Information Retrieval

Exercises session n°2: Pre-Processing & Dictionary

Exercise 1: Porter's stemmer

Concerning the Porter's stemmer rules (cf. lecture n°2):

1. Why does the rule `ss` → `ss` appears as it seems to have no effect?
2. Apply these rules to the words: `circus canaries ponies boss`.
3. What rule should be added to stem `pony` too?
4. The stemmer output for `ponies` looks strange as it does not belong to an usual dictionary. Is this noxious for retrieval?

Exercise 2: Online Porter's stemmer

Using an online Porter's stemming tool (cf. lecture n°2), stem the following words:

```
information
retrieval
retry
automation
automated
automatically
automobile
```

Exercise 3: Pre-Processing and inverted index

Consider the following three short documents:

> Glimpse is an indexing and query system that allows for search through a file system or document collection quickly. Glimpse is the default search engine in a larger information retrieval system. It has also been used as part of some web based search engines.
>
> Doc #1

> The main processes in an retrieval system are document indexing, query processing, query evaluation and relevance feedback. Among these, efficient updating of the index is critical in large scale systems.
>
> Doc #2

> Clusters are created from short snippets of documents retrieved by web search engines which are as good as clusters created from the full text of web documents.
>
> Doc #3

First remove stop words (choose a list, cf. lecture n°2) and punctuation, normalize, and apply Porter's stemming algorithm to the 3 documents (Note: use an online stemming application for this purpose).

Create an inverted index of the three documents, including the dictionary and the postings (cf. lecture n°1, slide "Indexer steps: Dictionary & Postings"). The dictionary should also contain (for each term) statistics: the total number of occurrences in the collection and the document frequency.

What are the search results for the following Boolean queries (in each case explain how you obtained them from the inverted index):

```
index AND query
index OR query
index AND (NOT query)
(search AND query) OR (search AND retrieve)
(index OR cluster) AND (web OR system)
```