Information Retrieval, Assignment 1 - Crawling
Autumn 2015

Matthias Fuhr, Andrei Pârvu

October 14, 2015

1 Results

Distinct URLs found: 5313
Exact duplicates found: 31
Near duplicates found: 395
Number of unique pages mostly in English: 2043
Term frequency of "student": 3645

2 Strategy & Assumptions

2.1 Explored URLs

We only explored urls which ended in *.html and we counted the ones from which we can successfully retrieve the html code.

2.2 Exact duplicates

We consider two pages to the exact duplicates if the hashes of their html source code is the same (we do not filter out anything at this step)

2.3 Near duplicates

For near-duplicates, we filter the html source such that we keep only the text that we find between tags which have a "contentMain" id. We consider either or <section> tags. After getting the text between these tags we eliminate further tags present in it (they don't really matter with respect to the meaning of the page).
We prefered using this technique in order to avoid marking two pages as similar based on the left-side or right-side menus.
We split the text in shingles formed out of 7 words. To compute the similarity we use a SimHash sketch having 3 tables and considering 25-bit prefixes of 32-bit hashes. One could adjust this limit, decreasing the prefix and increasing the number of tables for a more broader match.

2.4 Multi-threading

We implemented our solution in a multi-threading style to compensate the large IO overhead of reading lots of html pages. The number of used threads can be set from the THREADS variable (which is currently 1, because crawling from the ETH network runs pretty fast).