Datenanalyse im Web

Prof. Dr. Ingo Claßen
HTW Berlin

Entwicklung einer einfachen Suchmaschine

1 Aufgabenstellung

Entwickeln Sie eine einfache Suchmaschine in der Programmiersprache Python, in dem Sie folgende Teilaufgaben bearbeiten:

1. Crawling der Webseiten und Aufbau des Web-Graphen
2. Berechnung PageRank aller Seiten
3. Indizierung der Web-Seiten und Suche
4. Kombination von Seiten- und Index-Rang zur Verbesserung der Suchergebnisse

2 Teilaufgaben

2.1 Crawling

In einem ersten Schritt sollen Sie einen Crawler entwicklen, der die in den Folien angebene Struktur hat. Python stellt dafür verschiedene Komponenten bereit: Für den Downloader können `urllib2` und `urlparse` eingesetzt werden. Mit `Beautiful Soup` steht ein Parser für Html-Seiten zur Verfügung. Die Frontier kann über zwei Listen realisiert werden, wobei die eine die neuen, noch zu besuchenden URLs enthält, und die andere bereits besuchte.

Der Crawler wird mit folgender Liste von Start-URLs (Seed) initialisiert:

```
http://mysql12.f4.htw-berlin.de/crawl/d01.html
http://mysql12.f4.htw-berlin.de/crawl/d06.html
http://mysql12.f4.htw-berlin.de/crawl/d08.html
```

Die Steuerung über die Frontier soll sichergestellen, dass alle verlinkten Seiten (`d01` bis `d08`) erreicht werden und jede Seite nur einmal besucht wird.

Siehe `http://mysql12.f4.htw-berlin.de/crawl/` für einen Überblick über alle Seiten. Dokument `d08.html` ist eine Spam-Seite.

2.2 PageRank

Die Seitenstruktur ist Ausgangspunkt für die Berechnung der PageRanks aller Seiten. Die Berechnung kann über Matrizenrechnung erfolgen oder durch den Aufbau einer internen Graphstruktur. Verwenden Sie $(0.95, 0.05)$ für die Kombination Dämpfungsfaktor/Teleportationsrate und einen Delta-Wert von 0.04.

2.3 Suche

Zur Implementierung der Suche müssen Sie einen Index aufbauen, wie er in den Folien beschrieben ist. Verwenden Sie

- die vorgegebene tf-idf-Gewichtung,
- das Cosinusmaß für das Scoring
- und die auf der Aufgaben-Webseite angegebenen Stopwörter.

Führen Sie auf dem Index eine Suche mit den einfachen Suchbegriffen *tokens*, *index*, *classification* und dem zusammengesetzten Suchbegriff *tokens classification* durch. Man sieht, dass die Spam-Seite fast immer den höchsten Rang erhält.

2.4 Kombination von indexbasierter Suche und PageRank

Erweitern Sie die Suche, in dem Sie Indexrang und PageRank geeignet kombinieren. Diese Suche soll die Spam-Seite abwerten.