachricht schreckte Ungarn am vergangenen Mittwoch wie ein Paukenschlag auf.      |
| 0.8095 | Sarrazin erschwere die Diskussion über die Integration statt sie zu erleichtern.     |
| 0.8125 | Seine Frau habe sich einen Freund zugelegt und sei jedes Wochenende weggegangen.      |
| 0.8138 | Das spielt der neuen linken und sozialistisch geprägten Elite voll in die Hände.      |
+--------+-------------------------------------------------------------------------------+
```

Nach den besten 10%

Strafe	Satz
14.9093	Und zu Hause laufe ich natürlich auch nicht in hohen Schuhen rum.
14.9093	Verwenden Sie am Besten eine möglichst lange Kombination aus Buchstaben, Zahlen und Sonderzeichen.
14.9093	"Damit setzt sich die Bundesregierung dem Vorwurf der Käuflichkeit von Politik aus."
14.9093	Der Bürgermeister ließ daraufhin die Linde fällen, ohne Rücksprache mit dem Umweltministerium und ohne europaweite Ausschreibung.
14.9093	Die erste erkennbare Veränderung steht kurzfristig an:
14.9093	Die Schüler sollen keine Hausaufgaben mehr bekommen, wenn sie Ganztagsunterricht haben.
14.9093	Er starb nach einjähriger Ehe.
14.9093	«Er steht hundertprozentig im Kader», sagte Trainer Thomas Tuchel am Freitag.
14.9093	Ich war in New York, als ich erstmals von dem Erdbeben erfuhr, am Abend nach den ersten Erdstößen.
14.9093	Künftig lebt er mit seiner Familie in Hamburg, und diese Stadt will sich der leidenschaftliche Marathonläufer natürlich laufend erschließen.
14.9093	Möglicherweise handelt es sich um eine Überdosis.
14.9093	Ob melancholisch schön oder euphorisch, aus lauter Lebensfreude, Ralph versteht es in seinen Zuhörern die Emotionen zu wecken.
14.9093	Warum machen wir es nicht wie die Skandinavier?
14.9093	Was mich stutzig gemacht hat, sie sollen in Lissabon tätig sein, warum ausgerechnet Lissabon?
14.9093	Welch wichtige Entscheidung.
14.9093	Bundeskanzler Werner Faymann befürwortete dabei eine Diskussion über eine Neuorganisation des Bundesheeres.

Nach 50%

Strafe	Satz
37.1220	Anstelle der Fackel soll eine Aussichtsplattform für Besucher entstehen: Immerhin wollen jährlich bis zu 30.000 Menschen das Stadion besichtigen.
37.1220	Aus Sicherheitsgründen wird es in der Zeit von 20.30 bis 22.30 Uhr auf dem Schloßberg Zugangsbeschränkungen geben.
37.1220	Bizer wirft ein, dass der Staat Geld ausgeben müsse, um die Wirtschaft am Laufen zu halten und dadurch gezwungen sei, die Schuldengrenze auf später zu verschieben.
37.1220	Das deutsche Team war am Dienstag mit nur 15 Spielern zur WM-Endrunde gereist, drei Kicker sollen noch dazu stoßen.
37.1220	"Das ist eine hervorragende Leistung der gesamten Belegschaft und des Managements des VW-Konzerns", sagte Osterloh.
37.1220	Das Ministerium wehrte ab: Als Notfallmanager würden nur ausgebildete Fahrdienstleiter eingesetzt, ihre Ausbildung müsse nicht extra geprüft werden.
37.1220	"Der EURO ist tot, es lebe die D-Mark" klingt es inzwischen öfter aus den Gazetten und Diskussionsforen.
37.1220	Der Hauptschullehrer gerät ins Plaudern.
37.1220	Der Kluge gibt da schon mal nach und schiebt seinen Drahtesel ein wenig…
37.1220	Der wegen seines unklaren Verhältnisses zu einer 18-Jährigen unter Druck geratene italienische Regierungschef Silvio Berlusconi will die Veröffentlichung von pikanten Fotos aus seiner Luxusvilla auf Sardinien verhindern.

Nach 90%

Strafe	Satz
81.2586	Zur Finanzierung des neuen Rekord-Pakets will die Regierung dem Parlament voraussichtlich Ende dieses Monats einen Nachtragshaushalt für das am 1. April begonnene Fiskaljahr 2009 vorlegen.
81.2586	25.09.2009: SPD-Chef Franz Müntefering hat seinen Rückzug angedeutet.
81.2586	Bremen: In Bremen un Bremerhaben fehlen 150 Gendaarms.
81.2586	Jede Woche WerWoWas in Ihrem RGA:
81.2586	Seit über vierzig Jahren hält sich die britische Band («Smoke on the Water») auf dem Rock-Olymp – selbst eine mehrjährige Auszeit in den 1980er-Jahren hat den Legendenstatus nicht beschädigt.
81.2587	FPÖ-Interims-Chef Herbert Haupt habe sich "eher schwächlich präsentiert" und wenig Interesse für wichtige Sachthemen, wie die Sicherung der Pensionen, bekundet, so Bures abschließend.
81.2587	Die schöne Matinee vor vollem Haus gab den Rahmen für eine Ehrung der Ortsgruppe: Evelyn Faupel singt seit 30 Jahren im SAV-Chor mit, aber nicht nur das.
