COMP0084 Information Retrieval and Data Mining (2023/24)
Coursework 1

Anonymous ACL submission

1 Task 1

In this task, we obtained vocabularies of the entire passage collection. Passages are stored in the file `passage-collection.txt`. To obtain the vocabularies of each passage, a custom tokeniser was used. The tokeniser performs the following processes to split the text and format each token:

1. **Replacing non-alphanumeric**: Each character other than `a-z`, `A-Z` and `0-9` is replaced with a blank space.
2. **Splitting consecutive alphanumeric**: Words with consecutive alphanumeric are split. Hence, each token consist of only alphabets or numbers. For examples, "bm25" → ["bm", "25"], "25bm" → ["25", "bm"] and "bm25bm" → ["bm", "25", "bm"].
3. **Stemming**: Stemming convert each word into its root form, for example "meeting" and "meets" are both converted to "meet". The conversion is based on a set of rules. In this task, the Porter Stemming Algorithm. was adopted. The implementation of such algorithm in the Python package `gensim` was used. Note that the Porter stemmer also converts every token into its lowercase form.
4. **Split by whitespaces**: Tokens are obtained by splitting the processed text according to the whitespaces. These whitespaces include blank spaces, tabs and line breaks.

The index of terms (i.e. unique vocabularies) of the passage collection is therefore obtained. The size of of the index is reported in table 1. We also report the size of index after removing stopwords. Stopwords are commonly used words such as ’is’, ’am’ and ’are’, which are considered to contains low information about the text. We used the English stopwords list provided in `NLTK` for the removal.

Stopwords	Included	Excluded
Index Size	94363	94237

Table 1: Index size of the passage collection.

Furthermore, we verified if the probability of word occurrence f against it frequency ranking k follows the Zipf’s law defined as

$$f = \frac{1}{k \sum_{i=1}^{N} i^{-1}},$$

where N is the vocabularies size. The Zipf’s law suggests that the frequency times the rank of a term should be a constant $C = (\sum_{i=1}^{N} i^{-1})^{-1}$.

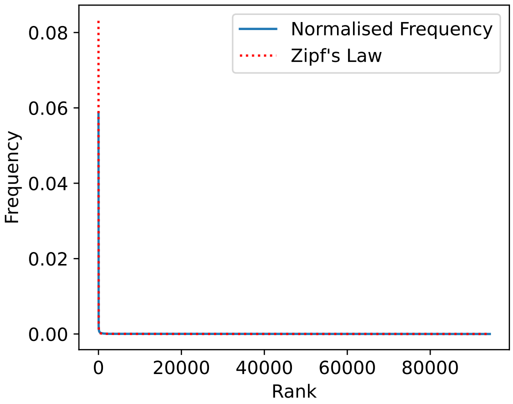

Figure 1: Empirical term frequency against frequency ranking (**including** stopwords) compared with Zipf’s law with parameter $s = 1$.

Figure 1 shows the comparison between the empirical and theoretical distribution. Figure 2 shows the same plot in log scale and provides a better comparison. From the log-log plot, it is noticed that the empirical curve deviates more from the theoretical one for terms with low frequency. This might be the effect of stemming, where different words are converted to the same form. Some rare forms of a word might have been stemmed to a common form during the tokenisation process, leading to the reduced frequency of rare words. This belief can be further confirm by Table 2, where the mean of frequency times ranking is much less than the expectation of Zipf's law. Hence, a large amount of terms have way less frequency as predicted by Zipf's law.

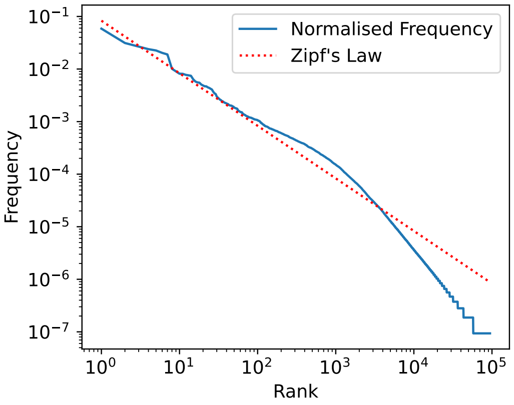

Figure 2: Empirical term frequency against frequency ranking (**including** stopwords) compared with Zipf's law with parameter $s = 1$ in log-log scale.

Expected constant C	0.083111
Mean of $f \times k$	0.018652 ± 0.024152

Table 2: Zipf's constant of the passage collection with stopwords **included**.

