



UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Universität Paderborn
Fakultät für Wirtschaftswissenschaften
Department Wirtschaftsinformatik
Lehrstuhl für Wirtschaftsinformatik, insb. Data Analytics

Seminararbeit

Kulinarische Trends in Yelp Reviews

von
Alexander Worona
7075164
Detmolder Straße 10, 32825 Blomberg
worona@mail.uni-paderborn.de

von
Kevin Zalipski
7073988
Wittenauer Straße 6, 32825 Blomberg
zalipski@mail.uni-paderborn.de

vorgelegt bei
Prof. Dr. Oliver Müller

Eingereicht am 21. Februar 2020

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken, Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

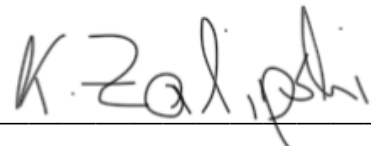
Blomberg, 21. Februar 2020



(Alexander Worona)

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken, Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Blomberg, 21. Februar 2020



(Kevin Zalipski)

Inhaltsverzeichnis

1. Business Understanding	1
2. Data Understanding	1
3. Data Preparation	3
3.1. „Food“-Reviews kategorisieren	3
3.2. Reviewtexte normalisieren	4
3.3. N-Gramm-Analyse	7
3.4. Nicht-englischsprachige Reviews ausfiltern	8
4. Modeling	9
4.1. Bag-of-Words-Modell	9
4.2. LDA-Modell	10
5. Evaluation	13
5.1. Topics	14
5.2. Trends	15
6. Deployment	19
Literaturverzeichnis	21
A. Digitale Abgabe	23

Abbildungsverzeichnis

2.1.	Anzahl der Sterne im Bezug zur Länge	2
2.2.	Verteilung der Sterne	2
2.3.	Reviewanzahl pro Monat	3
3.1.	Datensatzanteil unserer „Food“-Kategorisierung	4
3.2.	Entwicklung der Review-Textlänge nach Verarbeitungsschritten . .	5
3.3.	Top 15 verwendete Wörter inkl. Stopwords	5
3.4.	Top 15 verwendete Wörter ohne Stopwords	6
3.5.	Logarithmierte Reviewanzahl nach Sprache (Top 10)	9
4.1.	LDA-Visualisierung mit LDAvis	12
5.1.	Verlauf und Score des Eis-Topics	15
5.2.	Verlauf und Score des Steak-Topics	16
5.3.	Verlauf und Score mehrerer asiatischer Topics	16
5.4.	Verlauf und Score des Bar-Topics	17
5.5.	Verlauf und Score des mexikanischen Topics	18
5.6.	Verlauf und Score für verschieden-ethnische Küchen	18
5.7.	Verlauf und Score für Speisekarten-Optionen	19
6.1.	Interaktive Code-Zellen in Google Colab	20

Tabellenverzeichnis

3.1. Auszug irrelevanter <i>Food</i> -Unterkategorien	4
3.2. Lemmatisierung mit POS-Tags	7
3.3. Auszug erzeugter Bi- und Trigramme	7
3.4. Vergleich eines Reviews vor und nach der Verarbeitung	8
4.1. Beispiel-Bag-of-Words	10
4.2. Beispiel-Topic über mexikanische Küche	10
4.3. Coherence-Werte für verschieden-große LDA-Modelle	11
4.4. Topic-Score-Zuordnung für ein Beispiel-Review	13
5.1. Auszug der erzeugten Topics	14

1. Business Understanding

Forschungsthema dieser Studienarbeit ist es, einen bereitgestellten Datensatz nach trendenden Gerichten und Restaurants zu untersuchen. Wir möchten überprüfen, ob zu bestimmten Zeitpunkten Restaurants oder Gerichte häufiger erwähnt und bewertet wurden, ob also ein Trend zu erkennen ist. Wir betrachten sowohl einen Langzeittrend als auch saisonale Auswirkungen auf die Beliebtheit eines Gerichts, einer Küche oder ähnlicher Aspekte.

Anhand der Reviews, ihrer Anzahl und ihrer Sterne-Bewertung soll eine Analyse des zeitlichen Verlaufs ermöglicht werden.

Dabei ist es besonders für nicht-traditionelle Restaurants, und Restaurants, die noch eröffnet werden sollen, wichtig, Trends zu erkennen, um ihr Geschäft anzupassen oder in ein Geschäft aus- oder einzusteigen.

Vor allem in den USA ist ethnische Küche sehr beliebt (Buzalka, 2001), insbesondere die asiatische, mexikanische und italienische Küche (Roseman, 2006). Wir möchten daher den Verlauf dieser Küchen betrachten. Weiterhin kann es von Interesse dieser Restaurants sein, wie sich die Beliebtheit verschiedener Gerichte einer bestimmten Küche über einen Zeitraum verhält.

Interessant ist außerdem, wie sich das Verhalten von Restaurantbesuchern auf gesündere Optionen und Vielfalt der Speisekarte auswirkt. Nach DiPietro, Roseman und Ashley (2006) ist das Bewusstsein der Konsumenten gegenüber gesünderem Essen gestiegen; Schnellrestaurants passen ihre Speisekarten an und bieten gesündere Alternativen an, welche einen signifikanten Anstieg der Verkaufszahlen erlebten, und weniger gesunde Optionen einen Abstieg (DiPietro et al., 2006). Wir möchten den Verlauf dieses Trends an dem Yelp-Datensatz beobachten.

Mit Topic-Modellen, wie LDA (Blei, Ng & Jordan, 2003), sollen die verschiedenen Eigenschaften eines Restaurant-Reviews, wie Küche und Gerichte, erkannt werden. Die Reviews sollen mit diesem Modell analysiert und gelabelt bzw. gewichtet werden, sodass wir einen zeitlichen Verlauf der Reviews respektive der erkannten Topics analysieren können.

Wir evaluieren unser Topic-Modell, indem wir die Interpretierbarkeit, Nützlichkeit und Kohärenz der Topics nach Boyd-Graber, Mimno und Newman (2014) untersuchen. Weiterhin beurteilen wir unsere Zeitreihendaten, stellen an Ihnen die Machbarkeit unserer Ziele fest und betrachten, ob wir glaubwürdige und interpretierbare Zeitreihengraphen erhalten haben.

2. Data Understanding

Der zur Verfügung gestellte Datensatz stammt von der Website Yelp und beinhaltet Reviews von den Nutzern zu verschiedenen Unternehmen. Er wurde als SQL-Datenbank zu csv-Dateien exportiert, welche insgesamt eine Größe von 5,70 GB aufweisen.

Der Datensatz besteht aus mehreren Tabellen, wobei wir bei dieser Studienarbeit den Fokus auf die Tabellen *business* (Größe: 20 MB) und *review* (Größe: 3,25 GB) legen. Jede Zeile der beiden Datensätze hat eine eigene ID zur Zuordnung. Die *business*-Tabelle beinhaltet Spalten für die ID und den Namen des Businesses, die Adresse und Position, außerdem die durchschnittliche Sternebewertung, ob es aktuell geöffnet ist und die Anzahl der Bewertungen. Die Anzahl an Einträgen (Zeilen) dieses Datensatzes ist 156 639.

In einem *review* sind sowohl Business-ID, als auch User-ID und gegebene Sterne, sowie Datum und Text des Reviews angegeben. Der Text des Reviews wird über ein Freitext-Feld eingegeben, sodass die Länge der Reviews stark variiert. Man kann zusätzlich sehen, wie oft ein Review von anderen Nutzern als *useful*, *funny* oder *cool* bewertet wurde. Hier sind 4 736 897 Einträge vorhanden.

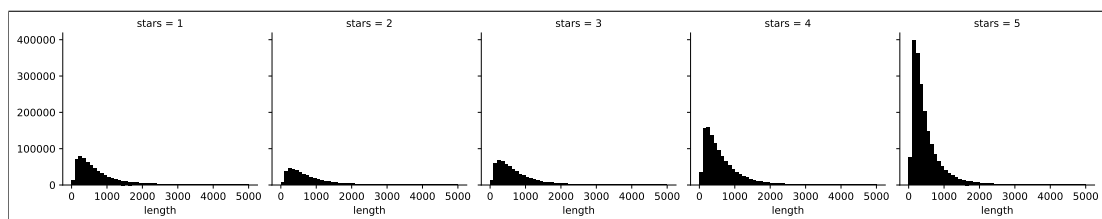


Abbildung 2.1.: Anzahl der Sterne im Bezug zur Länge

Abbildung 2.1 zeigt, dass die Reviews mit einem Stern in der Regel am kürzesten sind. Im Großen und Ganzen haben die meisten Reviews eine maximale Länge von 500 Zeichen, und sehr wenige sind 2 000–3 000 Zeichen lang.

