Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Information Engineering 1: Information Retrieval

Theorie: Indexierung/Vergleich

Kapitel 2

Martin Braschler

Inhalt

- Indexierung
  - Verschlagwortung
  - Volltextindexierung
- "Bag of Words"-Modell
- Rangierungsregeln
- Wortstatistiken
- Vektorraummodell
- Boole'sches Retrieval
- Probabilistisches Retrieval

Ausgangslage

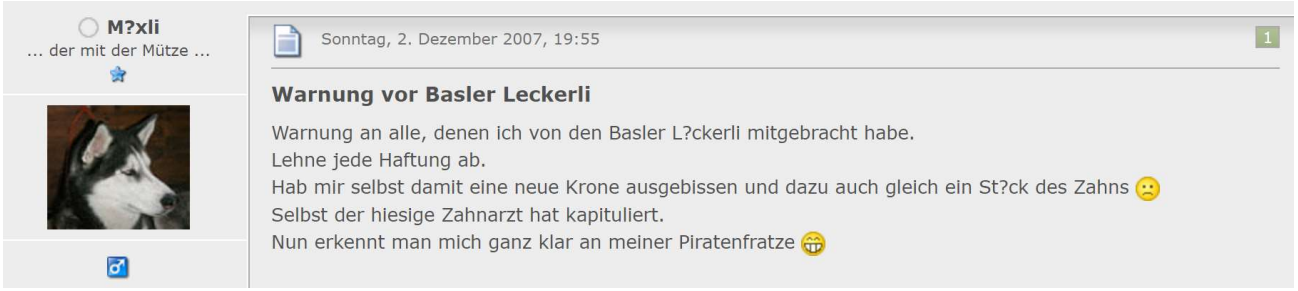
- Man nehme:
  - 1 vages Informationsbedürfnis
  - 1‘000‘000 unstrukturierte Volltexte
- Gesucht: Resultat
- Vorgehen:
  1. Anfrage und Dokumente vergleichbar machen
  2. Vergleich
- Datenbank: in Tabelle Apfel sind nur Äpfel
- Information Retrieval: Es werden Äpfel mit Birnen verglichen

Das Problem

- Relevante Dokumente können auch überhaupt nicht auf die Suchanfrage passen (= enthalten keine Suchbegriffe)
- Irrelevante Dokumente können sehr wohl alle Suchbegriffe beinhalten

A.- Max Michel brach sich am 26. September 1975 im Zivilschutzkurs beim Essen eines "Totenbeinli"-Bisquits den rechten oberen, bereits plombierten Eckzahn ab.

Durch Verfügung vom 5. März 1976 lehnte die Militärversicherung die Haftung für den Zahnschaden mit der Begründung ab, das Abbrechen eines vorbehandelten Zahnes beim normalen Kauakt sei eine dem Ergrauen der Haare vergleichbare Zerfallserscheinung. Das Abbrechen könne unter beliebigen Lebensumständen erfolgen und werde im Dienst nicht mehr gefördert als im Zivilleben. Die



Die Grundidee

- Wir stellen den Match über die gemeinsame Merkmale in den Dokumenten dar
- Wir haben schon gesehen, dass «Wörter» hierzu wohl nicht die ideale Lösung sind
- Trotzdem ist das Matching auf Wort-basierten Merkmalen die Grundlage des «klassischen Information Retrievals»
- Klassisches IR in seiner reifsten Form lierfert weiterhin sehr starke (State-of-the-Art) Retrievaleffektivität

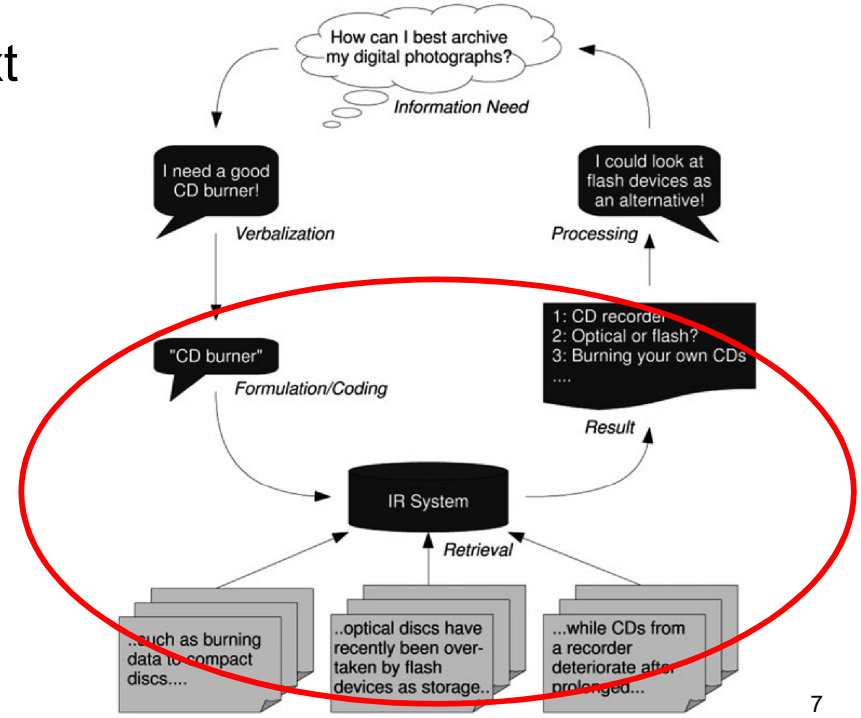
Warum Wort-basiertes Matching schwierig ist

