Worksheet 7: Document Analysis

Statistical Analysis and Document Mining

MSIAM 1st year / ENSIMAG 2nd year

Question 1

You are willing to analyse a document containing 1,000,000 words.

1. Let $k = 10$, $\beta = 0.5$. Following the Heaps' law, what is the number of distinct words in the document?
2. It was found that the 7% of all words are "the" article. Following the Zipf's law, estimate the value of λ. What is the frequency of the second most frequent word? The third most frequent one?

Question 2

We have the following pre-processed training documents:

- d_1 : {"cat", "sit", "mat", "cat", "jump", "bed", "cat", "good", "sit"}
- d_2 : {"cat", "dog", "jump", "sit", "dog", "cat", "animal"}
- d_3 : {"table", "sit", "write", "think", "good", "book", "dog", "sit"}

1. Can you guess the context of each document?
2. What is the tf-idf weights for the following words: "cat", "dog", "sit", "animal", "book"?
3. Comment the question above.

Question 3

1. Represent in the compressed sparse row (CSR) format the following matrix:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 0 & 0 & 4 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 2 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 3 & 0 & 1 & 1 \end{pmatrix}.$$

2. Write a pseudo-code to multiply a CSR sparse matrix A by a (dense) vector b.