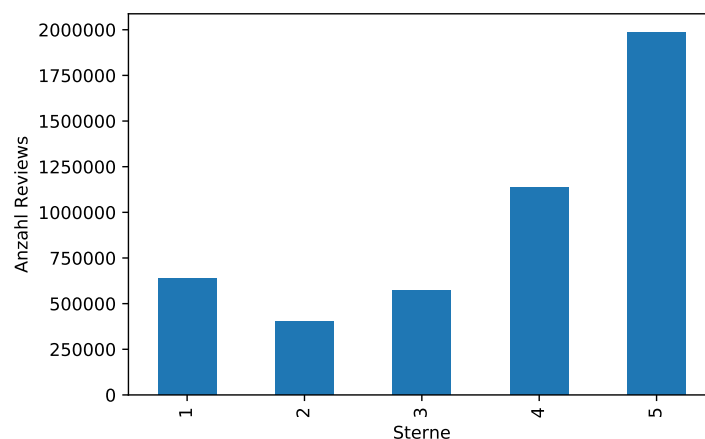


Abbildung 2.2.: Verteilung der Sterne

In Abbildung 2.2 ist zu sehen, wie oft die jeweiligen Sterne an Reviews vergeben wurden. Am häufigsten wurden dabei 5 Sterne vergeben, fast doppelt so viele wie die am zweithäufigst vergebenen 4 Sterne. Am wenigsten wurden 2 Sterne vergeben, was dazu führt, dass die Verteilung der typischen Hockeyschläger-Kurve gleicht.

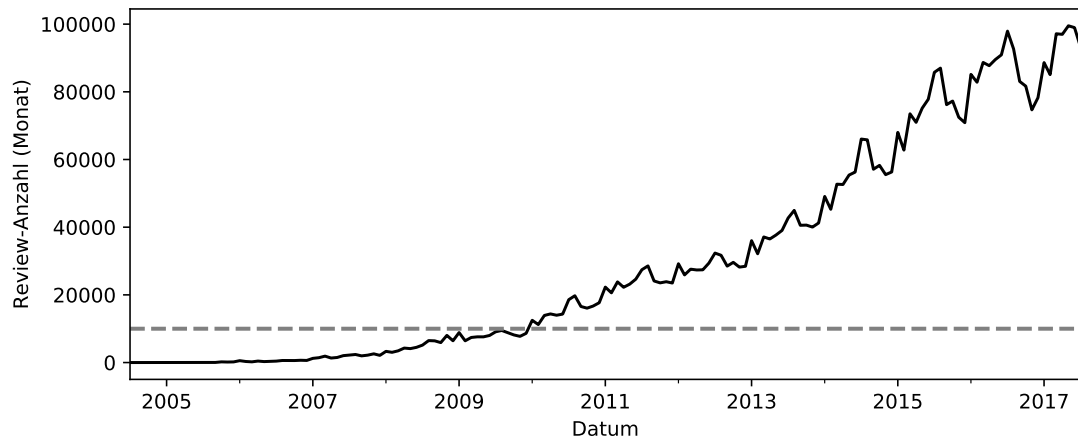


Abbildung 2.3.: Reviewanzahl pro Monat

In Abbildung 2.3 ist die Entwicklung der Anzahl an Food-Reviews pro Monat über die vorhandenen Jahre unseres Datensatzes zu sehen. Man kann erkennen, dass es im Zeitraum bis 2007 sehr wenige bis gar keine Reviews gab und die Anzahl danach erst anstieg.

Die im folgenden Teil dieser Arbeit gezeigten Zeitreihengraphen beziehen sich daher hauptsächlich auf Reviews ab dem Jahr 2010.

3. Data Preparation

3.1. „Food“-Reviews kategorisieren

Da wir uns in dieser Studienarbeit primär auf Restaurants, Imbisse etc. fokussieren, müssen wir vorab alle Reviews ausfiltern, welche auf andere Businesses bezogen sind. Da die jeweiligen Reviews nicht direkt mit einer Kategorie gelabelt sind, nutzen wir zur Filterung die Tabelle *category* und Yelps bereitgestellten Kategorien-Baum¹. Wir kategorisieren Datensatzeinträge als „Food“, wenn das bewertete Business der Oberkategorie *Restaurants* oder *Food* zugehörig ist.

¹https://www.yelp.com/developers/documentation/v3/all_category_list

Die *Food*-Kategorie beinhaltet jedoch auch einige, für uns uninteressante, Unterkategorien (Tabelle 3.1), welche wir zusätzlich ausfiltern.

Nach Filterung erhalten wir einen Teilsatz an „Food“-Businesses, welche 36% der Gesamtzahl an Businesses ausmachen.

Butcher
Convenience Stores
Farmers Market
Internet Cafes
Water Stores

Tabelle 3.1.: Auszug irrelevanter *Food*-Unterkategorien

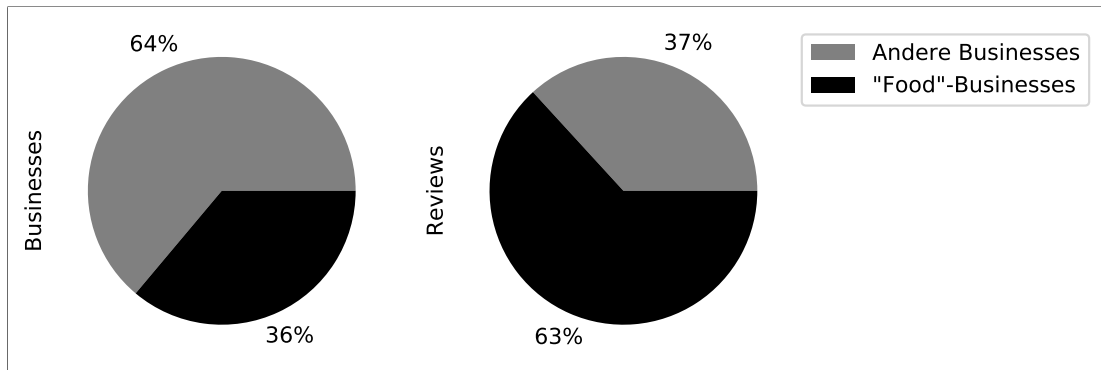


Abbildung 3.1.: Datensatzanteil unserer „Food“-Kategorisierung: Links der Anteil an „Food“-Businesses; rechts der Anteil an „Food“-Reviews.

Der Anteil an Reviews von Restaurants und ähnlichen Business-Typen beträgt 63%, die restlichen Reviews machen 37% aus. Dabei werden beispielsweise Einkaufs-, Beauty- oder Mode-Unternehmen bewertet. Auffällig ist auch, dass sich ca. zwei Drittel aller Reviews auf ca. ein Drittel aller Businesses beziehen.

Insgesamt erhalten wir ca. 3 Millionen Reviews für 56 000 Businesses. Für unsere weitere Verarbeitung und Modelle verwenden wir eine Teilmenge von 1 800 000 Reviews ab dem Jahr 2010.

3.2. Reviewtexte normalisieren

Wir normalisieren die Texte, um die benutzten Wörter der Reviews in einer einheitlichen Form zu haben, da in vielen Fällen verschiedene Versionen eines Wortes vorhanden sind (Manning, Raghavan & Schütze, 2008, p. 28).

Zur Normalisierung der Reviewtexte nutzen wir das Natural-Language-Processing-Tool SpaCy (Honnibal & Montani, 2017). SpaCy bietet zahlreiche Modelle verschiedener Sprachen zur Textverarbeitung an. Wir verwenden das 789-MB-große Modell *en_core_web_lg*², welches mit Online-Texten aus Blogs, Nachrichten und Kommentaren aus dem Datensatz *OntoNotes 5.0* (Weischedel et al., 2013) trainiert wurde.

