SSA 3 - 2017/18
Web Technology Coursework

Rob Powell

Hand in by 2pm on 9th March 2018 via DUO.

Coursework Description

This coursework involves developing two programs for multi-document summarisation, and writing a short report detailing how your programs work, and analysing the results obtained. You should submit your report as a pdf file, and submit both the report and the two pieces of code via DUO. The programming aspects should be completed in either Python, Java, or C/C++, and your report should be a maximum of 3 pages of A4 in length, including any diagrams, and must use a minimum of 10pt font. Any references used in the report should be clearly identified.

Basic Programming Objective

Implement a Term Frequency (TF) algorithm as seen in Lecture 5:

- Count the number of times each word appears in each document. I suggest storing this data as some form of matrix (each row is a word, each column a document, the entry being the number of times that word appears in that document), it will help you with the advanced objectives!
- Use a list of stop words of your choosing to eliminate certain words.
- Calculate the weight of each sentence in each document.
- Output a single summary document from the 8 input documents, with user configurable length. This should be taken as a runtime parameter (argument), and represents the maximum number of words in the summary. E.g. we might want both a 100 word short summary, and a 250 word longer summary.

Advanced Programming Objective

Implement a Term Frequency-Inverse Document Frequency (tf-idf) algorithm as seen in Lecture 5:

- Count the number of times each word appears in each document as before.
- Calculate the tf-idf score for each word in each document.
- Calculate the weight of each sentence in each document.
- Output a single summary document with user configurable length as before.

You should test each of your programs with the 8 documents provided on DUO. These will allow you to check your programs are working as expected, and give you results to analyse in your report.

Report

The report should include, *in your own words*, some background information on the problem area, an explanation of how your code works, and a comparison of the different techniques. You should then analyse the results of applying your programs to the documents on DUO. Think about whether the summaries produced seem accurate and read coherently for different lengths of summary. Finally conclude with some ideas of how you could change or adapt the algorithms you have implemented (given more time) to produce a better quality summary.

Marking Criteria

The breakdown of marks are as follows:

- Programming Objectives - 65%
  - Term Frequency algorithm - 30%
  - Term Frequency-Inverse Document Frequecy algorithm - 25%
  - Output quality compared to gold standard - 5%
  - Style, clarity and comments - 5%
- Report - 35%
  - Background - 5%
  - Implementation description and comparison of techniques - 10%
  - Presentation and analysis of results - 10%
  - Conclusion - 5%
  - Writing style, structure, grammar and presentation - 5%