Textdatenbanken – Übung 2

Segmentierung

Segmentierung

- In welche Einheiten wird zerlegt?

Segmentierung

- Welche Gründe gibt es für die Segmentierung?

Satzsegmentierung

- Regelbasiert
  - Welche Regeln sind sinnvoll?
  - Welche Probleme verursachen sie?

Satzsegmentierung

- Die Stadt liegt inmitten eines ehemaligen Binnendeltas, das durch die Anlage von Mühlgräben und Hochwasserschutzanlagen häufig umgestaltet wurde. So ist die Parthe ehemals ein Nebenfluss der Pleiße gewesen, während sie heute in die Weiße Elster mündet. In den 1950er Jahren wurden der Pleißemühlgraben und ein Teil des Elstermühlgrabens – im Mittelalter für den Betrieb von Mühlen teilweise künstlich angelegte Nebenarme der beiden Flüsse – wegen der Verschmutzung durch Industrieabwässer aus der Braunkohleverarbeitung südlich von Leipzig verrohrt oder verfüllt, so dass Leipzig seinen Charakter als Flussstadt teilweise verlor.
- Welche Regeln/Ressourcen sind nötig?

de.wikipedia.org/wiki/Leipzig

Satzsegmentierung

¿Beiñere euskera irakurri ez duanak lenengo aldiz asten denean parraparra irakurtzekoa?

加加林所乘坐的太空船在 108 分鐘内環繞地球一次。地面的控制人員，需要在太空船發射後的 25 分鐘，才能確認太空船已進入穩定軌道。回程的時候，在距離地面 7 公里的地方，東方一號會將他彈射出來，而太空船和太空人會有各自的降落傘。之所以要分開降落，原因是太空船的降落傘會為太空人帶來危險。最後，他安全返回地面。

Satzendezeichen - Beispiel

Latin

.

!

?

Arabic

-

؟

Chinese, Japanese and Korean

。

！

？

Devanagari

।

Armenian

՝

։

Ethiopian

።

፧

Satzsegmentierung - Thai

ที่เกิดเหตุพบศพชายอยู่ข้างศาลาพักผู้โดยสารทางเข้าอ่างเก็บน้ำชลประทานห้วยแกง สภาพสวมเสื้อเชิ้ตสีฟ้า นุ่งกางเกงยีนส์ขายาวสีน้ำเงิน ทราบชื่อคือ จ.ส.ต.สถาปัตย์ สุตนนท์ อายุ 40 ปี ตำรวจสังกัดงานป้องกันปราบปรามสภ.สมเด็จ จ.กาฬสินธุ์ ซึ่งย้ายมาช่วยราชการงานป้องกันและปราบ

Satzsegmentierung

- In Folge zogen 1409 etwa 1000 der dortigen deutschen Lehrkräfte und Studenten nach Leipzig, wo die Artistenfakultät den Lehrbetrieb aufnahm. Dieser wurde sofort von der Stadt ein Gebäude in der Petersstraße übereignet. Die Landesherren, Friedrich der Streitbare und Wilhelm der Reiche, bewilligten der Universität anfangs einen Jahresetat von 500 Gulden und stifteten zwei Kollegien, das große und das kleine Fürstenkolleg, für die zwei abgabefreie Häuser in der Ritterstraße bereitgestellt wurden. Noch 1409 wurde das „Studium generale“ durch Papst Alexander V. bestätigt. Am 2. Dez. 1409 werden Johannes Otto von Münsterberg zum Rektor gewählt und die Universitätssatzung verlesen.

  Anne Wills sonntagabendliche Runde zu den Plagiatsvorwürfen gegen Dr. Karl- Theodor zu Guttenberg im Rahmen seiner Dissertation offenbarte worin das wirkliche Problem liegt. Es handelt sich dabei nicht um die bloße Tatsache, dass Karl-Theodor zu Guttenberg sich weitläufig, ohne dies kenntlich zu machen, beim gedanklichen Gut anderer Autoren bedient hat, sondern viel mehr um den Umgang der Öffentlichkeit mit eben diesem Faktum.

- Welche Regeln/Ressourcen sind hier nötig?

