Lab 1: Crawling, indexation and webpage research

1 Crawler

2 Specific Indexation

3 Research

4 Theorical questions

4.1 Please explain what strategy should be adopted for indexing pages in several languages (each page is composed of only one language, but the corpus includes pages in several languages). What should you watch out for? Please explain the process you propose.

4.2 Solr allows by default to do a fuzzy search. Please explain what it is and how Solr implements it. Some first names may have a lot of spelling variations (eg Caitlin : Caitilin, Caitlen, Caitlinn, Caitlyn, Caitlyne, Caitlynn, Cateline, Catelinn, Catelyn, Catelynn, Catlain, Catlin, Catline, Catlyn, Catlynn, Kaitlin, Kaitlinn, Kaitlyn, Kaitlynn, Katelin, Katelyn, Katelynn, etc). Is it possible to use, while keeping a good performance, the fuzzy research made available by Solr to do research taking into account such variations? If not what alternative(s) do you see, please justify your answer.