Content Analysis: Topic Modelling & Word Field Analysis

Die Pandemie in Sprache und Text – Corona-Podcasts & co.

Jürgen Hermes, Lukas Mönch, Felix Rau, Nils Reiter

30. April 2021

Introduction

- Content vs. Language
- Content is encoded in words

Introduction

- Content vs. Language
- Content is encoded in words
- Ambiguity:
  - The same word can encode different content (depending on context)
  - The same content can be expressed by different words

Introduction

- Content vs. Language
- Content is encoded in words
- Ambiguity:
  - The same word can encode different content (depending on context)
  - The same content can be expressed by different words
- Two methods
  - Word field analysis (low tech)
  - Topic modelling (higher tech)

Section 1

Word Field Analysis

Introduction

- Word field: Collection of words with related meanings Trier (1931)
  - not strictly synonymous as WordNet synsets
- Defined by us

Example (Liebe)

Abschiedskuß adorieren anbeten Angebetete aufrichtig Aufrichtigkeit begehren Begierde begieren beweinen Beziehung brennen Brust Busen Ceremonie Copulation copulieren ehrlich empfinden Empfindsamkeit Engel Entzücken fühlen Funke Gefühl geliebt Geliebte Geliebte Geliebter Geliebteste Gemahl Gemahlin Glück Heirat Heirath Herz Herzen Hochzeit huldigen Jüngling küssen Kuß kuscheln Kuss lüstern Leidenschaft Liebchen Liebe liebeglühend lieben Liebende liebenswürdig Liebenswürdigkeit Liebesfest Liebhaber Liebhaberin liebkosen Liebkosung Liebliche Liebste Liebster Liebstes Lippe Lust Rose Schönheit Seele sehnen Sehnsucht Sinn sinnlich Sinnlichkeit streicheln Trauung treu Treue Umarmung Unschuld verehren Verehrung vergöttern Vergötterung Verlangen verlangen Verlieben vermählen Vermählung verzehren Wollust zärtlich Zärtlichkeit Zeremonie Willand/Reiter (2017)

How to

- Take text, count number of words from field. ✔

How to

- Take text, count number of words from field. ✔

Design decisions

- Lemmatisation of the text vs. full form lexicon
  - Easy to use lemmatiser: Spacy (Python) `https://spacy.io`; Stanford CoreNLP (Java): `https://stanfordnlp.github.io/CoreNLP/`
  - Imperfect lemmatisation
  - (Likely) incomplete full form lexicon

How to

- Take text, count number of words from field. ✔

Design decisions

- Lemmatisation of the text vs. full form lexicon
  - Easy to use lemmatiser: Spacy (Python) `https://spacy.io`; Stanford CoreNLP (Java): `https://stanfordnlp.github.io/CoreNLP/`
  - Imperfect lemmatisation
  - (Likely) incomplete full form lexicon
- Normalisation: Text length, word field length

How to

- Take text, count number of words from field. ✔

Design decisions

- Lemmatisation of the text vs. full form lexicon
  - Easy to use lemmatiser: Spacy (Python) `https://spacy.io`; Stanford CoreNLP (Java): `https://stanfordnlp.github.io/CoreNLP/`
  - Imperfect lemmatisation
  - (Likely) incomplete full form lexicon
- Normalisation: Text length, word field length
- Ambiguity
  - Words that belong to multiple/all fields
  - Only one sense belongs to a field at all

How to

- Take text, count number of words from field. ✔

Design decisions

- Lemmatisation of the text vs. full form lexicon
  - Easy to use lemmatiser: Spacy (Python) `https://spacy.io`; Stanford CoreNLP (Java): `https://stanfordnlp.github.io/CoreNLP/`
  - Imperfect lemmatisation
  - (Likely) incomplete full form lexicon
- Normalisation: Text length, word field length
- Ambiguity
  - Words that belong to multiple/all fields
  - Only one sense belongs to a field at all
- Weighting: TF-IDF, MI

How to establish word fields?

