BAN432
Applied Textual Data Analysis for Business and Finance

Preprocessing and cleaning textual data, part I

Christian Langerfeld and Maximilian Rohrer

16 September, 2022

Packages and data files needed in today's lecture

For today's lecture please make sure you have these packages installed:

- `tidytext`
- `readr`
- `stopwords`
- `wordcloud`

We will work with two data files today. You find them on Canvas:

- brown.txt
- data_for_lecture_05.Rdata

Overview

1	Introduction to course & basic R	Introduction
2	Introduction to R, specific to textual analysis	
3	Collecting textual data: APIs	Collecting data
4	Collecting textual data: EDGAR	
5	Preprocessing and cleaning, part I	Preprocessing data
6	Preprocessing and cleaning, part II	
7	Guest Lecture: Gisle Andersen (NHH)	
8	Regex-based application, Geography	Analyses
9	Regex-based application, Keyword in Context	
10	Automatic text summarization	
11	Sentiment: Twitter & Critical understanding	
12	Sentiment: Finance application	
13	Doc-Clustering: Cosine similarity & k-means	
14	Doc-clustering: Topic models	
15	Doc-Clustering: Multinominal Inverse Regression	
16	Guest Lecture: Vegard Larsen (Norges Bank)	
17	Contemporaneous papers in Finance	
18	Recap	

Today's lecture

- what is a corpus?
  - different kinds of corpora
- what is a word?
  - type/token
  - function word vs. content word
- frequency lists
  - frequency distributions in a corpus
  - Zipf's law
  - Heaps' law
- tokenization

Introduction

- What is a corpus?
  - machine-readable collection of texts (written and spoken)
  - text is produced in a natural communicative setting
  - the collection of texts should be as representative and balanced as possible with respect to language variety or genre

Introduction – Different kinds of corpora

- general vs. specific:
  - general corpora are compiled to cover a language as a whole e.g. American English
  - specific corpora cover a particular variety or genre

Introduction – Different kinds of corpora

- general vs. specific:
  - general corpora are compiled to cover a language as a whole
    e.g. American English
  - specific corpora cover a particular variety or genre
- raw vs. annotated:
  - raw corpora contain only the corpus texts itself
  - annotated corpora contain additional information for each text
    - a header with meta-information (author, date published . . .)
    - a body with the text itself and some additional information
      e.g. part-of-speech for each word

Introduction – Different kinds of corpora

- general vs. specific:
  - general corpora are compiled to cover a language as a whole e.g. American English
  - specific corpora cover a particular variety or genre
- raw vs. annotated:
  - raw corpora contain only the corpus texts itself
  - annotated corpora contain additional information for each text
    - a header with meta-information (author, date published . . .)
    - a body with the text itself and some additional information e.g. part-of-speech for each word
- static vs. dynamic corpora:
  - static corpus is compiled and remains unchanged
  - dynamic (or monitor) corpus is constantly extended with new material (e.g. the Norwegian newspaper corpus)

Introduction – Different kinds of corpora? (cont.)

- diachronic vs. synchronic:
    - diachronic corpora cover text material from a long time span
    - synchronic corpora cover contemporary language

Introduction – Different kinds of corpora? (cont.)

- diachronic vs. synchronic:
  - diachronic corpora cover text material from a long time span
  - synchronic corpora cover contemporary language
- monolingual vs. parallel corpora:
  - monolingual corpora cover just one language
  - parallel corpora contain the same texts in different languages (e.g. eur-lex.europa.eu)

Introduction

- the corpus itself does not contain meaning, it just contains frequencies of occurrence, i.e. how often words, grammatical patterns etc. occur in the corpus
- the analyst has to interpret the frequencies in a meaningful way

Frequency lists:

- most basic corpus linguistic tool
- generate a frequency list if you want to know how often a word occurs in the corpus
- usually two columns: (a) the word (b) the frequency in the corpus

word	*frequency*
the	62,580
of	35,958
and	27,789
...	...

