TEXT MINING

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Invited lectures @ UdelaR

Adrien Guille, Associate Professor of Computer Science
University of Lyon, France

INTRODUCTION

SOME REAL-WORLD APPLICATIONS OF TEXT MINING / NLP

- Sentiment analysis
  - Marketing
  - Trading
- Information retrieval
  - Searching
  - Recommendation
- Content curation
  - Spam filtering
  - Offensive message detection

COURSE PLAN

- Lecture 1 - Statistical properties of natural language: Today
- Lab 1 - Manipulating text with R: Wednesday Oct. 3rd
- Lecture 2 - Supervised text classification: Thursday Oct. 4th
- Lecture 3 - Topic modeling & unsupervised text classification: Monday Oct. 8th
- Lab 2 - Supervised and non-supervised text classification with R: Oct. 10th
- Lecture 4 - Representation learning for text mining: Thursday Oct. 11th

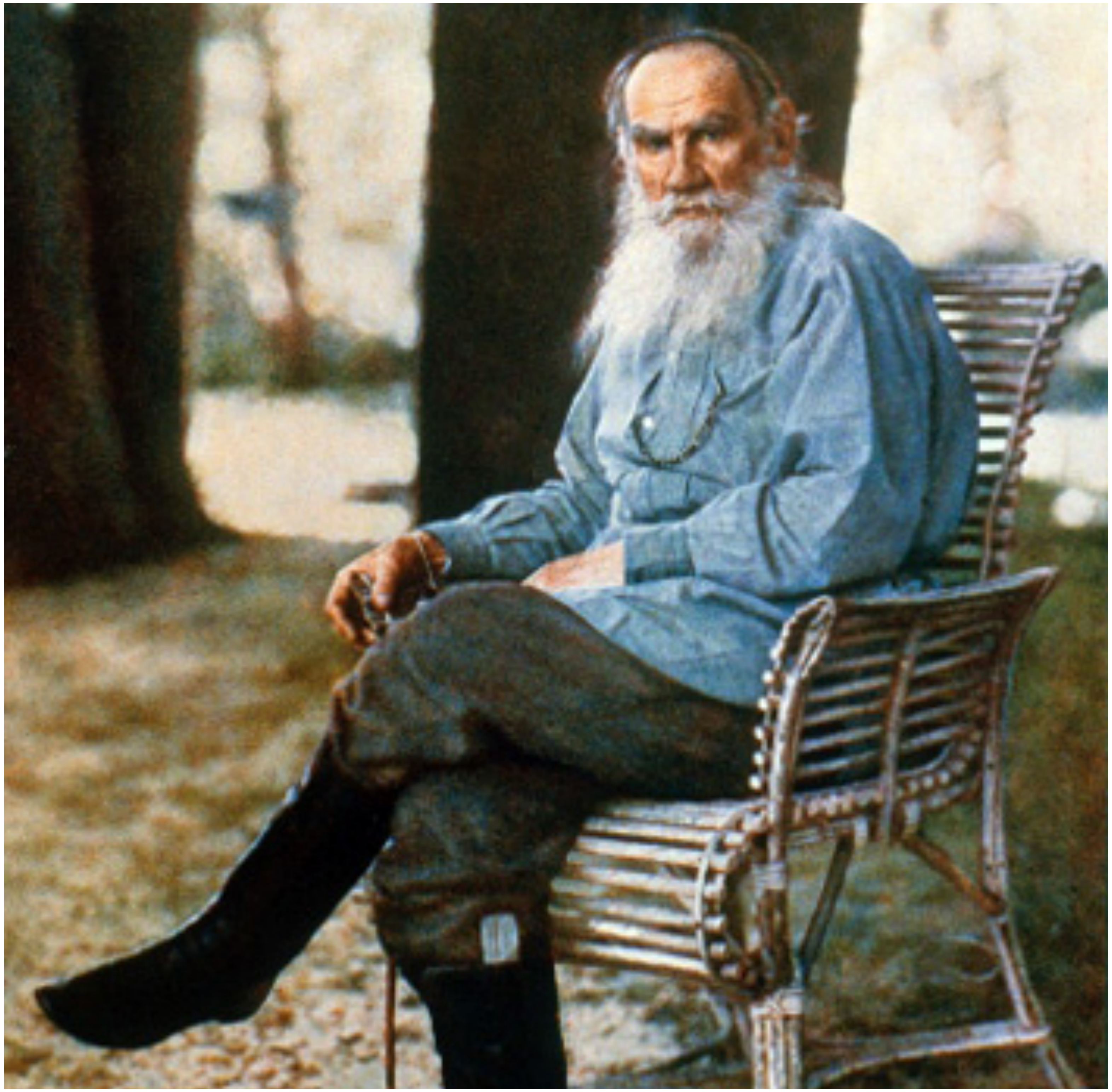
COURSE PROJECT: ANALYZING MOVIE REVIEWS

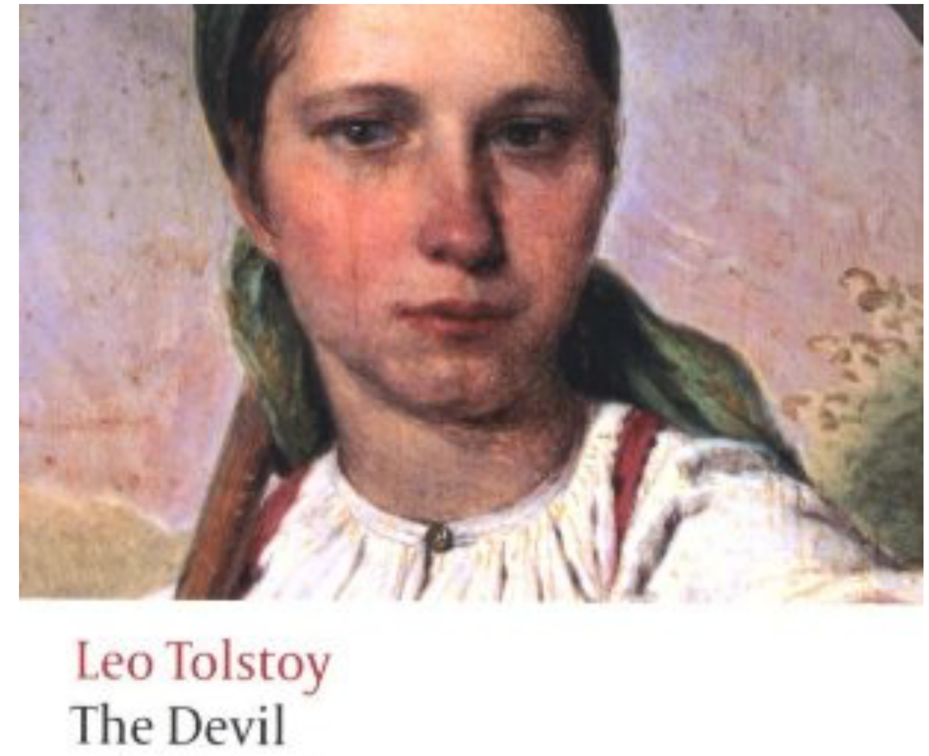
- Prepare a detailed R notebook, presenting your methodology, the results you obtained and your comments
- Plan
  - Data description
  - Task #1 Polarity prediction (*i.e.* determining wether a review is good or bad?)
  - Task #2
    A. Topic modeling (*i.e.* identifying the general themes underlying the reviews)
    B. Review clustering (*i.e.* grouping similar reviews)

