Information Retrieval - Fall Semester 2015

Group ID: 11
ollel@student.ethz.ch
eholmer@student.ethz.ch
pmichal @student.ethz.ch

January 16, 2016

Our crawler produced the following statistics:

- Distincs URLs found: 3970
- Exact duplicates found: 29
- Near duplicates found: 425
- Unique English pages found: 1597
- Term frequency of "student": 10319

Assumptions and implementation details

- **Visited URLs:** As instructed, we only follow links ending with ".html" not containing "?#" symbols, i.e the page */loginf545.html?resource=%2Fcontent%2Fmain%2Fde.html* would not be considered. Additionally, we exclude all valid pages with no content, i.e. a blank page.
- **Distinct URL count:** All links satisfying the above criteria are counted as a unique URL, regardless of duplicities.
- **URL Parsing:** We only extract paragraphs and headers as we found that these contain majority of relevant content, while not assuming any particular HTML structure, *e.g* . ⟨ section id= "content" ⟩.
- **Exact duplicate detection:** Whenever a new page has a *SimHash* fingerprint identical to an an already visited page, we verify if their textual content is identical. Only in such case do we consider these as identical, because two slightly different pages might be assigned the same fingerprint.
- **Near duplicate detection:** All pages with similarity at least $\frac{126}{128}$ that are not exact duplicates are considered near duplicates.