Introduction – types and tokens

- but: *word* is ambiguous
- how many words does the following example contain?

```
the word and the phrase
```

Introduction – types and tokens

- but: *word* is ambiguous
- how many words does the following example contain?

```
            the word and the phrase
```

- linguists make a difference between *types* and *tokens*
- the above example contains:
  - 5 (word)tokens: `"the" "word" "and" "the" "phrase"`
  - 4 (word)types: `"the" "word" "and" "phrase"`

word	*frequency*
the	62,580
of	35,958
and	27,789
...	...

- what we see in frequency lists are *types* and *token frequencies*

Introduction

What is a word?

- are *car* and *cars* the same word?
- *September* and *Sept*?
- *1960* and *25-year-old*?
- how many words are there in *don't* and *Gonna*?

Bear that in mind while working with frequency lists!

Corpora used in today's lecture

(1) Brown corpus

- complied in the 1960s
- one of the first corpora that were available electronically
- size about 1,000,000 words
- today's corpora are much lager
  - British National Corpus: 100,000,000 words
  - Corpus of Contemporary American English: 450,000,000 words
  - Norsk aviskorpus: 1,400,000,000 (in 2015, still growing)
- Brown corpus consists of samples of 500 texts from 15 genres
- is meant to be representative for the America English (written) language of 1961

Corpora used in today's lecture (cont.)

(2) Wikipedia corpus
   - Simple web crawler initialized at "Brown-corpus" wiki page
   - 250 pages with 599,377 words

(3) Earning calls corpus
   - Transscripts of the introduction part to quarterly earnings calls
   - 1000 calls with 1,889,256 words

The Brown corpus

Task 1:

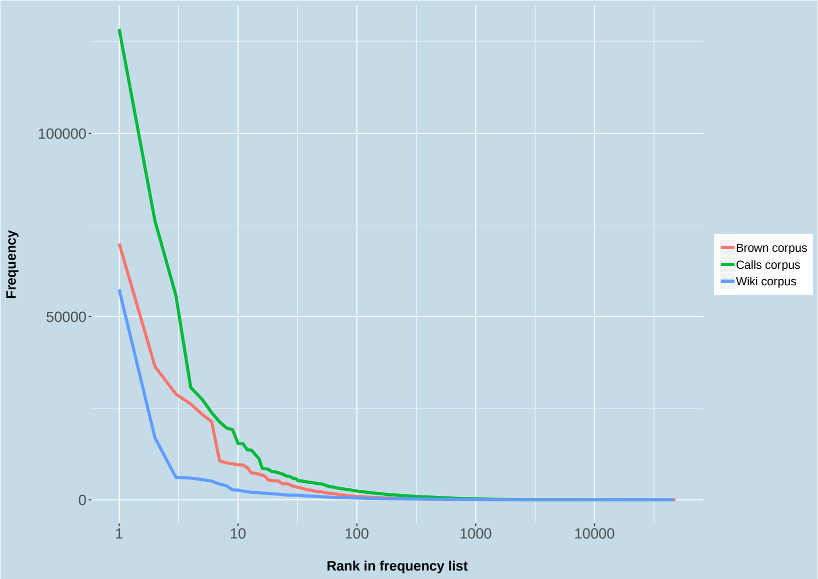
- research question: How are frequencies distributed in a corpus (and hence in natural language)?
- download the file `brown.txt` from Canvas
- steps:
  (1) load the Brown corpus in R
  (2) tokenize the text, i.e. split it into words
  (3) make a frequency list
  (4) plot the frequencies
- we develop an approach together in class

Distribution of word frequencies

Task 2

- download the file `data_for_lecture_05.Rdata` from Canvas
- the file contains 2 data frames
  - `wiki.freq`
  - `earning.calls.freq`
- make a plot of each frequency list (use logarithmic scaling of the x-axis)
- compare the two plots with the plot of the Brown corpus