81.2587	Die Überlebenden der Maquis von Mustapha Bouyali, denen sich die „Afghanen“13 anschlossen, erhielten die Unterstützung der Gegner von Madani, die beim Kongreß von Batna in der Minderheit waren.
81.2587	HÖINGEN – Von einem Firmengrundstück in der Harkortstraße in Höingen wurde zwischen Dienstag gegen 15.30 Uhr und Mittwoch gegen 6 Uhr Aluminiumschrott entwendet.
81.2587	Schon ab 2010 soll, so der Plan der «Bolivarischen Allianz für unser Amerika», eine virtuelle Währung mit dem symbolträchtigen Namen Sucre die Abwicklung des Handels unter 9 südamerikanischen Ländern regeln und den Dollar als Referenzwährung ablösen.
81.2587	Von Jo Nattermanns Landhaus Nattermann kamen 150 Euro.

Nach 99%

Strafe	Satz
100.8417	Das Übereinkommen wurde von den Ländern der Wirtschaftskommission der Vereinten Nationen für Europa (UN/ECE) ausgehandelt und am 25. Juni 1998 auf einer paneuropäischen Konferenz der Umweltminister um dänsichen Århus angenommen - daher sein Name.
100.8417	Die Finanzschuld des Bundes steigt auf 57,81 Prozent des BIP an - das sind 147 Milliarden Euro (zum Vergleich: Im Jahr 2000 betrug die Finanzschuld 57,37 Prozent).
100.8417	Laut einer Studie des Spektra-Instituts aus dem Jahr 2000 kaufen in Österreich bereits 110.000 Haushalte, ein Drittel der Computer-Haushalte, Waren und Dienstleistungen, insbesondere Bücher, elektronische Geräte und Reisen, im Internet ein.
100.8417	"Während den ersten drei Monaten des Jahres 2005 konnten 44.402 Zugriffe auf die Englischsprachigen Seiten verzeichnet werden.
100.8417	Das Unternehmen sicherte sich sowohl die Printausgabe wie auch die Online-Version BusinessWeek.com. "Die Akquisition wird unsere Online-, TV- und Mobile-Produkte stärken", erklärte Bloomberg-Vorsitzender Peter Grauer.
100.8418	Nach dem Galaterbrief kann der »Nachkomme Abrahams« nur einer sein; er findet sich in Christus und allen, die ihm angehören (Gal 3,16.29).
100.8419	Das auf die außeruniversitäre Forschung zugeschnittene bm:vit-Programm „FEMtech – Frauen in Forschung und Technologie“ etwa stellt jeden Monat eine „FEMtech-Expertin“ vor.

Elementare Abfragemöglichkeiten

Abfragen für Wörter

Statistisch

- Häufigkeit eines Wortes (im Korpus)

Analog zum Wörterbuch

- Grammatikangaben, Silbentrennung,
- Sachgebiet
- Bedeutungsbeschreibung, evtl. mehrere Bedeutungen
- Synonyme, Ober- und Unterbegriffe

Korpusbasierte Abfragen

- Gesucht sind Belegsätze für ein gegebenes Wort.
- Typisches gemeinsames Auftreten: Nachbarschafts- und Satzkookkurrenzen zu einem Wort

Mehr Abfragemöglichkeiten

Abfragen für Wörter

Analog zum Wörterbuch

- Wörter ähnlicher Bedeutung
- Sachgebiet
- Synonyme, Ober- und Unterbegriffe
- Wörter mit ähnlicher Schreibweise (kleiner Levenshtein-Abstand)

Korpusbasierte Abfragen

- Belegsätze für Kookkurrenzen
- Gesucht sind Belegsätze für die verschiedenen Bedeutungen eines Wortes.
- Kookkurrenzen sortiert nach den Bedeutungen des Stichwortes
- Kookkurrenzen sortiert nach Wortart

Abfragen für Sätze

- Satz mit POS-Tags
- Satz mit NER-Tags
- Satz mit Syntaxbaum
- Ähnliche Sätze
  - Große Ähnlichkeit: Quasidubletten
  - Geringere Ähnlichkeit: vergleichbarer Inhalt
  - Syntaktische Ähnlichkeit: nach Struktur

Abfragen unter Berücksichtigung der Zeit

- Häufigkeit über die Zeit
  - Neue Wörter
  - Aussterbende Wörter
  - Regelmäßig wiederkehrende Wörter
- Richtiges Zeitfenster auswählen (Tag / Monat / Jahr)
- Kookkurrenzen für einen Tag

Hilfreiche Vermutung: Heute ist ein Wort nur in einer Bedeutung auffällig, d.h. alle Kookkurrenzen gehören zu einer Bedeutung

Abfragen über mehrere Korpora

Welche Wörter kommen in Korpora verschiedener Sprachen gemeinsam vor?

- Eigennamen (George W. Bush, IBM)
- Internationalismen: Video, Computer
- Einige Stoppwörter: in, an, ja
- Falsche Freunde: war, die (de/en)

Welche Wörter sind verblüffend ähnlich: Cognates

- Große Mindestlänge und kleiner Levinshtein-Abstand,
- z.B. Präsident / presidente / president, Universtität / universidad / university
- Falsche Freunde: freedom / Frieden