- Sprache ist nicht „eindeutig“:
  - Synonyme (eine Bedeutung – mehrere Wörter)
  - Homonyme (mehrere Bedeutungen – ein Wort)
  - Umschreibungen
  - Metaphern
  - Wortformen (Singular, Plural, Verbformen, etc.)
  - Schreibfehler, Rechtsschreibevarianten, Transkriptionsvarianten
  - Abkürzungen
  - → Anfrage und Dokument „passen nicht zusammen“.
- Bedeutungen (und Relevanz) ändern mit der Zeit
  - (nine eleven=Notrufnummer oder Inbegriff einer Zeitenwende? Beides? Oder ein Porsche?)
- Menschen lernen dazu oder vergessen; was gestern irrelevant war, kann schon heute ins Schwarze treffen (relevant sein).
- Sprachen, Akronyme, Phrasen, Komposita, versch. Schreibweisen, etc.
- Unschärfe macht uns das Leben schwer (im Retrieval, automatische Sprachverarbeitung)
- *Unschärfe in der Sprache macht uns kreativ (Metapher, Analogien, alte Begriffe im neuen Kontext, Witze, Andeutungen, ich kann schon versuchen etwas zu sagen, bevor ich alles 100% verstanden habe.)*

Abgrenzung

Wir bewegen uns im IR-Kreislauf «ganz unten», d.h., wir betrachten das eigentliche IR-System, und wollen verstehen, wie die Rangliste zu Stande kommt.

- Anfrage: codiertes Informationsbedürfnis
- Dokumente: codierte Information
- Wir konzentrieren uns im folgenden auf Text



Frage:

wie wird in unstrukturierten Dokumenten (natürliche Sprache) Information "codiert"?

Inhaltstragende Wörter

- → Annahme: die Wörter in einem Text (und ihre Häufigkeiten) sagen etwas über die Relevanz dieses Textes aus (→ Alternativen?)
- Information wird in inhaltstragenden Wörtern codiert.
- Welche der folgenden Wörter sind "inhaltstragend"?

- "Dieser kurze Text behandelt das Verhältnis zwischen Information Retrieval und Datenbanken. Er befasst sich ***nicht*** mit Informationssystemen im Allgemeinen. Information Retrieval bezeichnet vor allem das Auffinden unstruktierter Information als Antwort auf ein Informationsbedürfnis. Datenbanken dienen typischerweise zum Wiederauffinden strukturierter Daten."

Inhaltstragende Wörter

- Manche Wörter tragen "wenig" Inhalt (lexikalische Bedeutung). Es sind dies die "Funktionswörter", wie Artikel, Partikel, Pronomen, Konjunktionen, ..
  - Aber: beachten sie die Bedeutung von "nicht" im vorhergehenden Beispiel!
- Manche Wörter sind inhaltstragend, helfen aber nicht, den Text zu charakterisieren
- Für das Retrieval besonders interessant sind diejenigen Wörter, welche inhaltstragend sind und den Text "auszeichnen"

Worthäufigkeiten

- Worthäufigkeiten sind nicht gleichverteilt
- Einige wenige Wörter sind sehr häufig (→ welche?)
- Sehr viele Wörter sind sehr selten

- Zipfsches Gesetz:

- Rang der Häufigkeit * Häufigkeit = konstant

Wortverteilung Frankfurter Rundschau

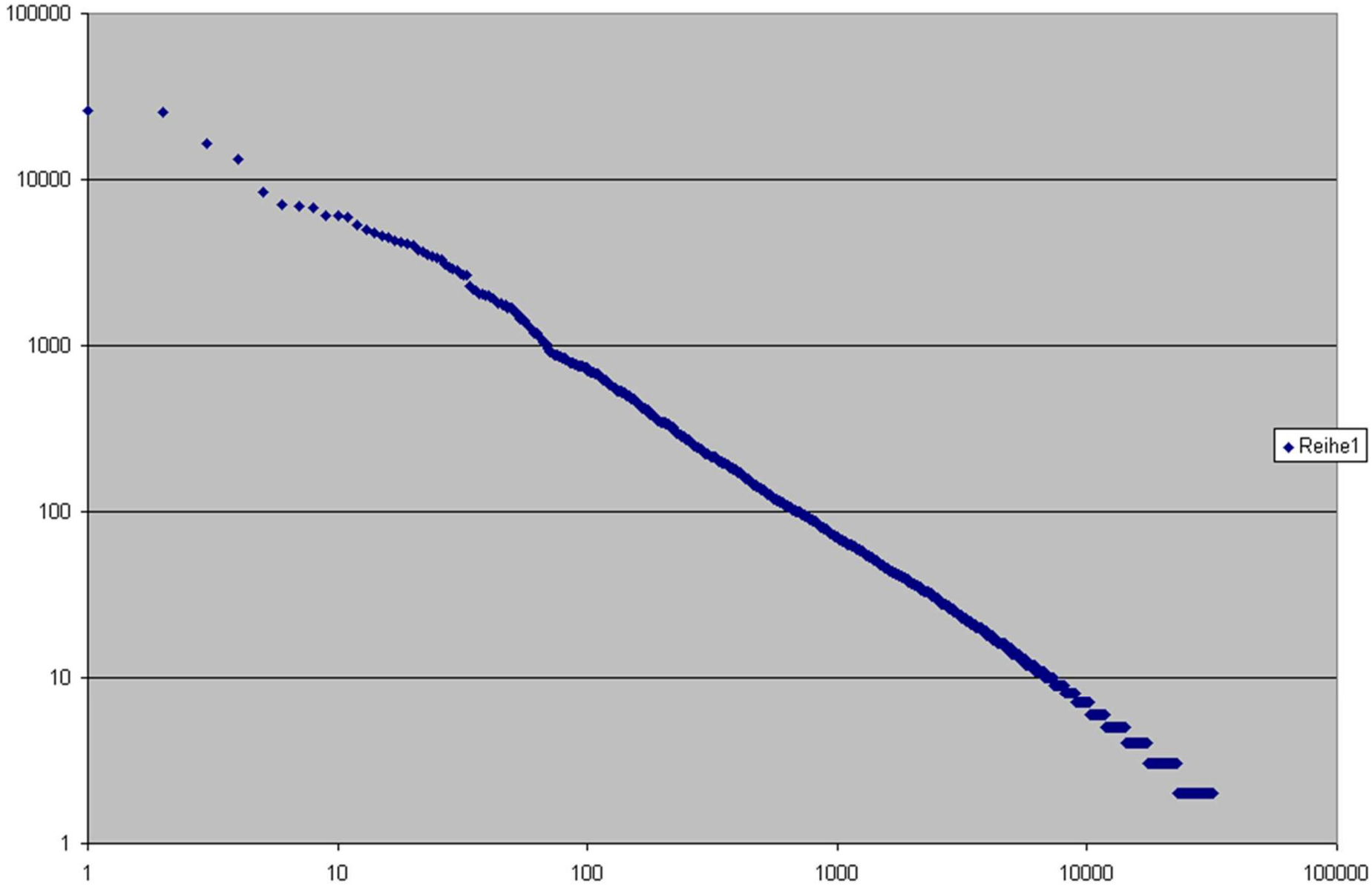
Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Logarithmischer Plot

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw



Worthäufigkeiten

26052 der	5275 des	3489 an
25633 die	4979 sich	3460 es
16571 und	4774 auf	3341 bis
13353 in	4552 dem	3274 als
8472 den	4473 nicht	3063 nach
7118 das	4257 ein	2957 sie
6921 von	4222 ist	2891 aus
6719 im	4096 uhr	2810 werden
6104 für	3984 eine	2689 daß
6015 mit	3763 am	2670 er
5990 zu	3673 auch	2630 bei

Anfrage und Dokument vergleichen

Krebse bekämpfen **mit** Gift

empoisonner **des** écrevisses

Da der Rote Sumpfkrebs **mit** Raubfischen bekämpft werden kann, ist diese Massnahme dem Gifteinsatz gegen Sumpfkrebse vorzuziehen.