Abkürzungsliste - Default

Dan.
dat.
Dec.
def.
Del.
Dem.
dept.
Dept.
Det.
Deut.
dia.
Dial.
Dict.
dim.
Diosc.
Disp.
Dist.
Disus.
Div.
Dom.
Dr.
Drs.
dud.
Dyn.
ebb.
Eccl.
Ecclus.
Eclec.
Econ.
ed.
Ed.
eff.
Egypt.
Elect.
elk.
ell.
em.
emph.
Encyc.
Eng.
Engin.

Abkürzungsliste - Deutsch

Adr.
ADSp.
adv.
Adv.
Advermg.
AdW.
Aecvd.
Aedb.
Aede.
Aen.
aero.
aeron.
Aeron.
Af.
AfA.
Afaik.
afghan.
AfNS.
afr.
Afr.
afrik.
afrikan.
Afu.
AFuG.
AfV.
AfW.
Afwe.
ag.
Ag.
Agaf.
agb.
AGDir.
Age.
Ägebo.
Agfa.
agg.
Aggr.
Agi.
Agitprop.
Agjj.
Agm.
Agmm.

Abkürzungsliste – Deutsch Wikipedia

A.
a.
Abb.
Abf.
Abg.
abg.
Abh.
Abk.
abk.
Abs.
Abw.
Adir.
Adj.
adj.
Adr.
Adv.
adv.
Afr.
Ag.
agg.
Aggr.
Ahg.
Akad.
akad.
Akk.
Alg.
allg.
alph.
altgr.
Am.
Amp.
amtl.
Amtsbl.
An.
anat.

Satzsegmentierung

- Es geht immer komplizierter:
  - Autor Gerhard Raff erinnert an den Lehrer Ludwig Amandus Bauer, der vor... vor wie vielen Jahren gestorben war?
  - Mir schien das unlogisch, aber dann dachte ich genauer nach... Datenbanken sind hier völlig unbrauchbar.
  - "Sind die Kuweiter bereit, alle anderen Muslime als ihresgleichen in Kuweit zu akzeptieren?" fragte Dajin und beantwortete sich seine Frage selbst: "Ich bezweifle es.“.
  - Die Christlichen Demokraten heißen jetzt Volkspartei; aber ist damit die Korruption ausgeschwitzt? fragt sich Herr Rossi.

- Lohnt der Aufwand?

Satzsegmentierung / Putzen

- Neujahrskonzert in der Mercatorhalle - Duisburg - DerWesten. Klassik : Mercatorhalle Duisburg im CityPalais, König-Heinrich-Platz, 0203 393060, 18-20 Uhr: "Das Neujahrskonzert 2009". Reit- und Fahrverein Ziethen e.V., Leutfeldstraße 18, Reithalle, 02151 409589, 11 Uhr: "Sprung ins Neue Jahr". LEGOLAND Discovery Centre Duisburg, Philosophenweg 23-25, 0203 570888. Wilhelm Lehmbruck Museum, Friedrich-Wilhelm-Straße 40, Ausstellungswerkstatt, 0203 2832630 / 3294, 11 Uhr: "Karin Hochstatter".
- Copyright © 1999-2007 Tripsoft, © 2007-2009 WebHouse, s.r.o., Trnava, Slovak Republic.

  - Lösung?

Problemfall: Zahlen

„Die partielle Sonnenfinsternis war am 4. Januar. Die Bedeckung der Sonne war mit 86% bei Skellefteå in Schweden am größten.“

„Am 3. habe ich einen Termin - beim Prof. Ich bin gespannt was er sagt.“

„Am 20. Dezember 1409 wurde Johannes Otto von Münsterberg zum Rektor gewählt.“

„An jeder 100. Autobahnabfahrt steht ein solches Schild.“

Lösung?

Problemfall: Zahlen

Die partielle Sonnenfinsternis war am 4. Januar. Die Bedeckung der Sonne war mit 86% bei Skellefteå in Schweden am größten.

Aber:

Die partielle Sonnenfinsternis war am 4. Januar 2011. Die Bedeckung der Sonne war mit 86% bei Skellefteå in Schweden am größten.

Lösung?

Problemfall: Zahlen

Erneutes aber:

Er belegte den 1050. Platz beim New-York-City-Marathon.

Dessau liegt an der Autobahn 9.