Frequency plots for the corpora (log-scaled)



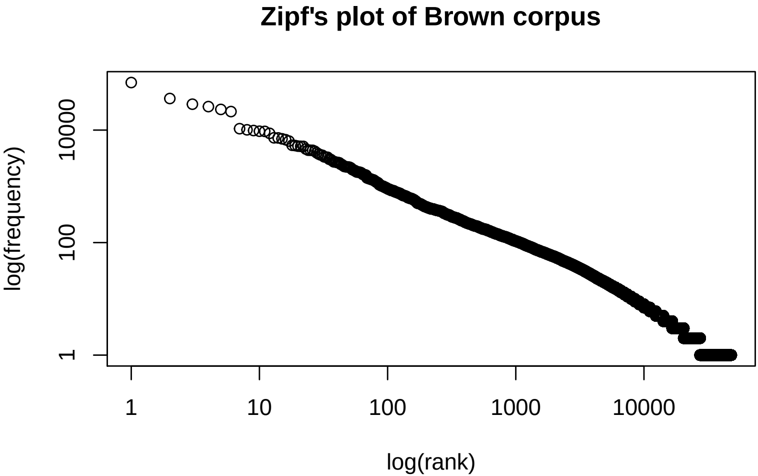
Zipf's law

- George Kingsley Zipf (1902–50) observed that frequency and rank are inversely related

Table 3: Data from the Brown corpus

rank (r)	type	frequency (f)	$r \times f$
1	the	70003	70003
2	of	36473	72946
3	and	28935	86805
4	to	26247	104988
5	a	23377	116885
100	way	924	92400
200	head	436	87200
1000	income	110	110000
2000	previously	58	116000
44706	zwei	1	44706

Zipf's plot



Words from ca. rank 27000 onward occur only once.

Zipf's law (cont.)

- word types have a very skewed distribution.
- in any larger corpus, almost 50% of the word types occur only once (hapax legomena)

Table 4: Frequencies of *hapax legomena*

Corpus	total *types* in corpus	hapax	%
Brown	44706	17779	39.8
Wiki	46077	21454	46.6
Earnings	35157	14304	40.7

Task 3: how can we use *R* to find the number of “hapaxes”, e.g. in the Brown corpus?

Zipfs' law (cont.)

- a few types are exceedingly common
- the top 50 types in a corpus account for 30 – 40% of the tokens

Table 5: Frequencies of the top 50 word types

Corpus	total *tokens* in corpus	top 50	%
Brown	1022006	413291	40.4
Wiki	599377	164351	27.4
Earnings	1889256	686987	36.4

Content words vs. function words

- content words: nouns, verbs, adjectives, adverbs
  - refer to objects, actions or properties
  - open class, new words can be added
- function words: determiners, prepositions, conjunctions, pronouns, auxiliary verbs, . . .
  - grammatical relationships between words
  - little substantive meaning
  - closed class
- for many text mining tasks, the most frequent function words are removed