*L'écrevisse rouge **des** marais pouvant être combattue par l'introduction de poissons prédateurs, il y a lieu de substituer cette mesure à l'empoisonnement projeté.*

- Anfrage und Dokument können nicht direkt verglichen werden. Anfrage und Dokument müssen zuerst vergleichbar gemacht werden

  (➔ Erschliessen = Indexing).

Definition Indexierung

- Zwei oft verwendete Begriffe sind Volltextindexierung und Verschlagwortung
- Wir betrachten zuerst die Verschlagwortung (traditionelles "Retrievalwerkzeug")
- Der Inhalt wird durch eine Menge von Schlagwörtern oder Deskriptoren (engl. "key words", "subjects" oder "descriptors") beschrieben

Manuelle und automatische Verschlagwortung

- Verschlagwortung geschieht
  - manuell
  - automatisch
  - in einer Kombination der beiden Prozesse ("computerunterstützt")
- Die Deskriptoren stammen aus
  - einem unlimitierten Vokabular (freie Indexierung) oder
  - einer autorisierten Liste von Deskriptoren (kontrollierte Indexierung)
- Beachten Sie jedoch, dass
  - die Qualität der Indexierung bestimmt massgebend die Qualität des Information Retrievals. D.h. selbst der beste Retrieval-Algorithmus hilft bei einem schlampigen Index nicht weiter.
  - die manuelle Indexierung sehr aufwendig und kostspielig ist
  - Dokumentenvokabular <-> Anfragevokabular!

Automatische Verschlagwortung

- Ein anschaulicher Weg, um einen Textinhalt automatisch zu beschreiben, ist die Häufigkeit der einzelnen Wörter zu untersuchen.
- Beispiel:

  A Bibliographic Search by Computer

  – Updating plasma-physics data was a chance to experiment with information and programs of the Technical Information Project at MIT. The computer searched for indicative words in titles of papers that shared bibliographic references and those that referred to papers that have become classics in plasma-physics.
- Die drei häufigsten Worte sind: plasma-physics, information, papers.
- Der Titel jedoch deutet folgendes an: bibliographic, search, computer
- → Folglich ist die automatische Indexierung nicht so einfach wie man glauben könnte.

Volltextindexierung

- Wir betrachten nun die Volltextindexierung
  - Sozusagen ein Sonderfall von automatischer Verschlagwortung mit freiem Vokabular
  - Historisch als grundsätzlich unterlegen betrachtet (→ Kommentar?)

Erschliessen

- Da der Rote Sumpfkrebs mit Raubfischen bekämpft werden kann, ist diese Massnahme dem Gifteinsatz gegen Sumpfkrebse vorzuziehen.
- *L'écrevisse rouge des marais pouvant être combattue par l'introduction de poissons prédateurs, il y a lieu de substituer cette mesure à l'empoisonnement projeté.*

Buchstabenumwandlung
Wortextraktion
Stoppwortelimination

- rote sumpfkrebs raubfischen bekaempft massnahme gifteinsatz sumpfkrebse vorzuziehen
- *écrevisse rouge marais pouvant combattue introduction poissons prédateurs lieu substituer mesure empoisonnement projeté*

Tokenisierung

- Die Tokenisierung (auch: Wortextraktion) ist der Prozess, der die einzelnen Wörter aus dem Text extrahiert

- Im Allgemeinen werden einige oder alle der folgenden Schritte ausgeführt:
  1. Dokumentenformate konvertieren
  2. Zeichencodierung anpassen (ISO-8859-1, Unicode, ..)
  3. Gross-/Kleinschreibung normalisieren
  4. Text entlang Trennzeichen in Tokens separieren (eigentliches Tokenisieren)

Tokenisierung

- Zeichencodierung: Umlaute werden nicht konsistent geschrieben
  - Im Deutschen, falls die Tastatur die entsprechenden Zeichen nicht bietet
  - Im Französischen bei Grossschreibung, …
- Gross-/Kleinschreibung
  - Grossschreibung am Satzanfang
  - «der gefangene Floh» vs. «der Gefangene floh»
  - Englisch: Namen vs. andere Nomina («Bush»/«bush»)
- Eigentliche Tokenisierung
  - Definition des «Worts» ist nicht ganz einfach; wir sprechen daher von Tokens: grob gesagt, das Auftreten eines Wortes als Zeichenkette, unabhängig einer Definition
  - Trennzeichen haben nicht immer die selbe Funktion: Satzende vs. Dezimalpunkt, Komma vs. Tausendertrenner, Bindestrich vs. Trennstrich

- Diese Phänomene werden weitgehend toleriert und sind in der durchschnittlichen Performance schlecht messbar, können aber in Einzelfällen wichtig sein

Stoppwortelimination

- Die häufigsten Wörter (siehe Folie 13) werden aus dem Text eliminiert
- Diese Wörter sind nicht inhaltstragend
- Treffer auf diesen Wörtern sind (meist!) nicht hilfreich, und "verdecken" echte, wertvolle Treffer
- Da die Stoppwörter etwa 40% des Textes ausmachen, lässt sich auch Speicherplatz sparen

Frage:

Diskutieren Sie Vor- und Nachteile der Stopwortelimination

Erschliessen (cont.)

