COMP 38120 Workshop 3: Exploring language identification, tokenization and stemming

1. Download from Blackboard

1.1 From COMP38120 BB space, Workshop 3 folder, download the file under the link "sample files for tokenizing". This is a zip file that will uncompress into a folder called "tokfiles". There are 5 texts in different formats. One is a recipe, two are scientific and the remaining two are the Japanese and Chinese examples. The different formats of the English texts are there a) for ease of use and b) so that you can experiment with copy/pasting from different formats into the demos.

1.2 Download from the same space the file under "URLs for demos". Open it, then click on, or copy/paste from, URLs in this file, as you prefer, to access the various demos.

2. Language identifier: how good is it for text in various languages that it knows about?

3. Tokenizers:

3.1 WITHOUT LOOKING AT THE SAMPLE ENGLISH TEXTS

Individually: Write down 3 rules that you think would be applicable in general for tokenizing English text. These essentially will express your (initial) definition of a token. Assume there is *no dependency* among rules. You can express rules informally as in:

Don't do X or *Do Y*
e.g*. Don't separate two consecutive single quotation marks* (as in LaTeX '')

Group: discuss these briefly to see if you have any rules in common. Merge the rules into 1 set.

3.2 USING THE SAMPLE ENGLISH TEXTS

Group: try out the sample English texts in the various tokenizers (NB: some may limit the length of the input – in this latter case, chunk your texts up, don't just process the first *n* characters).

Note issues you observe with the tokenization behaviour of each tokenizer for each text (including any problems with character sets/encodings)

Note to what extent your group rules have apparently been applied by each tokenizer.

What, if any, additional rules would you consider adding to your group rule set in the light of your tests – and why? Label each rule in your set according to the domain/text type it is useful for (use "GENERAL" for a generally applicable rule).

Note: I do **not** expect you to know precisely what e.g. a biologist, a chemist or a chef would consider to be a token. However, you should be able to have some thoughts about what might be reasonable and why.

Prepare points to report to the class.

4. Stemming

4.1. Group: Run your sample English texts through the Porter and Lancaster stemmers demo.

Observe the results, note down a variety of examples of good/bad stemming decisions (i.e. indicating what you consider to be incorrect stemmings, understemmings or overstemmings).

Incorrect: "just plain wrong" according to your group.
Understemming: e.g. stemming "divide" to "divid" and "division" to "divis"
Overstemming: e.g. stemming both "neutral" and "neutron" to "neutr"

Which stemmer do you think does a better job?

Prepare points to report to the class.

4.2 Group: Use the WPO search engine

Many popular search engines do not allow user control of stemming. Even specialised engines either apply stemming by default, or offer truncation or wildcard operators that are not as constrained as stemming. One engine that offers user-controlled stemming is Patentscope, from the WIPO. Patentscope's field combination advanced search facility is:
https://patentscope.wipo.int/search/en/structuredSearch.jsf
Ensure the 'stem' checkbox is ticked, and also it is helpful to select 'office=UK', then in the ENGLISH ABSTRACT field search with a query that is likely to demonstrate different results with stemming. E.g., searching for 'extents' should find you documents with also 'extent'. Searching for 'flies' will find documents with 'fly', and it should also find documents with 'flying'. If you investigate the icon of a tree diagram, next to the RSS one, you will be shown the query tree, so that you can see the details. In general, you should observe the differences in the number of hits with and without stemming. Also check to see if results are giving you unexpected (or wrong) items. Some other forms to try:
hooves
vertices
witness
paste
appendicitis
dying
admiral
engineer
racketeer
responsible
several
experiment
secretion
designate
colonize
positive
amenities
universities

Prepare some points to report to the class.