SEMINAR
KLASSIFIKATION & CLUSTERING

KORPUSAUFBEREITUNG
TRAININGS- UND TESTKORPORA

Stefan Langer

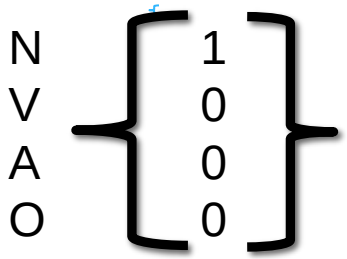
stefan.langer@cis.uni-muenchen.de

Typen von Merkmalen

- Man unterscheidet üblicherweise zwischen folgenden Merkmalstypen
  - Nominale Merkmalen
    - Ausprägungen lassen haben keine festgelegte Reihenfolge
    - Beispiel: Wortarten (parts of speech)
    - → Keine Mittelwertbildung o.ä.
  - Ordinale Merkmalen
    - Ordnung auf Ausprägungen möglich aber Abstände nicht systematisch
    - z.B. Bewertungen in der Sentiment-Analyse (gut mittel schlecht),
    - Relationen wie >, < (größer, kleiner) anwendbar; Mittelwertbildung problematisch
  - Metrische Merkmalen
    - z.B. Wortlänge
    - → Operationen wie Mittelwertbildung möglich

Merkmalsumwandlung zur Klassifikation

- Die Eingabe eines Datensatzes für einen Klassifikationsalgorithmus erfolgt in der Regel in Form eines Vektors mit numerischen Werten
  - Nominale Merkmale
    - Die verschiedenen Ausprägungen von nominalen Merkmalen werden jeweils einer Vektordimension mit binären Werten zu gewiesen (one hot encoding). Beispiel Wortarten.



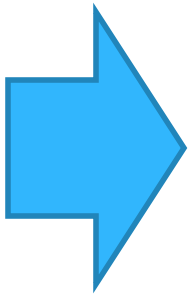
  - Ordinale Merkmale
    - da pseudometrisch, können sie eventuell als metrische Merkmale behandelt werden
    - Alternativ wie nominale Merkmale
  - Metrische Merkmale
    - Normalisiert oder original

Ziel – Transformation in Vektor / Tensor

Grimstad den 6 October 1844 Kjære Hanna! Tilgiv at jeg ikke før har besvaret Dit Brev, men det har været mig reent umuligt. I hele Sommer har jeg glædet mig til, at jeg skulde komme hjem til Din Confirmation; men det kan ikke blive noget af, da her fortiden er meget at bestille, og Reimann har just ikke synderlig Lyst til at gjøre noget selv; jeg kan saaledes ikke mundtlig aflægge Dig min Lykønskning til denne Dag; ...

Det var ret en Festdag øm Moder og kjærlig Veninde – Den kjære Afdøde, hvis gode Forstand, tilligemed flere Legemsfortrin, udmærkede den fremfor de fleste af dens Lige, var ogsaa en god Musekat, der fangede Rotter og Muus ligesaa let, som Mdm Reimann uddeler et Ørefigen. For denne Gang maa det være nok da Du veed jeg ikke kan Noget med at skrive lange Breve, og maa desuden skrive hjem iaften. Svar mig endelig saasnart Du faaer Tid dertil Din hengivne HJIbsen

