COMP0084 Information Retrieval and Data Mining Coursework 1

Abstract

This coursework explores text representation techniques and retrieval models. We examine various pre-processing techniques, including tokenisation, normalisation, lemmatisation, and stemming, and investigate the application of Zipf's Law in natural language processing, analysing word frequency distributions and the impact of stopword removal.

A key focus of this work is the construction of an inverted index. We evaluate three query likelihood language models, including Laplace, Lidstone, and Dirichlet smoothing, comparing their performance in ranking retrieval results.

1 Introduction

Only unigram text representations are used to solve this coursework's tasks. Vector graphics format (.pdf) is used for all figures in this report.

2 Task-1 Text Statistics

2.1 Text Pre-processing

Four types of pre-processing choices: parsing, tokenisation, normalisation and lemmatisation and stemming, are described and justified with examples in the following section. In total, 10,284,445 tokens and vocabulary size of 115,698 have been identified without stop words removal.

2.1.1 Parsing

When the input file is not raw text (e.g., JSON, HTML, XML...), parsing becomes an essential pre-processing step. The purpose is to extract meaningful text and identify structural elements (e.g. titles, links, headings) while filtering out irrelevant or non-textual elements like hashtags or metadata.

2.1.2 Tokenisation

Tokenisation is the process of chopping up a document unit into pieces called tokens. Take the first sentence in passage-collection.txt as an example:

Sentence: This is the definition of RNA along with examples of types of RNA molecules.

Tokens:[This][is][the][definition] [of] [RNA] [along] [with] [examples] [of] [types] [of] [RNA] [molecules] [.]

However, raw tokens need further processing, including lowercasing, removing punctuation, handling contractions to become **terms**, which are stored in the vocabulary index.

Terms:this,is,the,definition,of,rna, along,with,examples,types,molecules

Tokenisation is complex in certain languages. In English, handling contractions (e.g., "won't") and hyphenated words (e.g., old-fashioned, glycerol-1-phosphate) can be challenging. In Chinese and Japanese, words lack space delimiters and require segmentation.

2.1.3 Normalisation

Normalisation involves canonicalising tokens, which means transforming tokens into a standard form. The goal is to group tokens that are slightly different but semantically equivalent (e.g. punctuation, diacritics, accents, hyphens).

colour ⇔ color   resumé ⇔ resume
won't ⇔ will not ⇔ [won]['t]

2.1.4 Lemmatisation and Stemming

Lemmatisation and stemming are techniques to reduce words to their base form. Stemming applies rule-based truncation ("running" → "run", "studies" → "studi") and is fast but may produce non-words. Lemmatisation uses a dictionary to map words to their correct lemma based on context ("running" → "run", "better" → "good"). It is more accurate but slower and resource-intensive.

Stemming suits simple tasks like search, while lemmatisation is preferred for applications requiring linguistic precision.

2.2 Zipf's Law

Zipf's Law states that in natural language, the frequency of a word is **inversely proportional** to its rank. This means that a few words (e.g., "the," "is," "of") occur very frequently, while most words appear rarely.

In Figure 1, the observed distribution exhibits a steep initial drop followed by a long tail, which is characteristic of Zipf's Law. The steep decline at the beginning confirms that high-ranking words dominate the dataset, while the flattening of the curve indicates a large vocabulary of low-frequency words.

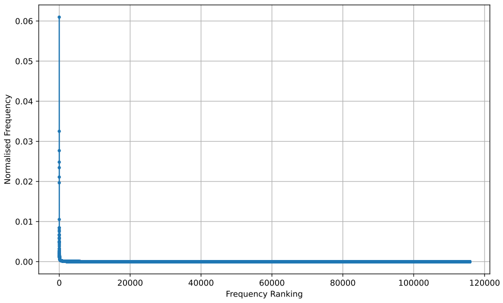

Figure 1: Normalised Frequency v.s. Ranking

To quantify the differences between the empirical distribution and the theoretical Zipf's law distribution, we refer to Equation 1.

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

The absolute error for each rank is computed and visualised in a log-log plot (Figure 2). This confirms that while mid-range frequencies align well with Zipf's prediction, the tail of the empirical distribution diverges from the expected $\frac{1}{k}$ trend.