²https://spacy.io/models/en#en_core_web_lg

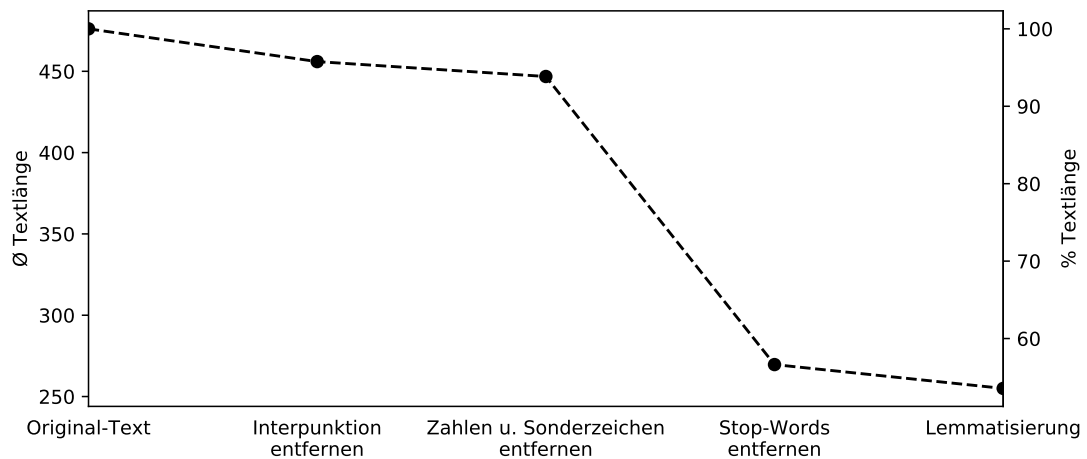


Abbildung 3.2.: Entwicklung der Review-Textlänge nach Verarbeitungsschritten

In Abbildung 3.2 sieht man die Entwicklung der durchschnittlichen Textlänge nach dem jeweiligen Verarbeitungsschritt. Am Ende beträgt die durchschnittliche Reduktion der Review-Textlänge 46,50%.

Um die Review-Texte zu normalisieren, haben wir Sonderzeichen und Zahlen entfernt. Sonderzeichen und Zahlen sind dann bspw. in Kombination 100\$ oder 50€, die in diesem Beispiel dazu benutzt werden, um Preise anzugeben. Weitere Anwendungsbeispiele von Zahlen sind Angaben von Wartezeiten oder um den Wert einer Maßeinheit zu nennen (bspw. Flüssigkeitsmenge eines Trinkbechers). Diese sind aber nicht relevant für uns, sodass wir sie entfernen können.

Wir haben auch Stopwords wie bspw. *a*, *and*, *but* und *or* mithilfe der *Stoplist* des SpaCy-Modells entfernt. Dort sind die üblichen Wörter vordefiniert, sodass man damit einfach die Reviews auf vorhandene Stopwords überprüfen und anschließend entfernen kann.

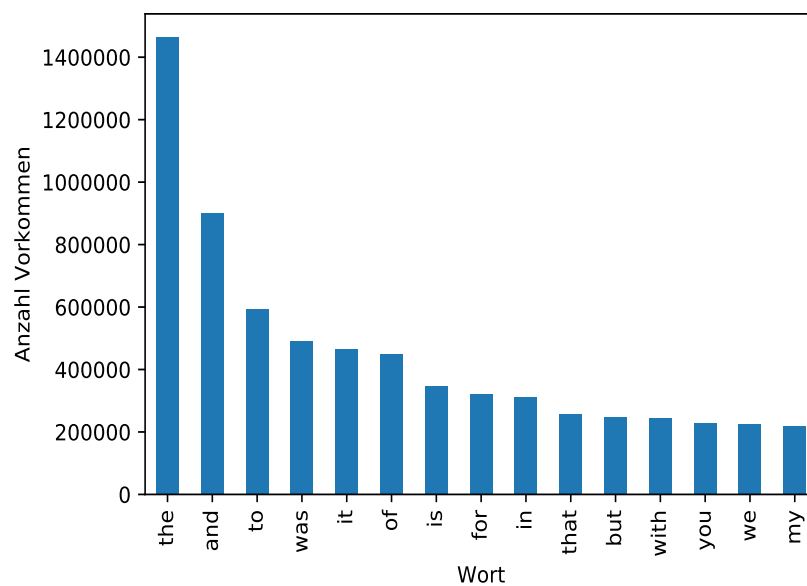


Abbildung 3.3.: Top 15 verwendete Wörter inkl. Stopwords

Abbildung 3.3 und Abbildung 3.4 zeigen die 15 am häufigsten verwendeten Wörter der ersten 200 000 Reviews³. In Abbildung 3.3 sind auch Stopwords enthalten, sodass erwartungsgemäß auch nur diese in den Top 15 enthalten sind.

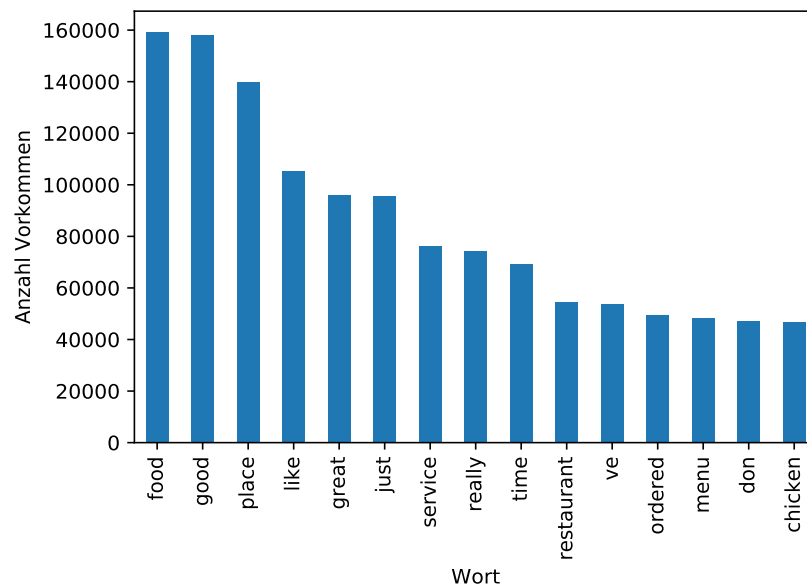


Abbildung 3.4.: Top 15 verwendete Wörter ohne Stopwords

Stopwords machen einen großen Teil der Texte aus und verringern damit zudem die Qualität von Topics in Topic Models. Für die Entfernung ist es am besten, keine korpuspezifische Stopwords Liste zu benutzen, sondern die am häufigsten verwendeten Wörter mit einer generischen Liste zu entfernen. (Schofield, Magnusson & Mimno, 2017)

Zur einfacheren Verarbeitung haben wir dann die Reviews auf Kleinschreibung geändert. Dass man mehrere Versionen mit unterschiedlicher Groß- und Kleinschreibung eines Wortes hat, wird so verhindert. Hinzu kommt, dass die Dictionaries im Normalfall nur die kleingeschriebene Version des Wortes beinhalten, weswegen die Veränderung auf Kleinschreibung ein sehr wichtiger Schritt ist.

Lemmatisierung wird verwendet, um Wörter von Ihrer grammatischen Form zu befreien und Sie in ihre „Wörterbuch“-Form – dem Lemma – zu bringen (Manning et al., 2008, p. 32).

Stemming schneidet das Ende eines Wortes hingegen rigoros ab, ohne eventuelle semantische Unterschiede, wie Kategorisierung von Verb oder Substantiv, unterscheiden zu können (Manning et al., 2008, p. 32).

³Aufgrund von Geschwindigkeitsvorteilen wurde hierfür nicht die komplette Anzahl an Reviews benutzt, 200 000 sollten aber repräsentativ genug sein.

„[I'm] at a meeting“		„[I] will be meeting“	
POS	Lemma	POS	Lemma
ADP	at	VERB	will
DET	a	AUX	be
NOUN	meeting	VERB	meet

Tabelle 3.2.: Lemmatisierung mit POS-Tags:

Das Wort *meeting* kann mit der Zuordnung der Wortart korrekt lemmatisiert werden:

Das Lemma des Substantivs lautet *meeting* und des Verbs *meet*.