- Introspection
- Corpus analysis
  - Search for seed words
  - Look in context of findings for additional words

How to establish word fields?

- Introspection
- Corpus analysis
  - Search for seed words
  - Look in context of findings for additional words
- Extract words from existing resources
  - Glossary of Coronavirus Update: `https://www.ndr.de/nachrichten/info/Das-Glossar-zum-Corona-Podcast,podcastcoronavirus146.html`
  - Franz Dornseiff (2010). *Der deutsche Wortschatz nach Sachgruppen*. 8th ed. De Gruyter

Patterns and Multi-Word Expressions (MWEs)

- Many concepts can only be added to a field as a MWE
  - E.g., „Britische Mutation“, „Kurve abflachen“

Patterns and Multi-Word Expressions (MWEs)

- Many concepts can only be added to a field as a MWE
  - E.g., „Britische Mutation“, „Kurve abflachen“
- Complex, because syntactic variation
  - „Kurve schnell abflachen“, „Kurve in Deutschland abflachen“, „Kurve ohne größere Schäden abflachen“, ...
- Theory: Matching of syntactic patterns
  - Parsing of the text
  - Word fields $\rightarrow$ partial syntactic tree fields
  - Match tree with partial tree

Results and Interpretation

- Very small numbers
- Best use: For comparison
  - „Number of words from field X is increasing over time“
  - „Drosten uses more words from field X than Kekulée“

Implementations

Implementations

Seriously: Regular Expressions

- Most common/powerful: Perl-compatible regular expressions (PCRE)
- Important expression: `\b`

Example

- RE `Hunde?` matches on „Hund“ and „Hunde“
- Unfortunately, it also matches on „Hundesteuer“
  (and everything else starting with this prefix)

Seriously: Regular Expressions

- Most common/powerful: Perl-compatible regular expressions (PCRE)
- Important expression: `\b`

Example

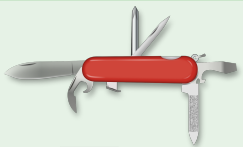
- RE `Hunde?` matches on „Hund“ and „Hunde“
- Unfortunately, it also matches on „Hundesteuer“ (and everything else starting with this prefix)
- `\b` matches on word boundaries
  - Not a real character
  - Matches on position between `\W` and `\w`
- `\bHunde?\b` matches on „Hund“ and „Hunde“, but not
  - „Hundesteuer“
  - „Blindenhund“

Seriously: Regular Expressions

- Most common/powerful: Perl-compatible regular expressions (PCRE)
- Important expression: `\b`

Example

- RE `Hunde?` matches on „Hund“ and „Hunde“
- Unfortunately, it also matches on „Hundesteuer“ (and everything else starting with this prefix)
- `\b` matches on word boundaries
  - Not a real character
  - Matches on position between `\W` and `\w`
- `\bHunde?\b` matches on „Hund“ and „Hunde“, but not
  - „Hundesteuer“
  - „Blindenhund“



Symbolbild

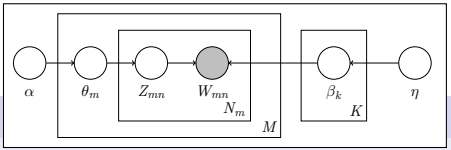
Section 2

Topic Modelling

Introduction

- Latent Dirichlet Allocation (LDA), „topic modelling“ Blei et al. (2003)
- Generative model to represent topical structures in documents
  - I.e., assumption: documents were created according to a generative process
  - Parameters in this process are *latent*, not observable
- The only observable thing we have is the words in each document

