Information Retrieval - Fall Semester 2015

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

October 14, 2015

Our crawler produced the following statistics:

- Distincs URLs found:
- Exact duplicates found:
- Unique English pages found:
- Near duplicates found:
- Term frequency of "student":

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.
- **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 two pages have identical *SimHash* fingerprint, we verify if their text content is identical. Only in such case do we consider these as identical, because two slightly different pages might be assigned the same fingerprint.