CS 6200 HW2

Zhuocheng Lin

June 10, 2020

1 Indexing Files

Stop words are removed in all indices.

Index Type	File Size
Stemming, Compressed	73.0MB (74,769KB)
Stemming, Uncompressed	152MB (156,656KB)
No Stemming, Compressed	77.5MB (79,447KB)
No Stemming, Uncompressed	158MB (161,864KB)

2 Model Performance

Average precision is shown in the table.

Index	TF-IDF	Okapi BM25	LM(Laplace)
Previous, Stemming	0.3478	0.3552	0.2638
Stemming, Compressed	0.3491	0.3564	0.2710
Stemming, Uncompressed	0.3491	0.3564	0.2710
No Stemming, Compressed	0.2382	0.2497	0.2001
No Stemming, Uncompressed	0.2382	0.2497	0.2001

3 Proximity Search

Add proximity to existing BM25 retrieval model and run the original queries. Related formula are shown below,

- Okapi BM25: $D = 84660, k_1 = 1.2, b = 0.5$

$$bm25(d,q) = \sum_{w \in q} \left[\log(\frac{D + 0.5}{df_w + 0.5}) \cdot \frac{tf_{w,d} + k_1 \cdot tf_{w,d}}{tf_{w,d} + k_1((1-b) + b \cdot \frac{len(d)}{avg(len(d))})} \cdot \frac{tf_{w,q} + k_2 \cdot tf_{w,q}}{tf_{w,q} + k_2} \right] \tag{1}$$

- Proximity Score: $\alpha = 0.01$

$$proximity(d,q) = \log(\alpha + \exp(-distance(d,q))) \tag{2}$$

- Final Score:

$$score = bm25(d,q) + proximity(d,q)$$

Note: $proximity(d,q)$ only exist when a document contains at least 2 query terms (no distance for only 1 query term)

Index	BM25 (no proximity)	BM25 (proximity)
Stemming	0.2255	0.2373
No Stemming	0.1813	0.1894