Generative Process



Pseudocode

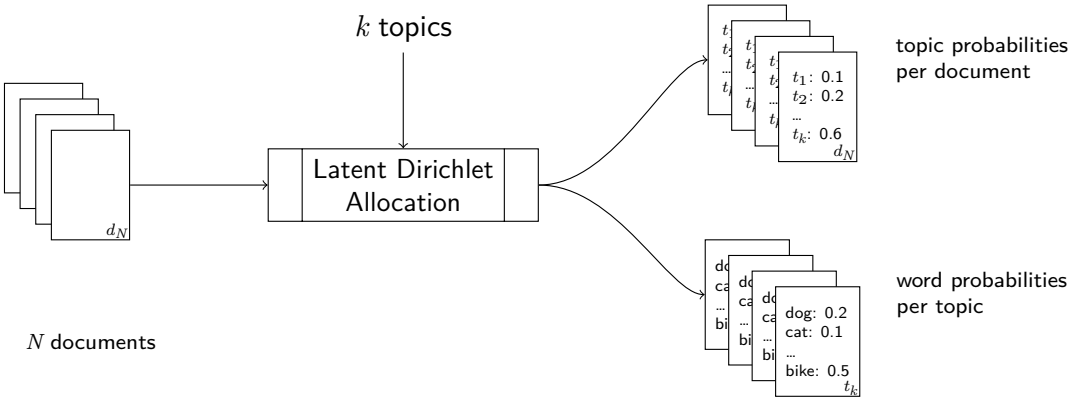
for $k = 0; k < K; k \to k + 1$ (for each topic)

1. Choose $\beta_k \sim \text{Dir}(\eta)$

for $m = 0; m < M; m \to m + 1$ (for each document)

1. Choose $N_m \sim \text{Poisson}(\zeta)$
2. Choose $\theta_m \sim \text{Dir}(\alpha)$
3. for $n = 0; n < N; n \to n + 1$ (for each word in document m)
   - 3.1 Choose a topic $Z_{mn} \sim \text{Multinomial}(\theta_m)$
   - 3.2 Choose a word $w_{mn} \sim \text{Multinomial}(\beta_{Z_{mn}})$

Input/Output



demo

Caveats

- Topics are probability distributions and do not have names
- Some topics are hard to interpret
- Manual labeling: interpretation step
- Automatic labeling: Bhatia et al. (2016) and Lau et al. (2011)

Implementations

- Java/CLI: Mallet `http://mallet.cs.umass.edu`
- Python: gensim `https://radimrehurek.com/gensim/`
- R: lda `https://cran.r-project.org/package=lda`, topicmodels `https://cran.r-project.org/package=topicmodels`

Implementations

- Java/CLI: Mallet `http://mallet.cs.umass.edu`
- Python: gensim `https://radimrehurek.com/gensim/`
- R: lda `https://cran.r-project.org/package=lda`, topicmodels `https://cran.r-project.org/package=topicmodels`
- GUIs
  - Dariah topics explorer: `https://dariah-de.github.io/TopicsExplorer/`
  - Topic Modeling Tool: `https://github.com/senderle/topic-modeling-tool`

Questions?

References I

- Bhatia, Shraey/Jey Han Lau/Timothy Baldwin (2016). „Automatic Labelling of Topics with Neural Embeddings“. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 953–963.
- Blei, David/Andrew Y. Ng/Michael I. Jordan (2003). „Latent dirichlet allocation“. In: *Journal of Machine Learning Research* 3, pp. 993–1022.
- Dornseiff, Franz (2010). *Der deutsche Wortschatz nach Sachgruppen*. 8th ed. De Gruyter.
- Lau, Jey Han et al. (2011). „Automatic Labelling of Topic Models“. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 1536–1545.
- Trier, Jost (1931). *Der deutsche Wortschatz im Sinnbezirk des Verstandes; die Geschichte eines Sprachlichen Feldes*. Heidelberg: Winter.

References II

 Willand, Marcus/Nils Reiter (2017). „ Geschlecht und Gattung. Digitale Analysen von Kleists ›Familie Schroffenstein‹ “. In: *Kleist-Jahrbuch* 2017. Ed. by Günter Blamberger et al., pp. 142–160.