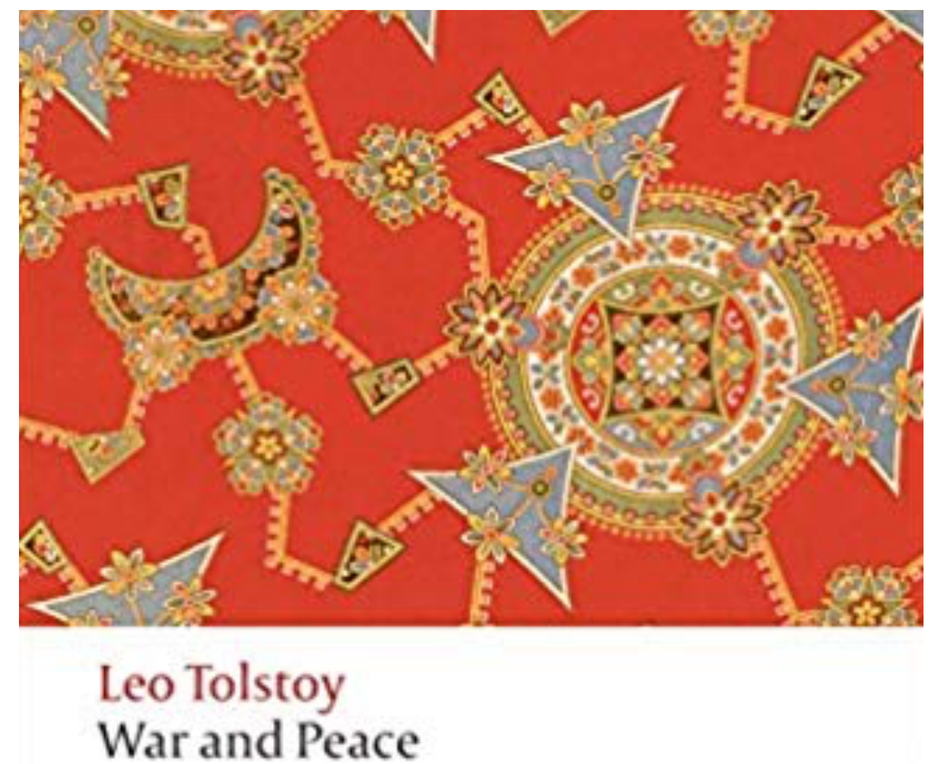
LECTURE 1

Statistical properties of natural language

WORD COUNTS







Left: Leo Tolstoy in 1908 (supposedly the first color portrait taken in Russia)

Right: Covers of «The Devil» and «War & Peace» (Oxford World's Classics)

BASIC PROPERTIES

- File size
  - The Devil: 102ko
  - War & Peace: 3.4mo
- Number of **words types**, *i.e.* distinct words, vocabulary (V)
  - The Devil: 2625
  - War & Peace: 18 261
- Number of **word tokens**, *i.e.* word occurrences
  - The Devil: 18 438
  - War & Peace: 569 129

MOST COMMON WORDS IN EACH CORPUS: STOP-WORDS

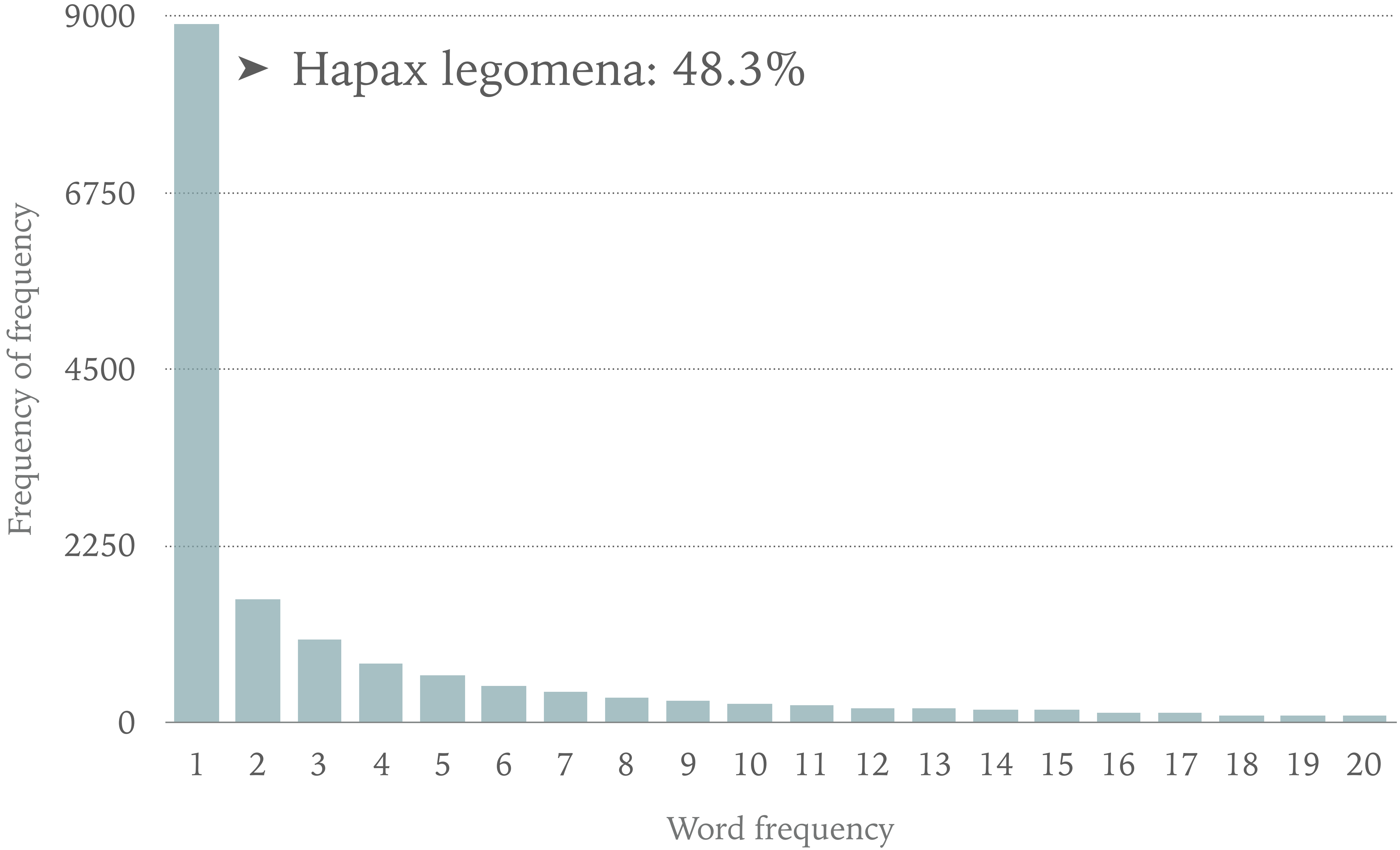
Word	Frequency
and	774
the	756
to	639
he	533
was	392
that	360
her	329
a	323
it	323
of	295

