UPPSALA
UNIVERSITET

Statistical analysis of textual data

Måns Magnusson
Department of Statistics
Uppsala University

June 10, 2024

UPPSALA
UNIVERSITET

Section 1

Text as data

UPPSALA UNIVERSITET

Why?

- Large textual corpora are (or are becoming) digital

UPPSALA UNIVERSITET

- Text as data
- Text representations

Why?

- Large textual corpora are (or are becoming) digital
- We want to study meaning and semantics to draw conclusions about society: treating text as data.
- Interest in inference from data

UPPSALA UNIVERSITET

- Text as data
- Text representations

Example of problems

- Concept history: The changing meaning of "information" and "propaganda"

Example of problems

- Concept history: The changing meaning of "information" and "propaganda"
- Sociology: How do public discourses on immigration form and change over time?

UPPSALA UNIVERSITET

Example of problems

- Concept history: The changing meaning of "information" and "propaganda"
- Sociology: How do public discourses on immigration form and change over time?
- Law: What affects the outcome of a court case? (Outcome prediction)

UPPSALA UNIVERSITET

- Text as data
- Text representations

The statistical problem with textual data

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?

The statistical problem with textual data

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?
- The structure of language and statistical inference
  - Hierarchical, discrete, sparse, high-dimensional
  - Long distant relations/context, syntax, noisy (errors)

UPPSALA
UNIVERSITET

The statistical problem with textual data

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?
- The structure of language and statistical inference
  - Hierarchical, discrete, sparse, high-dimensional
  - Long distant relations/context, syntax, noisy (errors)
- Very large corpora

UPPSALA UNIVERSITET

The statistical problem with textual data

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?
- The structure of language and statistical inference
  - Hierarchical, discrete, sparse, high-dimensional
  - Long distant relations/context, syntax, noisy (errors)
- Very large corpora
- Very large models

The statistical problem with textual data

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?
- The structure of language and statistical inference
  - Hierarchical, discrete, sparse, high-dimensional
  - Long distant relations/context, syntax, noisy (errors)
- Very large corpora
- Very large models
- Challenges in analysis of textual data today:
  - Scalability of statistical methods
  - Causal inference
  - Drawing conclusions about society from textual data

The statistical problem with textual data

UPPSALA UNIVERSITET

- Text as data
- Text representations

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?
- The structure of language and statistical inference
  - Hierarchical, discrete, sparse, high-dimensional
  - Long distant relations/context, syntax, noisy (errors)
- Very large corpora
- Very large models
- Challenges in analysis of textual data today:
  - Scalability of statistical methods
  - Causal inference
  - Drawing conclusions about society from textual data
- We know that all models are wrong.

The statistical problem with textual data

The quick brown fox jumps over the lazy dog.

- What is the difference regarding text as data?
- The structure of language and statistical inference
  - Hierarchical, discrete, sparse, high-dimensional
  - Long distant relations/context, syntax, noisy (errors)
- Very large corpora
- Very large models
- Challenges in analysis of textual data today:
  - Scalability of statistical methods
  - Causal inference
  - Drawing conclusions about society from textual data
- We know that all models are wrong.
  But some models can be useful.

How to measure and estimate meaning

UPPSALA UNIVERSITET

- Text as data
- Text representations

```
Closing arguments were heard yesterday in the
Federal bankruptcy fraud trial of Stephen J.
Sabbeth, whose legal problems have raised doubts
about his ability to continue as leader of the
Nassau County Democratic Party.
```

- The distributional hypothesis (see Sahlgren, 2008)

  a word is characterized by the company it keeps (Firth, 1957)

How to measure and estimate meaning

```
Closing arguments were heard yesterday in the
Federal bankruptcy fraud trial of Stephen J.
Sabbeth, whose legal problems have raised doubts
about his ability to continue as leader of the
Nassau County Democratic Party.
```

- The distributional hypothesis (see Sahlgren, 2008)

  a word is characterized by the company it keeps (Firth, 1957)

- The context of words:
  - word windows
  - "documents"
  - left context (predict the next word)

The most common model classes

- Latent semantic models
  - topic models (documents)
  - word embeddings (word windows)

The most common model classes

- Latent semantic models
    - topic models (documents)
    - word embeddings (word windows)
- Transformer Neural Networks
    - Encoder models (word windows, masked language models)
    - Decoder models (left context, next word prediction)

UPPSALA UNIVERSITET

- Text as data
- Text representations

How should we do it? The Box process

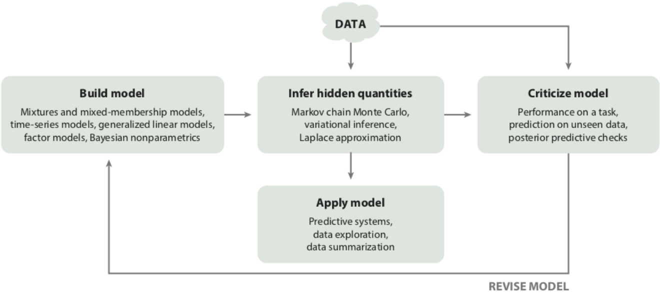

Figure: The Box approach (Box, 1976, Blei, 2014)

