Text Processing Assignment Feedback

Task: Document Retrieval

Student: Ruiqing Xu (acb19rx)

Submitted: Thursday, 4 November 2021 21:20:14 o'clock GMT

(Deadline: 3pm, Friday, 12 November, 2021)

Total Mark:	20/25
Final Mark (after any late penalty):	20/25

Automatic Test Results:

F-SCORES:

	stp-/stm-	stp+/stm-	stp-/stm+	stp+/stm+
TFIDF	0.18	0.19	0.24	0.24
TF	0.07	0.15	0.10	0.17
BINARY	0.06	0.12	0.08	0.15

TIMES:

	stp-/stm-	stp+/stm-	stp-/stm+	stp+/stm+
TFIDF	7.56	4.73	7.11	4.63
TF	4.97	3.29	4.70	3.12
BINARY	4.35	3.05	4.17	2.88

PRECISION:

	stp-	stp+/stm-	stp-	stp+/stm+

RECALL:

	stp-	stp+/stm-	stp-	stp+/stm+

	/stm-		/stm+	
TFIDF	0.20	0.22	0.26	0.27
TF	0.08	0.17	0.11	0.19
BINARY	0.07	0.13	0.10	0.16

	/stm-		/stm+	
TFIDF	0.16	0.17	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: 12/15

Functionality:

All configurations of the system have been successfully implemented and achieve good performance with recall, precision and f-measure scores at the level or close to the level of the best known scores.

Efficiency:

Reasonable times achieved for all settings, but substantially better times (i.e. < 1 second) can be achieved for all settings. Note that you can compute all IDFs in advance if the term-weighting scheme is TFIDF, and you can drop query vector sizes when calculating the cosine similarity.

Code Style:

Code is well-structured, with appropriate use of comments and good choice of variable names.

Report: 8/10

Description of implementation:

You have explained all term-weighting schemes. However, a description of the principal steps in processing and/or the principal data structures used (their structure and intended use) is more relevant for the report.

Results:

Full and relatively clear statement of results. Graphical and use of colours can hselp to make comparisons between numbers easier to grasp.

Discussion of results:

Good observations regarding the use of stoplists and stemming and the impact of different term weighting schemes.