

## Session 1: Introduction. Preprocessing. Text Statistics Exercise List, Fall 2021

---

**Basic comprehension questions.** Check that 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, which is a set of words, and returns the set of documents that contain *all* the words in the query.

Imagine that we configure the system in four different modes, and we ask four times the same query.

- Mode 1: We don't remove stopwords and we don't stem neither documents nor queries. Let  $A_1$  be the set of returned documents.
- Mode 2: We don't remove stopwords, but we stem both documents and queries. Let  $A_2$  be the set of returned documents.
- Mode 3: We remove stopwords, but don't stem. Let  $A_3$  be the set of returned documents.
- Mode 4: We remove stopwords, and then we stem both documents and queries. Let  $A_4$  be the set of returned documents.

What relations can you prove among  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$ ? For example, is  $A_1 = A_2$ ? Is  $A_2$  a subset of  $A_4$ ?, etc.

## Exercise 3

We have a document collection with a total of  $N$  word occurrences ( $N$  is large). We are told that it follows a Zipf's law of the form  $frequency = c \cdot rank^{-\alpha}$ .

1. What is  $c$  if  $\alpha = 2$ ?
2. And if  $\alpha = 1$ ?
3. Assume again  $\alpha = 2$ . What is the frequency of the most common term?
4. And what is the frequency of the 100th most frequent term?
5. And (roughly) how many words have frequency 1?

### Exercise 4

We have a document collection with a total of  $10^6$  term occurrences. Supposing that terms are distributed in the texts following a power law of the form

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

give estimates of (1) the number of occurrences of the most frequent term; (2) the number of occurrences of the 100-th most frequent term; (3) the number of words occurring more than 2 times. *Hint:*  $\sum_{i=11}^{\infty} \frac{1}{i^2} \cong 0.095$ .

### Exercise 5

We are given a random sample of 10,000 documents from a collection containing 1,000,000 documents. We count the different words in this sample, and we find 5,000. Supposing that the collection satisfies Heaps' law with exponent 0.5, give a reasoned estimate of the number of different words you expect to find in the whole collection.

### Exercise 6

Let us deduce Heaps' law from Zipf's law.

- Let a collection have  $N$  word occurrences, with the frequency  $f_i$  of the  $i$ -th most common word proportional to  $i^{-\alpha}$ ,  $\alpha > 1$ .
- Figure out (from previous exercises) the proportionality constant.
- Estimate the rank  $i$  such that  $f_i$  is likely to be less than 1.
- Explain why this should roughly be the number of distinct words we expect to see in the collection.
- Deduce that this number is  $k \cdot N^\beta$ . Tell the values of  $k$  and  $\beta$  as a function of  $\alpha$ .

[Note: The given formulation of Zipf's law cannot, for obvious reasons, be taken too literally: If for some large  $i$  we have  $c \cdot i^{-\alpha} = 0.03$ , it makes no sense to say that the  $i$ th word appears 0.03 times in the collection. More abstractly, one could imagine texts generated by some random process which assigns probability  $P(w)$  to the event that a random position in the text contains the word  $w$ . Then the word with rank 1 is the  $w$  with highest  $P(w)$ , etc. Zipf's law is a statement about the form of the probability distribution  $P$ . One can then compute rigorously the expected number of distinct words in

a text of length  $N$  according to this probabilistic model. Let us just say that we this way we obtain the same  $\beta$  but a different  $k$ .]

[Note 2: It is also possible but a bit more involved to deduce a power law for word frequencies (generalizing Zipf's law) from Heap's law]