Common words in «The Devil»

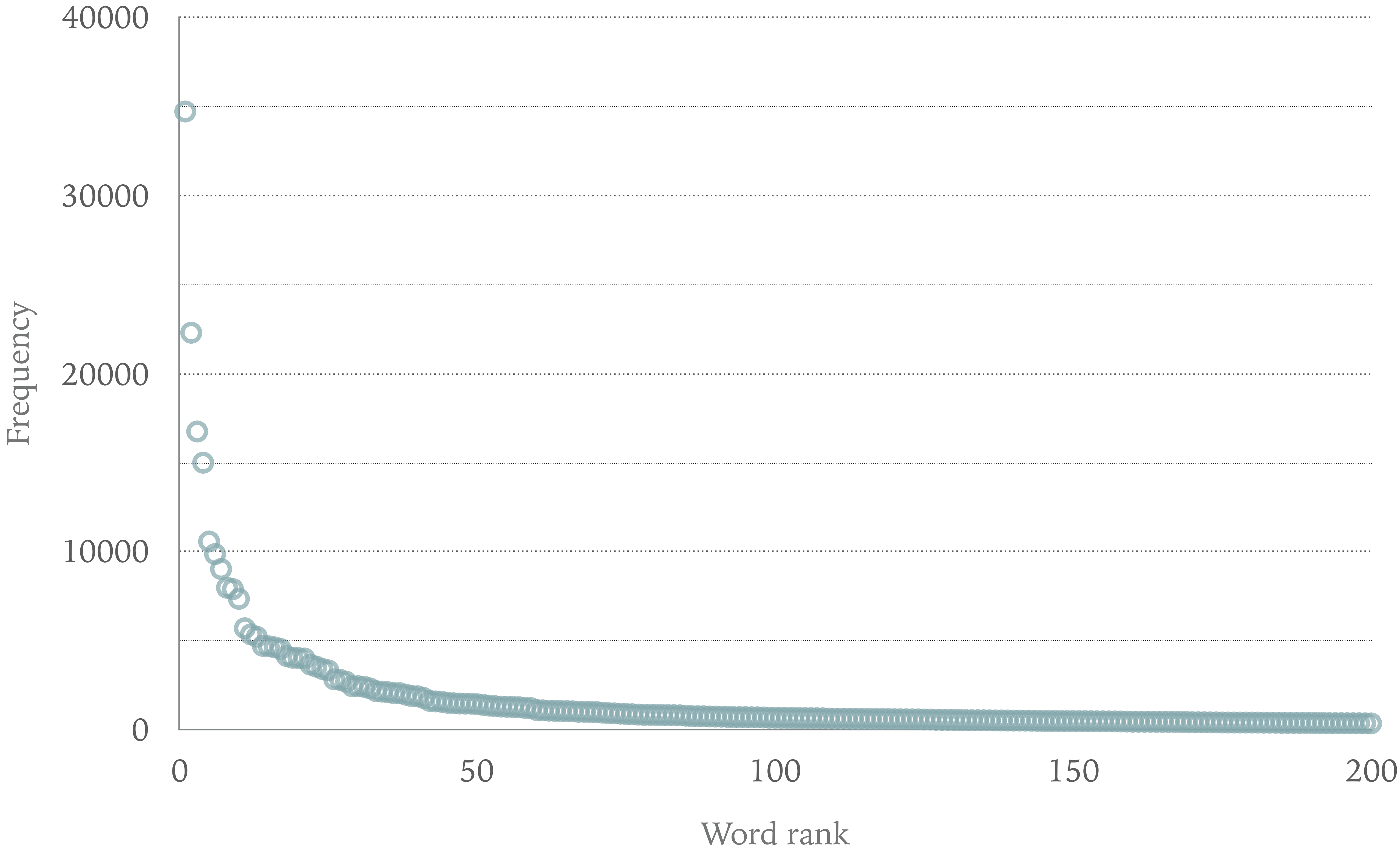
Word	Frequency
the	34721
and	22302
to	16755
of	15004
a	10580
he	9875
in	9036
his	7984
that	7908
was	7360

