ASSIGNMENT 2 | COL 764

Dipen Kumar
2018CS50098

Stemming done using porter stemmer, tokenized with `re.split('[.,:;\\\\\/\n\t\s\r\'\"\(\)\[\]\{\}]', text)`
I didn't used any stop words

Rocchio Reranking:

Approximation used for non-relevent documents: 2000 random choosen documents other than those mentioned in top100
Vocabulary used: All words in top100 and 2000 randomly choosen non-relevent documents
tf used: log(1+ word frequency)/(1+log(1+ total word count))
idf used: log(1+(total number of documents in collection/total number of documents with the term))

used rocchio for finding qm (modified query vector) from q0 (initial query vector), dr (average relevant document vector), and dn (average non-relevent document vector)

Did Grid search to find appropriate aplha, beta, and gamma

alpha = 0.95, beta = 0.7, gamma = 0.25
nDCG@5 = 0.46, nDCG@10 = 0.46, nDCG@15 = 0.46, MAP = 0.055

Grid Search:

alpha range = 0.90,0.95,1.00
beta values taken = 0.70,0.75,0.80
gamma values taken = 0.15,0.20,0.25

alpha	beta	gamma	nDCG@5	nDCG@10	nDCG@15	MAP
0.90	0.70	0.15	0.43	0.43	0.43	0.054
0.90	0.70	0.20	0.44	0.44	0.44	0.054
0.90	0.70	0.25	0.45	0.45	0.45	0.055
0.90	0.75	0.15	0.43	0.43	0.43	0.054
0.90	0.75	0.20	0.44	0.44	0.44	0.054
0.90	0.75	0.25	0.45	0.45	0.45	0.055
0.90	0.80	0.15	0.44	0.44	0.44	0.055
0.90	0.80	0.20	0.44	0.45	0.45	0.055
0.90	0.80	0.25	0.45	0.45	0.46	0.055
0.95	0.70	0.15	0.44	0.44	0.44	0.054
0.95	0.70	0.20	0.45	0.45	0.45	0.055
0.95	0.70	0.25	0.46	0.46	0.46	0.055
0.95	0.75	0.15	0.44	0.44	0.44	0.054
0.95	0.75	0.20	0.44	0.45	0.45	0.054
0.95	0.75	0.25	0.45	0.45	0.45	0.055
0.95	0.80	0.15	0.43	0.43	0.43	0.054
0.95	0.80	0.20	0.44	0.44	0.44	0.054

0.95	0.80	0.25	0.45	0.45	0.45	0.055
1.00	0.70	0.15	0.43	0.43	0.43	0.054
1.00	0.70	0.20	0.44	0.44	0.44	0.054
1.00	0.70	0.25	0.44	0.45	0.45	0.055
1.00	0.75	0.15	0.43	0.43	0.43	0.054
1.00	0.75	0.20	0.44	0.44	0.44	0.054
1.00	0.75	0.25	0.45	0.45	0.45	0.055
1.00	0.80	0.15	0.43	0.43	0.44	0.054
1.00	0.80	0.20	0.45	0.44	0.45	0.054
1.00	0.80	0.25	0.45	0.45	0.46	0.055