COMP0084 Coursework 1

Anonymous ACL submission

Abstract

None

1 Task 1: Text Statistics

1.1 D1

There are some text preprocessing method:

1. Parsing:
   - (a) Change the format of the file to raw text. No needed for the input file 'passage-collection.txt'.
   - (b) Identify structural elements: the contents in the input file are lines of sentences without titles, headings.
2. Normalization:
   - (a) Delete url from the document collection.
   - (b) All characters are converted to lowercase letters.
   - (c) Delete all non alpha characters.
3. Tokenization
   - (a) The input sentences are converted to tokens with the help of NLTK package.
4. Lemmatization
   - (a) Change a word to its base form.
5. Stemming
   - (a) Reduce inflected tokens to their root form.

Here we only choose Normalization and Tokenization to preprocess documents and queries. If all stop words kept, our algorithm returns 127512 tokens, that is, vocabulary size is 127512. The top 10 tokens with the highest occurrences in the collection are [('the', 626848), ('of', 334267), ('a', 283665), ('and', 255205), ('to', 240957), ('is', 216854), ('in', 202195), ('for', 108149), ('or', 86925), ('you', 86659)]. Figure 1 and Figure 2 display the occurrence probability and Zipf's distribution with $k = 1$ in different axis. The gap between empirical distribution and Zipf's distribution is small, especially when the rank is high. For the right end of the cuve, the Zipf's distribution is about 10 times larger than the empirical distribution, one reason could be that our tokens are too sophisticated.

Figure 3 and Figure 4 also show two distributions but the stop words are removed. In this case, the size of vocabulary is 127361 and the top 10 tokens with the highest occurrences are [('1', 43953), ('2', 33913), ('one', 27298), ('name', 25159), ('3', 22597), ('also', 21757), ('number', 21363), ('may', 20555), ('cost', 17127), ('used', 16540)]. According to the Zipf's law when $s = 1$, we have

$$f(k; s, N) = \frac{k^{-s}}{\sum_{i=1}^{N} i^{-s}} = \frac{1}{k \sum_{i=1}^{N} i^{-1}} \quad (1)$$

where N is the vocabulary size and k is the terms's frequency rank. If we remove stop words, that is, some tokens with a high frequency and high rank are excluded and thus for the current token with a high rank, its frequency is reduced, cause a low occurrence probability in Figure 3 and 4. For example, the term with rank 1, its probability drops by an order of magnitude from 10^{-1} to 10^{-2}.

2 Task 2: Inverted index

2.1 D3

We follow the method introduced in the task 1 and remove stop words in the file "passage-collection.txt" to generate our vocabulary and save it in the file "terms_removed.txt". We inverted index of terms in the passage in the file "candidate-passages-top1000.tsv" in the following steps:

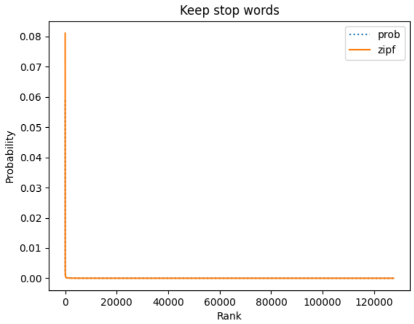

Figure 1: The occurrence probability against its occurrence rank for all terms and Zipf's distribution.

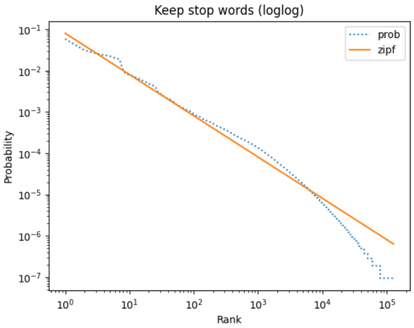

Figure 2: The occurrence probability against its occurrence rank for all terms and Zipf's distribution in loglog.

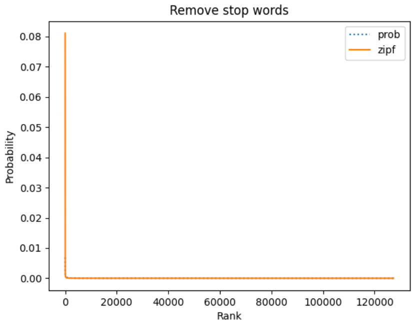

Figure 3: The occurrence probability against its occurrence rank for all terms and Zipf's distribution.

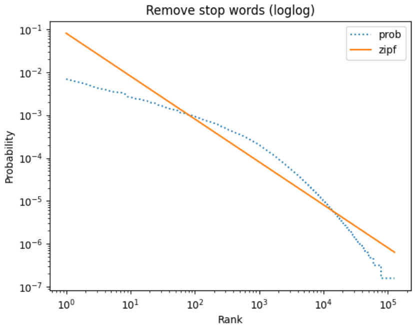

Figure 4: The occurrence probability against its occurrence rank for all terms and Zipf's distribution in loglog.