Satzsegmentierung

- Weitere hilfreiche Features?

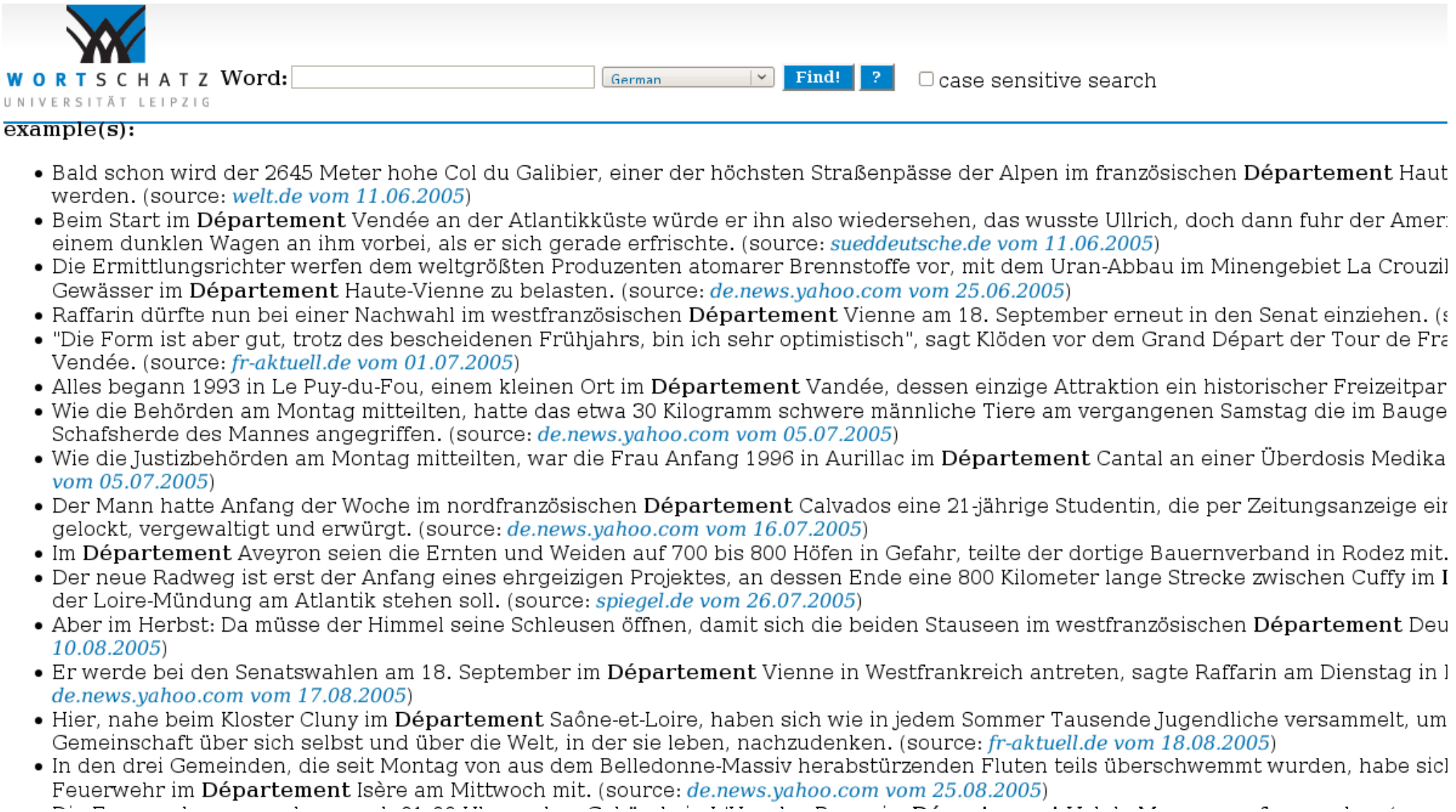
Satzsegmentierung

- Satzlänge:

Satzsegmentierung

- Typische Satzanfänge bzw. -enden:

Satzsegmentierung



WORTSCHATZ UNIVERSITÄT LEIPZIG Word: German Find! ? case sensitive search

example(s):

