CS 6200 HW1

Zhuocheng Lin

June 4, 2020

1 Model Performance

Model Name	Mean Average Precision	Precision at 10	Precision at 30
ES built-in	0.3552	0.4800	0.4227
Okapi TF	0.2901	0.4680	0.3707
TF-IDF	0.3478	0.4800	0.4173
Okapi BM25	0.3552	0.4800	0.4227
Unigram LM with Laplace smoothing	0.2638	0.4640	0.3760
Unigram LM with Jelinek-Mercer smoothing (0.9)	0.2834	0.3920	0.3560

2 Pseudo-relevance Feedback

2.1 Procedure

- Retrieve the top $k = 30$ documents using *TF-IDF* model.
- Rules to identify interesting words in the documents:
  - Rule 1:
    $$\frac{tf}{dl} - \frac{ttf}{total}$$
    where tf is the term frequency in the top k documents, dl is first k document length in words, ttf is total term frequency in the collection, $total$ is the total term frequency of all terms in the collection.
  - Rule 2:
    $$\frac{cf_w}{k} - \frac{df}{D}$$
    where cf_w is the number of documents containing the term in the top k documents, k is the retrieved relevant documents, df is the number of documents containing the term in the collection, D is the number of documents in the collection.

2.2 Returned interesting terms

- Query 85: "allegations corrupt public officials"
  - Rule 1:
    corrupt, offici, investig, alleg, salina, govern, charg, parti, public, prosecutor, bribe, former, offic, mexico, ivanov, accus, gdlyan, chun, ligachev, case, attorney, indict, arrest, de, law, andropov, convict, probe, prison, scandal
  - Rule 2:
    corrupt, alleg, public, investig, former, offici, offic, bribe, convict, accus, parti, prosecutor, polit, top, case, law, govern, charg, deni, probe, arrest, senior, power, campaign, son, leader, obtain, connect, report, wrongdo
  - ES-Builtin:
    corrupt, alleg, public, offici, bribe, racket, ziyang, zhao, democraci, charg, parti, investig, communist, xiaop, accus, attorney, indict, deni, court, case, prosecutor, govern, law, agenc, polit, presid, make, privat, call, peopl, state, spokesman, leader, say, report, minist, militari
- Query 59: "weather caused least"
  - Rule 1:
    weather, servic, wind, flood, caus, cold, least, tornado, storm, area, high, nation, temperatur, record, inch, rain, snow, damag, power, peopl, river, mph, degre, north, near, forecast, death, school, today, thunderstorm
  - Rule 2:
    weather, caus, least, storm, wind, servic, area, forecast, south, nation, temperatur, rain, damag, mph, low, north, counti, near, coast, blame, flood, death, northern, hour, meteorologist, power, cold, inch, night, high
  - ES-Builtin:
    weather, caus, least, forecast, temperatur, rain, wind, cold, snow, thunderstorm, wheat, crop, damag, peopl, kill, injuri, accid, offici, problem, two, injur, polic, diseas, three, mile, death, soldier

2.3 Modified model performance

2.3.1 Original

Model Name	Mean Average Precision	Precision at 10	Precision at 30
ES built-in	0.3552	0.4800	0.4227
Okapi TF	0.2901	0.4680	0.3707
TF-IDF	0.3478	0.4800	0.4173
Okapi BM25	0.3552	0.4800	0.4227
Unigram LM with Laplace smoothing	0.2638	0.4640	0.3760
Unigram LM with Jelinek-Mercer smoothing (0.9)	0.2834	0.3920	0.3560

2.3.2 Method 1 (add first 5 words)

Model Name	Mean Average Precision	Precision at 10	Precision at 30
ES built-in	0.3636	0.4960	0.3973
Okapi TF	0.3237	0.4520	0.3813
TF-IDF	0.3549	0.4760	0.3867
Okapi BM25	0.3636	0.4960	0.3973
Unigram LM with Laplace smoothing	0.2956	0.4480	0.3747
Unigram LM with Jelinek-Mercer smoothing (0.9)	0.3395	0.4360	0.3667

2.3.3 Method 2 (add first 5 words)

Model Name	Mean Average Precision	Precision at 10	Precision at 30
ES built-in	0.3156	0.4720	0.3973
Okapi TF	0.2841	0.4400	0.3760
TF-IDF	0.3016	0.4360	0.3760
Okapi BM25	0.3171	0.4720	0.3973
Unigram LM with Laplace smoothing	0.2574	0.4400	0.3720
Unigram LM with Jelinek-Mercer smoothing (0.9)	0.2651	0.3680	0.3253

2.3.4 ES Built-in (add first 5 words)

Model Name	Mean Average Precision	Precision at 10	Precision at 30
ES built-in	0.1904	0.3320	0.2627
Okapi TF	0.2162	0.3760	0.2907
TF-IDF	0.1927	0.3280	0.2653
Okapi BM25	0.1906	0.3320	0.2600
Unigram LM with Laplace smoothing	0.2093	0.3520	0.2893
Unigram LM with Jelinek-Mercer smoothing (0.9)	0.1659	0.2520	0.2147

2.3.5 ES Built-in (add first 1 word)

Model Name	Mean Average Precision	Precision at 10	Precision at 30
ES built-in	0.2996	0.4680	0.3813
Okapi TF	0.2901	0.4640	0.3653
TF-IDF	0.2927	0.4720	0.3720
Okapi BM25	0.2997	0.4680	0.3813
Unigram LM with Laplace smoothing	0.2640	0.4680	0.3600
Unigram LM with Jelinek-Mercer smoothing (0.9)	0.2455	0.3720	0.3187