Session 1: Introduction, Preprocessing, Text Statistics

CAIM: Cerca i Anàlisi d'Informació Massiva

Exercise list, Fall 2025

Basic Comprehension Questions. Make sure you can answer them before proceeding.

1. Tell five Information Retrieval Systems you frequently use.
2. Tell the typical sequence of transformations we apply to a text while preprocessing and before adding to the index.
3. Tell the difference between stemming and lemmatizing.
4. Zipf's law tells the relation between X and Y. What are X and Y?
5. Heaps' law tells the relation between X and Y. What are X and Y?

Exercise 1

Guess (without using any software) what a text preprocessor could give on this text if it performs stopword removal and stemming:

```
We found my lady with no light in the room but the reading-lamp.
The shade was screwed down so as to over-shadow her face. Instead of looking
up at us in her usual straightforward way, she sat close at the table, and
kept her eyes fixed obstinately on an open book.

"Officer," she said, "it is important to the inquiry you are conducting
to know beforehand if any person now in this house wishes to leave it?"
```

(William Wilkie Collins, The Moonstone, Chapter 16)

Exercise 2

Suppose that our document retrieval system lets us enter a query (a set of words), and it returns the set of documents containing **all** the words in the query.
We configure the system in four modes and ask the same query each time:

- **Mode 1:** No stopword removal, no stemming (docs and queries). Let result = `A1`.
- **Mode 2:** No stopword removal, but stemming applied (docs and queries). Let result = `A2`.
- **Mode 3:** Stopword removal, no stemming. Let result = `A3`.
- **Mode 4:** Stopword removal + stemming (docs and queries). Let result = `A4`.

What relations can you prove among `A1`, `A2`, `A3`, `A4`? (e.g., Is `A1 = A2`? Is `A2` $\subseteq$ `A4`?)

Exercise 3

We have a document collection with total of **N word occurrences** (N large). The collection follows Zipf's law:

$$frequency = c \cdot rank^{-\alpha}$$

1. What is c if $\alpha = 2$?
2. What is c if $\alpha = 1$?
3. Assume $\alpha = 2$. What is the frequency of the most common term?
4. What is the frequency of the 100th most frequent term?
5. Roughly, how many words have frequency 1?

Exercise 4

We have a document collection with a total of 10^6 term occurrences. Suppose terms follow a power law:

$$f_i \approx \frac{c}{(i+10)^2}$$

Estimate: (1) Number of occurrences of the most frequent term. (2) Number of occurrences of the 100th most frequent term. (3) Number of words occurring more than 2 times.

Hint: $\sum_{i=11}^{\infty} \frac{1}{i^2} \approx 0.095$

Exercise 5

A random sample of **10,000 documents** is taken from a collection of **1,000,000 documents**. In the sample, we count **5,000 distinct words**. Assuming the collection follows **Heaps' law** with exponent 0.5, give a reasoned estimate of the total number of different words in the whole collection.

Exercise 6

Derive **Heaps' law** from **Zipf's law**: - Let a collection have N word occurrences, with frequency f_i of the *i-th most common word* proportional to $i^{-\alpha}, \alpha > 1$.
- Figure out the proportionality constant.
- Estimate the rank i such that $f_i < 1$.
- Explain why this rank is roughly the number of distinct words expected in the collection.
- Deduce that this number is $k \cdot N^{\beta}$.
- Give k and β as functions of α.