- Bald schon wird der 2645 Meter hohe Col du Galibier, einer der höchsten Straßenpässe der Alpen im französischen **Département** Haut werden. (source: *welt.de vom 11.06.2005*)
- Beim Start im **Département** Vendée an der Atlantikküste würde er ihn also wiedersehen, das wusste Ullrich, doch dann fuhr der Amer einem dunklen Wagen an ihm vorbei, als er sich gerade erfrischte. (source: *sueddeutsche.de vom 11.06.2005*)
- Die Ermittlungsrichter werfen dem weltgrößten Produzenten atomarer Brennstoffe vor, mit dem Uran-Abbau im Minengebiet La Crouzil Gewässer im **Département** Haute-Vienne zu belasten. (source: *de.news.yahoo.com vom 25.06.2005*)
- Raffarin dürfte nun bei einer Nachwahl im westfranzösischen **Département** Vienne am 18. September erneut in den Senat einziehen. (s
- "Die Form ist aber gut, trotz des bescheidenen Frühjahrs, bin ich sehr optimistisch", sagt Klöden vor dem Grand Départ der Tour de Fra Vendée. (source: *fr-aktuell.de vom 01.07.2005*)
- Alles begann 1993 in Le Puy-du-Fou, einem kleinen Ort im **Département** Vandée, dessen einzige Attraktion ein historischer Freizeitpar
- Wie die Behörden am Montag mitteilten, hatte das etwa 30 Kilogramm schwere männliche Tiere am vergangenen Samstag die im Bauge Schafsherde des Mannes angegriffen. (source: *de.news.yahoo.com vom 05.07.2005*)
- Wie die Justizbehörden am Montag mitteilten, war die Frau Anfang 1996 in Aurillac im **Département** Cantal an einer Überdosis Medika *vom 05.07.2005*)
- Der Mann hatte Anfang der Woche im nordfranzösischen **Département** Calvados eine 21-jährige Studentin, die per Zeitungsanzeige ei gelockt, vergewaltigt und erwürgt. (source: *de.news.yahoo.com vom 16.07.2005*)
- Im **Département** Aveyron seien die Ernten und Weiden auf 700 bis 800 Höfen in Gefahr, teilte der dortige Bauernverband in Rodez mit.
- Der neue Radweg ist erst der Anfang eines ehrgeizigen Projektes, an dessen Ende eine 800 Kilometer lange Strecke zwischen Cuffy im **I** der Loire-Mündung am Atlantik stehen soll. (source: *spiegel.de vom 26.07.2005*)
- Aber im Herbst: Da müsse der Himmel seine Schleusen öffnen, damit sich die beiden Stauseen im westfranzösischen **Département** Deu *10.08.2005*)
- Er werde bei den Senatswahlen am 18. September im **Département** Vienne in Westfrankreich antreten, sagte Raffarin am Dienstag in *de.news.yahoo.com vom 17.08.2005*)
- Hier, nahe beim Kloster Cluny im **Département** Saône-et-Loire, haben sich wie in jedem Sommer Tausende Jugendliche versammelt, um Gemeinschaft über sich selbst und über die Welt, in der sie leben, nachzudenken. (source: *fr-aktuell.de vom 18.08.2005*)
- In den drei Gemeinden, die seit Montag von aus dem Belledonne-Massiv herabstürzenden Fluten teils überschwemmt wurden, habe sich Feuerwehr im **Département** Isère am Mittwoch mit. (source: *de.news.yahoo.com vom 25.08.2005*)

corpora.informatik.uni-leipzig.de

Satzsegmentierung

- Vorkommen „Département“ in de07: 372 („Departement“: 708)
- Durchschnittliche Satzposition: 11,3
- Positionsverteilung:

Position	8	7	2	9	10
Frequenz	31	30	29	27	21

Satzsegmentierung

- **Der** Historiker Karl Schlögel singt ein Loblied auf einen neuen Menschen: den Migranten, den Nomaden, der "sich anschickt, d Sesshaftigkeit und des Ackerbaus antrainiert hat, abzustreifen". (source: *spiegel.de vom 01.01.2005*)
- "**Der** Schwung, mit dem sie in ihren Texten Literatur, Theater, Kunst oder Kino auseinander nahm und in den Kontext kulturell Ausnahmefigur der US-amerikanischen Kritik gemacht. (source: *spiegel.de vom 01.01.2005*)
- **Der** Theaterregisseur Andreas Kriegenburg hat mit Wilhelm Roth über seine Frankfurter Theaterfassung des Lars-von-Trier-R *01.01.2005*)
- **Der** Tod Susan Sontags wird auch in der FAZ groß aufgemacht. (source: *spiegel.de vom 01.01.2005*)
- **Der** nekrophilen Neigung dieser Zeitung entspricht man mit einer nochmaligen Verbeugung vor den Toten des Jahres. (source
- **Der** Gigant ist um das Automobil herum entstanden, ist nach ihm konzipiert. (source: *spiegel.de vom 01.01.2005*)
- **Der** für Personenverkehr zuständige Unternehmenssprecher Achim Stauß bestätigte am Dienstag die Meldung, gegenüber SP
- **Der** entsprechende Tarifantrag der Bahn muss noch genehmigt werden, Mitte Januar solle die Werbekampagne beginnen. (so
- Alle Informationen laufen in der Leitstelle zusammen: **Der** Standort und die Geschwindigkeit der Güterwaggons werden mit Hi Radaranlagen und Gewichtsmesseinrichtungen ermittelt. (source: *spiegel.de vom 01.01.2005*)
- **Der** Bürgermeister der argentinischen Hauptstadt, Aníbal Ibarra, sagte auf einer Pressekonferenz, 618 Menschen seien verlet
- **Der** Bundeskanzler setzt sich nach der verheerenden Flutkatastrophe in Südostasien für langfristige Hilfen für die betroffener engagieren. (source: *sueddeutsche.de vom 01.01.2005*)
- **Der** US-Jazzklarinettist, Big-Band-Leader und Komponist starb im Alter von 94 Jahren in Los Angeles. (source: *sueddeutsche.de*
- **Der** im Mai 1910 geborene Shaw wuchs in Connecticut auf und machte sich schon in frühen Jahren einen Namen als Saxophoni *01.01.2005*)

- Vorkommen „Der“ in de07: 3066062
- Durchschnittliche Satzposition: 1,9
- Positionsverteilung:

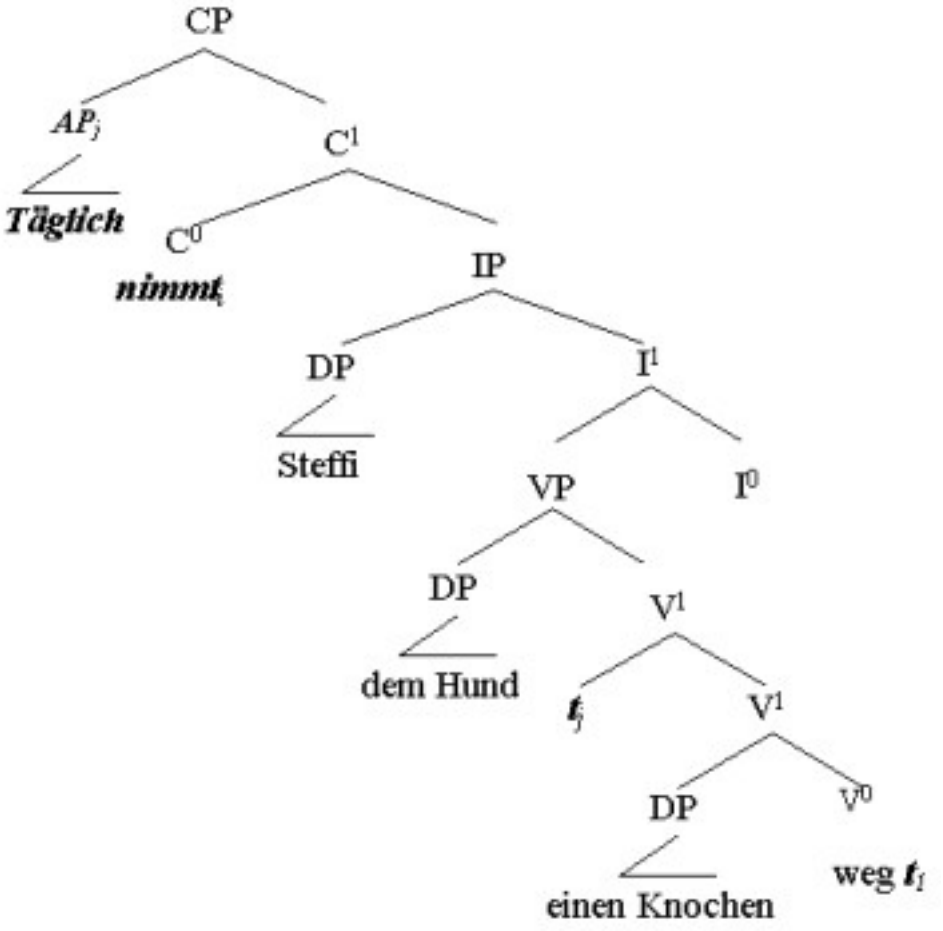
Position	1	2	3	4
Frequenz	2658945	93000	55709	27806

