LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

SMO Workshop:

Natural Language Processing

A brief introduction

Gregor Wiedemann | g.wiedemann@leibniz-hbi.de
Media Research Methods Lab
Leibniz-Institute for Media Research | Hans-Bredow-Institut

Workshop schedule

- 1st half: Intro lecture:
  - Text as data / pre-processing
  - Lexicometrics: Frequency, Keyness, Co-occurrence
- 2nd half: 2 Tutorials:
  - preprocessing basics
  - lexicometric tweet analysis

Defintion

- „Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data." (Wikipedia 2022)
- NLP utilizes
  - linguistic knowledge
  - statistical knowledge

Text as data

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- **symbol** ← meaning
  - character: linguistic unit (meaning representation)
    - 豊 δ A ♣ …
  - glyph: graphical representation of a symbol
    - a ← {a **a** a a A A A A}
- **alphabet:** fixed set of characters
  - {a, b, c, d, e, …}
- **encoding**
  - unambigous assignment of characters from an alphabet to {bit patterns, octets, …}
  - standards („a“): ASCII (61), ISO-8859-1 (61), UTF-8 (U+0061), …

Text as data

- **String:** concatenation of alphabet elements
  - „Hello world!“, „“, „0001011110001010 1“, „To be or not to be…“
  - essential, elementary data type in computer linguistics
  - common operations: e.g.
    - concatenation: „Hello“ + „World“ + „!“ → „Hello World!“
    - splitting: split(„Hello World!“, „ “) → {„Hello“, „World!“}
    - case conversion: uppercase(„Hello“) → „HELLO“
    - substring: substr(„Hello“, start = 0, length = 4) → „Hell“
- **Document:** compound data type
  - (collection of) strings (e.g. title, body) [+ Metadata]
- **Corpus:** collection of documents

Text as data

- **Type** (cp. class)
  - (abstract) string representing a meaningful concept, e.g. words
- **Token** (cp. object)
  - (concrete) string as instance of a meaningful concept

{ disciplines, distinction, concept, … }

„In disciplines such as knowledge representation and philosophy, the type–token distinction is a distinction that separates a concept from the objects which are particular instances of the concept."

(Wikipedia → Type–token distinction)

- **Vocabulary**
  - complete set of all types occuring in a [document | collection]

Text as data

- Transformation of text into numerical objects

$$\left\{ \begin{array}{l} \text{disciplines} \\ \text{distinction} \\ \text{concept} \\ \ldots \end{array} \right\} \qquad \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 4 & \ldots \\ 2 & 0 & 0 & \ldots \\ 2 & 5 & 1 & \ldots \end{pmatrix}$$

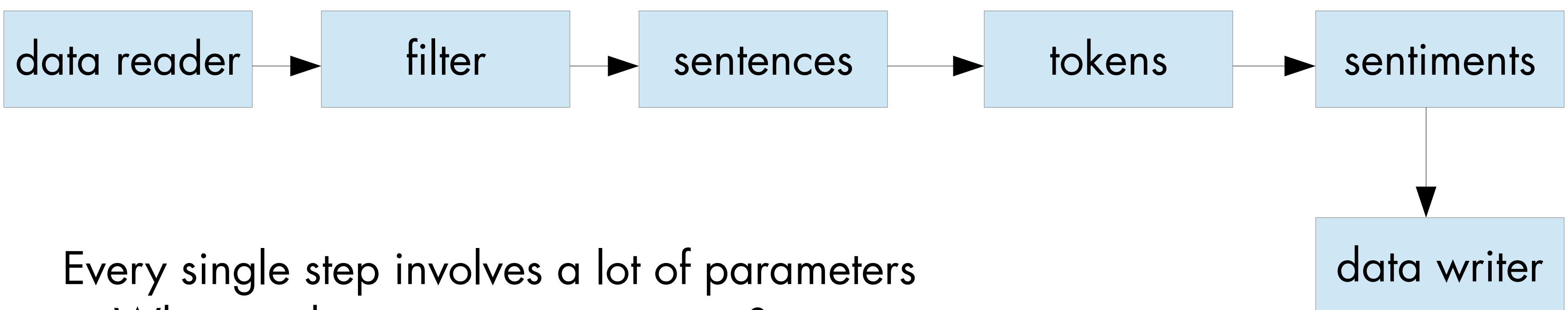
List of strings | vector | matrix

- Transformed objects → Data Mining
  - process of discovering patterns in large data sets

NLP piplines

- **PIPELINE**: application of different data manipulation procedures in row
  - preprocessing
  - actual analysis
  - output format

e.g.