Grimstad den 5te Januar 1850. Min kjære Ven! Gjennem din sidste Skrivelse har jeg modtaget «Catilinas» Dødsdom, – det gjør mig ondt, men det kan ikke nytte at tabe Modet. Du har virkelig Ret i at dette tilsyneladende Nederlag i Grunden ikke er at betragte som et Saadant. C. var jo kuns bestemt til at være en Forløber for de Planer vi i denne Retning have aftalt, og den kan endnu ligefuldt opfylde sin Bestemmelse. Jeg er ganske af samme Mening som du, nemlig at det er rigtigst at sælge Stykket, og jeg troer at Stykkets Afviisning, ifølge Directionens Skrivelse snarere vil gavne end skade, da det ikke synes som om Mangel paa indre Gehalt har bevirket at det ikke blev antaget. Salget af Stykket afhandler du naturligviis efter eget Forgodtbefindende, kuns vil jeg bemærke at det forekommer mig bedre at sælge Forlagsretten end at lade det trykke paa eget Forlag, da vi i sidste Tilfælde baade maatte rykke ud med en Deel Penge, forat bestride Trykningen, og desuden først lidt efter lidt vilde fra Thelemarken har jeg benyttet til nogle mindre Digte, afpassede efter bekjendte Folkemelodier

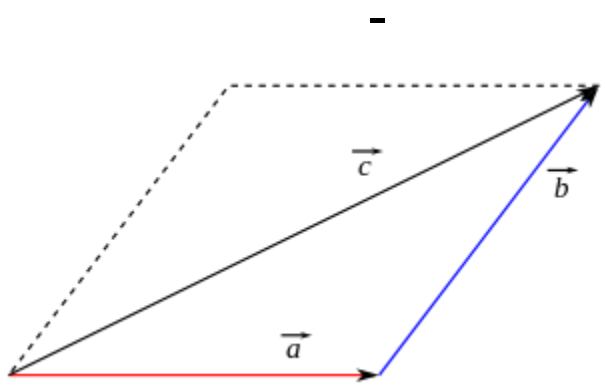


	F1	F2	F3	F4
D1	0.3	0	0.2	0
D2	…			
D3				
D4				
D5				
…				

Vektoren

Ein Vektor ist eine geordnete Menge von numerischen Werten mit N Dimensionen (N-Tupel reeller Zahlen).

$$\vec{x} = \begin{pmatrix} 7 \\ 3 \\ 0 \end{pmatrix}$$



Vektoren

- Ein Vektor ist eine geordnete Menge von numerischen Werten in N Dimensionen
- $\vec{x} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$
- In der Darstellung eines Textes in einem dünn besetzten Vektor repräsentieren die Dimension in der Regel Wörter oder längere Ausdrücke:
- *Das Kind sieht das Rind*
- $\vec{x} = \begin{pmatrix} das & 2 \\ kind & 1 \\ sieht & 1 \\ rind & 1 \end{pmatrix}$

Wiederholung: Vektoren

- Ein normalisierte Vektor ist ein Vektor der Länge 1
- Die Länge eines Vektors

$$|\vec{x}| = \sqrt{a^2 + b^2 + \ldots}$$

Ein Vektor kann einfach normalisiert werden, indem man ihn durch seine Länge teilt.

Vektoren und Wortlisten

- Dokumente können als Wortlisten dargestellt werden
- Wortlisten können als Vektoren dargestellt werden. Die Dimensionen des Vektors sind die Wörter/bzw. Wortformen
- Die Werte des Vektoren können sein
  - Wortvorkommen (→ binärer Vektor)
  - Worthäufigkeiten/relative Häufigkeiten/TF-IDF
  - Zusammenfassungen von Worthäufigkeiten/reduzierte Matrix

→ Sparse Vector

Binäre Vektoren

Spezialfall eines Vektors sind die sogenannten binären Vektoren – sie sind eine geordnete Folge von Nullen und Einsen (d.h. die Werte können nur 0 und 1 sein):

$$\vec{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Frequenz & Häufigkeit: Übersicht

- Absolute Häufigkeit
- Relative Häufigkeit
- Dokumentfrequenz
- IDF (inverse document frequency)
- TF (term frequency)
- TF/IDF

Absolute Häufigkeit und relative Häufigkeit

- Absolute Häufigkeit: f (einfaches Abzählen)
- Gesamtzahl der zählbaren Einheiten (z.B. Wörter): N
- Relative Häufigkeit: h = f/N