- Rote Sumpfkrebs Raubfischen bekaempft Massnahme Gifteinsatz Sumpfkrebse vorzuziehen
- écrevisse rouge marais pouvant combattue introduction poissons prédateurs lieu substituer mesure empoisonnement projeté

Wortzerlegung
Wortnormalisierung

- rot sumpf krebs raub fisch kaempf massnehm gift einsetz sumpf krebs vorzieh
- écrevisse rouge marais pouvant combattue introduction poisson prédateur lieu substituer mesure empoisonner proje

Stemming/Wortnormalisierung

- In Sprachen werden mehr oder weniger Wortformen verwendet, um Dinge wie Kasus, Numerus, Genus etc. anzuzeigen
- Da wir ja auf Dokumenten von potentiell beliebigen Autoren suchen, ist die Verwendung dieser Wortformen im Prinzip frei.
- Dieses Phänomen ist je nach Sprache unterschiedlich stark ausgeprägt.
  - Im Englischen gibt es relativ wenig verschiedene Formen für die meisten Wörter
  - Im Deutschen sieht das schon anders aus (bis zu 144 Formen für ein Verb!)
  - noch stärker ausgeprägt im Finnischen und anderen Sprachen...

Indikativ	
Präsens	*Perfekt*
ich gehe	ich bin gegangen
du gehst	du bist gegangen
er/sie/es geht	er/sie/es ist gegangen
wir gehen	wir sind gegangen
ihr geht	ihr seid gegangen
sie gehen	sie sind gegangen
Präteritum	*Plusquamperfekt*
ich ging	ich war gegangen
du gingst	du warst gegangen
er/sie/es ging	er/sie/es war gegangen
wir gingen	wir waren gegangen
ihr gingt	ihr wart gegangen
sie gingen	sie waren gegangen
Futur I	*Futur II*
ich werde gehen	ich werde gegangen sein
du wirst gehen	du wirst gegangen sein
er/sie/es wird gehen	er/sie/es wird gegangen sein
wir werden gehen	wir werden gegangen sein
ihr werdet gehen	ihr werdet gegangen sein
sie werden gehen	sie werden gegangen sein

Quelle: dict.leo.org

Stemming/Wortnormalisierung

- Information Retrieval ist **kein linguistischer Schönheitswettbewerb**
- Wir wollen Dokumente auffinden unabhängig von einzelnen Wortformen
- Aber: die Dokumente sollten relevant sein, d.h., der gesuchte Sachverhalt sollte "abgebildet" werden
- «Light Stemmer»: entfernt Inflektionsendungen (Zahl, Geschlecht, Zeit)
- Voll ausgebauter Stemmer: entfernt auch Derivationsendungen (ändert die Wortart)
- Aber Achtung: Wörter können linguistisch verwandt sein, aber für die Suche wäre ein Treffer unnütz: Bildung <-> Bild

Stemming/Wortnormalisierung

- Wenn möglich werden im Information Retrieval **einfache, regelbasierte Verfahren** verwendet, um die Wörter zu normalisieren
- Englisch: Porter-Stemmer, Wirkung umstritten
  - «Some form of stemming is almost always beneficial. [..] The average absolute improvement due to stemming is small, ranging from 1-3%» (Hull, 1996)
- Deutsch: verschiedene Möglichkeiten, klare Wirkung
  - «Stemming is useful for German text retrieval in most cases. […] we obtained performance gains measured in mean average precision of up to 23% […] (Braschler & Ripplinger, 2004)
- Für manche Sprachen kaum möglich (zu viele Formen, zu irregulär)

Porter-Stemmer

- Auszüge aus dem Regel-Set (Porter, M.: An Algorithm for Suffix Stripping, 1980), je nach Implementation ca. 60 Regeln

Step 1a

```
SSES -> SS                      caresses  ->  caress
IES  -> I                       ponies    ->  poni
                                ties      ->  ti
SS   -> SS                      caress    ->  caress
S    ->                         cats      ->  cat
```