Every single step involves a lot of parameters
→ What are best parameter settings?
→ Reproducibility?

Vector Space Model

- Idea: Encode textual
  - documents in **vectors**
  - collections in **matrices**
- **data = event counts**
- dimensionality of vector space
  - |vocabulary of collection|

- D1: Kim is leaving home.
- D2: Kim is at home.
- D3: Karen is leaving.

Kim	is	leaving	home	.	at	Karen
1	1	1	1	1	0	0
1	1	0	0	1	1	0
0	1	1	0	1	0	1

Document-Term-Matrix

$$\text{n documents} \quad \begin{pmatrix} 1 & 0 & 5 & 1 & \cdots \\ 0 & 0 & 0 & 2 & \\ 1 & 3 & 0 & 1 & \\ 0 & 0 & 0 & 0 & \\ \vdots & & & & \ddots \end{pmatrix}$$

m word types

- may get very large!
- events: frequency counts of word types in each document
- bag of words
- very sparse (contains mostly zeros)
- variations:
  - binary event counts
  - paragraphs as documents
  - sentences as documents
  - additional n-grams (n > 1) as events
  - …

n – size of collection
m – size of vocabulary

Stop words

- stop words = list of words considered as no meaningful for specific NLP task
- → can be filtered out of global vocabulary to reduce data / improve performance

- am
- among
- amongst
- amoungst
- amount
- an
- and
- another
- any
- anyhow
- anyone
- anything
- anyway
- anywhere
- are
- around
- as
- at
- back
- be
- became
- because
- become
- becomes
- becoming
- been
- before
- beforehand
- behind
- ...

N between ~100 … ~1000

Stop words

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

Nach den Ungarn hatten auch die Comecon-Behörden bereits die Hand nach Kapitalisten-Dollars ausgestreckt. Vor wenigen Monaten hatte die Internationale Bank für Wirtschaftliche Zusammenarbeit in Moskau, das Finanzzentrum des Comecon, bei einer europäischen Bankengruppe elf Millionen Dollar ausgeliehen.

In Wall Street werden indes noch Anleihen gehandelt, die eine russische Schuld aus dem Jahre 1916 verbürgen. Damals lieh sich Zar Nikolaus II 75 Millionen Dollar. Doch nach der Revolution verweigerten die Kommunisten die Rückzahlung der Schuld. Heute sind 1000 Dollar von 1916 in Wall Street nur noch um 40 Dollar wert. Zur Zeit der Konferenz von Jalta im Jahre 1945 hatte dagegen die späte Hoffnung auf eine Erstattung der Russenschulden die Papiere immerhin auf einen Kurs von 230 Dollar hinaufgetrieben. smi

Stop words

[redacted] Ungarn [redacted] auch [redacted] Comecon-Behörden bereits [redacted] Hand [redacted] Kapitalisten-Dollars ausgestreckt. [redacted] Monaten [redacted] Internationale Bank für Wirtschaftliche Zusammenarbeit [redacted] Moskau, [redacted] Finanzzentrum [redacted] Comecon, [redacted] europäischen Bankengruppe elf Millionen Dollar ausgeliehen.

[redacted] Wall Street [redacted] indes noch Anleihen gehandelt, [redacted] russische Schuld [redacted] Jahre 1916 verbürgen. Damals lieh [redacted] Zar Nikolaus II 75 Millionen Dollar. [redacted] Revolution verweigerten [redacted] Kommunisten [redacted] Rückzahlung [redacted] Schuld. Heute [redacted] 1000 Dollar [redacted] 1916 [redacted] Wall Street nur noch [redacted] 40 Dollar wert. [redacted] Zeit [redacted] Konferenz [redacted] Jalta [redacted] Jahre 1945 [redacted] späte Hoffnung [redacted] Erstattung [redacted] Russenschulden [redacted] Papiere immerhin [redacted] Kurs [redacted] 230 Dollar hinaufgetrieben. smi

