

Web Mining: Übung 2

Lösungsvorschlag



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Aufgabe 1: Spracherkennung Theorie



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Zwei Listen: Liste mit ermittelten Buchstaben/-paaren und Referenzliste(n)
- ▶ Vergleiche Position der Buchstaben(-paare) mit denen der Referenzlisten
- ▶ Ermittlung eines Scores für jede Referenzsprache:

- ▶ Ermittle Differenz d der beiden Positionen
- ▶ addiere

$$\frac{1}{1 + d}$$

zum Score

- ▶ Höchster Score hat die größte (?) Übereinstimmung von Buchstaben/-paaren

Aufgabe 1: Spracherkennung Praxis



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- 1 german
- 2 english
- 3 english
- 4 spanish
- 5 german
- 6 spanish
- 7 spanish
- 8 english
- 9 german
- 10 german

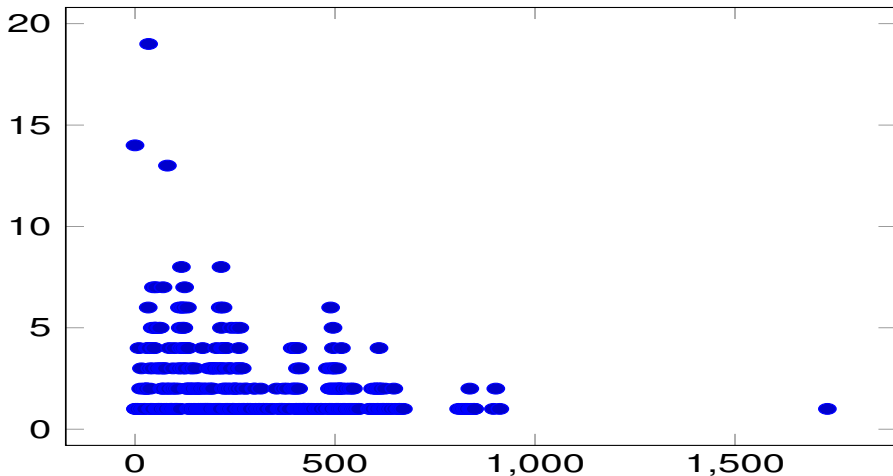
Referenzlisten für Monogramme: <https://de.wikipedia.org/wiki/Buchstabenhäufigkeit>

Referenzlisten für Bigramme:

<http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/>

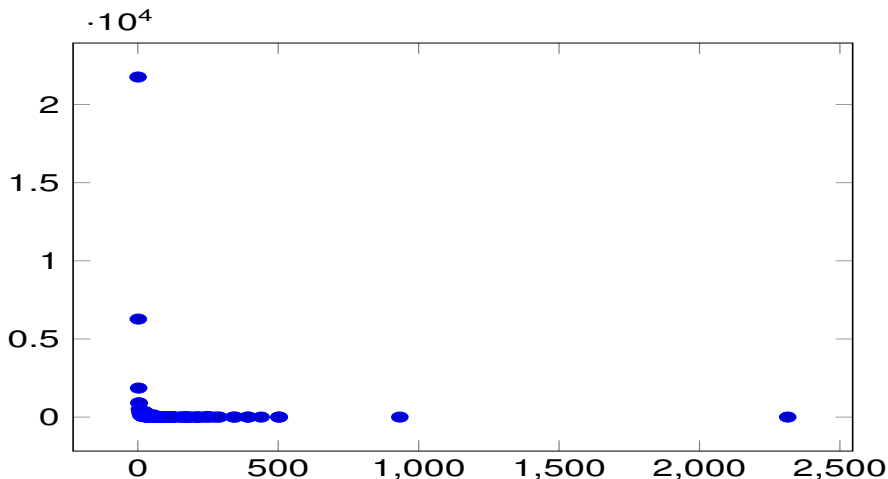
Aufgabe 2: Entwicklung eines Crawlers

Histogramm: Es gibt y Seiten mit x Links



Aufgabe 2: Entwicklung eines Crawlers

Histogramm: Es gibt x Links, die y mal auftraten



Aufgabe 2: Entwicklung eines Crawlers

Erkennung wiederkehrender Links



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- Entfernung aller Listen:

```
1 <ul>
2   <li><a href="start.html">Start</a></li>
3   <li><a href="news.html">News</a></li>
4   <li><a href="impressum.html">Impressum</a></li>
5 </ul>
```

- Entfernung mit CSS-Selektor: `[class*='nav']`:

```
1 <div class="main-navigation">
2   ...
3 </div>
```

