Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Week 8: Crawling

Assignment 1: Crawling the Web

a) What is the advantage of using HEAD requests instead of GET requests during crawling? When would a crawler use a GET requests instead of a HEAD request?

The challenges of IR in regards to the web lie in the following areas:

A HEAD request only requests the header of a resource, i.e. the information about the content size, the last update, the content type etc., without actually downloading the resource. This can be used to determine whether a resource needs to be downloaded again or not while at the same time not using to many resources (disk space, processing power) on both the crawler and the crawlee.

b) Why is it better to distribute hosts (rather than individual URLs) between the nodes of a distributed crawl system?

A crawler needs to take care of not crawling one host too often, because this is impolite (wasting the server's resources on a crawl task) and, in the extreme case, can cause the server to fail, because there are too many requests (denial of service). So a crawler needs to keep track about when he last crawled a server. This information is best kept in memory for one machine, instead of distributing it (which would be an unnecessary overhead for this task).

c) How does BigTable handle hardware failure?

It justs restarts another server. The data can be read from the transaction log, as per definition, all changes to a tablet are recorded in a transaction log.

Assignment 2: (Programming) Crawling

Print the ranking of Wikipedia titles and snippets in Wikipedia together with its NDCG value for the queries:

a) " smithee"

```
query: smithee
NDCG@10: 0.3876511272273169
Alan Smithee:
Alan Smithee steht als Pseudonym für einen fiktiven Regisseur der Filme
verantwortet bei denen der eigentliche Regisseur seinen Namen nicht
Fahr zur Hölle Hollywood:
Film DT Fahr zur Hölle Hollywood OT An Alan Smithee
Goldene Himbeere 1999:
und Ben Myron Fahr zur Hölle Hollywood An Alan Smithee Film Burn Hollywood
Burn
Catchfire:
OS Englische Sprache Englisch REG Dennis Hopper als Alan Smithee
Hellraiser IV – Bloodline:
Indizierung indiziert LEN OS Englische Sprache Englisch REG Alan Smithee
```