COMP4650 Assignment 1 Answers

Jieli Zheng

u6579712

Q1 & Q2

TF scores:
Mean Average Precision: 0.16925271708009806
R-Precision: 0.19803030303030303
Reciprocal rank: 0.37175925925925923
Precision at 5: 0.19999999999999998
Precision at 10: 0.14666666666666667
Precision at 15: 0.09777777777777778

TF-IDF scores:
Mean Average Precision: 0.28087837129801413
R-Precision: 0.29668109668109666
Reciprocal rank: 0.5202777777777777
Precision at 5: 0.30000000000000004
Precision at 10: 0.21666666666666665
Precision at 15: 0.14444444444444446

Q3

The system performs better with TF-IDF instead of TF similarity.
I think the Reciprocal rank is the most useful measure for evaluating this government system.
Firstly I assume that this information retrieval system is designed for government office to search for the previous official government documents. Officer usually has clearly targets so that only the exact document will be needed. And in this case it's a precision-critical task and there is only one relevant document in the system. As a result, the Reciprocal rank should be the most important measure depending on our demand.

First setting:
PorterStemmer (Default), PunktSentenceTokenizer
Scores:
Mean Average Precision: 0.7524007936507937
R-Precision: 0.6950000000000001
Reciprocal rank (chosen): 0.9
Precision at 5: 0.6
Precision at 10: 0.42000000000000004
Precision at 15: 0.27999999999999997

Firstly, we only change the Tokenizer to PunktSentenceTokenizer, which tokenize text by using punctuations(';', ':', ',', '.', '!', '?'). Compared to the default setting in Q2, we find that Reciprocal rank of the first setting is much better than the default one in Q2. The main reason is that WhitespaceTokenizer divides text by space, tab and newline, which is less informative than punctuations because government documents usually have more punctuations than whitespaces and punctuations are designed for spliting meanings.

Second setting:
LancasterStemmer, PunktSentenceTokenizer:
Scores:
Mean Average Precision: 0.5935119047619046
R-Precision: 0.575
Reciprocal rank (chosen): 0.7
Precision at 5: 0.4800000000000001
Precision at 10: 0.32
Precision at 15: 0.21333333333333337

In the second setting, both the Tokenizer and the stemmer have been changed. We find that the Reciprocal rank of the second setting is worse than the first one when we change default PorterStemmer to LancasterStemmer. LancasterStemmer is based on Lancaster (Paice/Husk) stemming algorithm, which performs worse than PorterStemmer which is based on the Porter stemming algorithm. LancasterStemmer is less effective than PorterStemmer because LancasterStemmer only converts suffixes.

Third setting:
SnowballStemmer with language set to English, WhitespaceTokenizer(Default)
Scores:
Mean Average Precision: 0.22738894328775283
R-Precision: 0.24219696969696972
Reciprocal rank (chosen): 0.5253703703703703
Precision at 5: 0.25333333333333335
Precision at 10: 0.1833333333333333
Precision at 15: 0.12222222222222225

In the third setting, we use SnowballStemmer in English and default tokenizer. And we find the Reciprocal rank is relatively the same as the default setting. SnowballStemmer is designed by the suffix stripping algorithm, which is similar to PorterStemmer.

Forth setting:
ISRIStemmer, WhitespaceTokenizer(Default)
Scores:
Mean Average Precision: 0.09838395863395863
R-Precision: 0.15106060606060603
Reciprocal rank (chosen): 0.21888888888888888
Precision at 5: 0.14666666666666667
Precision at 10: 0.10666666666666666
Precision at 15: 0.07111111111111111

In the forth setting, we use ISRIStemmer and default tokenizer. ISRIStemmer is based on Arabic Stemming without a root dictionary Algorithm and the Reciprocal rank of this setting is much worse than the default setting. The main reason is that ISRIStemmer sometimes converts words to wrong word stems that changes the vector form of the document.

Q4