Semantische Suche auf einem Web-Korpus

Philipp Bausch

April 25, 2014

Übersicht

Broccoli

- Verwendet Wikipedia als Korpus der Textsuche

Broccoli

- Verwendet Wikipedia als Korpus der Textsuche
- Was findet man hier nicht?

Was Wikipedia ist

Was Wikipedia ist

Link

News-Seiten

- *triviale* Informationen

News-Seiten

- *triviale* Informationen
  - Verlauf von Ereignissen
  - Details von Ereignissen

News-Seiten

- *triviale* Informationen
  - Verlauf von Ereignissen
  - Details von Ereignissen
- Meinungen von Authoren
- Debatten

Übersicht

Datensätze

- ClueWeb12

Datensätze

- ClueWeb12
  - nicht Kostenlos
  - nur englische Texte
  - nur kleine Texte ($<$ 10MB)
  - 2.820.500 URLs

Datensätze

- ClueWeb12

- Commoncrawl Web Corpus 2012

Datensätze

- ClueWeb12

- Commoncrawl Web Corpus 2012
  - Kostenlos
  - 3.8 Milliarden Web Dokumente
  - mehr als 100TB Datenmenge
  - 61 Millionen Domains

Commoncrawl

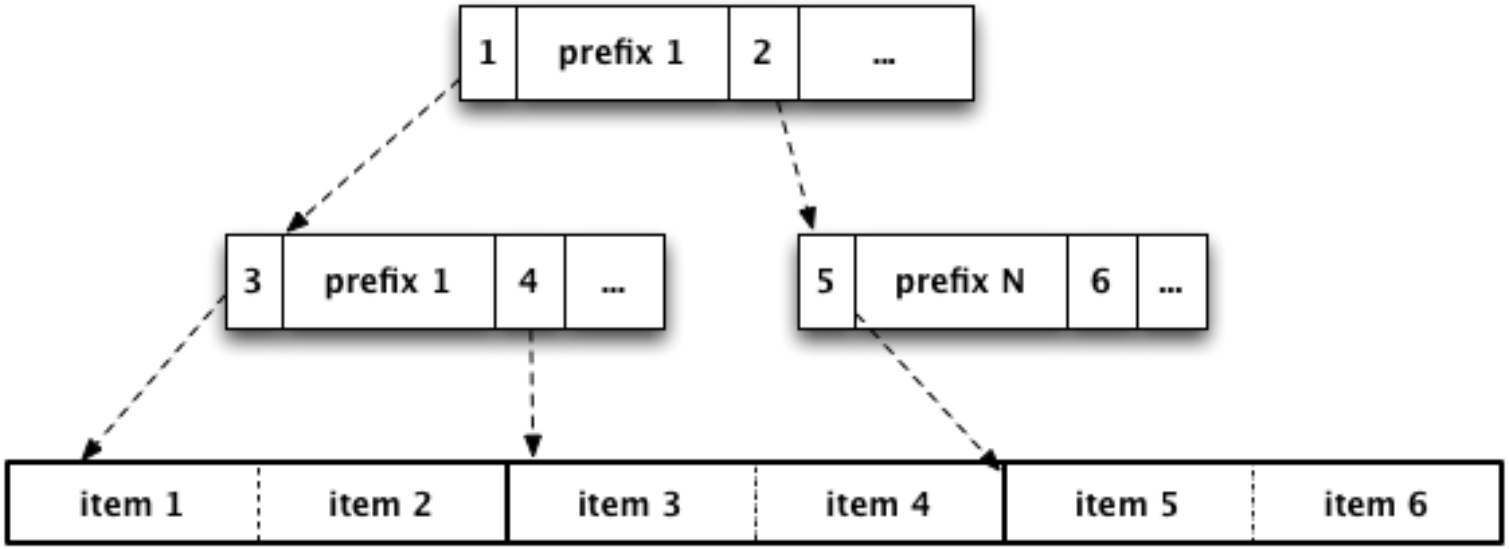
- Web Korpus von 2012 auf AmazonS3

Commoncrawl

- Web Korpus von 2012 auf AmazonS3
  - per EC2 Hadoop/HTTP abrufbar

Commoncrawl

- Web Korpus von 2012 auf AmazonS3
  - per EC2 Hadoop/HTTP abrufbar
  - URL-Index für ARC Dateien (Scott Robertson)



URL-Index Aufbau [IMG01]