Pruning

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Pruning = filtering the vocabulary of a collection by minimum / maximum thresholds of occurrence
- very useful preprocessing step to reduce vocabulary size:
  - Count occurrence of types in the complete collection
  - keep only those terms which occur above / below a defined threshold
- Caution: distribution of language data → see chapter *„frequency analysis"*

- term frequency:
  - sum all term occurrences in all documents
  - filter terms which occur e.g.
    count(term) > 1 AND count(term) < 100
- document frequency:
  - for each term count number of documents in which it is contained
  - allows for filters like: terms which occur e.g. in
    more than 99% AND
    less than 1%

    of documents

Unification: Stem vs. Lemma

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Unification:
  - observation: similar semantic types share similar orthographic forms
  - connect -ion, -ions, -ive, -ed, -ing
  - Idea: map variants to reduced form
    - → reduce vocabulary
    - → reduce data sparsity

- Two methods:
  - Stemming: cut of endings by language specific rules
  - Lemmatization: mapping of types to linguistic its lemma by dictionary lookup (external resource)

Stemming

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

- Standard appraoch: Porter Stemmer (1980) / Snowball
- separation of suffixes by rules, e.g.
  - SSES → SS caresses → caress
    IES → I ponies → poni
  - (if m>1) EED → EE feed → feed
    agreed → agree
    m = number of syllables
- Problems:
  - overstemming: artificial ambiguity
    - {organization, organ} → organ
  - understemming: unification fails
    - European → european, Europe → europ

Lemmatization

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Lookup of canonical / dictionary form
- usually retrieved by long dictionary files which contain
inflected type	lemma type
European	Europe
Europe	Europe
Organizations	Organization
- Problems:
  - getting external resources (e.g. ASV Leipzig list of > 600.000 type-lemma-relations for German)
  - incomplete lists

Example

McLean Industries Inc's United
States Lines Inc subsidiary said it has agreed in principle to
transfer its South American service by arranging for the
transfer of certain charters and assets to <Crowley Mariotime
Corp>'s American Transport Lines Inc subsidiary.
U.S. Lines said negotiations on the contract are expected
to be completed within the next week. Terms and conditions of
the contract would be subject to approval of various regulatory
bodies, including the U.S. Bankruptcy Court.

PREPROCESSING

mclean industri inc united
st line inc subsidiari said agre principl
transfer south american servic arrang
transfer certain charter asset crowley mariotime
corp american transport line inc subsidiary
line said negoti contract expected
complet within next week term condit
contract subject approv various regulatory
bodies includ bankruptci court

- lowercase
- remove punctuation
- remove stop words
- stemming
- strip white spaces

Sentence detection / Tokenization

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- → Essential preprocessing step!
  Badly tokenized text data may lead to bad results
- frequent errors:
  - intra-word dashes: 'front-end' → 'front end' OR 'front-end'
  - quotation marks '„Hello“' → '„ Hello “' OR '„Hello“'
  - dots for abbreviation: 'Mr.' → 'Mr .' OR 'Mr.'
  - colon / semicolon: 'Monday' → 'Monday:' OR 'Monday :'
  - apostrophe:

    „O'Neill“ → „Neill“ OR „ONeill“ OR „O'Neill“ OR „O ' Neill“ OR „O' Neill“
    „aren't“ → „aren t“ OR „aren't“ or „arent“ or „are n't“

Part-of-Speech

- Task PoS-Tagging = Assign a word type label to each token in a sentence
  - The cat barks at the dog .
  - DET NN VF PRE DET NN $
- Ideal task for machine learning classifiers on annotated training data!
  - e.g. Conditional Random Field classifier: most probable sequence of outcome labels to an input sequence
- label sets are called „tag sets" → different sets for different languages / tasks
  - English: Penn Treebank POS tags (36 labels)
  - German: STTS Stuttgart/Tübingen Tagset (57 labels)
  - Translingual: Universal POS tags (17 labels)

Summary

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Linguistic Preprocessing
  - shall reduce / unify data for application specific purpose
  - may contain various steps in row
    - Encoding
    - Spelling correction
    - Removing uninformative data: noise, duplicates, stopwords, low/high frequent terms (pruning), dictionaries
    - Sentence detection, tokenization, Part-of-Speech tagging
    - Unification: punctuation, capitalization, stemming, lemmatization
  - best setup usually has to be identified experimentally (or by experience)
  - caution: order of steps may influence result!

