SENTIMENT LEXIKON (ANALYSIS)

SENTIMENT ANALYSIS
AUSGANGSPUNKT

- Aufgabe: Analyse von Texten zur Erkennung der Haltung (positiv/negativ)

- Typische Analyseschritte:
  - Gegeben: z.B. Wörterbuch mit bekannten Sentiments, Bsp. SentiWS
    (+ mögliche grammatikalische, … Regeln)
  1. Abgleich mit unbekanntem Dokument
  2. Aggregation zu finalem Sentiment Wert

- Frage:
  - Wie wurde das initiale Sentiment Lexikon erstellt?
  - Falls automatisch, wie?

SENTIMENT ANALYSIS
WÖRTERBUCHERSTELLUNG

Ideen:

- Manuell
- Existierende Wörterbücher → Übersetzung, Erweiterung (an Domäne)
- Neu? → aus Dokumenten mit NLP ⛏

SENTIMENT ANALYSIS
WÖRTERBUCHERSTELLUNG

Gegeben:

Dokumente mit Bewertungen, z.B. Reviews von Amazon, IMDb, Yelp, …

Gesucht:

Bewertungen der Wörter / Einfluss der Wörter auf die Dokumenten-bewertung

Ansatz: **?**

WÖRTERBUCHERSTELLUNG – ANSATZ

- Feature Importance (Selection); Koeffizienten
  - Evtl. bekannt aus anderen ML Verfahren, z.B. Lineare Modelle, Regression
  - Mutual Information
  - PMI (→ SentiWS, Stärke von semantischen Assoziationen)

„This one trick NLPler don‘t want you to know …“

- Wort-Bewertung-Korrelation
  - Pearson’s r (standard correlation coefficient),
  - Spearman’s rho (rank correlation),
  - Kendall’s tau
  - …

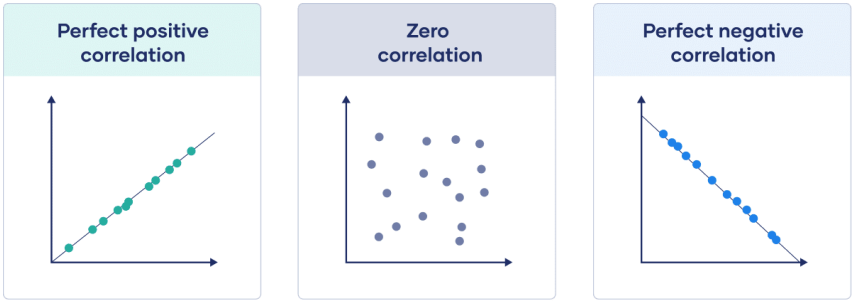
Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Quelle: https://www.scribbr.com/statistics/correlation-coefficient/

PEARSON-KORRELATIONSKOEFFIZIENT

- „quantitatives Maß zur Beurteilung der Stärke der Beziehung zwischen zwei stetigen Merkmalen“ Korrelationskoeffizient nach Pearson
- Wertebereich: [-1, +1]
  - Vorzeichen: +1 (-1) zeigt exakt positiven (negativen) linearen Zusammenhang
  - 0, wenn kein linearer Zusammenhang
- Voraussetzung:
  - Linearer Zusammenhang
  - Normalverteilung (?)

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$



Quelle: https://www.scribbr.com/statistics/correlation-coefficient/

BEISPIEL
DOKUMENTE MIT TEXT

Sentiment	Text
10	Ein richtig tolles Spiel.
1	Ganz schlechte Leistung.
9	Sehr leckeres Gericht. Zu empfehlen.
2	Ganz schlechtes Spiel. Nicht zu empfehlen.

Generell sinnvoll, Daten zu putzen:

- Entfernen von Stopworten, Zeichensetzung, Zahlen, etc.
- Möglich: Filterung nach Minimal-/Maximalfrequenz einzelner Wörter

+ Tokenisieren

BEISPIEL
DOKUMENTE MIT TEXT

Sentiment	Text
10	Ein **richtig** tolles Spiel.
1	**Ganz** schlechte Leistung.
9	**Sehr** leckeres Gericht. Zu empfehlen.
2	**Ganz** schlechtes Spiel. **Nicht** zu empfehlen.