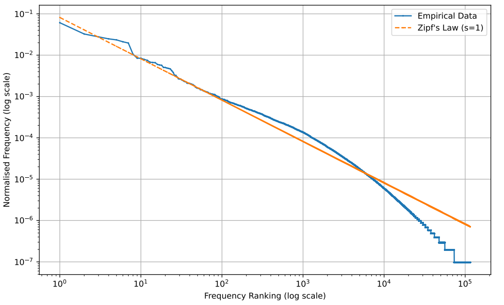

Figure 2: Empirical v.s. Zipf's law distribution

Under the normaslied Zipf's law, k represents the rank (where $k = 1$ is the most frequent word and $k = N$ is the least frequent), and s is typically close to 1 for natural language. When $s = 1$, Zipf's law predicts $f(k) \propto 1/k$. Empirically, the text follows this trend in the mid-range of ranks, but deviates in the tail (low-frequency words).

This deviation occurs due to two main factors:

1. **Statistical fluctuations**: Rare words appear infrequently, leading to higher variance in their observed frequencies. Since Zipf's expected frequency scales as $\frac{1}{k}$, words with high k are more sensitive to these fluctuations, leading to deviations from the theoretical curve.
2. **Finite sample effects**: The normalisation factor $\sum_{i=1}^{N} i^{-s}$ assumes a complete vocabulary, but in practice, some words predicted by Zipf's law are missing or underrepresented in the dataset, causing a drop below the expected $\frac{1}{k}$ trend.

2.3 Effects of Stopwords Removal

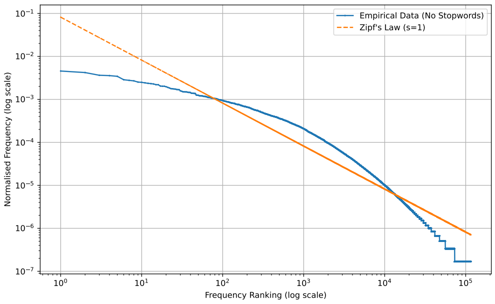

Figure 3: Empirical v.s. Zipf's law distribution with stop words removal

In Figure 3, the comparison is made after the removal of stop words. The vocabulary size reduced from to 115,698 to 115,547. To quantify the effect, the L_1 difference and log-scale Mean Squared Error (MSE) are computed, as shown in Table 1.

	L_1 Difference	Log-scale MSE
Stop Words	0.3450	2.943125
No Stop Words	0.7324	1.552659

Table 1: Comparison With and Without Stop Words

The L_1 norm measures the total absolute deviation between the two distributions. The increase in L_1 after stop word removal indicates that the empirical frequencies were more closely aligned with Zipf's law before their removal. In contrast, the lower value of Log-scale MSE after stop word removal suggests a better fit for mid-to-low-frequency words.

3 Task 2 – Inverted index

The inverted index was constructed by processing the passages from the `candidate-passages-top1000.tsv` file, where each passage is first tokenized after applying necessary text preprocessing (lowercasing, removal of non-alphabetic characters, selective stemming, and optional stop word removal). In the index, each unique term is mapped to a posting list

Model	Passage ID	Score	Passage Text
Laplace	3647358	-29.027364	*"Blood Flow. Blood flow refers to the movement of blood through the vessels from arteries to the capillaries and then into the veins. Pressure is a measure of the force that the blood exerts against the vessel walls..."*
	3899060	-29.432932	*"By placing a radioactivity counter over the head, one can measure the amount of blood flow into the brain. (See How Nuclear Medicine Works .) The cerebral blood flow study takes 20 to 30 minutes to perform. If there is no blood flow to the brain..."*
	7919347	-29.450423	*"...The rate, or velocity, of blood flow varies inversely with the total cross-sectional area of the blood vessels...The rate, or velocity, of blood flow varies inversely with the total cross-sectional area of the blood vessels."*
Lidstone	2068541	-23.563301	*"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 ..."*
	6707713	-23.670896	*"Best Answer: Blood vessels become damaged, you bleed (which washes out debris) and it fills the wound, the blood vessels constrict & blood flow slows down..."*
	5553584	-23.737568	*"Esophageal varices are swollen blood vessels in the esophagus (swallowing tube), the tube that connects the mouth to the stomach. Esophageal varices may appear in people with serious liver disease. Esophageal varices occur when normal blood flow to your liver is slowed..."*
Dirichlet	5553585	-10.132441	*"What causes esophageal varices? Esophageal varices occur when normal blood flow to your liver is slowed. Liver disease may create scar tissue..."*
	6980266	-10.243578	*"Causes. About 25% of adults with cirrhosis have portal vein thrombosis, usually because blood flow through the severely scarred liver is slow. When blood flow is slow, blood is more likely to clot. Any condition that makes blood more likely to clot can cause portal vein thrombosis."*
	5163856	-10.362413	*"1 This means that more blood rubs against the walls of the vessel and it slows blood flow. 2 In any one capillary, this resistance is an advantage because the slowed blood flow has more time for gas exchange to occur. 3 When an arteriole dilates, the diameter almost doubles."*

Table 2: Top three retrieval results for a sample query using different smoothing methods.

containing detailed information: for each passage in which the term appears, the index stores both the frequency of the term and the exact positions (token offsets) at which it occurs.