Lexicometrics

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

1. Frequency analysis
2. Key term extraction
3. Cooccurrence analysis

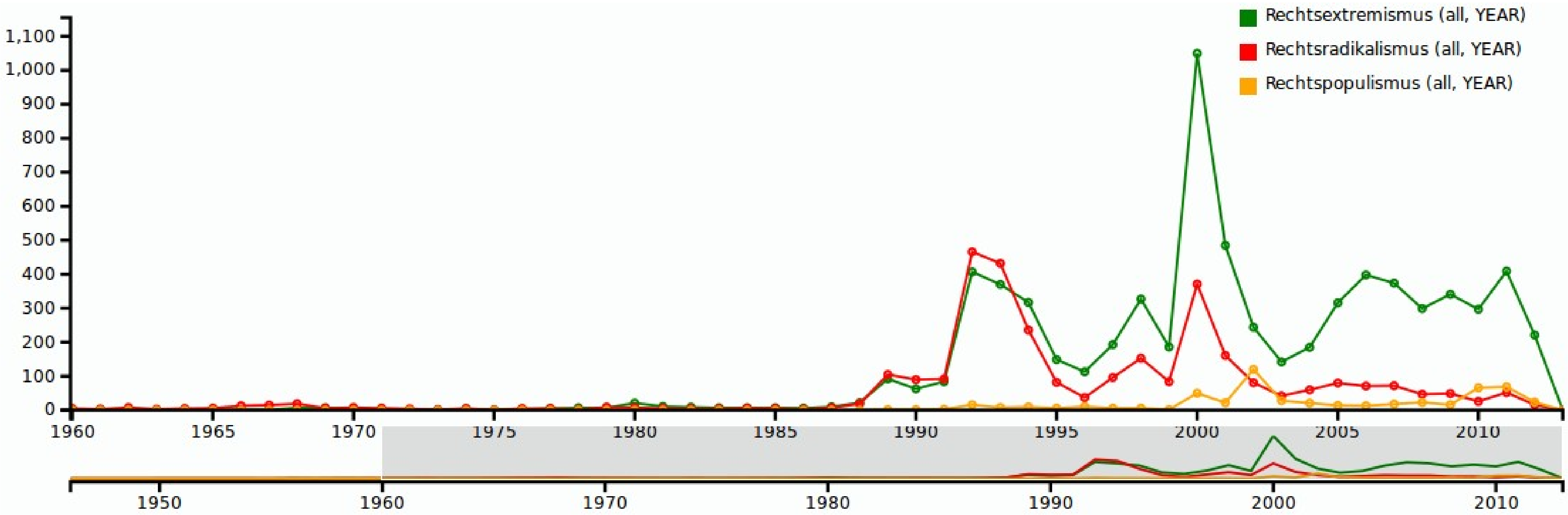
1. Frequency analysis

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Motivation: Analysis: comparing frequencies of units of analysis per context
  - 1) between different UoA
  - 2) in different collections
  - 3) over time
- Possible Units of Analysis (UoA):
  - **terms** → in CA we often will concentrate on those
  - concepts (set of terms), ...
  - documents, paragraphs, ...
  - linguistic units (sentences, punctuation marks, vowels, ...)
- Context Units
  - term frequency: frequency of a term within a document / entire collection
  - document frequency: frequency of documents containing a term