10	tolles Spiel
1	schlechte Leistung
9	leckeres Gericht empfehlen
2	schlechtes Spiel empfehlen

BEISPIEL
DOKUMENT-TERM-MATRIX (TF VS. TF-IDF)

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment
0	0	0	0	0	0	1	1	10
0	0	0	1	1	0	0	0	1
1	1	1	0	0	0	0	0	9
1	0	0	0	0	1	1	0	2

0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.79	0.00
0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00	0.00
0.49	0.62	0.62	0.00	0.00	0.00	0.00	0.00	0.49
0.53	0.00	0.00	0.00	0.00	0.67	0.53	0.00	0.53

** Nicht die Standard TF-IDF Formel aus der Vorlesung (!), sondern die Berechnung durch scikit-learn. Hier zusätzliches +1 gegen Division durch 0.*
Plus zeilenweise Normalisierung (L2) und gerundete Werte!

BEISPIEL
DOKUMENT-TERM-MATRIX + BEWERTUNG → KORRELATIONEN

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment	Sentiment
0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.79	0.00	**10**
0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00	0.00	**1**
0.49	0.62	0.62	0.00	0.00	0.00	0.00	0.00	0.49	**9**
0.53	0.00	0.00	0.00	0.00	0.67	0.53	0.00	0.53	**2**

Nutzung des Pearson Korrelationskoeffizienten

- Korrelation jeder Worttyp-Spalte mit den Sentiment
  → Einfluss von Wort auf Sentiment-Wert
- Andere Korrelationsmaße möglich → ähnliche Ergebnisse

BEISPIEL KORRELATIONEN

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment	Sentiment
0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.79	0.00	10
0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00	0.00	1
0.49	0.62	0.62	0.00	0.00	0.00	0.00	0.00	0.49	9
0.53	0.00	0.00	0.00	0.00	0.67	0.53	0.00	0.53	2

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment
-0.033768								

BEISPIEL KORRELATIONEN

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment	Sentiment
0.00	**0.00**	0.00	0.00	0.00	0.00	0.62	0.79	0.00	**10**
0.00	**0.00**	0.00	0.71	0.71	0.00	0.00	0.00	0.00	**1**
0.49	**0.62**	0.62	0.00	0.00	0.00	0.00	0.00	0.49	**9**
0.53	**0.00**	0.00	0.00	0.00	0.67	0.53	0.00	0.53	**2**

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment
-0.033768	**0.50128**							

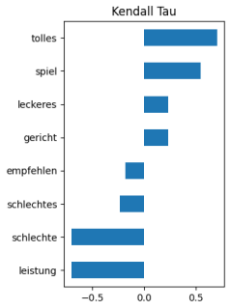
BEISPIEL KORRELATIONEN

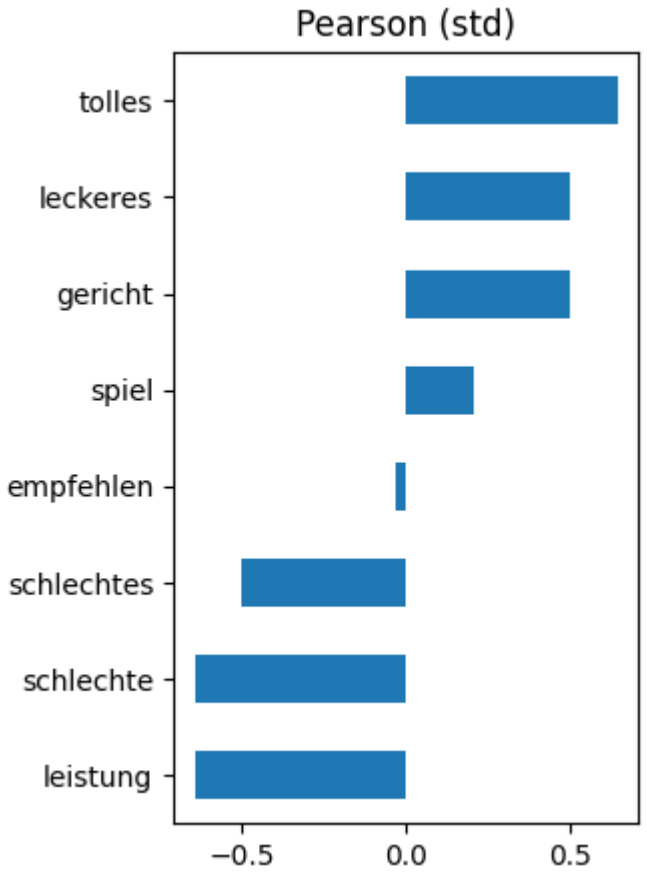
empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment	Sentiment
0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.79	0.00	10
0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00	0.00	1
0.49	0.62	0.62	0.00	0.00	0.00	0.00	0.00	0.49	9
0.53	0.00	0.00	0.00	0.00	0.67	0.53	0.00	0.53	2

empfehlen	gericht	leckeres	leistung	schlechte	schlechtes	spiel	tolles	sentiment
-0.033768	0.50128	0.50128	-0.644503	-0.644503	-0.50128	0.203028	0.644503	-0.033768