Task 4: Make wordclouds from Earning Calls corpus with and without stopwords

Stopword removal – Earning calls corpus

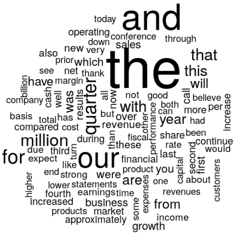


Figure 1: stopwords not removed

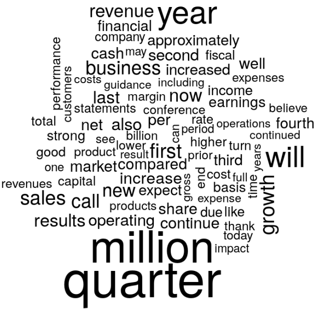


Figure 2: stopwords removed

Corpus size: Representativeness

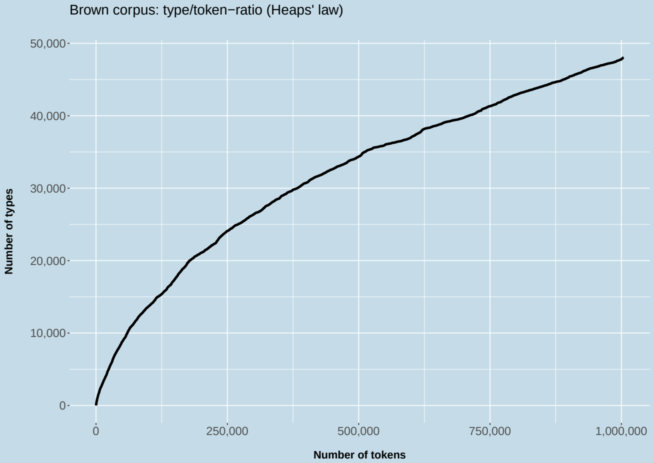
- in linguistics, a corpus is meant to be a representative sample of the language under investigation
- for some subsets of the language it is easy to find a representative sample
    - if you study the language in annual reports of firms that operate in the US, you can compile a corpus of *all* 10-Ks (finite number of documents)
    - if you study business language in general your sample has to contain texts from other business related text genres, e.g. Marketing, Management, Economics etc. (infinite number of documents)

Corpus size: Representativeness (cont.)

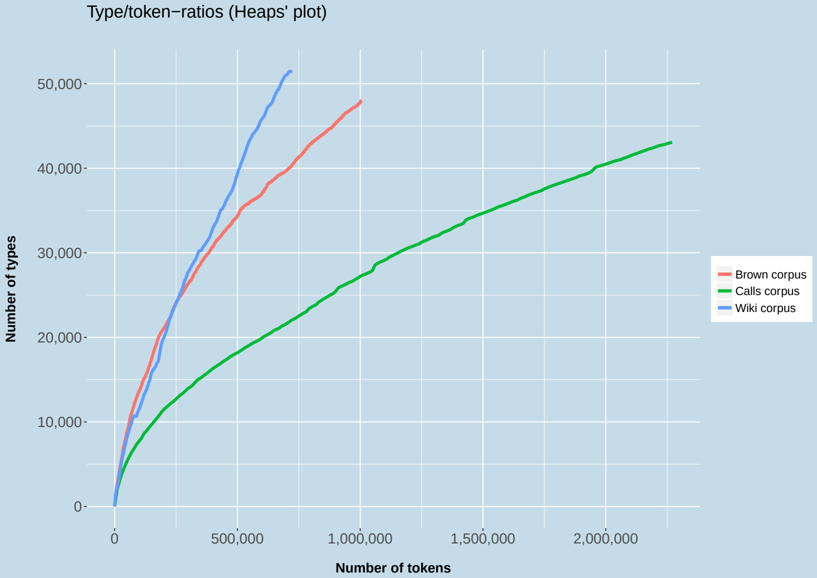
- it is impossible to compile a representative corpus of general language
- how big should a corpus be?
- can we discover all the words that belong to the (sub)language under investigation?
- if we double the size of a corpus, do we double the number of types (unique words) as well?

Heaps' law

- As more text (tokens) are gathered, diminishing returns of new vocabulary (types)
- Lexical closure (saturation): the curve of lexical growth has become asymptotic



Heaps' law (cont.)



Preprocessing tasks

- so far, in this lecture, we have talked about frequencies
- compiling of frequency lists is an important preprocessing task
- other tasks are:
  - sentence splitting
  - compile n-gram lists
  - convert encoding
  - stemming
  - part-of-speech tagging

Summary of today's lecture

- what is a corpus?
- different kinds of corpora (general vs. specific, etc.)
- what is a word?
- function words vs. content words
- type/token
- frequency distributions
- Zipf's law
- Heaps' law