UNIB20005: Language and Computation

Project 2: Corpus Analysis

A corpus is a representative sample of language use which can be used to support empirical research on language. The web is an excellent source of language data. In this project you will create your own modest-size corpus, in a genre of your choice, and carry out a series of small programming tasks to analyse its contents.

Identify a genre and download a collection of HTML files containing 10,000-50,000 words of text in this genre; explain your choice of data and document the source of the data in your written report.

Next, complete the following programming tasks:

1. *Getting the data:*
   (a) Write a function `clean` which processes the raw data and strips out the HTML, and saves the result in a single file `corpus.txt`. Briefly discuss any issues you encountered in your report.
   (b) Write a function `tokens` which processes `corpus.txt` using your own regular expression tokenizer; include comments to explain the purpose of each part of your regular expression (cf §3.7 of the textbook). The function should return a list of tokens.
2. *Basic processing:*
   (a) Write a function `lexical_diversity` which calculates the ratio of word types to word tokens, and returns a floating-point number. Take care to normalize the tokens.
   (b) Write a function `distinctive` which uses information about word length and/or frequency and/or any other properties you like, to identify the words that are most characteristic of this genre. The function should return a list of words. Note 10 characteristic words in the report, and discuss.
3. *Advanced processing:*
   (a) Write a function `collocations` which reports pairs of words that are found adjacent to one another more often than one would expect based solely on word frequency. The function should return a list of pairs of words. Select some characteristic collocations and discuss them in your written report.
   (b) Write a function `average_polysemy` to calculate the average polysemy of the nouns in your corpus, where the polysemy of a noun is taken to be the number of WordNet synsets it has. Note the 10 most polysemous words in your report, and discuss.
4. Implement any other corpus processing task of your choice, and discuss it in the report, be sure to cover motivation and findings (100-200 words). (This task should have similar complexity to the last two tasks.)

Add comments to your code to explain any aspects of your program which are not immediately obvious to the reader. Discuss any other insights you have based on this work in your written report.

Your work will be assessed for correctness and clarity. The project must be original work. The report should be about 400 words long. Your submission is worth 10% of the total marks for this subject.

Please submit three files `corpus.py`, `corpus.txt`, and `report.pdf` as email attachments, and email them to `sbird@unimelb.edu.au`, using the subject line `L&C Project 2`. All submissions will be acknowledged. If you do not receive an acknowledgement by the end of Monday 17 September, please resend.

Submit your work by the end of week 8 (10pm on Friday 14 September).