Deshalb verwendet das SpaCy-Modell für die Lemmatisierung ihren am Textdatensatz trainierten Part-of-Speech-Tagger (kurz: POS-Tagger). Dabei wird jedes Wort eines Dokumentes mit der jeweiligen Wortart, sprich Verb, Substantiv, Artikel etc., gelabelt (Manning & Schütze, 1999, p. 341). Tabelle 3.2 demonstriert die Lemmatisierung mit POS-Tags.

3.3. N-Gramm-Analyse

Um häufige Phrasen in Texten zu erkennen, werden Wort-N-Gramme gebildet, ihre Häufigkeit gemessen und die häufigsten gruppiert. Ein Wort-N-Gramm bezeichnet hier eine Aufteilung eines Textes in Gruppen aus N Wörtern. Für $N = 2$ erhält man für „either way worth [every]⁴ penny [...]“ eine Folge aus Wort-Bigrammen $\{(either, way), (way, worth), (worth, penny)\}$. Da die Phrasen „either way“ und „worth [every] penny“ häufig in dem Textdatensatz vorkommen, werden sie zu „either_way“ und „worth_penny“ zusammengefügt.

Das Bewerten des benachbarten Auftretens – der Kollokation – wird über eine *Scoring*-Funktion bestimmt. Das von uns verwendete *Phrase-Model*⁵ von Gensim (Řehůřek & Sojka, 2010) implementiert ihre *Scoring*-Funktion nach Mikolov, Sutskever, Chen, Corrado und Dean (2013):

$$\text{score}(w_i, w_j) = \frac{(\text{count}(w_i w_j) - \delta) |\text{vocab}|}{\text{count}(w_i) \cdot \text{count}(w_j)}$$

δ bezeichnet die minimal geforderte Anzahl an Vorkommnissen für ein Wort w oder Bigramm $w_i w_j$, $|\text{vocab}|$ die Größe des Gesamtvokabulars der Texte. Weiterhin muss der *Score* einen bestimmten Schwellenwert überschreiten, damit ein Bigramm zusammengefügt wird.

mac_n_cheese
onion_ring
french_toast
egg_roll
ice_cream

Tabelle 3.3.: Auszug erzeugter Bi- und Trigramme

⁴„every“ wurde als Teil der Stoplist bereits vorher entfernt.

⁵<https://radimrehurek.com/gensim/models/phrases.html>

Für unsere Analyse gelten die (standardmäßig vorgegebenen) Parameter $\delta = 5$ und $\text{score}_{\min} = 10$ für den Schwellenwert.

Wir bilden für unseren Datensatz sowohl Bi- als auch Trigramme.

Tabelle 3.4 zeigt die Auswirkung der kompletten Textnormalisierung an einem Beispiel.

Beispiel-Review	
... vor Verarbeitung	... nach Verarbeitung
This is probably the 6 th time me and my boyfriend coming here.	probably th time boyfriend come
There hamburgers are awesome and so are there wings.	hamburger awesome wing
We found this place by going on yelp.	find place go yelp
Boy am I glad that we choose this place and the waitress Jess is GREAT!!!	boy_glad choose place waitress jess great
They have the fishbowl drink specials on Fridays \$10.	fishbowl drink special fridays
And they are normally \$15.	normally
Either way worth every penny not watered down.	way worth_penny water
They are 84 oz.	oz
always a great time	great time

Tabelle 3.4.: Vergleich eines Reviews vor und nach der Verarbeitung

3.4. Nicht-englischsprachige Reviews ausfiltern

Da im Datensatz auch Reviews vorhanden sind, die nicht auf Englisch geschrieben wurden, klassifizieren wir die Texte nach der Sprache. Danach fügen wir dem Datensatz die zusätzliche Sprachen-Spalte *language* hinzu, um die Reviews nun nach der Sprache filtern zu können.

Der Vorgang lässt sich leicht in den Textnormalisierungs-Prozess integrieren, da die Textverarbeitung mit SpaCy über Plug-Ins erweiterbar ist. Wir verwenden das Plug-In *spacy-langdetect*⁶, basierend auf einem Python-Port⁷ der Spracherkennungsbibliothek von Nakatani (2010), welche eine Genauigkeit von über 99% für 53 Sprachen bietet.

Der Datensatz besteht zu ca. 98,73% aus englischsprachigen Texten.

⁶<https://spacy.io/universe/project/spacy-langdetect>

⁷<https://github.com/Mimino666/langdetect>

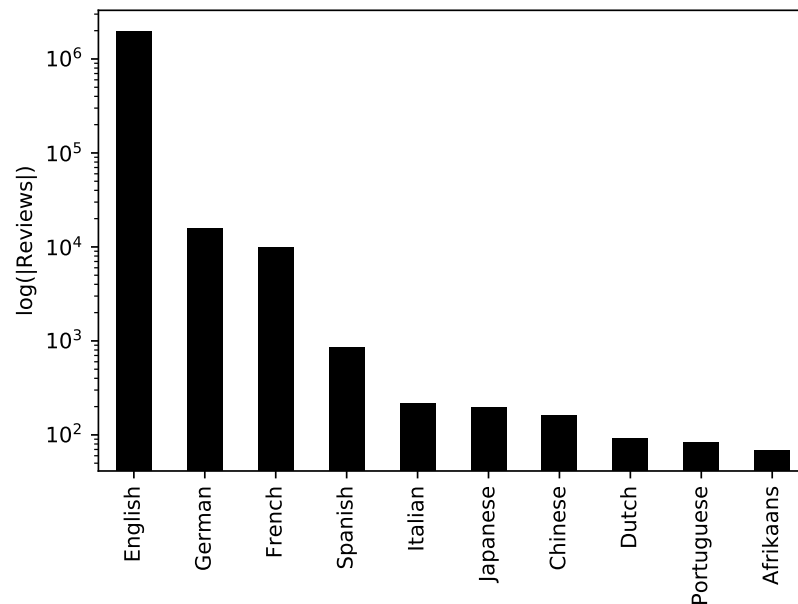


Abbildung 3.5.: Logarithmierte Reviewanzahl nach Sprache (Top 10)

Die (logarithmierte) Reviewanzahl nach Sprache ist in Abbildung 3.5 abgebildet. Englisch ist bei weitem die am häufigsten verwendete Sprache in den Reviews, weshalb wir uns in dieser Studienarbeit auf die englischen Reviews beschränken.

4. Modeling

4.1. Bag-of-Words-Modell

Zusätzlich zu der Entfernung von Stopwords betrachten wir nun auch keine Wörter mehr, die extrem häufig oder selten vorkommen, indem wir ein *Dictionary*¹ bilden, welches die Wortvorkommnisse des kompletten Datensatzes zählt.

Wir entfernen anschließend alle Wörter, die nicht mindestens 60-mal vorkommen oder über 40% des Dictionarys ausmachen.

¹<https://radimrehurek.com/gensim/corpora/dictionary.html>

Text	Wort ^a	Anzahl
„excellent bubble_tea good service good food try bubble_tea place area rose_tea good authentic_taiwanese_cuisine“	service	1
	try	1
	excellent	1
	area	1
	bubble_tea	2
	rose_tea	1

Tabelle 4.1.: Bag-of-Words eines Reviews:

Die Wortreihenfolge des Reviewtextes wird nicht mehr beibehalten; die Wörter „good“ und „place“ wurden als zu häufig aus dem Dictionary gefiltert; „authentic taiwanese cuisine“ kam weniger als 60-mal im Datensatz vor.

^aIn dem eigentlichen Bag-of-Words-Modell wird ein Wort-Index oder eine -ID anstelle des tatsächlichen Wortes verwendet.

Zusätzlich ist für unser weiteres Vorgehen die Struktur eines Reviews beziehungsweise die Reihenfolge der Wörter des Reviews unwichtig. Wir können mit dem Dictionary ein *Bag-of-Words*-Modell für die Reviewtexte erzeugen. Dabei zählt das Bag-of-Words-Modell die Anzahl der Wörter, die in dem jeweiligen Review vorkommen (Manning et al., 2008, p. 117) (Tabelle 4.1).