We also studied the effect of removing stopwords from the index and the result is reported in Figure 3. It is observed that the empirical curve become flatten at the higher rankings portion. As stopwords with usually high rankings are removed, less frequent terms move up the rankings and hence high rankings part are not as steep as before. Other other hand, around the ranking of order 10^3, the frequency went up much higher than the Zipf's frequency. The overall effect of removing stopwords is shown in Table 3. It is noticed that the empirical Zipf's constant becomes closer to the theoretical one. Though the standard deviation of the empirical mean is also larger, one may, in some sense, claim that the distribution after removing stopwords adheres more to Zipf's law.

Expected constant C	0.083120
Mean of $f \times k$	0.029006 ± 0.035261

Table 3: Zipf's constant of the passage collection with stopwords **excluded**.

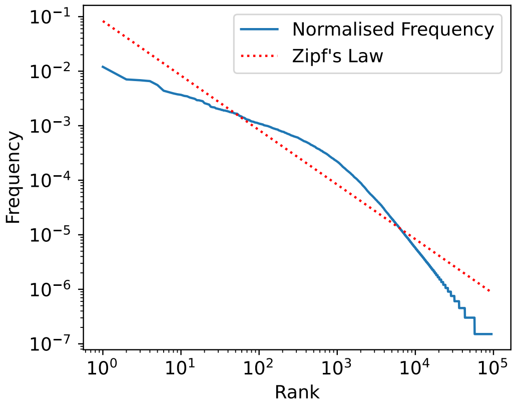

Figure 3: Empirical term frequency against frequency ranking (**excluding** stopwords) compared with Zipf's law with parameter $s = 1$ in log-log scale.

The index obtained in this task is stored as a python dictionary with the full passage string as key and a list of token string as value. The dictionary was used in later tasks.

2 Task 2

In this task, we built an inverted index for the passage collection. The index with stopwords **excluded** was used to create the inverted index as the distribution is more adhere to Zipf's law.

We first parsed the PID map of passage from the file candidate-passages-top1000.tsv. With this map, the inverted index is then constructed with the following format:

```
inverted_index = {
    <term> : {
        <pid> : <term_frequency>
    }
}
```

where the term was used as the key and the value was the passage's ID that contains the term and the term frequency in that passage. In addition to the inverted index, the length of each passage (number of tokens) was also stored in another map using the passage's ID as the key and the length as the value. The information stored in these two maps allows computing the following quantities for later tasks:

- Term frequency in each document;
- Term frequency in the corpus;
- Length of each document;
- Average length of documents in the corpus.

3 Task 3

In this task, we compute the top 100 candidate passages (some queries has less candidates) for each query provided in the file `test-queries.tsv`. See files `tfidf.csv` and `bm25.csv` for the query results obtained using TF-IDF vectors' cosine similarity and BM25 respectively.

4 Task 4

In this task, we repeat task 3 using Query Likelihood Language Model. In particular, we applied the following models:

- Laplace smoothing;
- Lidstone correction to Laplace smoothing with $\epsilon = 0.1$;
- Dirichlet smoothing with $\mu = 50$.

See the files `laplace.csv`, `lidstone.csv`, and `dirichlet.csv` for the corresponding results.

4.1 Discussion

- **Which language model do you expect to work better?**

  The Dirichlet smoothing is expected to work better than the other two models.

  All three models aim to ameliorate the problem of assigning zero probability to words that are in the vocabularies but not seen in the document. However, both Laplace smoothing and Lidstone correction would assign a constant probability to every unseen terms. This does not seem to be realistic as some unseen words might be more probable than others. Dirichlet smoothing addresses this issue by incorporating the vocabulary knowledge of the corpus, probabilities of unseen words in a document are assigned according to their occurrence frequency in the corpus.

  On the other hand, Dirichlet smoothing also considers the length of a document to determine the amount of smoothing. This extra adjustment is not designed in the other two models.

  Table 4 show the top 5 results of the query `"socioemotional process definition"` obtained by each model. In this example, the Laplace smoothing model and Lidstone correction model get the same top 5 passages. It is obvisous that the results of these two model are irrelevant to the query, as compared to the Dirichlet model. The Dirichlet model sucessfully assigned a high probability to the word `"socioemotional"`, which capture the key semantic meaning of the query text, as this word is much rare than `"process"` and `"definition"` in the collection. This example clearly shows the shortcoming of the other two models that they failed to incorporate the frequency of words in the entire collections.