Step 1b

```
(m>0) EED -> EE                 feed      ->  feed
                                agreed    ->  agree
(*v*) ED  ->                    plastered ->  plaster
                                bled      ->  bled
(*v*) ING ->                    motoring  ->  motor
                                sing      ->  sing
```

Step 4

```
(m>1) AL    ->                  revival        ->  reviv
(m>1) ANCE  ->                  allowance      ->  allow
(m>1) ENCE  ->                  inference      ->  infer
(m>1) ER    ->                  airliner       ->  airlin
(m>1) IC    ->                  gyroscopic     ->  gyroscop
(m>1) ABLE  ->                  adjustable     ->  adjust
(m>1) IBLE  ->                  defensible     ->  defens
```

Führt durchaus zu Fehlern!

Bsp.: organize/organ

Beispiel Porter-Stemmer

- **Original text:**
  - Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales
- **Porter Stemmer:**
  - market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

(Quelle: J. Savoy)

Diskussion Stemming

- Diese "Stemmer" produzieren teilweise unsinnige (abgeschnittene) Wortformen («police» → «polic»)
- Problem «understemming»: zwei Wörter/Wortformen werden nicht auf einen gemeinsamen Stamm zurückgeführt, obwohl für das Matching wünschenswert («woman»/«women»)
- Problem «overstemming»: zwei Wörter/Wortformen werden auf einen gemeinsamen Stamm reduziert, obwohl für das Matching nicht wünschenwert («executive»/«execute»)

- → Viele dieser Probleme sind eine Frage der Gewichtung, resp. sind transparent für den/die Anwender/in

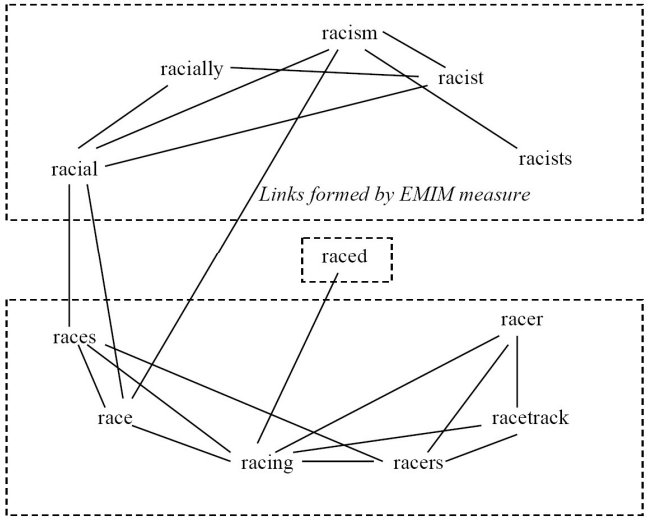
Alternativen zu Stemming

- Lemmatisierung: «korrekte» linguistische Analyse der Wortformen.
  - Aber Achtung: dies löst nicht per se das «under-/overstemming»-Problem!
  - Interessant, wenn die Stems angezeigt oder für weitere Schritte verarbeitet werden

- Corpus-based Stemming: zum Beispiel «Co-occurrence» Analyse
  - Fehlerhafte/unsinnige Reduktionen
  - Angepasst an einen spezifischen Korpus, aufwändig

- N-Gram (→ siehe später in diesem Kapitel)

- Semantische Lösungen, Expansionen, …



Quelle: J. Savoy

Komposita

- Im Deutschen (und in anderen Sprachen, z.B. Schwedisch) können (unendlich viele) Komposita gebaut werden, indem man Wörter zusammenfügt
- Diese Komposita können nicht lexikalisch aufgezählt werden
- Die Komposita sind eigentlich sehr gute Suchbegriffe, aber:
- Der gleiche Sachverhalt lässt sich immer auch als Nominalphrase umschreiben (z.B. mit den Einzelbegriffen)

Krasses, aber reales, Beispiel

"Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz" (aber: siehe auch die phrasale Umschreibung!) (Gesetz aus Mecklenburg-Vorpommern)

Auf Englisch könnte man etwa sagen: "beef labelling surveillance task transfer law".

Komposita

- Es ist sehr hilfreich, solche Komposita zu zerlegen (weitere ca. +30% Effektivität)
  - «decompounding contributes more to peformance improvement than stemming» (Braschler & Ripplinger, 2004)
- Aber: Wettbewerb, Frühstück → nicht alle Komposita sollten getrennt werden
- Wie trennen? Fussballweltmeisterschaft → Fussball|Weltmeisterschaft oder Fuss|Ball|Welt|Meisterschaft?
- → wenige Systeme bieten eine solche Zerlegung!