1. Frequency analysis

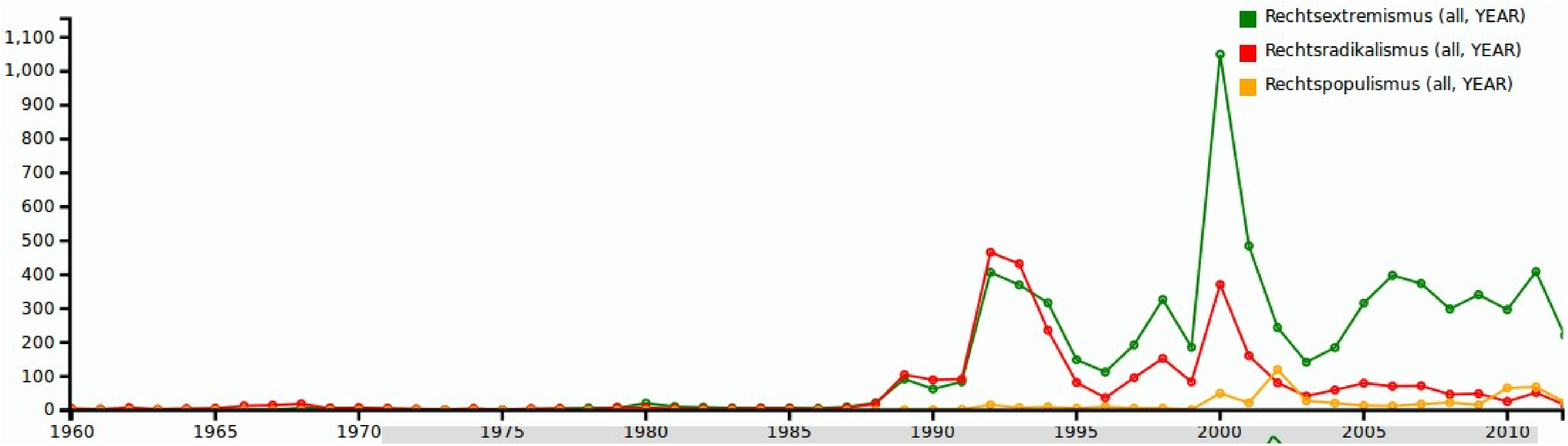
LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT



- **Problems of „term as events“:**
  - distribution of language data
  - „burstiness of terms“
  - varying collection sizes → normalize frequencies by collection size!

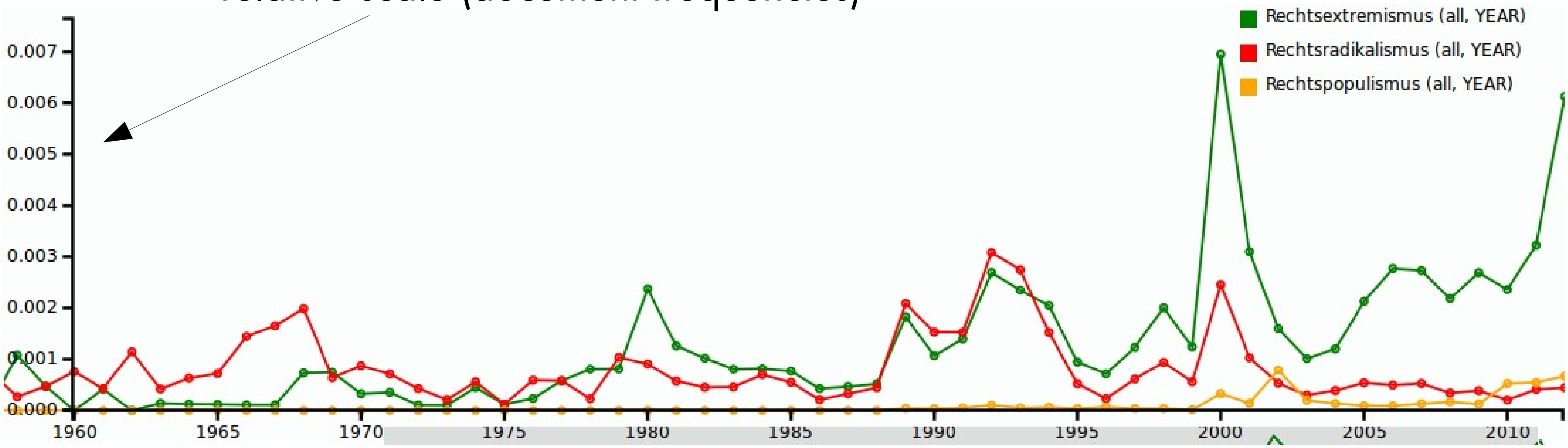
1. Frequency analysis

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT



1. Frequency analysis

relative scale (document frequencies)



1. Frequency analysis

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

1. Frequency analysis

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- **Dictionaries** (curated list of words) can be compiled to count conceptual events
  - e.g. basic approach of *sentiment analysis* → identification of subjective mood in source materials
    - positive terms: {good, awesome, brilliant, gorgious, …}
    - negative term: {bad, aweful, horrifying, devastating, …}