4.2. LDA-Modell

Wir verwenden, um Reviews unter Zuhilfenahme unseres Bag-of-Words-Modells nach ihrem Inhalt gruppieren zu können, das Topic-Modell *Latent Dirichlet Allocation* (LDA) nach Blei et al. (2003). Das Modell geht davon aus, dass ein Dokument aus einer Verteilung verschiedener Topics und ein Topic wiederum aus einer Verteilung verschiedener Begriffe besteht (Boyd-Graber et al., 2014). Das Wort „taco“ in dem Beispiel der Tabelle 4.2 hat den höchsten Wert der Wahrscheinlichkeitsverteilung des mexikanisch orientierten Topics.

Begriff	Wahrsch.
taco	0,102
mexican	0,046
burrito	0,042
salsa	0,031
chip	0,024
margarita	0,023
cheese	0,017
nacho	0,017
order	0,015
guacamole	0,015

Tabelle 4.2.: Beispiel-Topic über mexikanische Küche: Die 10 häufigsten Begriffe der Wahrscheinlichkeitsverteilung für das Topic.

Für das Topic Modeling mit LDA verwenden wir die Implementierung² der Gensim-Bibliothek.

Bei dem LDA-Modell wird die Zahl der Topics von dem „Modellierer“ bestimmt (Boyd-Graber et al., 2014). Und obgleich LDA Topic Models beliebt sind, um eine große Anzahl an Dokumenten strukturiert zu analysieren, existieren wenige Empfehlungen für die Bestimmung der Modell-Parameter (Wallach, Mimno & McCallum, 2009). Nach Wallach et al. ist die Wahl der Topic-Anzahl die problematischste Entscheidung eines Topic-Modells.

Weiterhin ist die Validierung von Topic Models und ihrer Kohärenz (*Coherence*) daher vor allem durch manuelle Inspektion von Menschen geprägt, wie Chang, Gerrish, Wang, Boyd-Graber und Blei (2009) demonstrieren.

Röder, Both und Hinneburg (2015) entwickelten in ihrer Arbeit Coherence-Maße, die streng mit den Bewertungen von Menschen korrelieren. Die Maße sind in Gensim implementiert³.

Wir untersuchen unser LDA-Modell manuell, verwenden aber ebenso das Coherence-Maß für die Optimierung unseres Modells und das Fine-Tuning der Anzahl an Topics (Tabelle 4.3).

Für die weiteren benötigten Parameter des LDA-Modells orientieren wir uns an den Trainings-Empfehlungen für LDA-Modelle von Mortensen und Konstantinovskiy. Im Speziellen wird der komplette Datensatz während des Trainings 40-mal durchlaufen.

Topics	C_v^a
35	0,4902
40	0,4955
45	0,4673
50	0,4482
55	0,4397

Tabelle 4.3.: Coherence-Werte für Modelle unseres Datensatzes mit unterschiedlicher Topic-Anzahl. Für die Wahl von 40 Topics erhalten wir den besten Wert.

^a C_v ist nach Röder et al. das am stärksten mit menschlichen Bewertungen korrelierende Coherence-Maß.

²<https://radimrehurek.com/gensim/models/ldamodel.html>

³<https://radimrehurek.com/gensim/models/coherencemodel.html>

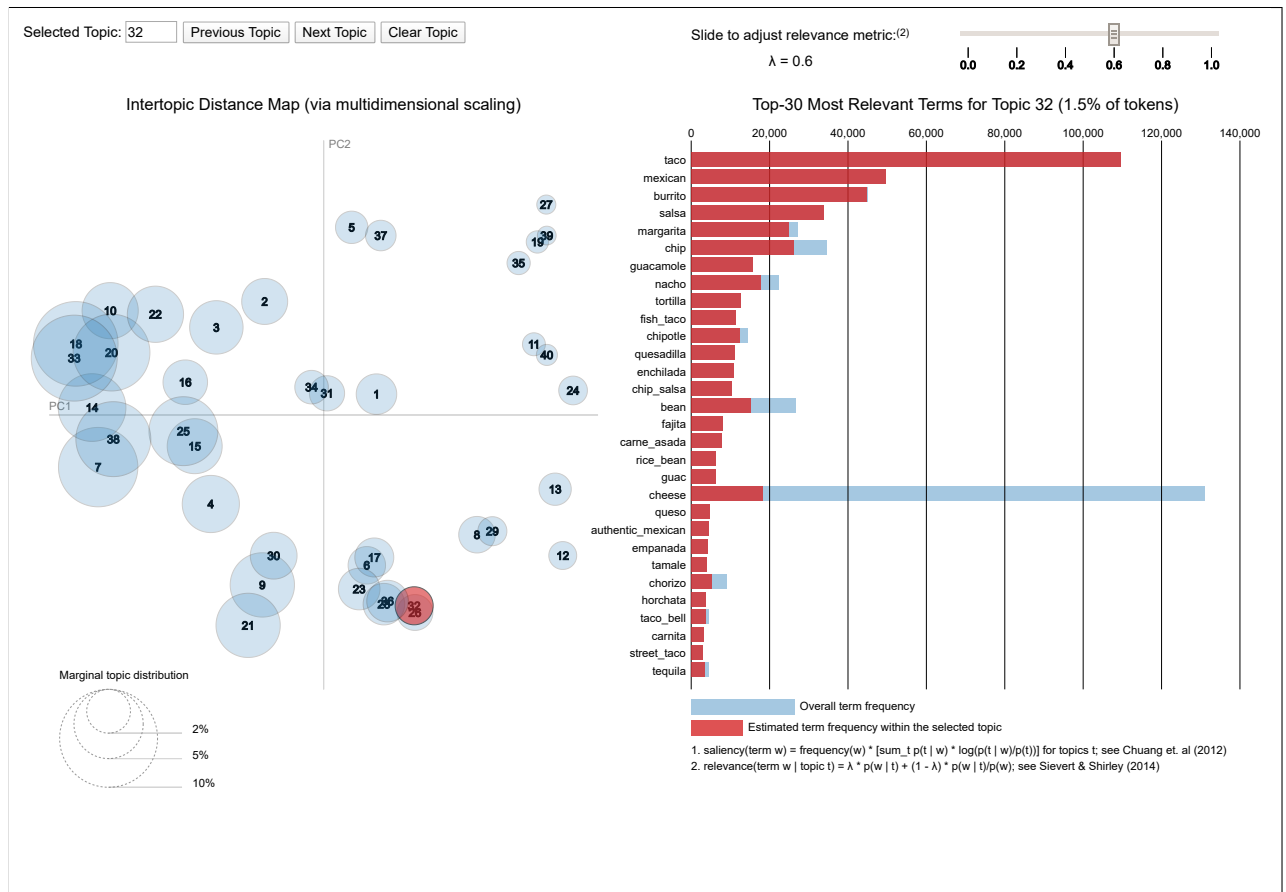


Abbildung 4.1.: LDA-Visualisierung mit LDavis:

Das „mexikanische Topic“ ist ausgewählt; auf der linken Seite ist die Distanz zu anderen Topics visualisiert; auf der rechten Seite werden die Topic-Begriffe angezeigt. Der Begriff „cheese“ ist trotz großer Wahrscheinlichkeitszuordnung für das Topic weiter unten eingeordnet, da „cheese“ auch ein allgemein verbreiteter Begriff für andere Topics ist.

Zur besseren manuellen Analyse der generierten Topics verwenden wir das Tool LDavis (Sievert & Shirley, 2014), für welches ein Gensim-kompatibler Python-Port⁴ existiert.

Das Tool erlaubt eine visuelle Analyse der generierten Topics, in der die Ähnlichkeit von Topics abgebildet ist und die Topic-Begriffe nach einem anderen Relevanz-Maß sortiert sind (Abbildung 4.1, vgl. Tabelle 4.2).

Mit dem trainierten LDA-Modell können wir jedem Review eine Wahrscheinlichkeitsverteilung über die Topics als Topic-Score zuordnen.

⁴<https://github.com/bmabey/pyLDavis>

„[...], this cute vegan restaurant has what every vegan wants most, flavour!!
The food is delicious and healthy, and they have cookies for anyone with a
sweet-tooth. [...]“

Topic	Topic-Score
dessert, cake, flavor, sweet	0,209
menu, option, item, choice	0,131

Tabelle 4.4.: Topic-Score-Zuordnung für ein Beispiel-Review:
Erwähnung von veganem Essen ist häufig korreliert mit der Erwähnung über verfügbare Menü-Optionen.

