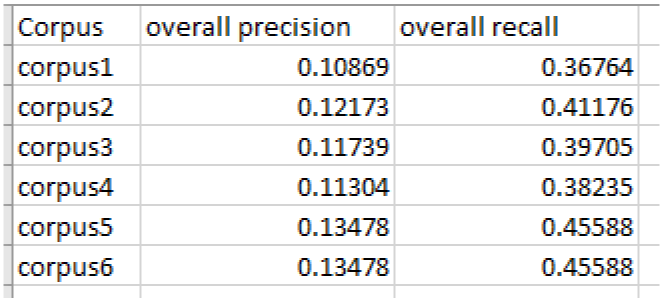
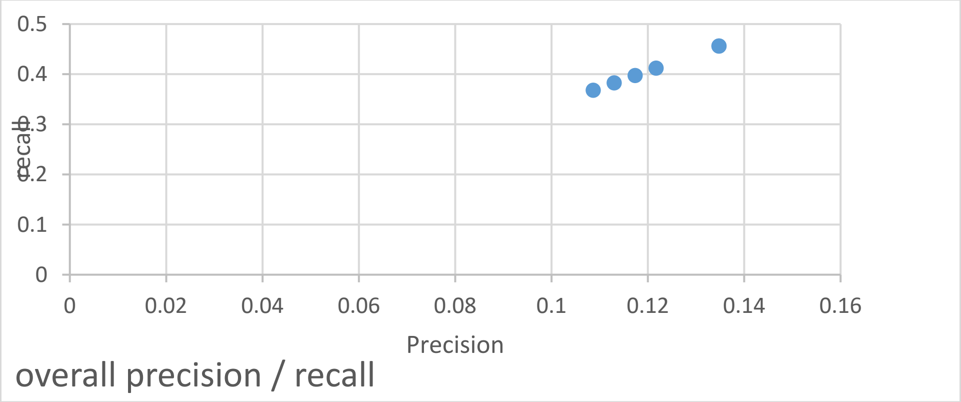
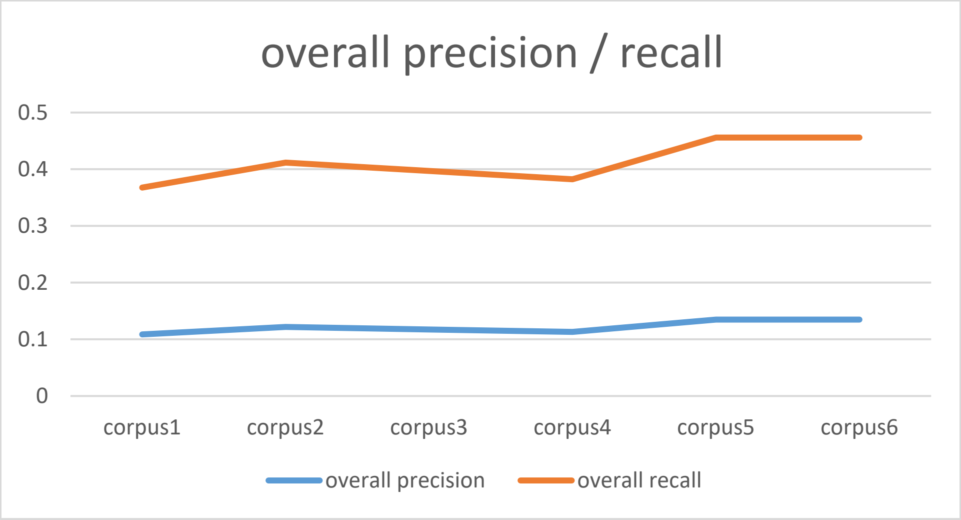
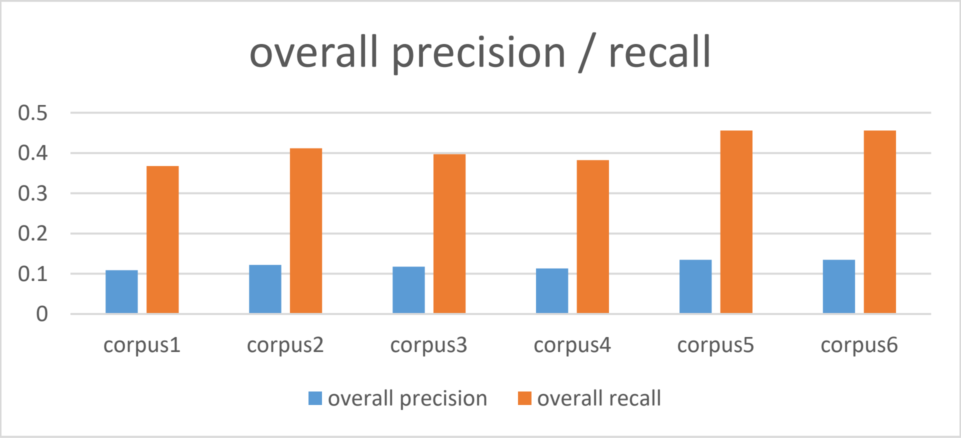


Corpus	overall precision	overall recall
corpus1	0.10869	0.36764
corpus2	0.12173	0.41176
corpus3	0.11739	0.39705
corpus4	0.11304	0.38235
corpus5	0.13478	0.45588
corpus6	0.13478	0.45588







Discussion:

Corpus1: {'tf': "raw", 'idf': "inverse", 'stopword': "none", 'stemmer': "none"},

Corpus2: {'tf': "log", 'idf': "smoothed", 'stopword': "none", 'stemmer': "none"},

Corpus3: {'tf': "raw", 'idf': "inverse"},

Corpus4: {'tf': "raw", 'idf': "smoothed"},

Corpus 5: {'tf': "log", 'idf': "inverse"},

Corpus 6: {'tf': "log", 'idf': "smoothed"}

1. From the data of (corpus 3 and corpus 4), (corpus 5 and corpus 6), the difference is the idf method. But the precision and recall are same or very similar. So we can know the method of idf has very little to do the precision and recall.

2. Form the data of (corpus 3 and corpus 5) and corpus (4 and corpus 6), the difference is the tf method. The precision and recall increase a lot. So we can know the method of tf has an influence on the precision and recall. If you use log tf method it will be better than raw.
   For corpus 3 and 5, it increase precision by 14.81% and recall by 14.82%. For corpus 4 and 6, it increase precision by 19.23% and recall by 19.23%.

3. From the data of corpus 1 and corpus 3, corpus 2and corpus 5, the difference is to use stemming and removing stopword or not. We can see the precision will increase a lot by using the stemming and removing stopword. For corpuse1 and corpus 3, it increase precision by 12% and recall by 8%. For the corpus 2 and 5, it increase precision by 10.72% and recall by 10.71%.

4. But for the corpus 2 and corpus4, the difference are whether they use the stemming or not and the tf method. Corpus 2 does not use stemming and use log tf method, is still higher than corpus 4 which use raw and stemming in precision and recall. So the tf method is more important than the stemming. Also we can see from the increasing percentage from point 2 and point3.

5. Conclusion:

   Stemming can give better precision (10-12 percent) and recall (8-11 percent) in information retrieval for short queries (1-3words) on short documents (500 words) than no stemming at all for languages as English. But tf log method is more important and idf method is not important.

   We are convinced that the cost in creating a stemmer is proportional to the gain when using the stemmer. Stemming is the right strategy to improve the precision and recall.