→ Tutorial 3

  - intersection of discoursive fields:
    - war terminology: {blitz, bomb, formation, neutral zone, red zone, kamikaze, …} measured in articles about soccer v american football
  - operationalization of theoretical hypthothesis
    - TINA rethorics: {no alternative, no other possibility, impossible, indispensable, …} with respect to different policy fields
- **Caution:**
  - Should all events count equally? (e.g. sentiments)
  - does occurrence match appropriate context? (feature-/aspect based sentiments)

Applying frequency analysis

- Context matters!
  - counting simple occurrence usually neglects contexts
  - but, right contexts can be assured by previous selection strategies
    - e.g. counting „no alternative“ in documents on European politics compared to a general corpus
  - ← Applying filter beforehand increases chances to generate informative data
- Utilization of frequency data for description / identification of
  - content shares → e.g. pie chart
  - trends / time series → e.g. line chart
- consider normalization strategies

Key term extraction

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- One task, many names:
  - „Terminology mining, term extraction, term recognition, or glossary extraction, is a subtask of information extraction. The goal of terminology extraction is to *automatically* extract relevant terms from a given corpus." [Wikipedia]
- Evaluation:
  - judgements on relevancy done by human experts

- Approaches based on:
  - Frequency
    - Frequency
    - TF-IDF
  - Comparison corpus
    - Log likelihood
    - Characteristic elements diagnostics

Frequency

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

- Assumption
  - the more frequent, the more important
  - removing stop words helps to identify more relevant terms
- Evaluation
  - language is Zipf distributed
  - raw frequency does not cover relevancy well

- Example:
  - protest data TAZ (2000-2009)
- Approach to get n most relevant terms
  1) create DTM from corpus
  2) compute vector v of column sums
  3) order v in decreasing order
  4) output item 1 to n of v

type/frequency		type/frequency (sw removed)	
die	22807	polizei	2072
der	19938	menschen	1492
und	11426	demonstration	1105
den	7180	berlin	982
das	5560	uhr	968
von	5145	demonstranten	961
auf	5013	kundgebung	700
mit	4957	samstag	666
sich	4668	neonazis	651
dem	4205	worden	632
ein	4188	straße	625
nicht	3909	npd	572
für	3858	berliner	558
eine	3625	jahr	537
ist	3486	jahren	534
des	3308	teilnehmer	532
sie	3299	rechten	484
auch	3115	seien	477
gegen	3070	demo	472
als	2521	motto	459

TF-IDF

- Basic assumption:
  - relevancy is correlated with term frequency and inversed document frequency

$$N = |D|$$

$$idf_w = \log\left(\frac{N}{n_w}\right)$$

$$weight_w = tf_{wd} \cdot idf_w$$

polizei	7.089975
rund	6.731556
neonazis	6.309970
uhr	6.270761
samstag	6.236580
kundgebung	6.021211
menschen	5.990043
npd	5.892383
sie	5.598942
gestern	5.563909
ist	5.556133
etwa	5.476461
berlin	5.457175
aufmarsch	5.453003
demonstranten	5.437748
hatten	5.436246
teilnehmer	5.336355
rechten	5.246831
nicht	5.204938
unter	4.991625

Corpus comparison

- Difference based Term Extraction methods follow a different approach:
  - comparing frequencies in a target corpus T with frequencies in a general comparison corpus C
  - significant deviation in T from expected term distribution measured in C is considered as relevancy criterion
- Tests used in CA
  - Log Likelihood (Dunning 1993; Rayson/Garside 2000)
  - Characteristic elements diagnostics (Lebart/Salem 1994)

Log Likelihood

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Contingency Table

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

- Log Likelihood
  - E1 = c*(a+b) / (c+d)
  - E2 = d*(a+b) / (c+d)
  - LL = 2*((a*log (a/E1)) + (b*log (b/E2)))