1. Create a dict to store our information. Keys are terms in "terms_removed.txt" and update values one by one.

2. Iterate each line in the file "candidate-passages-top1000.tsv". Processing the passage in the same way.

3. If a token in the passage also occurs in our vocabulary, get its occurrence in this passage and store occurrence with pid and qid into the previous dict.

4. Output a dict in the form "term 1":[(qid1,pid1,frequency), (qid1,pid1,frequency)], "term 2": ....

We can also return a result in a List type for creating a pd.DataFrame in the future.

3 Task 4: Query Likelihood Language Models

3.1 Compare of three models

We select some query-passage pairs with a high rank for each model.

3.1.1 qid=1108939: what slows down the flow of blood

Corresponding passage with the highest score.

Laplace: pid=2555774, score=-1497107
"Capable of undergoing vasoconstriction or vasodilation to influence blood flow and blood pressure"

ă Lidstone: pid=2068541, score=-24.69
"An aortic aneurysm can also lead to other problems. Blood flow often slows in the bulging

section of an aortic aneurysm, causing clots to form. If a blood clot breaks off from an aortic aneurysm in the chest area, it can travel to the brain and cause a stroke.Blood clots that break off from an aortic aneurysm in the belly area can block blood flow to the belly or legs.n aortic aneurysm can also lead to other problems. Blood flow often slows in the bulging section of an aortic aneurysm, causing clots to form. If a blood clot breaks off from an aortic aneurysm in the chest area, it can travel to the brain and cause a stroke."

Dirichlet: pid=3130232, score=-10.55
"Blood flow often slows in the bulging section of an aortic aneurysm, causing clots to form. If a blood clot breaks off from an aortic aneurysm in the chest area, it can travel to the brain and cause a stroke.Blood clots that break off from an aortic aneurysm in the belly area can block blood flow to the belly or legs.lood flow often slows in the bulging section of an aortic aneurysm, causing clots to form. If a blood clot breaks off from an aortic aneurysm in the chest area, it can travel to the brain and cause a stroke."

3.1.2 qid=1121986: what are the effects of having low blood sugar

Corresponding passage with the tenth highest score.

Laplace: pid=1188148, score=-1497113
An insulin overdose can cause low blood sugar levels (hypoglycemia), which can be quite dangerous. Some of the early symptoms of low blood sugar include:

Lidstone: pid=7518992, score=-33.91
Hypoglycemia refers to blood sugar levels that drop below the normal range. When blood sugar becomes too low, the body releases a hormone called epinephrine, which causes the body to release stored sugar into the blood. Epinephrine produces symptoms such as hunger, sweating, and shaking. As blood sugar drops even more, the body cannot get enough sugar to the brain, and additional symptoms develop due to the decrease in sugar to the brain. This causes dizziness, confusion, and weakness. As blood sugar continues to drop, and the brain does not have enough sugar to function properly, more severe effects occur, including permanent brain damage, seizures, coma, and death.

Dirichlet: pid=2899538, score=-14.74
Tresibaő may cause serious side effects that can be life-threatening, including: 1 Low blood sugar (hypoglycemia). Signs and symptoms that may indicate low blood sugar include anxiety, irritability, mood changes, dizziness, sweating, confusion, and headache. Low potassium in your blood (hypokalemia)

3.1.3 Analysis

Dirichlet Model is expected to be the better one among three models.
As Laplace Smoothing consider too much weight about unseen terms, the term "slow" does not occur in the passage 3130232. On the other, the vocabulary size (10^6) is much large than a normal query (10^1) thus the score of different passage for a query are controlled by unseen terms and has low variance.

$$P_{laplace}(w \mid D) = \frac{tf_{w,D} + 1}{|D| + |V|} \tag{2}$$

Lidstone model is an adapted model of Laplace model by changing 1 to a empirical parameter ε. What they do is add some weight to unseen weights and Lidstone model use a parameter to control the level of weights.

$$P_{lidstone}(w \mid D) = \frac{tf_{w,D} + \varepsilon}{|D| + \varepsilon|V|} \tag{3}$$

Dirichlet Model is a better estimator for short queries. It makes smoothing depends on sample size and works better in this case.

3.2 Similarity of models

Laplace model and Lidstone model are similar because they use the same method to share weights on the unseen words. Laplace model is a special case of Lidstone model ($\varepsilon = 1$). Both of them do not consider the statistical knowledge about the whole vocabularies.

3.3 Choice of ϵ in the Lidstone model

$\varepsilon = 0.1$ is a good choice in our data collection. In our collection, the size of vocabulary is about 10^6 and the average length of a passage is about 3×10^1. A small ε can reduce the weights on the unseen tokens.

3.4 Change μ to 5000 in Dirichlet Smoothing

$\lambda = \frac{N}{N+\mu}$, in our case, the average length of a passage is about 3×10^1. If we set $\mu = 5 \times 10^3$, the λ will reduce to 0 nearly, that is, we put too much confidence in background information and do not consider the information in a specific passage. $\mu = 5 \times 10^1$ could be a good parameter in this dataset.