

Project #2 for the Biomedical Information Retrieval Course

1. Project Overview

在 Information Retrieval 系統中，文字資料並不是平均分布的。少數詞彙會出現非常頻繁，而大量詞彙只出現少數幾次，這種現象與 Zipf's Law 密切相關。

本專案的目標，是利用約 1,000 篇 scientific abstracts 建立一個小型文字語料庫，實際分析：

Does the word-frequency distribution of a scientific document collection follow Zipf's Law?

你不能只引用教科書的結論，而必須透過實驗資料回答這個問題。

本專案也將進一步探討：

- Term Frequency
- Collection Frequency
- Document Frequency
- Vocabulary
- Zipf Distribution
- Log-Log Analysis
- Text Preprocessing
- TF-IDF

2. Learning Objectives

完成本專案後，你應該能夠：

1. 建立一個小型 scientific document collection。
2. 將非結構化文字轉換成可分析的 terms。
3. 計算 Collection Frequency (CF) 與 Document Frequency (DF)。
4. 建立 Rank-Frequency distribution。
5. 使用 log-log plot 分析 Zipf's Law。
6. 使用 regression 估計 Zipf exponent。
7. 分析 preprocessing 對 Zipf distribution 的影響。
8. 解釋 Zipf distribution 與 Information Retrieval 的關係。
9. 分析 CF、DF 與 TF-IDF 的關係。
10. 從實驗結果提出合理的研究性解釋。

Background: Zipf's Law

假設一個 collection 中的 terms 按照出現頻率由高到低排序：

Rank Term Frequency

1 the 8,532

2 of 6,921

3 and 5,832

4 data 4,120

... ...

Zipf's Law 可以簡化表示為：

$$f(r) \propto \frac{1}{r^k}$$

其中：

- $f(r)$: rank r 的 term frequency
- r : term rank
- k : Zipf exponent

理想化的 Zipf distribution 通常具有：

$$k \approx 1$$

取 logarithm 後：

$$\log f(r) = \log C - k \log r$$

因此，如果資料近似符合 Zipf's Law，在 log-log plot 上通常會呈現近似線性的關係。

Dataset

本專案使用約 1,000 篇英文 scientific abstracts。

資料來源可以是：

- PubMed
- Kaggle scientific abstract datasets

Dataset Requirements

你的 collection 必須：

- 約 1,000 篇 documents
- 以英文為主
- 每篇至少包含 abstract
- 每篇具有唯一 Document ID (PMID)

Implementation:

Part I — Text Preprocessing

首先建立你的 preprocessing pipeline

至少實作：

Step 1 — Tokenization

Step 2 — Case Folding 例如：Data, data, DATA 轉換成：data

Step 3 — Punctuation Handling 處理：data, data. data; 等 punctuation

Step 4 — Stopword Handling 建立兩種版本：保留 stopwords 移除 stopwords

Step 5 — Stemming or Lemmatization 至少選擇一種：Stemming or Lemmatization

Part II. — Build the Vocabulary 建立 collection vocabulary 請說明：

- Number of documents
- Total number of tokens
- Number of unique terms
- Average tokens per document

Part III — Collection Frequency

對每一個 term 計算：

$$CF(t) = \text{number of times term } t \text{ occurs in the entire collection}$$

將 terms 按照 CF 由高至低排序，至少列出 Top 50 terms

Part IV — Document Frequency

另外計算： $DF(t) = \text{number of documents containing term } t$

Part V — Plot the Zipf Distribution

Experiment 1 — Rank vs Frequency

請使用 appropriate scale，使整個 distribution 能夠清楚呈現

Experiment 2 — Log-Log Plot

建立第二張圖：

進行 linear regression：

$$\log(CF) = a - b \log(r)$$

報告：slope, intercept, Zipf exponent, R^2 , RMSE

Required Analysis 請回答以下問題

Question 1 Does your document collection follow Zipf's Law?

Question 2 What is your estimated Zipf exponent?