- **Which language models are expected to be more similar and why?**

  The Laplace smoothing model and Lidstone correction model are expected to be similar, as the Lidstone model is a slight variant of the Laplace model. The Laplace model assign a non-zero probability to every term in the vocabularies, where the probability of a term that is not seen in a document is given by

  $$\frac{1}{D + |V|},$$

  where D is the length of the document and $|V|$ is the vocabularies size. The Lidstone correction adjusts the probability to

  $$\frac{\epsilon}{D + \epsilon|V|}$$

  where $\epsilon \in (0, 1)$ is a parameter for the correction. It aims to reduce the probabilities assigned to unseen term. Hence, if ϵ is set to one, the Lidstone correction model reduce to the Laplace model.

  As aforementioned, Table 4 show an example where both the Laplace smoothing model and Lidstone correction model perform poorly. Here, we show another example in Table 5, where these two models outperform the Dirichlet model. The query text in the example concerned is `"wat is dopamine"`. Here, the Dirichlet model is confused by the word `"wat"` which is rare in the collection but does not capture the key semantic of the query. Nonetheless, it is shown that the performance of the Laplace smoothing model and Lidstone correction model are similar.

- **Comment on the value of $\epsilon = 0.1$ in the Lidstone correction. Is this a good choice (yes/no), would there be a better setting (if so, please provide a range of values), and why?**

  The Lidstone correction assign probability to a term T given that it is seen or unseen in a document with the following ratio:

  $$\frac{P(T|\text{seen in doc})}{P(T|\text{unseen in doc})} = \frac{T_f + \epsilon}{\epsilon}$$

  where T_f is the frequency of term T in the document. Hence, the total probability assigned for unique terms in a document d over that assigned for unseen term are given by:

  $$\frac{L_d + N_d\epsilon}{(|V| - N_d)\epsilon}$$

  where L_d is the length of document d and N_d is the number of unique term in document d. Intuitively, one would like the ratio be greater than 1, such that more probability is assigned to terms in the document. Solving the inequality yields

  $$\epsilon < \frac{L_d}{|V| - 2N_d},$$

  suggesting that the unseen terms should be given a lower probability for a shorter document, such that each term in the document would contain more information. Replacing L_d with the average document length of our collection (36.3 tokens) and N_d with the average number of unique terms in passages (25.1 tokens), we obtain $\epsilon < 0.0004$.

  If we also take into account the fact that our queries set are short, with an average length of 5.8 tokens, matching terms between the query and the document become more important. Thus, ϵ should be even smaller.

  Therefore, $\epsilon = 0.1$ is not a good choice. It should be much smaller and in particular should be smaller than 0.0004.

- **If we set $\mu = 5000$ in Dirichlet smoothing, would this be a more appropriate value (yes/no), and why?**

  It is not appropriate to set $\mu = 5000$. As the parameter μ carry some sense of sampling an additional passage from the collection, it should be of the same scale as the average passage length in the collection. Using the developed index of the collection, the average passage length is 36.3 tokens. Hence, the original choice of $\mu = 50$ seems to be a better approach. However, in practice, one should choose the value of μ using a validation set and locate the optimal value.

Query: `socioemotional processes definition`
(Tokens: `socioemot`, `process`, `definit`)

Rank	Query results of Laplace smoothing or Lidstone correction with $\epsilon = 1$	Query results of Dirichlet smoothing
1	Related Flashcards. 1 PMP 47 Processes Definitions PMBOK 5t... 2 PMP 47 Processes Inputs/Outputs Tools... 3 PMP - 47 Processes Definitions. 4 PMBOK Guide, 5th ed, Ch 5 - Definitions. 5 PMP 47 Processes. 6 47 PMP PROCESSES. 7 PMBOK Guide, 5th ed: Earned *[...TRIMMED]*	The first concept includes the three developmental processes – biological, cognitive, and socioemotional —that are the interacting and overlapping processes that influence periods of development. 1 Biological processes are those processes that include changes in an individual's *[...TRIMMED]*
2	Adiabatic Processes in the Atmosphere. The traditional definition of an adiabatic process is one in which heat is neither added to nor removed from the system. In the atmosphere, the system is an air parcel. This definition is patently incorrect because heat is not a quantity, *[...TRIMMED]*	The second concept that helps provide a framework for understanding the complexities of human development is that of periods of development, which are produced by the interplay of the biological, cognitive, and socioemotional processes.
3	Homeostasis refers to metabolic balance maintained by several processes. The human body has several examples of homeostasis. Learning about these processes makes it easier to understand how the body maintains its normal functions. slide 1 of 7. First, let's start with the definition of homeostasis.Below is the medical definition from Merriam-Webster *[...TRIMMED]*	From Bales' Interaction Process Analysis System and Fisher's Decision Proposal Coding System, Poole proposes 36 clusters of group activities for coding group interactions and 4 cluster-sets: proposal development, socioemotional concerns, conflict, and expressions of ambiguity.
4	process management-Computer Definition. The execution and monitoring of repeatable business processes that have been defined by a set of formal procedures. See BPM and knowledge-driven process management.rocess management-Computer Definition. *[...TRIMMED]*	Words near socioemotional development in the dictionary. 1 socioeconomies. 2 socioeconomist. 3 socioeconomists. 4 socioeconomy. 5 socioemotional development. 6 socioethical. 7 socioevolutionary. 8 sociofugal. 9 sociofunctional. 10 sociogeneses.
5	A computer central processing unit (CPU) is the brain of the computer. This chip runs all processes on a computer. A CPU... 1 Definition of Processor *[...TRIMMED]*	The developmental period of transition from childhood to early adulthood; it involves biological, cognitive, and socioemotional changes.