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

	LL	Frq
NPD	7867,59	1157
Demonstration	7789,70	1295
Demo	7098,27	829
Demonstranten	5463,47	1042
Kundgebung	5306,27	790
Neonazis	5224,08	751
Polizei	5165,54	2262
Aufmarsch	3704,68	468
Gegendemonstranten	2811,12	320
Neonazi	2565,88	305
taz	2438,61	474
Anti	2380,52	232
Antifa	2237,23	243
Demonstrationen	1841,17	400
Teilnehmer	1722,93	582
Teilnehmern	1464,79	335
Bündnis	1390,89	396
Nazis	1377,34	359
Motto	1370,38	468
Protest	1324,34	412

Cooccurrence Analysis

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- Structuralist semantics (F. de Saussure):
  - syntagmatic relation: signifiers which occur conjointly complement w.r.t function and content
  - paradigmatic relation: signifiers which occur in similar contexts have similar function w.r.t. grammar and content → cp. distributional hypothesis
- Computing cooccurrences
  - local context C(w): set of words that occur in the same 'window' as *w*
  - global context G(w): set of words which occur conjointly with *w* in a *statistically significant* manner
  - windows: sentences, paragraphs, documents, headlines, k left/right neighbour words

Cooccurrence Analysis

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

The sun is shining.	$C_{sentence}(sun) = \{The, is, shining\}$
The sun is burning.	$C_{sentence}(sun) = \{The, is, burning\}$
The light is shining.	$C_{sentence}(light) = \{The, is, shining\}$

$$G(sun) = \{The, is, shining, burning\}$$

$$G(sun) \sim G(light)$$

Cooccurrence Analysis

LEIBNIZ-INSTITUT FÜR MEDIENFORSCHUNG HANS-BREDOW-INSTITUT

- Counting co-occurrence
  - => focus on high frequent events in text data (Zipf's law!)
  - maximal frequency pair: „the – of“
- Determine significance of co-occurrence
  - statistical test measuring „surprise“
  - => better captures semantic characteristics of a text
  - there is not *the* single measure

Cooccurrence Analysis

- statistical significancy
  - measure of deviation from random conjoint occurrence
- measurements
  - n_a – windows containing A
  - n_b – windows containing B
  - n_{ab} – windows containing A and B
  - n – number of all windows

(bag of words within windows)

- significance measures
  - Frequency (baseline*)
  - Dice
  - Mutual Information
  - Log Likelihood

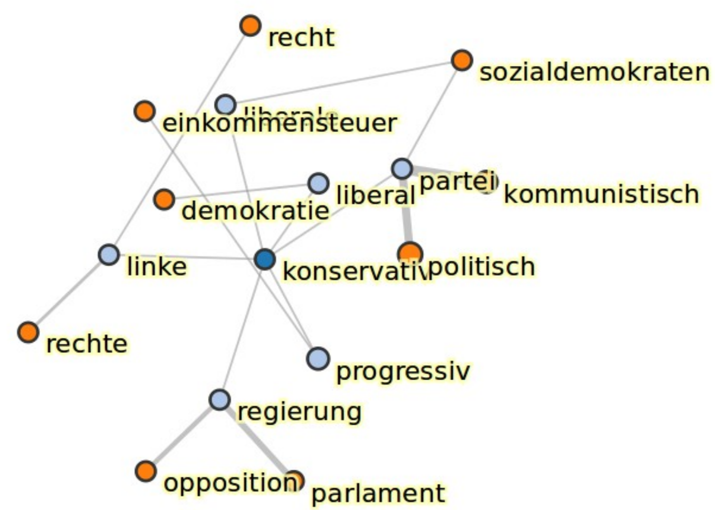
* remember Zipf!

Application in Social Science

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

- (change of) meaning may be inferred from cooccurrence results
- cooccurrence analysis → comparison of different result sets
  - change of context units (neighbours, sentence, document, …)
  - filter terms by POS-/NE-types
  - tracking change of global contexts by comparing time ranges

- Visual analytics:
  - tables
  - graphs
  - KWIC-Lists



NLP on Twitter

- Challenges:
  - short texts (280 chars. max)
  - special tokens: @mentions, #tags and URLs
  - special types: posts, replys, retweets
  - non-standard language variety (slang, „nooooo!“)
  - complex context (reply, conversation, debate)
  - complex metadata

- Solutions:
  - considerate decisions for handling special cases in text preprocessing
  - identification of fitting models, correction for their potential errors
  - **validate! validate! validate!**