De Data: Practical

Lecture 2: Text Analytics for Big Data

Mark Keane, Insight/CSI, UCD

Tokenizing
Q1

Q1.

- Try to get used to using then **nltk** package. Create a text file with a good number of words in this that should challenge the tokenizer. It should have at least, 200 words in it and items that challenge the tokeniser (e.g., IBM, and such like things). But, you need to make these up yourself, so you think about what it is doing.

Q1. Part (a)

- Load the file in and use nltk.word_tokenizer() on it. Report the list of tokens that are produced from it and note any oddities that arise. Comment on these oddities and how they might be handled.

Q1. Part (b)

- Now, do normalization on it, and report this output as your answer.

Q1. Part (c)

- Now, take the output from normalization step and run it through a pos-tagger. Report this output as your answer and highlight any inaccuracies that occur at this stage.

Stemming & Lemmatising Q2

Q2. Part (a)

- Tokenize a new text-file (200 words) and the stem it using Porter Stemming. Report your answer and some of weird things that Porter Stemming does.

Q2. Part (b)

- Tokenize the new text-file and then lemmatize it using WordNet Lemmatizer; note you may have to pos-tag the sentences first and then convert the tags to make this work. Report the result of these steps and point out some of the things that look wrong.

Q2. Part (c)

- Compare the outputs from Porter Stemming and the Lemmatisation of the same file. Which do you think is the best to use and why?

Webpage Loading
Q3

Q3.

- Finally, choose a remote webpage and extract key text content from it. Install the packages you need and then parse it using BeautifulSoup. Try to get to a point where you can extract one of its XML/HTML parts (e.g., title, summary, body)