Common words in «War & Peace»

FREQUENCY DISTRIBUTION IN «WAR & PEACE»



WORD FREQUENCY VERSUS RANK IN «WAR & PEACE»

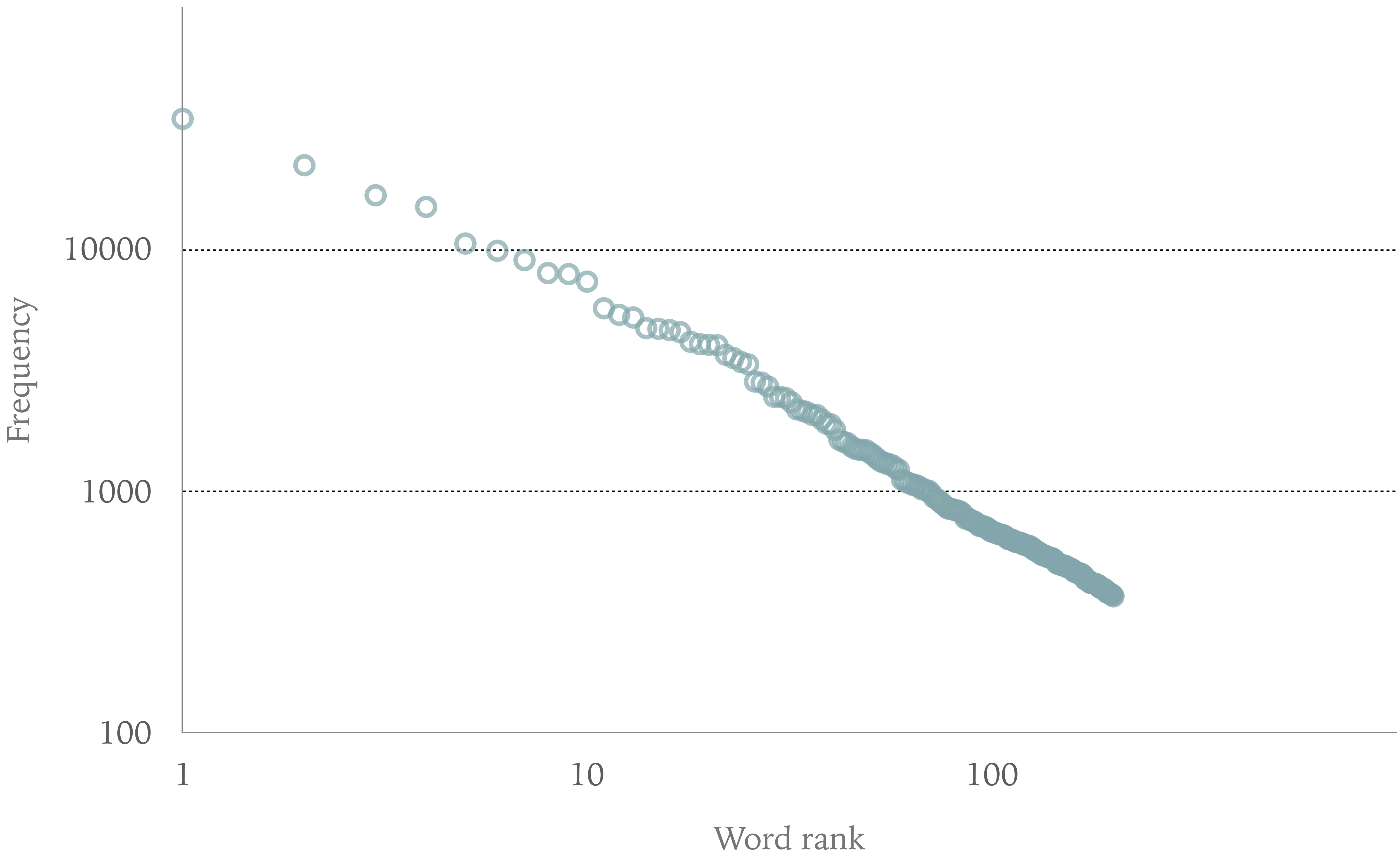


ZIPF'S LAW

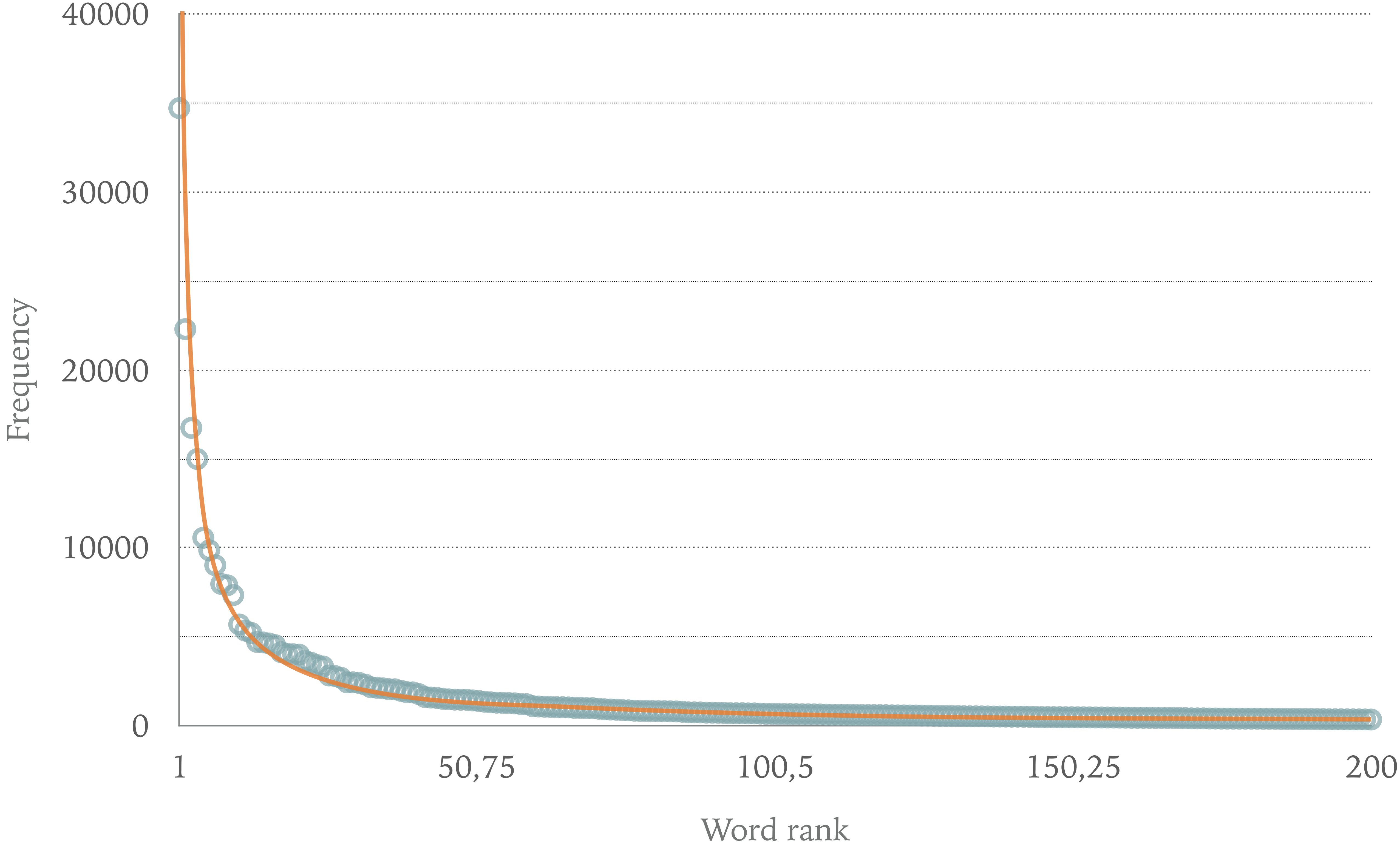
MODELING THE RELATION BETWEEN RANK AND FREQUENCY

- Zipf's assumptions
  - The frequency of a word depends
    - on its rank
    - on the frequency of the most common word
- Power law
  - Model the frequency as a function of rank
    - $f_r \simeq f_{max} \frac{1}{r^{-k}}$
  - Find parameter k via least-square fitting
    - $\log(f_r) \simeq \log(f_{max}) + k \log(r)$

WORD FREQUENCY VERSUS RANK IN «WAR & PEACE» (LOG AXES)



WORD FREQUENCY VERSUS RANK IN «WAR & PEACE»



YOU'LL SEE FOR YOURSELF DURING LAB 1

« Random texts exhibit Zipf's-law-like word frequency distribution.

-Wentian Li

IEEE Transactions on Information Theory, 1992

COLLOCATIONS

DEFINITION OF A COLLOCATION

- A contiguous sequence of two or more words
  - Syntactic unit
  - Semantic unit
  - Meaning cannot be unambiguously derived from from the meaning of its components