- Speichern der HTML-Dateien als `HASH(Kanonisierte URL).html`

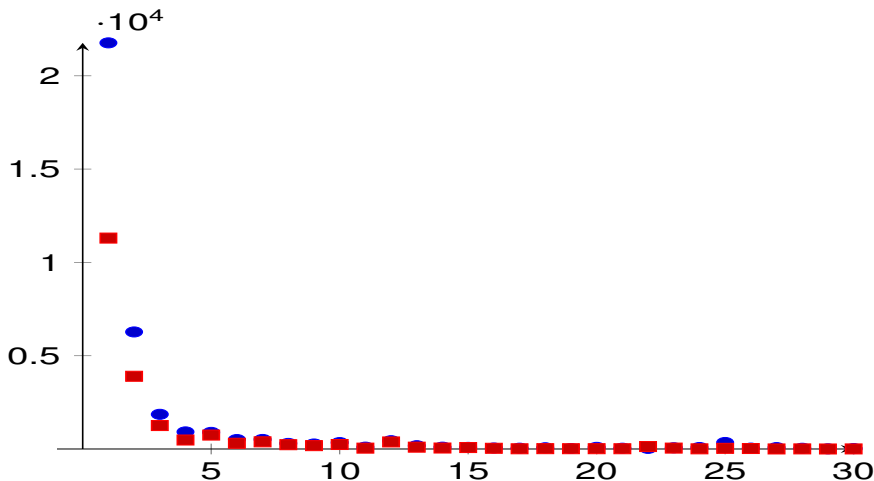
Aufgabe 2: Entwicklung eines Crawlers

Wirksamkeit der Duplikaterkennung

Es gibt x Links, die y mal auftraten



TECHNISCHE
UNIVERSITÄT
DARMSTADT





tbd

Aufgabe 2: Entwicklung eines Crawlers Erfahrungen & Probleme



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Suchstrategie funktioniert; die besuchten Hosts streuen
- ▶ Darstellung der Verlinkungen zwischen den Hosts wäre spannend („Wie vernetzt sind Hochschulen?“)
- ▶ Extrem häufig verlinkte Social Media Seiten erschweren solche Analysen. Bsp.: Anteil Twitter in offenen Links zu einem fortgeschrittenen Zeitpunkt:

$$\frac{8736}{24877} \approx 35\%$$

- ▶ Technische Lösung skaliert erwartungsgemäß nicht. Datenverwaltung nimmt mehr Zeit in Anspruch als Download:
 - ▶ Zu besuchende Links und Statistiken in Textdateien
 - ▶ Speichern eines jeden Datums direkt nach Erhebung (Persistenz)

Aufgabe 4: Größe des Webs Idee



TECHNISCHE
UNIVERSITÄT
DARMSTADT

1. Abschätzung des Index über Suchbegriff „a“ als häufigstes Wort im Englischen in beiden Suchmaschinen
2. Suche nach „Darmstadt ESOC“ (relativ wenige Ergebnisse)
3. Crawlen des Suchergebnisses um die gemeinsame Menge zu bestimmen
4. Ergebnisse:

Name (<i>i</i>)	s_i („Index“)	n_i (Ergebnisse)	n_0 (gem. Ergebnisse)
Google (<i>g</i>)	25.270.000.000	117.000	?
Bing (<i>b</i>)	140.000.000	72.800	

5. Größe des Webs:

$$N \approx s_g \frac{n_b}{n_0} \approx s_b \frac{n_g}{n_0}$$

Aufgabe 4: Größe des Webs

Problem: Googles Heuchelei oder Crawlern ohne gecrawled zu werden



TECHNISCHE
UNIVERSITÄT
DARMSTADT

► „About this page

Our systems have detected unusual traffic from your computer network. This page checks to see if it's really you sending the requests, and not a robot. Why did this happen?“

⇒ n_0 schätzen

Name (i)	s_i („Index“)	n_i (Ergebnisse)	n_0 (gem. Ergebnisse)
Google (g)	25.270.000.000	117.000	$1 \leq n_0 \leq 72.800$
Bing (b)	140.000.000	72.800	

$$\begin{array}{lcl} & \text{Minimal} & \text{Maximal} \\ N \approx s_g \frac{n_b}{n_0} = & [0,000253 \cdot 10^{15} & ; 1,84 \cdot 10^{15}] \\ N \approx s_b \frac{n_g}{n_0} = & [0,00000225 \cdot 10^{15} & ; 0,0164 \cdot 10^{15}] \end{array}$$

Aufgabe 4: Größe des Webs

Lösung & Interpretation



TECHNISCHE
UNIVERSITÄT
DARMSTADT

$$\begin{array}{rcl} N \approx s_g \frac{n_b}{n_0} = & \begin{array}{cc} \text{Minimal} & \text{Maximal} \end{array} \\ & [\quad 0,000253 \cdot 10^{15} \quad ; \quad 1,84 \cdot 10^{15} \quad] \\ N \approx s_b \frac{n_g}{n_0} = & [\quad 0,00000225 \cdot 10^{15} \quad ; \quad 0,0164 \cdot 10^{15} \quad] \end{array}$$

In Worten:

$$\begin{array}{rcl} & \begin{array}{cc} \text{Minimal} & \text{Maximal} \end{array} \\ N \approx s_g \frac{n_b}{n_0} = & [\quad 25 \text{ Milliarden} \quad ; \quad 1,84 \text{ Milliarden} \quad] \\ N \approx s_b \frac{n_g}{n_0} = & [\quad 225 \text{ Millionen} \quad ; \quad 16 \text{ Billionen} \quad] \end{array}$$

Laut <http://www.worldwidewebsize.com/> sind es etwa 50 Milliarden Seiten. Diese Zahl liegt im Rahmen dieser Schätzung. Aber Präzision sieht anders aus.