TF-IDF

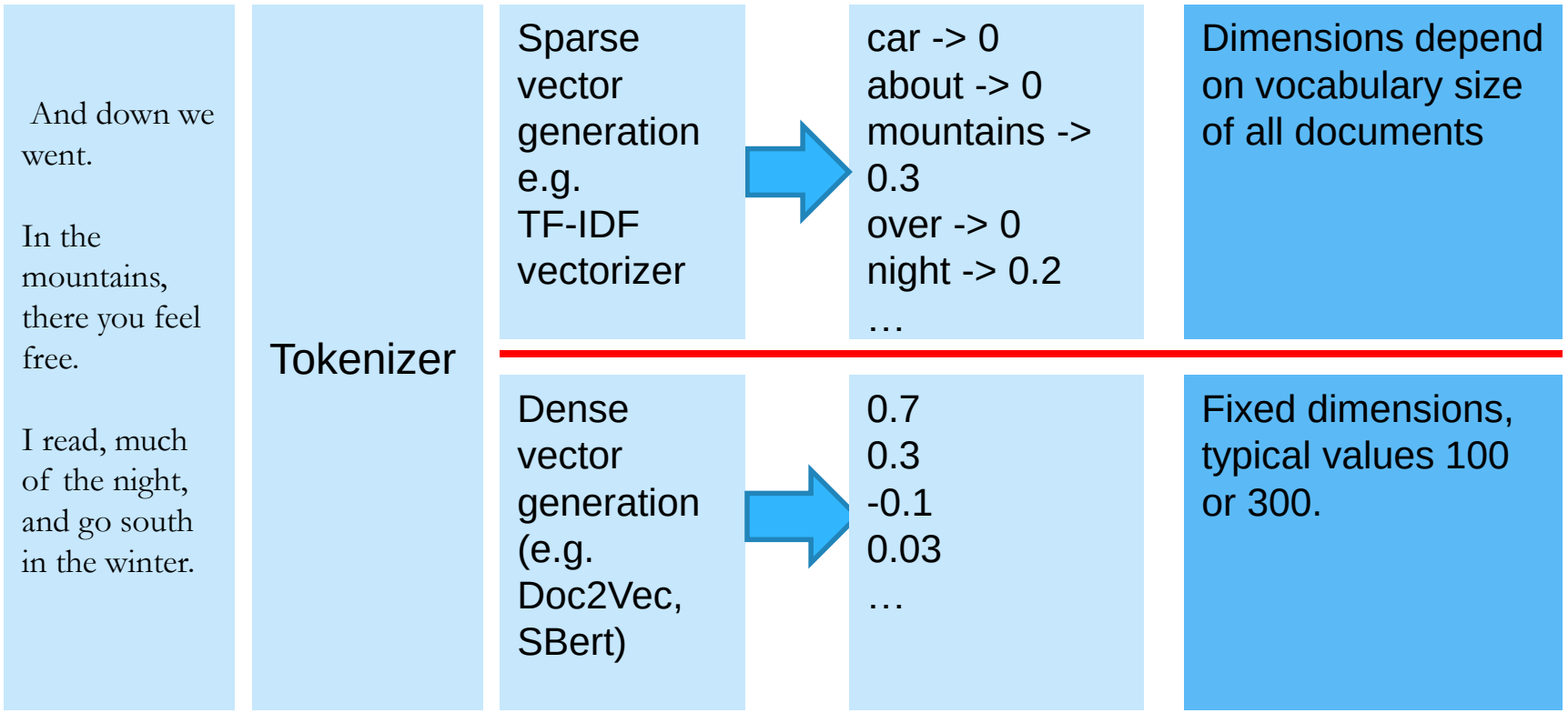
Maß für die Signifkanz eines Terms in Bezug auf ein Dokument

- tf (= h) is die relative Häufigkeit eines Terms in einem Dokument

- df (document Frequency)
  - $df = /d{:}t \in d\,/$ (Anzahl der Dokumente, die den Term t enthalten)