This design was chosen to support efficient and flexible retrieval, as it not only enables the rapid identification of documents containing a query term but also provides the positional data required for more advanced operations such as phrase queries and proximity-based ranking. By including both term counts and positional information, the index offers richer metadata that can be leveraged by subsequent retrieval models (e.g. TF-IDF, BM25, or query likelihood models) to improve scoring and ranking. The complete index is saved in `JSON` format, ensuring that it is both easily accessible and interoperable with other components of the system.

4 Task 4 - Query likelihood language models

4.1 Better Language Model

In our experiments, the Dirichlet-smoothed query likelihood model consistently returned higher (less negative) log–likelihood scores compared to the Laplace and Lidstone models. For instance, as shown in Table 2, the top three scores for Dirichlet smoothing (`-10.132441`, `-10.243578`, `-10.362413`) are much higher than Laplace (`-29.027364`, `-29.432932`, `-29.450423`) and Lidstone (`-23.563301`, `-23.670896`, `-23.737568`). This indicates that Dirichlet smoothing is less "surprised" by the presence of rare or specialized query tokens because it dynamically adjusts the amount of smoothing based on both document length and the overall collection frequency. This means that for longer documents (more evidence), the model relies more on the document counts, whereas for shorter documents it "backs off" more to the collection model.

Laplace Example (PID 7919347). By adding 1 to every term count, Laplace over-smooths passages, therefore, rare but discriminative phrases are not given sufficient weight, and the model may end up overemphasizing frequent patterns and repeating similar content. This results in redundant retrievals where the uniqueness of the document content is lost.

Lidstone Example (PID 5553584). Even though Lidstone uses a smaller additive constant ($\epsilon = 0.1$), it still applies the same uniform increment across all tokens. This weakens the effect of any single low-frequency term and can result in repeated or overly similar passages among the top retrievals. As a consequence, common background terms—such as "oesophageal varices" in this context—can dominate the score, leading to repetitive retrievals.

Dirichlet Example (PID 5163856). In contrast, Dirichlet smoothing allows certain specialized tokens to remain influential. It scales the smoothing to each document's length and to the rarity of terms in the collection, thus preserving more unique or context-specific information. This yields more varied and precise passages in the top results, avoiding the over-smoothing pitfalls seen with Laplace and Lidstone.

The complete passage text is appended as Table 3 at the end of the report.

4.2 Similar Language Models

Laplace and Lidstone smoothing are expected to yield more similar outcomes. Both methods work by adding a constant to term counts to avoid zeros—Laplace adds one, while Lidstone adds a fractional constant (ϵ). When ϵ is near one, the two methods essentially become equivalent. In our experiments, for general queries with frequent terms, the rankings produced by Laplace and Lidstone smoothing showed a high degree of overlap. For example, the top 100 retrieved passages for many queries were nearly identical under both methods, whereas Dirichlet often produced a different ranking because it incorporates the collection frequency in a more nuanced way.

4.3 More Assumptions

Using $\epsilon = 0.1$ is a moderate choice in the Lidstone correction. It avoids the extreme over-smoothing of add-one (Laplace) while still ensuring that unseen events receive a nonzero probability. However, the optimal value of ϵ is data–dependent. In corpora with very large vocabularies or very short documents, a smaller value ($\sim$0.01 to 0.05) might be preferable to avoid giving too much probability mass to unseen words. Conversely, in more uniform collections, a value closer to 0.1 might be adequate. Thus, while $\epsilon = 0.1$ is reasonable for our experiments, tuning on a development set might reveal that a slightly lower value could further improve discrimination.

Setting $\mu = 5000$ would generally be too high for our candidate passages. In Dirichlet smoothing, μ controls the balance between the document model and the collection model; a larger μ means the model relies more on the collection's statistics. For our dataset—where candidate passages are of moderate length—a μ value in the tens or low hundreds (like 50 or 100) typically provides a good balance. With $\mu = 5000$, nearly all documents would be "smoothed" to look like the overall collection, which would reduce the model's ability to distinguish between relevant and non–relevant passages.