Frage:

Wofür (= welches Ziel wird unterstützt) sind Stemming/Kompositazerlegung letztlich nützlich? Was sind die Nachteile?

Warum sind Websuchmaschinen in dieser Hinsicht oft eher konservativ?

Anfrage

Krebse bekämpfen mit Gift

empoisonner des écrevisses

Dokumente

- Da der Rote Sumpfkrebs mit Raubfischen bekämpft werden kann, ist diese Massnahme dem Gifteinsatz gegen Sumpfkrebse vorzuziehen.
- *L'écrevisse rouge des marais pouvant être combattue par l'introduction de poissons prédateurs, il y a lieu de substituer cette mesure à l'empoisonnement projeté.*

Indexierung

krebs kaempf gift

empoisonner écrevisse

- rot sumpf **krebs** raub fisch **kaempf** massnehm **gift** einsetz sumpf **krebs** vorzieh
- ***écrevisse*** *rouge marais pouvant combattue introduction poisson prédateur lieu substituer mesure* ***empoisonner*** *proje*

"Bag of words"

- Das Dokument wird als ungeordneter Sack von Wörtern betrachtet
- Dominantes Information Retrieval-Prinzip

Ein Beispiel des Outputs des Indexierungsprozesses

5 x canada

4 x august

3 x GDP, industry, output, service, statistic

2 x decline, good, index, price, product, report, rose, said, september

1 x adjust, agency, consecutive, domestic, dropp, earlier, federal, grew, gross, growth, increas, industrial, inflation, july, level, mainly, material, monthly, nation, producing, raw, result, separately, total, value

→ Für was steht dieser “Bag”? (Beispiel nach J. Savoy)

Das Originaldokument

<DOCNO> WSJ891101-0145 </DOCNO>
<HL> Canada's GDP Grew in August </HL>
<SO> WALL STREET JOURNAL (J) </SO>
<TEXT>
Canada's gross domestic product rose an inflation-adjusted 0.3% in August, mainly as a result of service-industry growth, Statistics Canada, a federal agency, said.

The August GDP was up 2.4% from its year-earlier level. GDP is the total value of a nation's output of goods and services.

Statistics Canada said service-industry output in August rose 0.4% from July. Output of goods-producing industries increased 0.1%.

Separately, Statistics Canada reported that its industrial-product price index dropped 0.2% in September, its third consecutive monthly decline.
It also reported a 2.6% decline in its raw-materials price index for September.

</TEXT>
</DOC>

Definition des "Terms"/"Merkmals"

- Es wird bei den folgenden Gewichtungsverfahren häufig von Term und Merkmal gesprochen.

- Gebräuchlicherweise: Term = eindeutiges Wort
- Token = Auftreten eines Terms
- Merkmal = Term oder andere eindeutige Einheiten (Phoneme, Linienkanten, Histogramme etc.) (Verallgemeinerung)

- Die Granularität der Merkmale/Terme kann variieren: Wortteile, Wörter, Phrasen etc.

Indexierung unterschiedlicher Granularität

- Informationsspuren (z.B n-grams)
  - Im Falle, dass einzelne Terme als Indexierungselemente ungeeignet sind (Texte mit grammatikalischen oder typografischen Fehlern, unbekannte Sprache des Textes etc.) müssen kleinere Elemente für die Indexierung herangezogen werden.
- Ein n-Gram ist eine Sequenz von n (3, 4, ...) Buchstaben
- Information

  _In

  Inf

  nfo

  for

  orm

  rma

  mat

  ati

  tio

  ion

  on_

N-Gram Indexierung

- Anwendungen
  - Sprachunabhängige Algorithmen (keine Normalisierung nötig)
  - Erkennung der Sprache eines Dokuments
  - (Spellchecking)
  - Fehlertoleranter Vergleich (OCR-Text)
  - Vergleich der Ähnlichkeit ganzer Dokumente