Satzsegmentierung

- Groß- und Kleinschreibung:
  - Siehe letzte Folie

Satzsegmentierung

- POS-Muster bzw. Syntaktische Struktur



Quelle: www.lingutrain.de

Segmentierung

Paretoprinzip

Segmentierung

Paretoprinzip

„Das Paretoprinzip, benannt nach Vilfredo Pareto (1848–1923), auch Pareto-Effekt, 80-zu-20-Regel, besagt, dass 80 % der Ergebnisse in 20 % der Gesamtzeit eines Projekts erreicht werden. Die verbleibenden 20 % der Ergebnisse benötigen 80 % der Gesamtzeit und verursachen die meiste Arbeit.“

http://de.wikipedia.org/wiki/Paretoprinzip

Tokenisierung

Tokenisierung

BYOT
Build Your Own Tokenizer

Tokenisierung – BYOT

Naiver Ansatz

Ich bin ein Berliner

Worttrenner

ASCII	Name
32	space
160	non-breaking space
8194	Leerzeichen; Breite n
8195	Leerzeichen; Breite m
8201	Leerzeichen schmal
8205	Null breiter Verbinder
…	

Tokenisierung - BYOT

Wir stehen gesellschaftlich auf der dritten Stufe.

Da werden wir zwar verprügelt, aber es gibt einen Grund dafür!

Zitat: Milhouse

Tokenisierung - BYOT

Seit 15 Jahren und 700 Sendungen kämpft Peter Escher im MDR-Ratgeber „Escher“ für die Rechte des kleinen Mannes.

In mittendrin spricht der „Robin Hood des Ostens“ über das Jubiläum und seine Tangoleidenschaft.

Alle weiteren Hefte im Verlag Dachauer Hefte, Dachau, Einzelpreis 22,- Mark; im Abonnement 19,80 Mark.

¿Cuál fue la mejor decada musical?

Tokenisierung

- „Fehler“ im Text:
  - Hier ein Überblick über die Trends bei der neunten Sportmesse "Golf Europe" im M,O,C Sports and Fashion Center in München-Freimann.
  - Hier ist die Luft prickelnd wie Champagner", wirbt die Kurverwal- tung.

Tokenisierung - BYOT

Finden Sie alle Daten, Nachrichten und Unternehmensmeldungen für International Business Machines (IBM) bei Yahoo!

Pläne für eine bloße Gesellschaft mit beschränkter Haftung mit dann 1300 Gesellschaftern seien verworfen worden.

MWU – Multi Word Units (Mehrworteinheiten)

- MWU in de07: ~ 800K
- Längenverteilung (Anzahl Wörter):

2	475818
3	160454
4	43607
5	23095
6	13198
7	6518
8	3240
9	1575
10	706
11	283
12	149
13	65