Model	Passage ID	Score	Passage Text
Laplace	3647358	-29.027364	*“Blood Flow. Blood flow refers to the movement of blood through the vessels from arteries to the capillaries and then into the veins. Pressure is a measure of the force that the blood exerts against the vessel walls as it moves the blood through the vessels.Like all fluids, blood flows from a high pressure area to a region with lower pressure. Blood flows in the same direction as the decreasing pressure gradient: arteries to capillaries to veins.The rate, or velocity, of blood flow varies inversely with the total cross-sectional area of the blood vessels. As the total cross-sectional area of the vessels increases, the velocity of flow decreases.Blood flow is slowest in the capillaries, which allows time for exchange of gases and nutrients. Resistance is a force that opposes the flow of a fluid.ressure is a measure of the force that the blood exerts against the vessel walls as it moves the blood through the vessels. Like all fluids, blood flows from a high pressure area to a region with lower pressure. Blood flows in the same direction as the decreasing pressure gradient: arteries to capillaries to veins.”*
	3899060	-29.432932	*“By placing a radioactivity counter over the head, one can measure the amount of blood flow into the brain. (See How Nuclear Medicine Works .) The cerebral blood flow study takes 20 to 30 minutes to perform. If there is no blood flow to the brain as demonstrated by this study, the brain is dead.A negative cerebral flow study is indisputable evidence of a dead brain. Normal cerebral blood flow study showing cranial space filled with blood. Cerebral blood flow study showing no blood entering the brain.Another confirmatory test is chemical: The patient can be given 1 mg of atropine IV.In the patient with an intact brain, atropine will dramatically increase the patient’s heart rate. In a brain-dead patient, atropine will not influence heart rate.y placing a radioactivity counter over the head, one can measure the amount of blood flow into the brain. (See How Nuclear Medicine Works .) The cerebral blood flow study takes 20 to 30 minutes to perform. If there is no blood flow to the brain as demonstrated by this study, the brain is dead.”*
	7919347	-29.450423	*“Blood Flow. Blood flow refers to the movement of blood through the vessels from arteries to the capillaries and then into the veins. Pressure is a measure of the force that the blood exerts against the vessel walls as it moves the blood through the vessels.Like all fluids, blood flows from a high pressure area to a region with lower pressure. Blood flows in the same direction as the decreasing pressure gradient: arteries to capillaries to veins.The rate, or velocity, of blood flow varies inversely with the total cross-sectional area of the blood vessels. As the total cross-sectional area of the vessels increases, the velocity of flow decreases.lood flows in the same direction as the decreasing pressure gradient: arteries to capillaries to veins. The rate, or velocity, of blood flow varies inversely with the total cross-sectional area of the blood vessels.”*

Continued on next page

Model	Passage ID	Score	Passage Text
Lidstone	2068541	-23.563301	*"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."*
	6707713	-23.670896	*"Best Answer: Blood vessels become damaged, you bleed (which washes out debris) and it fills the wound, the blood vessels constrict & blood flow slows down. Platelets in the blood stick to collagen fibres that make up the vessel wall and this acts as a plug.From this point on enzymes (proteins called clotting factors) trigger a chain of events takes place to form a clot: Prothrombin converts to thrombin, calcium is required for this.est Answer: Blood vessels become damaged, you bleed (which washes out debris) and it fills the wound, the blood vessels constrict & blood flow slows down. Platelets in the blood stick to collagen fibres that make up the vessel wall and this acts as a plug."*
	5553584	-23.737568	*"Esophageal varices are swollen blood vessels in the esophagus (swallowing tube), the tube that connects the mouth to the stomach. Esophageal varices may appear in people with serious liver disease. Esophageal varices occur when normal blood flow to your liver is slowed. Liver disease may create scar tissue in the liver which slows the flow of blood. When the blood to your liver is slowed, it begins to back up, leading to an increase of pressure in the major vein (portal vein) that carries blood to your liver."*
Dirichlet	5553585	-10.132441	*"What causes esophageal varices? Esophageal varices occur when normal blood flow to your liver is slowed. Liver disease may create scar tissue in the liver which slows the flow of blood. When the blood to your liver is slowed, it begins to back up, leading to an increase of pressure in the major vein (portal vein) that carries blood to your liver."*
	6980266	-10.243578	*"Causes. About 25% of adults with cirrhosis have portal vein thrombosis, usually because blood flow through the severely scarred liver is slow. When blood flow is slow, blood is more likely to clot. Any condition that makes blood more likely to clot can cause portal vein thrombosis."*
	5163856	-10.362413	*"1 This means that more blood rubs against the walls of the vessel and it slows blood flow. 2 In any one capillary, this resistance is an advantage because the slowed blood flow has more time for gas exchange to occur. 3 When an arteriole dilates, the diameter almost doubles."*

Table 3: Complete top three retrieval results for a sample query using different smoothing methods.