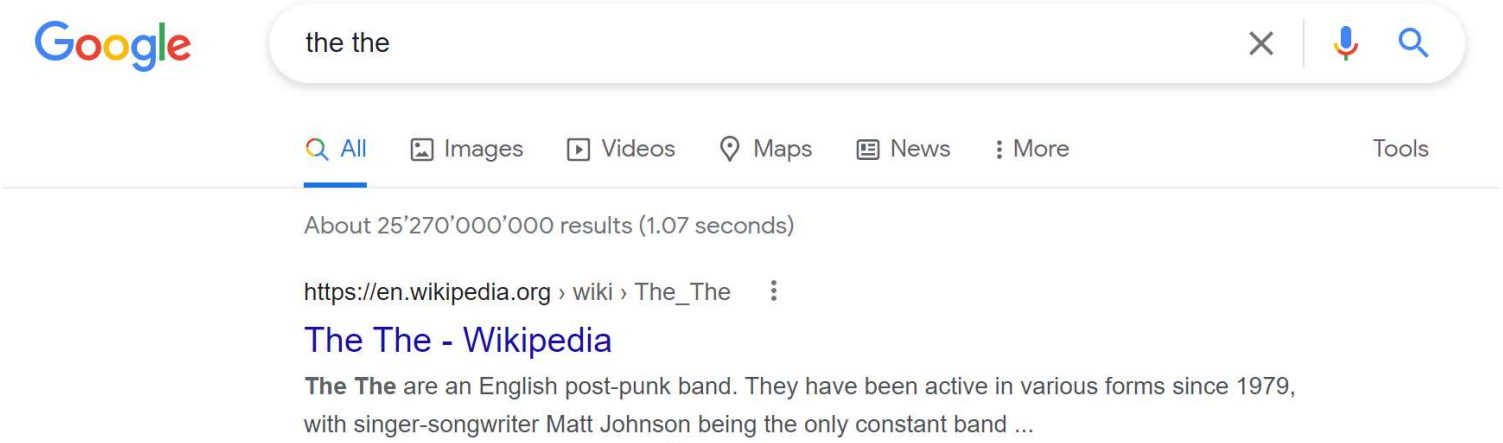
- Effektivität
  - «language-neutral methods can achieve accuracy comparable to language-specific methods» (McNamee & Mayfield, 2003)

Diskussion N-Gram

- Relativ ineffizient, Indexgrösse wird aufgeblasen
- Viele Matches auf einzelnen N-Grammen («Information» <-> «gratis»)
  → Gewichtung wichtig
- Sehr interessant für Sprachen mit wenigen Sprechern (und deshalb tendenziell wenig verfügbaren Sprachressourcen)
- Kann die Rolle des Decompounding ersetzen
  - «languages with greater mean word length fared relatively better with n-grams than with words» (McNamee & Mayfield, 2003)
- N muss bestimmt werden. Ideales N ist oft im Bereich {4,5}

Definition "Wort"

- Worte im Information Retrieval sind typischerweise Einheiten zwischen zwei Trennzeichen bei der Tokenisierung
- Probleme:
  - Nicht jede Sprache verwendet Trennzeichen zur Darstellung von Worten (Chinesisch, Japanisch, ...)
  - Was ist mit Begriffen wie "F/A-18", "Coca-Cola"?
  - Sind Phrasen evtl. bessere IR-Einheiten?
  - → Phrasenerkennung (hier nicht weiter behandelt, hilft u.a. bei manchen Problemen der Stopwortelimination)



Weitere Probleme: Name Matching

- Die folgenden Formen kommen in der Library of Congress (LOC) vor: (Auszug)

Al Qathafi, Mu'ammar Al Qathafi, Muammar El Gaddafi, Moamar
El Kadhafi, Moammar El Kazzafi, Moamer El Qathafi, Mu'Ammar
Gadafi, Muammar Gaddafi, Moamar Gadhafi, Mo'ammar
Gathafi, Muammar Ghadafi, Muammar Ghaddafi, Muammar
Ghaddafy, Muammar Gheddafi, Muammar Gheddafi, Muhammar
Kadaffi, Momar Kad'afi, Mu`amar al- Kaddafi, Muamar
Kaddafi, Muammar Kadhafi, Moammar Kadhafi, Mouammar
Kazzafi, Moammar Khadafy, Moammar Khaddafi, Muammar
Moamar al-Gaddafi Moamar el Gaddafi Moamar El Kadhafi
Moamar Gaddafi Moamer El Kazzafi Mo'ammar el-Gadhafi
Moammar El Kadhafi Mo'ammar Gadhafi Moammar Kadhafi
Moammar Khadafy Moammar Qudhafi Mu`amar al-Kad'afi
Mu'amar al-Kadafi Muamar Al-Kaddafi Muamar Kaddafi
Muamer Gadafi Muammar Al-Gathafi Muammar al-Khaddafi

Name Matching

- Probleme:
  - Verschiedene Transkriptionssysteme für fremde Schreibsysteme
  - Unterschiedliche, optionale Namensbestandteile
  - Lexikalisch nicht abschliessend aufzählbar

- Lösungsansätze
  - «Spelling correction»: Edit distance etc.
  - Phonetisches Matching
  - N-Gramme
  - Namenslisten (aber: konstant neue, wichtige Personen!)