01. Lexical Statistics

Computational Methods for Text Analysis

Пестова Алена

НИУ ВШЭ Санкт-Петербург

12 сентября 2022 г.

Assessment

$$0.3 * \text{Final project} + 0.3 * \text{Homeworks} + 0.4 * \text{In-class participation} \tag{1}$$

Topics

- Text Preprocessing (preparing text for analysis)
- Contrastive Analysis (comparing texts
- Text Classification
- Word Embeddings (presenting word as a vector, how can we do that and why)
- (*) Language Models

Introduction

Computational Linguistics - is a field at the intersection of applied linguistics and computer science.
Tasks: automated processing of text. Very different tasks and methods - from counting words to text generation.
In our course: methods of text analysis that can help to extract some information from texts for further research.

How to count words?

If we want to estimate the word frequency distribution, we need to calculate the number of occurrences of each word (token) in the text.
Tokenization - dividing text into tokens/words.
Questions:

- What is a token? (What to count and what not to count?)
- Which tokens are considered as the same word?

Tokenization

How many tokens?

Ой какие фотки<smile006><smile006><smile006>
А разве роды в 38недель не считаются нормой?

Tokenization

11? (divide by spaces)

Ой какие фотки<smile006><smile006><smile006> А разве роды в 38недель не считаются нормой?

Tokenization

11? (take only words)

Ой какие фотки <smile006><smile006><smile006> А разве роды в 38 недель не считаются нормой ?

Tokenization

13? (also considering punctuation)

Ой какие фотки <smile006><smile006><smile006> А разве роды в 38 недель не считаются нормой ?

Tokenization

14? (delete the typo)

Ой какие фотки <smile006><smile006><smile006> А разве роды в 38 недель не считаются нормой ?

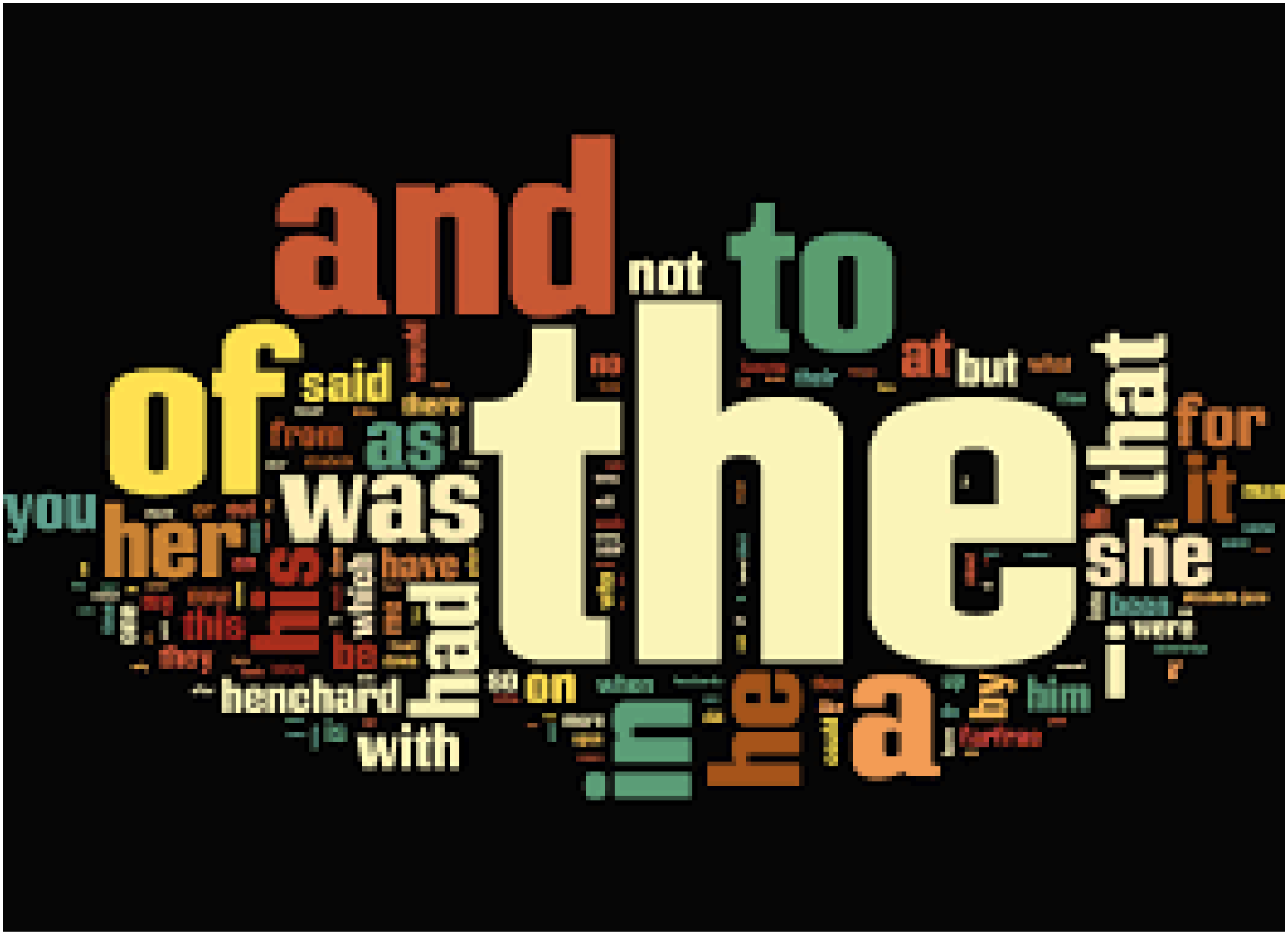
Tokenization

16? (count smiles separately)

Ой какие фотки <smile006> <smile006> <smile006> А разве роды в 38 недель не считаются нормой ?