Commoncrawl

- Web Korpus von 2012 auf AmazonS3
  - per EC2 Hadoop/HTTP abrufbar
  - URL-Index für ARC Dateien (Scott Robertson)
  - ARC Dateien enthalten vollen HTTP-Response

http://www.spiegel.de/international/0,1518,druck-345720,00.html 195.71.11.67 20120214055058 text/html 24686
HTTP/1.0 200 OK
Date:Tue, 14 Feb 2012 05:48:36 GMT
Server:Apache-Coyote/1.1
X-Powered-By:Servlet 2.4; JBoss-4.0.3SP1 (build: CVSTag=JBoss_4_0_3_SP1 date=200510231054)/Tomcat-5.5
Cache-Control:max-age=120
Expires:Tue, 14 Feb 2012 05:50:37 GMT
X-Host:lnxp-2885
X-Robots-Tag:noindex, nofollow, noarchive
Content-Type:text/html;charset=ISO-8859-1
Vary:Accept-Encoding
Content-Encoding:gzip
Content-Length:8341
X-Cache:MISS from lnxp-3958.srv.mediaways.net
X-Cache-Lookup:MISS from lnxp-3958.srv.mediaways.net:100
Via:1.1 www.spiegel.de, 1.0 lnxp-3958.srv.mediaways.net (squid/3.1.4)
Connection:close
x-commoncrawl-DetectedCharset:ISO-8859-1

Datenbeschaffung

Schritte:

1. Für alle gegebenen URLs durch den Index gehen

Eingabe:

www.cnn.com

www.news.google.com

...

Datenbeschaffung

Schritte:

1. Für alle gegebenen URLs durch den Index gehen
2. Alle zutreffenden locations merken

Ergebnis bis hier:

(ARCDatei1, Offset1, Größe1)
(ARCDatei2, Offset2, Größe2)
...

Datenbeschaffung

Schritte:

1. Für alle gegebenen URLs durch den Index gehen
2. Alle zutreffenden locations merken
3. Auf die location Zugreifen

Ergebnis bis hier:

http://www.spiegel.de/intern...
HTTP/1.0 200 OK
Date:Tue, 14 Feb 2012 05:48:...
...

Datenbeschaffung

Schritte:

1. Für alle gegebenen URLs durch den Index gehen
2. Alle zutreffenden locations merken
3. Auf die location Zugreifen

vorhanden? lokal laden

ansonsten herunterladen und speichern

Ergebnis bis hier:

http://www.spiegel.de/intern...
HTTP/1.0 200 OK
Date:Tue, 14 Feb 2012 05:48:...
...

Datenbeschaffung

Schritte:

1. Für alle gegebenen URLs durch den Index gehen
2. Alle zutreffenden locations merken
3. Auf die location Zugreifen

vorhanden? lokal laden

ansonsten herunterladen und speichern

4. **Whitespaces** entfernen

Ergebnis bis hier:
Wie gerade nur ohne
Whitespaces

Datenbeschaffung

Schritte:

1. Für alle gegebenen URLs durch den Index gehen
2. Alle zutreffenden locations merken
3. Auf die location Zugreifen

vorhanden? lokal laden

ansonsten herunterladen und speichern

4. **Whitespaces** entfernen
5. TSV-File erstellen

Ergebnis bis hier:

Eine TSV Datei mit einem Dokument pro Zeile:
URL<TAB>TITEL<TAB>INHALT<NL>

Übersicht

Boilerpipe

- HTML $\rightarrow$ inhaltlichen Texten
- Nutzt Features wie durchschnittliche Satzlänge

Boilerpipe

- HTML $\rightarrow$ inhaltlichen Texten
- Nutzt Features wie durchschnittliche Satzlänge
- Entfernt fast die gesamte Navigation
- Schnell [KFN10]

Probleme mit der UIMA-Pipeline

Beim Verwenden der Commoncrawl Daten:

- Probleme mit dem vorhandenen Entitätenerkenner:

Probleme mit der UIMA-Pipeline

Beim Verwenden der Commoncrawl Daten:

- Probleme mit dem vorhandenen Entitätenerkenner:
  - Titel Entitäten nur bei genauer Übereinstimmung
  - arbeitet mit Sektionen und Links (nur auf wikipedia)

Probleme mit der UIMA-Pipeline

Beim Verwenden der Commoncrawl Daten:

- Probleme mit dem vorhandenen Entitätenerkenner:
  - Titel Entitäten nur bei genauer Übereinstimmung
  - arbeitet mit Sektionen und Links (nur auf wikipedia)
- Probleme mit Kontexten:

Probleme mit der UIMA-Pipeline

Beim Verwenden der Commoncrawl Daten:

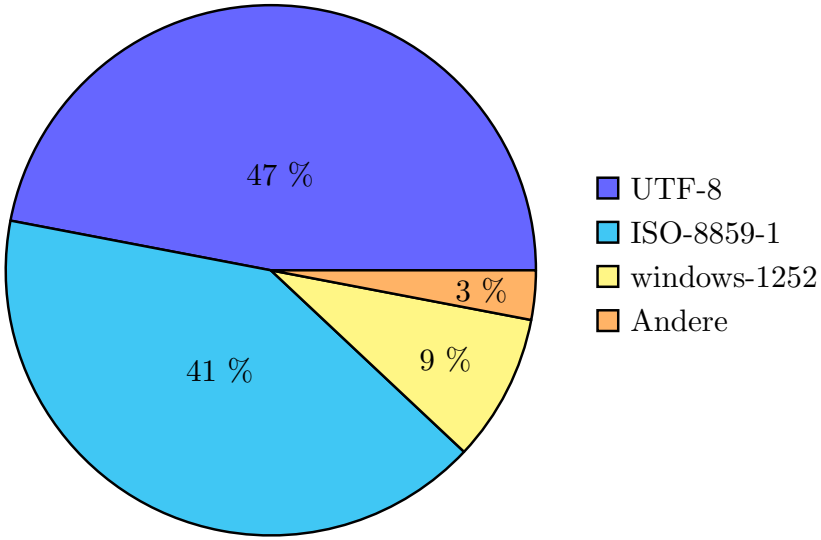
- Probleme mit dem vorhandenen Entitätenerkenner:
  - Titel Entitäten nur bei genauer Übereinstimmung
  - arbeitet mit Sektionen und Links (nur auf wikipedia)
- Probleme mit Kontexten:
  - manche Seiten erzeugen zu große Kontexte (keine Sätze)

Veränderung der UIMA-Pipeline

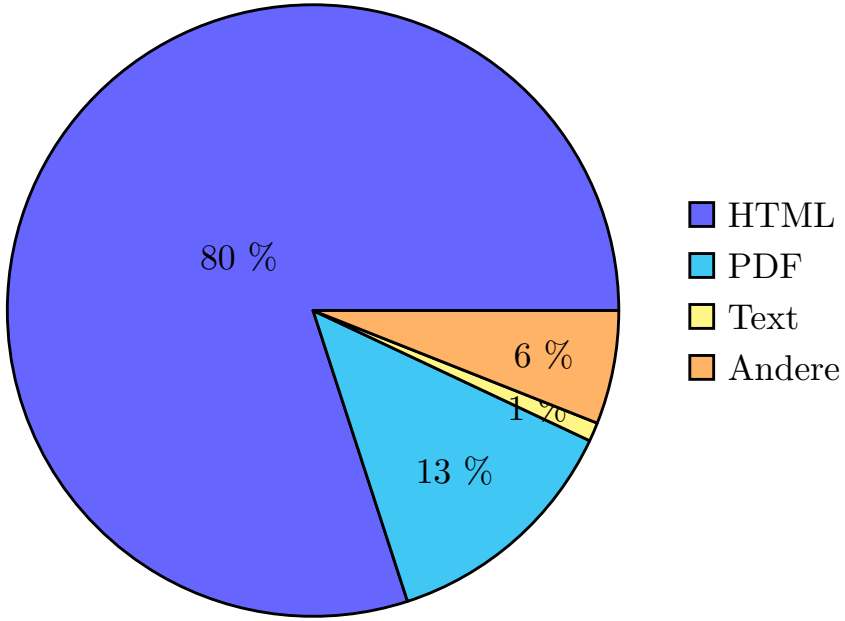
- *verallgemeinerter* Entitätenerkenner (mit und ohne POS-Tags)
- maximale Kontextgröße im Decomposer
- Parser zum Einlesen der TSV-Datei
- Boilerpipe zum Extrahieren des Textes

Übersicht

Coodierung und Typ



Verteilung der Coodierungen



Verteilung der Typen

Demo

Siehe Browser

Referenzen

KFN10 Kohlschütter, Christian ; Fankhauser, Peter ; Nejdl, Wolfgang: Boilerplate detection using shallow text features. In: Proceedings of the third ACM international conference on Web search and data min- ing ACM, 2010, S. 441–450

IMG01 `https://raw.githubusercontent.com/trivio/common_crawl_index/master/docs/tree.png`

Extras

- Seite die als HTML erkannt wurde.
- Komische Seite (Vorsicht!!)
- NULL-Byte im Titel
- RSA-Key
- RNA