Table 4: Top 5 query results obtained by each model for the query text "`socioemotional process definition`". Relevant words in the passages are highlighted with the corresponding color. Note that for this query, the top 5 results of Laplace smoothing model and the Lidstone correction model are exactly the same. Trimmed passage is indicted by *[...TRIMMED]* as the end.

Query: `wat is dopamine`

(Tokens: `wat`, `dopamin`)

Rank	**Query results of Laplace smoothing or Lidstone correction with $\epsilon = 1$**	**Query results of Dirichlet smoothing**
1	Dopamine and addiction. Cocaine and amphetamines inhibit the re-uptake of dopamine. Cocaine is a dopamine transporter blocker that competitively inhibits dopamine uptake to increase the presence of dopamine. Amphetamine increases the concentration of dopamine in the synaptic gap, but by a different mechanism.ocaine and amphetamines inhibit the re-uptake of dopamine. *[...TRIMMED]*	Royal temples Special class. Wat Phra Sri Rattana Satsadaram (Wat Phra Kaew), Bangkok; First class. Wat Phra Chetuphon Vimolmangklaram (Wat Pho), Bangkok; Wat Mahathat Yuwarajarangsarit, Bangkok; Wat Suthat Thepwararam, Bangkok; Wat Bowonniwet Vihara, Bangkok; Wat Rajapradit Sathitmahasimaram, Bangkok; Wat Rajabopit Sathitmahasimaram, Bangkok
2	A dopamine agonist is a compound that activates dopamine receptors in the absence of that receptor's physiological ligand, the neurotransmitter dopamine. Dopamine agonists activate signaling pathways through *[...TRIMMED]*	White adipose tissue (WAT) is composed of subcutaneous WAT and visceral WAT. The main functions of WAT have been described as storing and releasing fatty acids (FAs) that supply fuel to the organism during fasting periods.
3	Re-uptake is something that happens soon after a neuron transmits the dopamine meaning it actually sucks it all back. This is to prevent too much dopamine being transmitted from one neuron to another. Reuptake (by the neuron giving the dopamine) reduces the amount of dopamine in the brain. *[...TRIMMED]*	After the fall of Ayutthaya to the Burmese, King Taksin moved the capital to Thonburi where he located his palace beside Wat Arun on the opposite side of the river from Wat Pho, and the proximity of Wat Pho to this royal palace elevated it to the status of a wat luang (royal monastery).
4	Dopamine antagonists are most commonly used for psychiatric conditions. Photo Credit Purestock/Purestock/Getty Images. A dopamine antagonist is a chemical, medication or drug that prevents the actions stimulated by dopamine.Dopamine is *[...TRIMMED]*	Thus a wat chin is a Chinese temple (either Buddhist or Taoist), wat khaek is a Hindu temple and wat khrit or wat farang is a Christian church, though Thai *[...TRIMMED]*
5	Royal temples Special class. Wat Phra Sri Rattana Satsadaram (Wat Phra Kaew), Bangkok; First class. Wat Phra Chetuphon Vimolmangklaram (Wat Pho), Bangkok; Wat Mahathat Yuwarajarangsarit, Bangkok; Wat Suthat Thepwararam, Bangkok; Wat Bowonniwet Vihara, Bangkok; Wat Rajapradit Sathitmahasimaram, Bangkok; Wat Rajabopit Sathitmahasimaram, Bangkok	and wat if everyone else in the class hates the teacher and all our other teachers say were a good class and the teacher is only mean to our class wat do u do. Add your answer.nd wat if everyone else in the class hates the teacher and all our other teachers say were a good class and the teacher is only mean to our class wat do u do. Add your answer.

Table 5: Top 5 query results obtained by each model for the query text "`wat is dopamine`". Relevant words in the passages are highlighted with the corresponding color. Note that for this query, the top 5 results of Laplace smoothing model and the Lidstone correction model are exactly the same. Trimmed passage is indicted by *[...TRIMMED]* as the end.