Words Frequency

Words Frequency



Zipf's Law

Zipf's Law (1949) predicts the frequency of the word by irs rank in the frequency list:

$$f(w) = \frac{C}{r(w)^a} \tag{2}$$

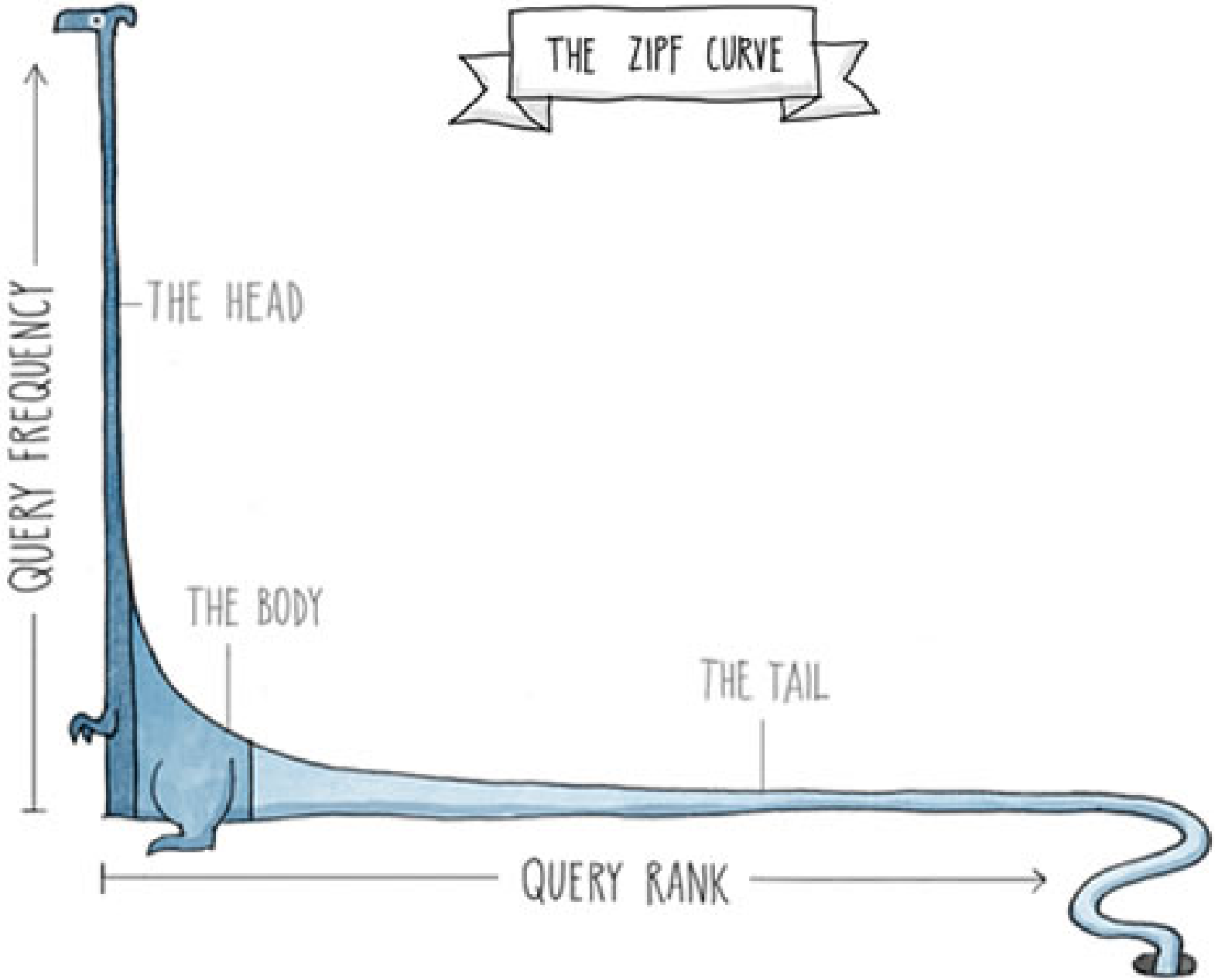
$f(w)$ — frequency of the word w

$r(w)$ — rank of the word w in the frequency list

C — constant

a — constant (close to 1)

Zipf's Law


THE ZIPF CURVE
QUERY FREQUENCY
THE HEAD
THE BODY
THE TAIL
QUERY RANK

Predictions of the Zipf's Law

If $a = 1$, $C = 60000$ then by Zipf's Law:

$$f(w) = \frac{60000}{r(w)}$$

- we will meet the most frequent word $f(w) = C/1 = 60000$ times
- the second - $C/2 = 30000$ times
- the third - $C/3 = 20000$ times
- 100th - $C/100 = 600$ times
- 101st - $C/101 = 594,06$ times
- and we will have the long tail of 80000 words with the frequency between $1,5$ and $0,5$.

Stop-words

The simplest way to decrease the number of lexical features is to delete the least informative words.

- Static List:
- Dynamic List:
  - Too frequent (N most frequent; frequency greater than k)
  - Too rare (frequency less than k)
  - Too short (less than M letters)
  - According to document frequency (present in more than k% texts or less than k texts)

Normalized Word Frequency

It is useful to represent counts on a normalized scale. A conventional unit for word frequencies in corpus linguistics is IPM (Instances Per Million).

$$\text{IPM} = \frac{\text{word frequency (count)}}{\text{number of words in the text / 1000000}}$$