- idf (inverse document frequency) ist der Logarithmus aus der invertierten relativen Dokumentenfrequenz des Terms
  - $idf = log \frac{|D|}{|d{:}t \in d|}$
  - Kombination aus tf und idf.
    - $tf\text{-}idf = tf{*}idf$

Dense / sparse vectors



Vektoren und Abstandsmaße

- Zahlreiche Klassifikations- und Clusteringalgorithmen arbeiten mit Distanz- bzw. Ähnlichkeitsmaßen auf Featurevektoren
  - z.B.
    - K-nearest neighbour
    - Single Link und Complete Link Clustering

Länge eines Vektors (vgl. Pythagoras)

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Vektornormalisierung

- **Normalisierter Vektor**
- Ein Vektor mit der Länge 1. Einige Rechenoperationen auf Vektoren erfordern normalisierte Vektoren. Ein Vektor kann normalisiert werden, indem der Wert aller Dimensionen durch die Vektorlänge geteilt wird

Vergleich zweier Vektoren

- Um zwei Vektoren vergleichen zu können, müssen sie dieselbe Anzahl von Dimensionen besitzen.

Wiederholung Euklidische Distanz

$$\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

- Weitere Optionen für Vektordistanz:

Allgemeine Form: Minkowski-Metrik

$$A = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Tchebyshev (= max $|x_i - y_i|$):

$$A = \lim_{p \to \infty} \left(\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \right)$$

Vektorenabstandsmaße

Maß	Definition	Einschränkungen	Anmerkungen
Einfache Übereinstimmung	$\vec{x} \cap \vec{y}$	nur binäre Vektoren	Summe aller Einträge in beiden Vektoren, die 1 sind.
Dice-Koeffizient	$\frac{2\lvert\vec{x} \cap \vec{y}\rvert}{\lvert\vec{x}\rvert + \lvert\vec{y}\rvert}$	nur binäre Vektoren	Länge der Schnittmenge beider Vektoren durch die Summe der Länge beider Vektoren
Jaccard-Koeffiziernt	$\frac{\lvert\vec{x} \cap \vec{y}\rvert}{\lvert\vec{x} \cup \vec{y}\rvert}$	nur binäre Vektoren	
Euklidische Distanz	$\lvert\vec{x} - \vec{y}\rvert = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$		Distanzmaß!
Manhattan-Metrik, L1-Metrik	$\sum_{i=1}^{n}\lvert x_i - y_i\rvert$	nach Manning-Schütze v.a. für Vektoren, die bedingte Wahrscheinlichkeiten enthalten	Distanzmaß! Senkrechte Striche sind hier Betragsstriche
Cosinus	$\frac{\vec{x} \cdot \vec{y}}{\lvert\vec{x}\rvert\lvert\vec{y}\rvert}$		Sind die Vektoren normalisiert (Länge 1) ist dieser Wert gleich dem Skalar-produkt
Skalarprodukt	$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$	= Cosinus, nur für normalisierte Vektoren	

Euklidische Distanz ergibt selbe Abfolge wie Cosinus

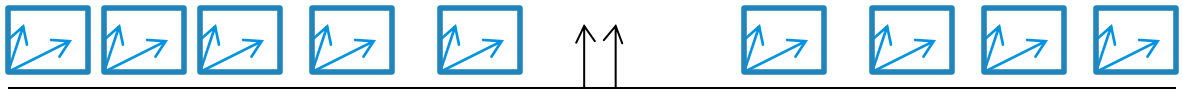
Embeddings

- Word embedddings sind ein Verfahren um Wörter oder Textsequenzen als Vektoren darzustellen (danneben gibt es andere, wie etwa Latent Semantic Indexing (LSI, a. LSA). Dazu wird der Kontext der Wörter (als ungeordnete Menge (bag of words) oder gewichtet nach Abstand herangezogen, um die Vektoren zu generieren
- Bekannster Algorithmus: word2vec (frei verfügbare Implementationen).