- Try to identify them by scanning a corpus with a 2-word window
  - A simple heuristic is based on a simple measure of association, the pointwise mutual information
  - Repeat the procedure to identify longer sequences

POINTWISE MUTUAL INFORMATION

- Increase in information about the occurrence of word *j* given *i*
  - $I(w_i, w_j) = \log \dfrac{P(w_i, w_j)}{P(w_i)P(w_j)}$
    $= \log \dfrac{P(w_i \mid w_j)}{P(w_i)}$
    $= \log \dfrac{P(w_j \mid w_i)}{P(w_j)}$
- Maximum likelihood estimates of the probabilities
  - $P(w_i) = \dfrac{\#(w_i)}{\sum_{k \in V} \#(w_k)}$
  - $P(w_i, w_j) = \dfrac{\#(w_i, w_j)}{\sum_{(w_{i'}, w_{j'}) \in C} \#(w_{i'}, w_{j'})}$

CORPUS REPRESENTATION

BAG-OF-WORD MODEL

- Assume word order is irrelevant
  - Describe a document as the multi-set of the words it contains
  - Preserve multiplicity, *i.e.* word frequency
- Vector space representation of a corpus (C)
  - A dimension per word, a vector per document
  - Linear operations make sense
    - Merging two documents = sum
    - Measuring the similarity between two documents = dot product, or norm of the difference
- Document-term matrix $X \in \mathbb{R}^{|C| \times |V|}$

BAG-OF-N-GRAMS

- Shallowly capture word ordering
  - Consider short contiguous word sequences as terms, *e.g.*
    - 2-gram, *i.e. b*igrams
      - New York, bad luck
    - 3-grams, *i.e. t*rigrams
      - Not so good, Dulce de leche
- N-grams largely increase the vocabulary size and make *X* increasingly sparse
  - Keep N small and filter out rare N-grams

TF-IDF WEIGHTING

- The more frequent a word is in a document, the more relevant it might be to this document
  - However, relevancy doesn't increase linearly with frequency
  - Also, this word might be frequent in all documents
- By Zipf's law, few words are responsible for most occurrences
  - More informative words are rare
- Compute a score for each word in each document
  - $Score_{d_i,w_j} = tf_{i,j} \times idf_j$
  - Where $tf_{i,j} = \log(1 + X_{ij})$ and $idf_j = \log\left(\frac{|C|}{df_j}\right)$