Question 3 Is R^2 sufficient to prove that the distribution follows Zipf's Law? 請說明你的理由

Question 4 Which portion of the rank-frequency curve fits Zipf's Law best?

(請觀察 High-frequency terms, Middle-frequency terms, Low-frequency terms)

Part VI — Effect of Preprocessing

至少比較以下四種條件：

Condition A — Basic preprocessing (Tokenization, Lowercase)

Condition B — Remove punctuation

Condition C — Remove stopwords

Condition D — Lemmatization or stemming

Analysis Question 回答：How does preprocessing affect the Zipf distribution?

請至少討論：

1. Vocabulary size
2. High-frequency terms
3. Zipf exponent
4. Shape of the distribution

Part VII — CF vs DF

選擇至少 20 個 terms，比較：

$CF(t)$

與

$DF(t)$

例如：

Term CF DF

data

model

algorithm

Term CF DF

cancer

learning

回答：

Question 1 Why can a term have a high CF but relatively low DF?

Question 2 Which measure tells us whether a term is widely distributed across documents?

Question 3 Why are CF and DF both important in Information Retrieval?

Part VIII — Connection to TF-IDF

Information Retrieval 常使用 inverse document frequency：

$$idf(t) = \log \frac{N}{df(t)}$$

其中：

- N ：number of documents
- $df(t)$ ：document frequency

請選擇至少 10 個 terms，計算：

$$IDF(t)$$

建立：

Term CF DF IDF

回答：Why do very common terms generally receive lower IDF values?

並說明 Zipf distribution 與 TF-IDF weighting 之間的關係。

Part IX — Information Retrieval Implications

請撰寫約 300–500 words：Why does Zipf's Law matter to an Information Retrieval system?

至少討論以下三個議題：

1. Stopword Removal

為什麼大量高頻詞可能對 retrieval effectiveness 沒有太大的 discriminative value？

2. Inverted Index

Zipf distribution 對：

- Dictionary size
- Postings lists

- Index storage
- Index compression

有什麼影響？

3. TF-IDF

為什麼高頻 term 通常具有較低的 discriminative power？

Optional Challenge — Compare Two Domains

如果你希望進一步挑戰自己，可以將 corpus 分成兩個不同領域。

例如：

500 Medical abstracts

500 Computer Science abstracts

比較：

Measure	Domain A	Domain B
---------	----------	----------

Documents		
-----------	--	--

Tokens		
--------	--	--

Vocabulary		
------------	--	--

Zipf exponent		
---------------	--	--

R^2		
-------	--	--

Top 10 terms		
--------------	--	--

回答：Does domain specificity affect the Zipf distribution? 請使用實驗結果支持你的觀察。

Final Challenge

最後請思考：

If Zipf's Law is only an approximate empirical property rather than an exact law of language, why is it still useful for Information Retrieval?

這不是要求你找到唯一正確答案。

你應該利用：

- 實驗結果
- 課堂內容
- 教科書
- Information Retrieval concepts

提出自己的 technical argument

One-Page Executive Summary

最後請另外提交一頁 Executive Summary 回答以下問題：

What did you learn from analyzing 1,000 scientific abstracts that you could not learn simply by reading the textbook?

建議包含：

Finding 1 你的 corpus 是否符合 Zipf's Law ?

Finding 2 你的 Zipf exponent 是多少 ?

Finding 3 Preprocessing 對結果造成什麼影響 ?

Finding 4 這個結果對 Information Retrieval 有什麼意義 ?

Research Questions 你的 final report 必須至少回答以下核心問題：

RQ1 Does the frequency distribution of terms in the 1,000 scientific abstracts follow Zipf's Law?

RQ2 What is the empirical Zipf exponent of the corpus?

RQ3 How does text preprocessing affect the observed Zipf distribution?

RQ4 What is the relationship between Collection Frequency and Document Frequency?

RQ5 Why is Zipf's Law relevant to Information Retrieval?