BEISPIEL
FINALES SENTIMENT LEXIKON

empfehlen	-0,03
gericht	0,50
leckeres	0,50
leistung	-0,64
schlechte	-0,64
schlechtes	-0,50
spiel	0,20
tolles	0,64



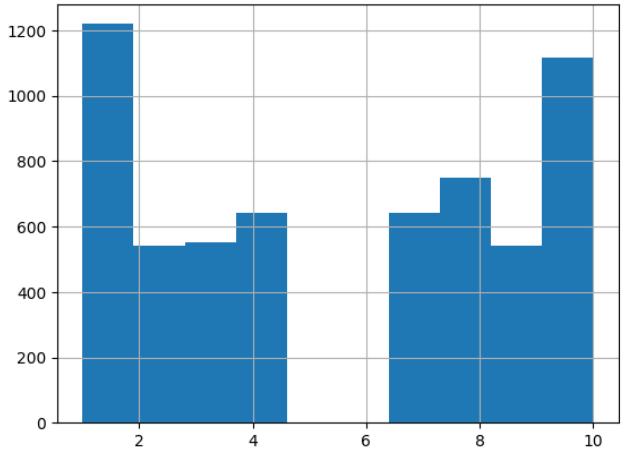


- TODO: Skalierung finaler Werte + Filterung?

SENTIMENT LEXIKON – „ECHTE DATEN“

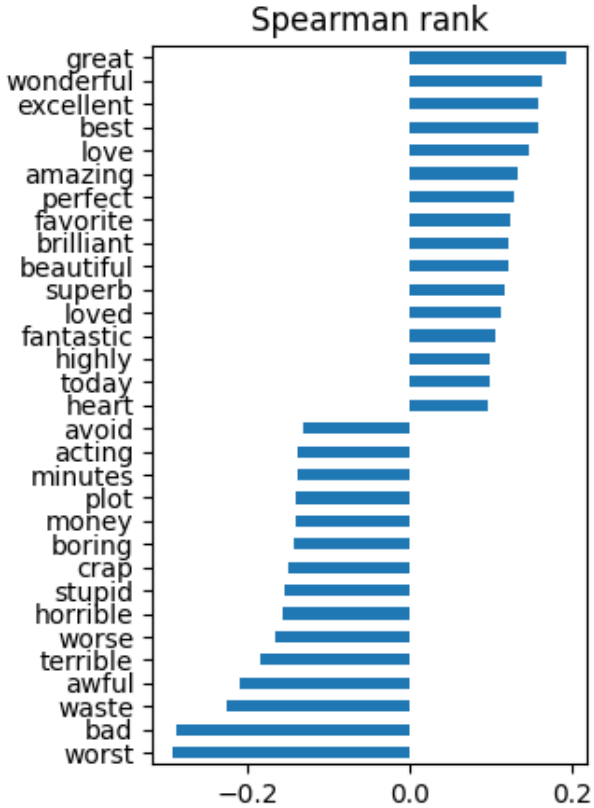
- Daten: IMDb – Large Movie Review Dataset,
  25.000 Texte mit Bewertung von 1 – 4 *(negativ)* , 7 – 10 *(positiv)*
  @ https://ai.stanford.edu/~amaas/data/sentiment/

- Schritte:
  1. Bereinigung mit spaCy
  2. TF-IDF Berechnung mit max. 5.000 Features
  3. Begrenzung auf 10.000 Dokumente
  4. Berechnung von Pearson Korrelation (mit pandas)
  5. Auswahl von 30 Top-Features nach Stärke
     → ½ positiv und ½ negativ



Histogramm von Bewertungen für Ausschnitt von 10k Dokumenten (Schritt 3)

ERGEBNIS: KORRELATIONEN



BEISPIEL ROHDATEN – POSITIV (9/10), „2_9.TXT“

Rohdaten: Bromwell High is nothing short of brilliant. Expertly scripted and perfectly delivered, this searing parody of a students and teachers at a South London Public School leaves you literally rolling with laughter. It's vulgar, provocative, witty and sharp. The characters are a superbly caricatured cross section of British society (or to be more accurate, of any society). Following the escapades of Keisha, Latrina and Natella, our three "protagonists" for want of a better term, the show doesn't shy away from parodying every imaginable subject. Political correctness flies out the window in every episode. If you enjoy shows that aren't afraid to poke fun of every taboo subject imaginable, then Bromwell High will not disappoint!