Section 2

Text representations

Computational text representation

- The tidy data format
  - Each variable is a column
  - Each observation is a row
  - Each type of observational unit is a table
- For text: a table with one-token-per-row
- A token is discrete unit of interest

Example:

The quick brown fox jumps over the lazy dog.

pos	*word_type*	*sentence*
1	*The*	1
2	*quick*	1
3	*brown*	1
...	...	...

UPPSALA UNIVERSITET

- Text as data
- Text representations

Other representations and aggregations

- Other representations
  - characters
  - vectors/lists of characters

UPPSALA UNIVERSITET

- Text as data
- Text representations

Other representations and aggregations

- Other representations
  - characters
  - vectors/lists of characters
- Aggregating text
  - document-term matrices
  - term-term matrices
- Commonly used in simple text classification

Example:

The quick brown fox jumps over the lazy dog. That was quick! ...

	*sentence*1	*sentence*2	...
the	2	0	...
quick	1	1	...
brown	1	0	...
...	...	...	...

UPPSALA UNIVERSITET

- Text as data
- Text representations

Aggregation using tidy text

The quick brown fox jumps over the lazy dog. That was quick! ...

pos	*word_type*	*sentence*	*n*
1	*The*	1	2
2	*quick*	1	1
3	*brown*	1	1
...	...	...	

UPPSALA UNIVERSITET

- Text as data
- Text representations

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based

UPPSALA UNIVERSITET

- Text as data
- Text representations

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)
  - Generate large vocabularies with many different words

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)
  - Generate large vocabularies with many different words
  - Out-of-vocabulary problem

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)
  - Generate large vocabularies with many different words
  - Out-of-vocabulary problem
- Training tokenizers on a given corpus:

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)
  - Generate large vocabularies with many different words
  - Out-of-vocabulary problem
- Training tokenizers on a given corpus:
- Byte-pair encoding (BPE):
  Combine most common characters until predefined vocabulary is reached
  The quick brown fox jump ##s over the 13 lazy dog ##s

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)
  - Generate large vocabularies with many different words
  - Out-of-vocabulary problem
- Training tokenizers on a given corpus:
- Byte-pair encoding (BPE):
  Combine most common characters until predefined vocabulary is reached
  The quick brown fox jump ##s over the 13 lazy dog ##s
- Wordpiece encoding (WE):
  Simplary to BPE, but uses a language model to score which to merge

A note on tokenization

The quick brown fox jumps over the 13 lazy dogs.

- Simple tokenization: Word-based
  - Difficult in some languages (e.g. Chinese)
  - Generate large vocabularies with many different words
  - Out-of-vocabulary problem
- Training tokenizers on a given corpus:
- Byte-pair encoding (BPE):
  Combine most common characters until predefined vocabulary is reached
  The quick brown fox jump ##s over the 13 lazy dog ##s
- Wordpiece encoding (WE):
  Simplary to BPE, but uses a language model to score which to merge
- Unigram tokenizer:
  Like WE, but start with a large vocabulary and trim it down.

Practical (vocabulary) curation

The quick brown fox jumps over the 13 lazy dogs.

- Lemmatizing/stemming (text normalization)

  The quick brown fox jump over the 13 lazy dog.

UPPSALA
UNIVERSITET

Practical (vocabulary) curation

The quick brown fox jumps over the 13 lazy dogs.

- Lemmatizing/stemming (text normalization)
  The quick brown fox jump over the 13 lazy dog.
- Lower casing
  the quick brown fox jump over the 13 lazy dog.

UPPSALA UNIVERSITET

- Text as data
- Text representations

Practical (vocabulary) curation

The quick brown fox jumps over the 13 lazy dogs.

- Lemmatizing/stemming (text normalization)
  The quick brown fox jump over the 13 lazy dog.
- Lower casing
  the quick brown fox jump over the 13 lazy dog.
- Remove numbers and punctuation
  the quick brown fox jump over the NN lazy dog

Practical (vocabulary) curation

The quick brown fox jumps over the 13 lazy dogs.

- Lemmatizing/stemming (text normalization)
  The quick brown fox jump over the 13 lazy dog.
- Lower casing
  the quick brown fox jump over the 13 lazy dog.
- Remove numbers and punctuation
  the quick brown fox jump over the NN lazy dog
- Stop
  words and rare words (Zipfs distribution/law, freq vs. rank)
  quick brown fox jump NN lazy dog