Wir führen die Topic-Zuordnung für jedes Review aus und bilden eine Topic-Matrix, wobei die Spalten für die Topic-Nummern und die Reihen für die Reviews stehen.

Nach Wieringa (2019)⁵ normalisieren wir die Topic-Scores für ein Review, so dass die Summe der Topic-Scores einer Reihe insgesamt 1 ergibt. Als weitere Anpassung des Vorgehens erstellen wir auch eine nach der Bewertung gewichtete Topic-Matrix, indem wir die Reihen der normalisierten Topic-Matrix mit der jeweiligen Sterne-Bewertung der Reviews multiplizieren. Für das 5-Sterne-Review der Tabelle 4.4 erhalten wir damit die Scores 3,076 (*Dessert etc.*) und 1,924 (*Menü-Optionen etc.*).

Für unsere Visualisierungen aggregieren wir die Reviews nach dem Veröffentlichungsdatum in Monatsgruppen und bilden davon den Mittelwert.

Wir haben weiterhin die Möglichkeit, unser Modell erneut auf gefundene Topics anzuwenden, um detailreiche LDA-Modelle und Zeitreihengraphen für spezielle Themen zu erhalten. Mit diesem Vorgehen können wir einzelne Topics genauer inspizieren und somit „Intra-Topic-Trends“ finden. Dies können beispielsweise die Trends verschiedener Gerichte, welche alle derselben Küche angehören, sein. Hierzu wählen wir die Reviews aus, bei denen das Gewicht des ausgewählten Topics am größten ist. Für das Topic „dessert, cake, flavor, sweet“ wäre beispielsweise das Review der Tabelle 4.4 enthalten. Für die ausgewählten Reviews können wir erneut unser Modeling unter angepassten Parametern anwenden.

5. Evaluation

In diesem Teil gehen wir nun auf unsere erhaltenen Topics ein und wollen den Verlauf der Topics anhand einiger Beispiel-Abbildungen verdeutlichen.

⁵<https://jeriwieringa.com/2017/06/21/Calculating-and-Visualizing-Topic-Significance-over-Time-Part-1/>

5.1. Topics

Nr.	Topic
0	menu option item choice choose offer vegan healthy try vegetarian
1	bar drink great happy_hour bartender patio sit cocktail nice friend
3	steak dinner restaurant meal service great order appetizer dessert salad
5	sushi roll fresh fish salmon order rice japanese chef sashimi
7	sandwich lunch salad bread today soup fresh cheese sub order
11	meat veggie wrap grill salad gyro fresh platter plate lamb
12	chicken fry rice order crispy meal piece tender grill dry
15	coffee drink cafe work store shop nice friendly starbucks location
16	breakfast egg brunch bacon pancake potato bagel waffle coffee order
18	tea water cup lemonade sweet lemon hot iced_tea ice_tea pot
23	shrimp fish seafood oyster fry lobster crab calamari fresh fish_chip
25	soup noodle bowl ramen beef broth pork rice dumpling come
27	burger fry cheese order bun poutine dog shake bacon onion_ring
28	bbq rib side mac_cheese meat brisket smoke sauce mac_n_cheese pull_pork
29	dessert cake flavor sweet try chocolate cookie like taste delicious
30	beer wine great selection bar beer_selection bottle pub local tap
31	taco mexican burrito salsa chip margarita cheese nacho order guacamole
33	ice_cream not lol like boba s n u be get
34	wing game tv ranch monday tuesday special wednesday hot buffalo
35	pizza italian pasta cheese order crust slice sauce topping delivery
38	donut slider dozen old_fashioned john glaze half_dozen deviled_egg x donut_shop
39	pho waffle smoothie vietnamese cash spring_roll bubble_tea doughnut fresh egg_roll

Tabelle 5.1.: Auszug^a der erzeugten Topics

^aTopics, die sich nicht auf Küche, Gerichte oder das Essen beziehen sind in der Tabelle ausgelassen. Das sind unter anderem Topics über Dekoration, Las Vegas, Standort, Preis, Sauberkeit usw.

Wir können in der Tabelle 5.1 sehen, dass wir größtenteils sehr passende und kohärente Topics, wie 1, 5, 16, 23, 31, 35, erhalten haben.

Wie bei vielen Topic Models üblich, existiert in unserem Modell auch ein Topic, 33, in welchem sehr generische (Stop¹-)Wörter (Boyd-Graber et al., 2014) gesammelt sind.

Topic 11 ließe sich als gemischt (Boyd-Graber et al., 2014) oder uneindeutig bezeichnen. Sowohl „veggie“, „fresh“ und „salad“ als auch „meat“, „gyro“, „platter“, sind untereinander kohärent, tauchen aber gemeinsam in einem Topic auf.

¹„but“, „not“ und „get“ sind in der Stoplist enthalten. Es kann aber sein, dass diese Wörter falsch geschrieben wurden und damit nicht als solche erkannt wurden. Im Lemmatisierungsprozess könnten diese dann zu „but“, „not“ oder „get“ umgeformt worden sein.

5.2. Trends

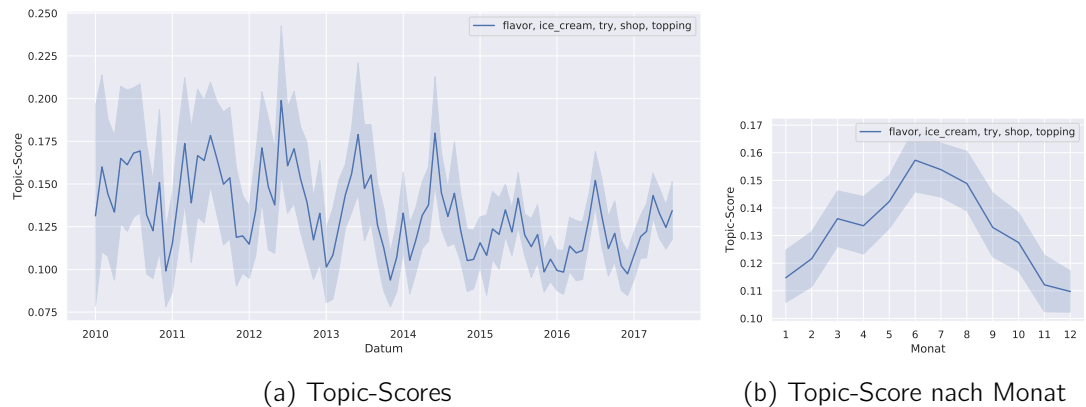


Abbildung 5.1.: Verlauf und Score des Eis-Topics

Abbildung 5.1 zeigt den Score des Eis-Topics über die Jahre und nach den Monaten. Die Schwankungen des Topics zwischen dem höchsten und dem niedrigsten Wert waren von 2010 bis 2015 stärker als ab 2017. Jedoch kann man immer noch gut erkennen, dass der Score jährlich im Sommer den höchsten Punkt erreicht, während der niedrigste Punkt im Winter liegt. Mithilfe von Abbildung 5.1b kann man genau sehen, dass der Juni der beste Monat ist, und es danach wieder rapide abfällt. Unternehmen die Eis anbieten, bspw. Eisdielen, sollten also damit rechnen, dass in den Sommermonaten eine höhere Nachfrage nach Eis vorliegt als in den restlichen Monaten des Jahres.

Abbildung 5.2 zeigt den Score des Steak-Topics nach Monaten, über die gesamte Zeit und zusätzlich die Trend-Regression des Topic-Scores über die Jahre. Man kann in Abbildung 5.2a erkennen, dass der Topic-Score im Februar seinen höchsten Wert erreicht, bevor er stark absinkt und im April, sowie von Juni bis zum August die niedrigsten Werte erzielt werden. Zum Ende des Jahres steigt der Score dann wieder an. Mithilfe von Abbildung 5.2b und Abbildung 5.2c offenbart sich, dass der Score des Steak-Topics über die Jahre hinweg stark gesunken ist. Für Restaurants gilt also, dass die Nachfrage im Sommer am geringsten, und im Winter am höchsten sein wird.