MWU

- Arten von MWU?
  -
  -
- Erstellung von MWU-Listen?
  -
  -
  -

MWU - Default

Accelerated Graphics Port
Accident of Birth
Ace Frehley
Acela Express
Ace of Base
Ace of Spades
Aces High
ACF Fiorentina
A Change of Seasons
Achillea millefolium
Achille Compagnoni
Achille Emana
Achille Lauro
Achim von Arnim
Achinoam Nini
AC Horsens
A Christmas Carol
Achtung Baby
Aci Bonaccorsi
Aci Castello
Aci Catena
Acid Drinkers
Acid Eaters
Acid house
Acid jazz
Aci Sant'Antonio
A Clockwork Orange
AC Milan
Acorn Archimedes
A Coruña
Acquanegra Cremonese
Acquanegra sul Chiese
Acquarica del Capo
Acquasanta Terme
Acquaviva Collecroce
Acquaviva delle Fonti
Acquaviva d'Isernia
Acquaviva Picena
Acquaviva Platani
Acqui Terme
Across the Universe
AC Siena
Acta Apostolicae Sedis
Action Comics
Action painting
Active Desktop
Active Directory
Active Server Pages

32k Einträge

MWU – Deutsche Wikipedia

Clyde Tombaugh
Cascading Style Sheets
Clara Zetkin
Carl von Linné
Charles Messier
Christian Morgenstern
Charles Darwin
Codex Manesse
Christiane von Goethe
Charles Lindbergh
Chinesische Sprachen
Carl Barks
Charles de Gaulle
Castel del Monte
Chemisches Element
Carl Friedrich Gauß
Claire Danes
Chemische Waffe
Chemische Energie
Cosimo de' Medici
Christiaan Huygens
Claus Schenk Graf von Stauffenberg

MWU - Bulgarisch

22 декември
23 декември
24 декември
25 декември
26 декември
27 декември
28 декември
29 декември
30 декември
31 декември
5 декември
Списък на писатели
Христо Ботев
Добри Войников
Софроний Врачански
Васил Друмев
Райко Жинзифов
Любен Каравелов
Константин Миладинов
Васил Попович
Григор Пърличев
Георги Раковски
Стефан Стамболов
Паисий Хилендарски
Добри Чинтулов
Константин Преславски
Мара Белчева
Димитър Бояджиев
Иван Вазов
Константин Величков
Димчо Дебелянов
Иван Йончев
Алеко Константинов
Официални празници в България
Стоян Михайловски
Велико Търново
Единен граждански номер
Алън Тюринг

Tokenisierung - BYOT

Die International Business Machines Corporation ist ein US-amerikanisches IT- und Beratungsunternehmen mit Sitz in Armonk bei North Castle im US-Bundesstaat New York.

Probleme?

Tokenisierung - Problemfälle

Vor- und Nachteile

IT- und Beratungsunternehmen

Lösungsvorschläge?

Tokenisierung

- Chinesisch

  „ 加加林所乘坐的太空船在 108 分鐘內環繞地球一次。地面的控制人員，需要在太空船發射後的 25 分鐘，才能確認太空船已進入穩定軌道。回程的時候，在距離地面 7 公里的地方，東方一號會將他彈射出來，而太空船和太空人會有各自的降落傘。之所以要分開降落，原因是太空船的降落傘會為太空人帶來危險。最後，他安全返回地面。 “

Tokenisierung

- „DerAtomunfallinJapan, steigendeStrompreise, dasvielerortskritisierteVerhaltenderStromanbietersowiederWunschnachUnabhängigkeitvonden MonopolistensinddiewesentlichenTriebfedern, diedasInteresseprivaterBauherrenanSolarstromanlagensprunghaftansteigenlassen.“
- Vorschläge?

Tokenisierung

- „deratomunfallinjapan, steigendestrompreise, dasvielerortskritisierteverhaltenderstromanbieters owiederwunschnachunabhängigkeitvondenmonopolistensinddiewesentlichentriebfedern, diedasinteresseprivaterbauherrenansolarstromanlagensprunghaftansteigenlassen.“
- Vorschläge?

Tokenisierung - Wörter

- Wie definieren WIR ein Wort?
  - 2 Ansätze

Tokenisierung - Wörter

- Wie definieren WIR ein Wort?
- 1. Ansatz: Die erlaubten Zeichenketten beschreiben
  - Buchstaben a-z, A-Z, ÄÖÜäöüß, Bindestrich
  - Ziffern? Weitere Sonderzeichen?
  - Nur „natürliche“ Reihenfolge oder auch Abweichungen? (pH-Wert)
- Problem: Es gibt immer Ausnahmen

- 2. Ansatz:
  - Wortgrenzen definieren
  - Nur Muster definieren, die ausgeschlossen werden