Spacy: Bromwell High short brilliant Expertly scripted perfectly delivered searing parody students teachers South London Public School leaves literally rolling laughter vulgar provocative witty sharp characters superbly caricatured cross section British society accurate society Following escapades Keisha Latrina Natella protagonists want better term shy away parodying imaginable subject Political correctness flies window episode enjoy shows afraid poke fun taboo subject imaginable Bromwell High disappoint

TFIDF: accurate: 0.148, afraid: 0.143, away: 0.089, better: 0.072, brilliant: 0.112, british: 0.123, characters: 0.069, cross: 0.144, delivered: 0.151, disappoint: 0.175, enjoy: 0.100, episode: 0.113, flies: 0.178, following: 0.130, fun: 0.093, high: 0.193, laughter: 0.153, leaves: 0.124, literally: 0.134, london: 0.143, parody: 0.155, perfectly: 0.127, political: 0.133, protagonists: 0.166, provocative: 0.182, public: 0.131, rolling: 0.159, school: 0.107, scripted: 0.170, section: 0.157, sharp: 0.156, short: 0.101, shows: 0.095, shy: 0.170, society: 0.257, south: 0.141, students: 0.147, subject: 0.251, superbly: 0.169, teachers: 0.185, term: 0.161, vulgar: 0.182, want: 0.082, window: 0.151, witty: 0.149

Sentiment: brilliant: 0.113, perfectly: 0.078, shows: 0.070, fun: 0.064, enjoy: 0.063, superbly: 0.050, witty: 0.046, society: 0.042, sharp: 0.034, episode: 0.034, south: 0.029, provocative: 0.022, british: 0.021, vulgar: -0.020, flies: -0.022, students: -0.025, away: -0.031, school: -0.036, want: -0.054, better: -0.085 → *Summe: +0.421 (+0.393, Top-20)*

BEISPIEL ROHDATEN – NEGATIV (1/10), „1_1.TXT“

Rohdaten: Robert DeNiro plays the most unbelievably intelligent illiterate of all time. This movie is so wasteful of talent, it is truly disgusting. The script is unbelievable. The dialog is unbelievable. Jane Fonda's character is a caricature of herself, and not a funny one. The movie moves at a snail's pace, is photographed in an ill-advised manner, and is insufferably preachy. It also plugs in every cliche in the book. Swoozie Kurtz is excellent in a supporting role, but so what?

Equally annoying is this new IMDB rule of requiring ten lines for every review. When a movie is this worthless, it doesn't require ten lines of text to let other readers know that it is a waste of time and tape. Avoid this movie.

Spacy: Robert DeNiro plays unbelievably intelligent illiterate time movie wasteful talent truly disgusting script unbelievable dialog unbelievable Jane Fonda character caricature funny movie moves snail pace photographed ill advised manner insufferably preachy plugs cliche book Swoozie Kurtz excellent supporting role annoying new IMDB rule requiring lines review movie worthless require lines text let readers know waste time tape Avoid movie

TFIDF: annoying: 0.129, avoid: 0.135, book: 0.115, character: 0.078, cliche: 0.196, deniro: 0.214, dialog: 0.138, disgusting: 0.173, excellent: 0.108, fonda: 0.205, funny: 0.091, ill: 0.164, imdb: 0.141, intelligent: 0.146, jane: 0.164, know: 0.079, let: 0.105, lines: 0.234, manner: 0.154, moves: 0.146, movie: 0.177, new: 0.091, pace: 0.145, photographed: 0.188, plays: 0.106, require: 0.202, review: 0.134, robert: 0.133, role: 0.099, rule: 0.178, script: 0.098, supporting: 0.131, talent: 0.130, tape: 0.172, text: 0.186, time: 0.122, truly: 0.113, unbelievable: 0.305, unbelievably: 0.190, waste: 0.117, worthless: 0.189

Sentiments: excellent: 0.156, role: 0.060, supporting: 0.053, plays: 0.048, new: 0.036, know: -0.037, imdb: -0.043, lines: -0.044, worthless: -0.045, unbelievably: -0.045, dialog: -0.054, review: -0.059, disgusting: -0.061, unbelievable: -0.064, let: -0.072, annoying: -0.100, movie: -0.120, script: -0.126, avoid: -0.133, waste: -0.210 → *Summe: -0.776 (-0.861, Top-20)*