5. Evaluation

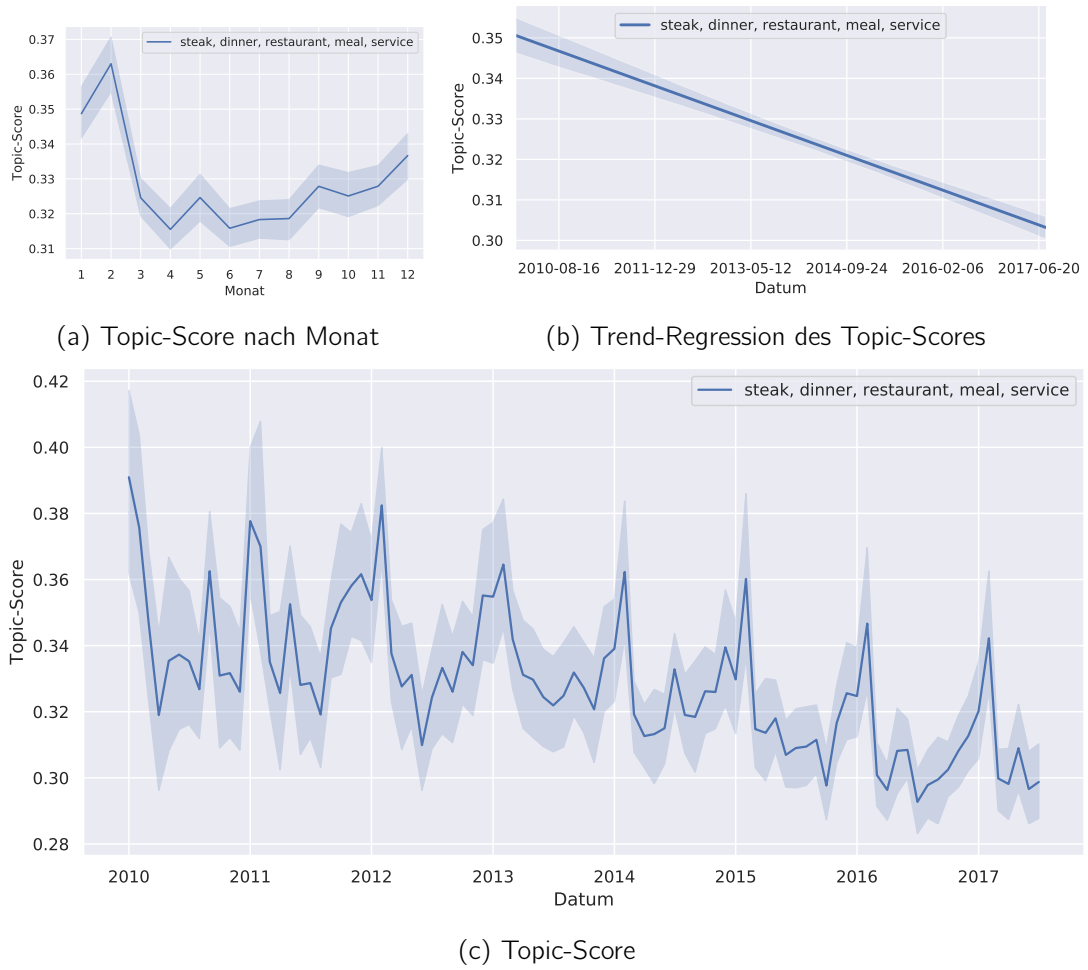


Abbildung 5.2.: Verlauf und Score des Steak-Topics

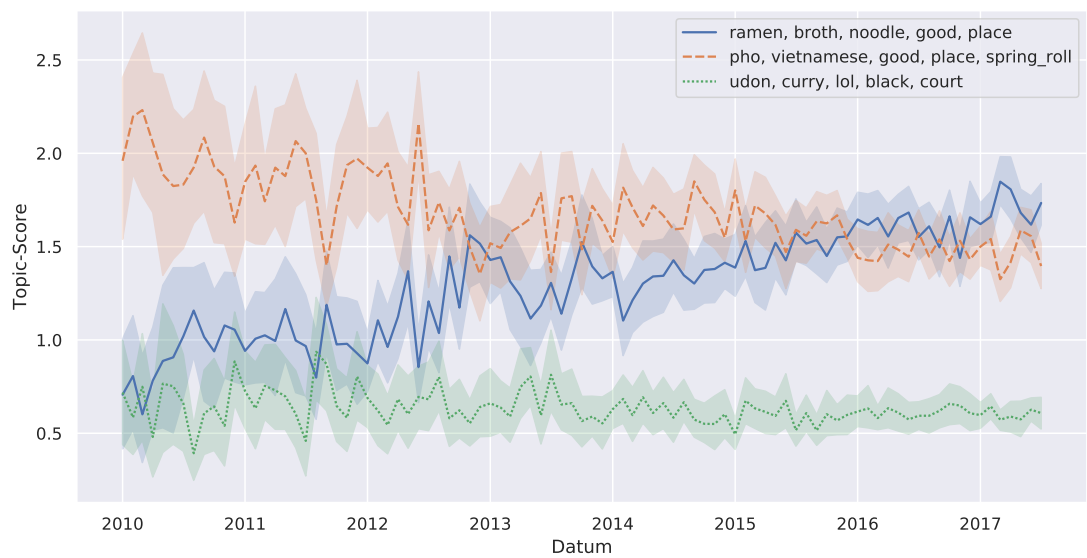


Abbildung 5.3 zeigt, wie mehrere asiatische Topics in Relation zueinander stehen. Das blaue Topic mit Ramen und ist das einzige der drei mit einem Wachstum, das grüne bleibt relativ gleich auf demselben Score und das orangene vietnamesische Topic sinkt im Gesamtverlauf etwas. Man kann daraus also schließen, dass Ramen beliebter wurden, und die anderen bleiben gleich beliebt bzw. haben etwas an Beliebtheit verloren.

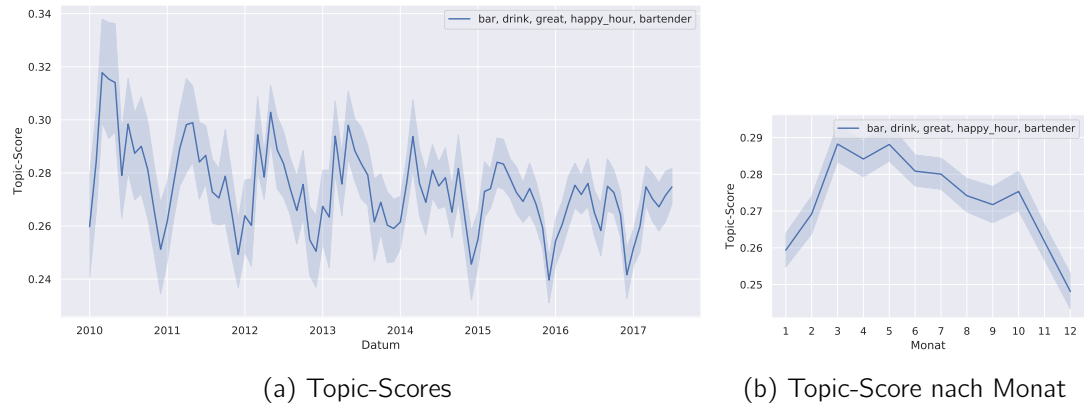


Abbildung 5.4.: Verlauf und Score des Bar-Topics

Man kann durch den immer weiter sinkenden Score in Abbildung 5.4a von Abbildung 5.4 erkennen, dass Bars an Beliebtheit verlieren. Außerdem zeigt Abbildung 5.4b, dass die Frühlingsmonate von März bis Mai für die höchsten Scores verantwortlich sind, vor allem in den Wintermonaten ist der Score deutlich geringer und hat regelmäßig im Dezember die niedrigsten Werte. Für Bars und sonstige Gaststätten liegt die Hauptsaison also zwischen März und Oktober.

Abbildung 5.5 zeigt, dass mexikanische Topics an Beliebtheit gewannen. Die besten Monate sind März bis Juli, und der höchste Wert wird im Mai erzielt, die schlechtesten Monate sind auch hier wieder November, Dezember und Januar.