Embeddings – word2vec

- Unterschiedliche Verfahren:
- CBOW trains prediction of target words from source context words (sum of the source context word vectors)

Der Mensch ist ein Abgrund, es schwindelt einem, wenn man hinunterschaut



- SkipGram trains prediction of the source context-words from the target word

Der Mensch ist ein Abgrund, es schwindelt einem, wenn man hinunterschaut

Sentence and document embeddings

Options:

- Calculated on bases of word embeddings (e.g. weighted average)
- Doc2Vec and similar approaches
- Transformers (e.g. SentenceBERT)
- Large language models - GPT et al

Besondere Eigenschaften von Textdaten

- Potentiell sehr viele Features/Dimensionen (Wortformen, Wort-Bigramme / Trigramme)
- Sparse (viele Wörter / Wortfolgen sehr selten)
- Tokenisierung – Satzzeichen …
- Morphologie – unterschiedliche Formen desselben Wortes
- Synonymie

Weitere Eigenschaften eines Textkorpus

- Sprache
- Dokumententyp / Format
  - PDF, Html, plain text, Bilddatei (-> OCR)
- Qualität der Textdaten (z.B. wegen <- OCR)
- Größe des Gesamtkorpus
- Größe der einzelnen Dokumente
- Thema/Themen
- Autor(en)

Textaufbereitung - Featurebestimmung

- Terme: N-Gramme / Token / Wörter / Phrasen
- Termbasiert - Vektoren:
  - Binäre Vektoren
  - Relative Häufigkeit
  - TF-IDF
- Verfahren zur Reduzierung von Vektoren:
  - Latent semantic indexing (LSI, auch … analysis, LSA)
- Semantische Aufbereitung
  - word embeddings
    - word2vec
    - Glove
    - ...
  - document embeddings

Textbeispiel

Unendlicher Spaß (engl. Originaltitel: *Infinite Jest*) ist ein Roman von David Foster Wallace aus dem Jahr 1996. Das im englischen Original 1079 Seiten starke Buch wurde vom TIME-Magazin zu einem der 100 einflussreichsten Romane seit 1923 gewählt.[1] Der Titel ist ein Zitat aus dem Shakespeare-Stück *Hamlet*.

Die FAZ bescheinigte dem Buch „endlose Sätze und ein apokryphes Fachvokabular“.

- David Foster Wallace: *Infinite jest : a novel.* 1st ed. Auflage. Little, Brown and Company, Boston 1996, ISBN 0316920045.
- David Foster Wallace: *Unendlicher Spaß.* 3. Auflage. Kiepenheuer & Witsch, Köln 2009 (Originaltitel: *Infinite jest*, übersetzt von Ulrich Blumenbach), ISBN 978-3-462-04112-5.

Textfeatures A: Ohne linguistische Aufbereitung

- Buchstaben
- Buchstaben-N-Gramme
- Großschreibung, Kleinschreibung
- Token / Wörter / Zahlen
- Wortfolgen (Bigramme, Trigramme …)
- Wortlänge
- Textlänge
- Textstruktur
  - Abschnittslänge

Textfeatures B: Mit linguistischer Aufbereitung

- Wortarten; selektive Wortauswahl
- Ergebnisse von Entity-Extraktion:
  - Eigennamen (Personen, Firmen, Orte)
  - Zeitausdrücke, Adressen, Emails
  - Ereignisse (Events)
  - Anwendungsspezifische Entities (z. B. Bezeichnungen von Maschinenteilen)
  - Relationen
- Sätze, Satzlänge
- Word embeddings, character embeddings

Extraktion von Daten aus dem Korpus

Textaufbereitung:

- Lesen im der korrekten Zeichensatzkodierung
- Behandlung von Groß- und Kleinschreibung
- Punktuation
- Weitere Normalisierung
  - Sonderzeichen
  - Zahlennormalisierung
  - Lemmatisierung
  - Abkürzungen

Corpora for text classification

- http://trec.nist.gov/data/reuters/reuters.html

- Reuters-21578
- RCV1/ RCV2

- 20 newsgroups
- Enron Email

- IMDB Movie Review Sentiment Classification (stanford). Sentiment Analysis
- News Group Movie Review Sentiment Classification (cornell). Sentiment analysis

Korpora: Der Klassiker - Reuters

- Reuters-21578
- 21,578 Texte der Nachrichtenagentur
- Gesammelt und klassifiziert von der Carnegie Group und Reuters für das CONSTRUE Textklassifikationssystem.
- 118 Themen
- 0-N Themen pro Dokument, aber meist 1 Thema
  - Download:
  - http://www.daviddlewis.com/resources/testcollections/reuters21578/
  - http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
- Nicht mehr ganz aktuell, jetzt eher Reuters-RCV1 (nächste Folie)

Reuters-RCV1

- Larger than original Reuters corpus (800 000 Documents)
- XML format
- Available on request
  - http://trec.nist.gov/data/reuters/reuters.html

Korpora - Wikipedia

- Wikipediaartikel sind in ein Kategoriensystem eingeordnet.
  - http://de.wikipedia.org/wiki/Wikipedia:Kategorien
  - http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Kategorien/Einordnung_von_Kategorien

Multihierarchisch

Benutzt zum: Training/Test z.B.

https://blog.codecentric.de/en/2013/03/how-to-use-wikipedias-full-dump-as-corpus-for-text-classification-with-nltk/

Als Wissensquelle fürs Training:

http://www.cs.ubbcluj.ro/~studia-i/2007-kept/207-BodoMinierCsato.pdf

Other corpora

English

- Reuters Corpus: A widely used corpus in text categorization research, containing news articles from the Reuters newswire service.
- 20 Newsgroups: A collection of approximately 20,000 newsgroup posts, partitioned into 20 different categories.
- Amazon Product Reviews: Customer reviews of products sold on Amazon, sentiment analysis and product categorization.
- Enron Email Corpus: Emails from the Enron corporation - email classification and text mining.
- Movie Review Data: Movie reviews - sentiment analysis and classification of movie reviews based on genre.

Multilingual and non-English

- Wikipedia Corpus: A large corpus of articles from Wikipedia - categorization and classification of articles based on subject matter.
- Europarl Corpus: A corpus of proceedings from the European Parliament, containing texts in 21 different European languages.
- MultiUN Corpus: A multilingual corpus of United Nations documents, available in six official languages: Arabic, Chinese, English, French, Russian, and Spanish.
- Pan-Unicode Corpus: A multilingual corpus of web pages in 107 languages, designed for cross-lingual text classification and retrieval.
- BabelNet: A multilingual lexical network that integrates information from many different sources, including Wikipedia and WordNet, and provides annotations for over 270 languages.
- The Corpus del Español: A large corpus of Spanish text, containing over 100 million words from a wide variety of genres, including news articles, literature, and spoken language.
- The Oslo Corpus of Tagged Norwegian Texts: A corpus of Norwegian text, containing over 10 million words of tagged text from a variety of genres, including news articles, literature, and scientific texts.

Our corpora

Corpus	Description	Languages	Purpose
Letters	Letters from various authors, different languages	en, de, da/no, others	language identification, author identification
Huffington post	News headlines and abstract from Huffington post, slightly adapted	en	Topic classification
IMBD movie	Movie reviews	en	Sentiment analysis

Inputformate

- PDF – sehr verbreitet, notorisch schwierig zu verarbeiten
- HTML
- Office-Formate (Word, Powerpoint, Excel) und Open-Source-Äquivalente
- Bilder (mit Text)