Text Processing Assignment Feedback

Task: Document Retrieval

Student: Bobby Williams (acb18bw)

Submitted: Friday, 12 November 2021 13:19:29 o'clock GMT
(Deadline: 3pm, Friday, 12 November, 2021)

Total Mark:	17/25
Final Mark (after any late penalty):	17/25

Automatic Test Results:

F-SCORES:				
	stp-/stm-	stp+/stm-	stp-/stm+	stp+/stm+
TFIDF	0.18	0.19	0.23	0.24
TF	0.07	0.15	0.10	0.17
BINARY	0.06	0.11	0.08	0.15

TIMES:				
	stp-/stm-	stp+/stm-	stp-/stm+	stp+/stm+
TFIDF	10.07	7.85	7.48	5.60
TF	7.40	4.81	6.09	3.93
BINARY	8.43	4.88	6.41	3.99

PRECISION:				
	stp-/stm-	stp+/stm-	stp-/stm+	stp+/stm+
TFIDF	0.21	0.22	0.26	0.27
TF	0.08	0.17	0.11	0.19
BINARY	0.07	0.13	0.09	0.16

RECALL:				
	stp-/stm-	stp+/stm-	stp-/stm+	stp+/stm+
TFIDF	0.17	0.18	0.21	0.22
TF	0.06	0.13	0.09	0.15
BINARY	0.06	0.10	0.08	0.13

[stp +/- → stoplist used/not] [stm +/- → stemming used/not] ["." → timeout (300s)] ["x" → code crashed]

Implementation and Code Style: 9/15

Functionality:
All configurations of the system have been successfully implemented and achieve excellent performance with recall, precision and f-measure scores at the level of the best known scores.

Efficiency:
Your code runs significantly slower than the known best performing approaches: times of under 1 second are possible for all conditions. In several conditions your code takes > 10 seconds to complete, implying it contains inefficiencies. For example, you appear to be creating a vector for each document where the vector contains all terms in the document, while only the terms shared with the query are relevant for the score. For short queries and long documents this is hugely inefficient.

Code Style:
Your code is reasonably well structured, with sensible variable names but it is poorly commented in places. Commenting all functions and important variables would improve your code.

Report: 8/10

Description of implementation:
Useful description of your implementation with references to key functions and pre-computations made in the interest of efficiency.

Results:
Full tables of results are presented. Graphs could be better utilised to facilitate comparison accross weighting schemes.

Discussion of results:

Good discussion of results. Good observations regarding the use of stoplists and stemming and the impact of different term weighting schemes. Correct speculation about the reduced additional value of stop word removal when TFIDF as opposed to TF and binary term weightings is used.