5. Evaluation

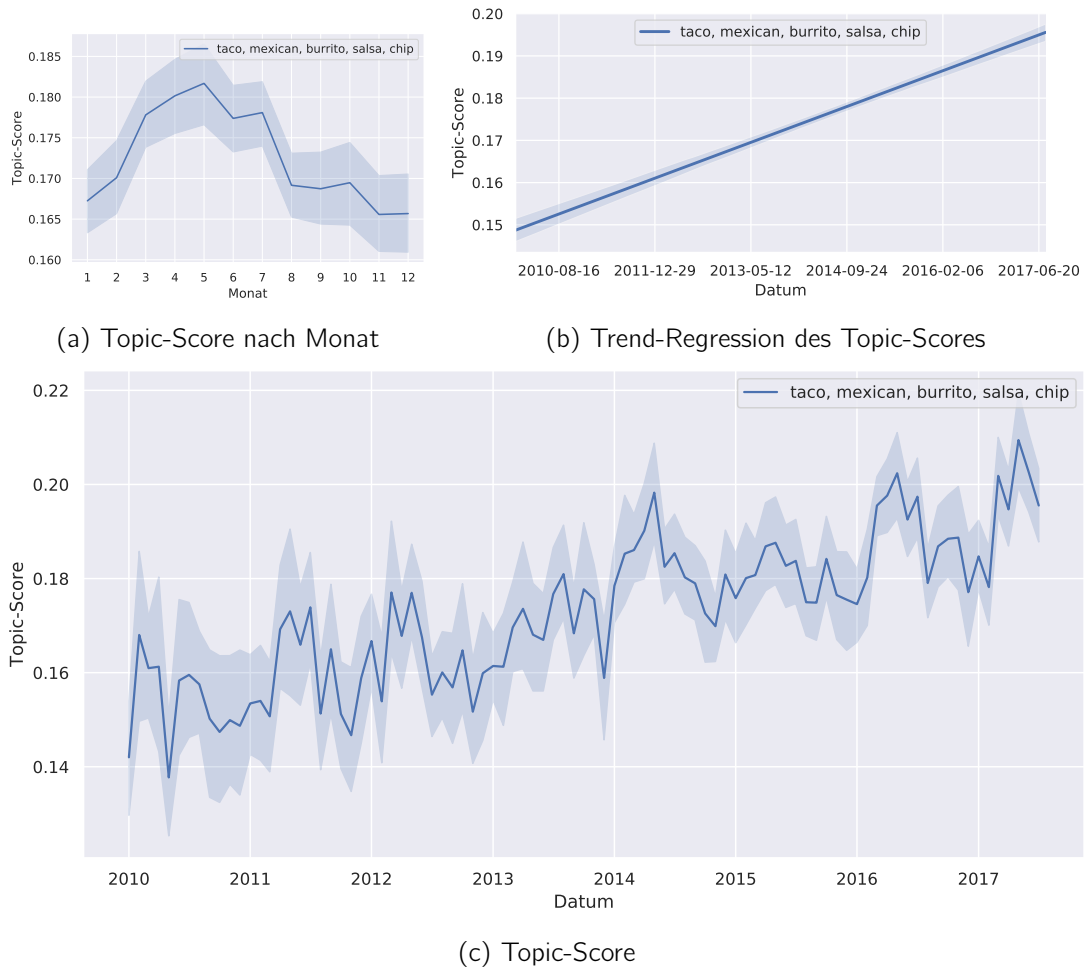


Abbildung 5.5.: Verlauf und Score des mexikanischen Topics

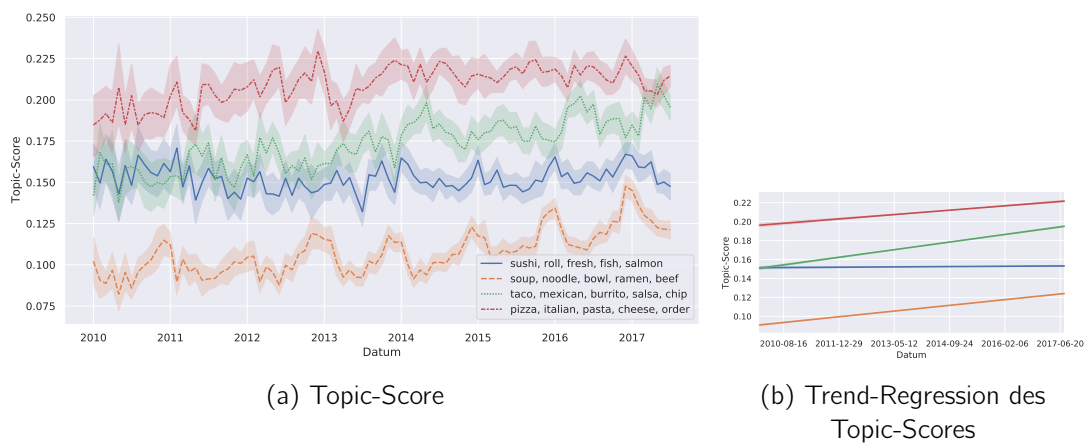


Abbildung 5.6.: Verlauf und Score für verschieden-ethnische Küchen

Wir sehen in Abbildung 5.6, dass asiatische, mexikanische und italienische Küche weiterhin einen positiven Trend aufweisen. Auffällig ist, dass Sushi- und Fisch-Gerichte stagnieren, während asiatische Suppen-Gerichte den stärksten Trend aufweisen.

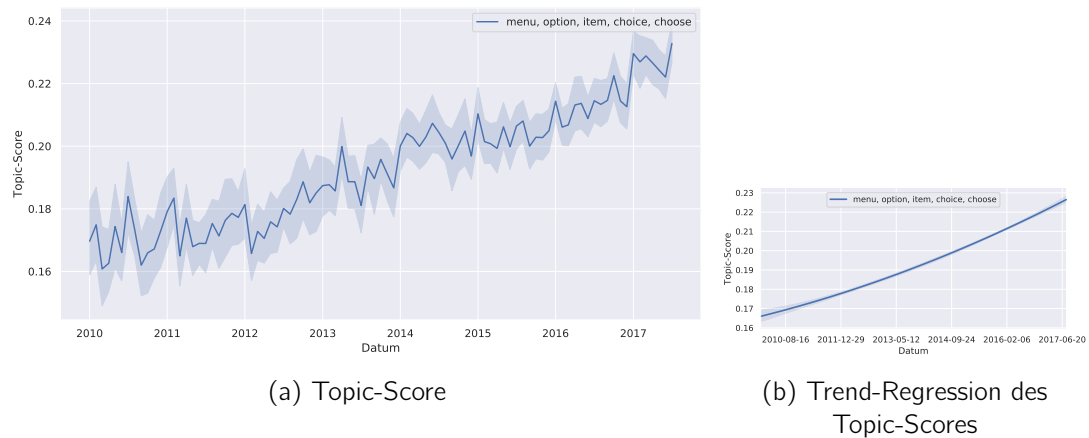


Abbildung 5.7.: Verlauf und Score für Speisekarten-Optionen

In der Tat erkennen wir in dem Datensatz einen positiven Trend im Bezug zu variationsreicheren Optionen in Speisekarten wieder. Dieses Topics enthält weiterhin Begriffe wie „offer“, „vegan“ und „healthy“ (Tabelle 5.1).

Wir haben auch bereits gesehen, dass Reviews korrekt mit diesem Topic zugeordnet werden (Tabelle 4.4).

6. Deployment

Der Code dieser Studienarbeit kann zur Bearbeitung anderer Datensätze benutzt werden, um dort die Topics, deren Score usw. erkennen zu können.

Da die verwendete LDA-Implementation¹ das erweiterte LDA-Modell „Online LDA“ (Hoffman, Bach & Blei, 2010) implementiert, kann das Topic-Modell laufend aktualisiert werden und somit mit neuen Reviews weitertrainiert werden.

In Jupyter Notebooks (Kluyver et al., 2016) können interaktive Tools (wie LDAvis: Abbildung 4.1) verwendet werden, um Experten ohne technischen Hintergrund die Handhabung mit dem Modell zu erleichtern.

In Online-Notebooks wie Google Colaboratory² können Teams gemeinsam Notebooks verwenden und bearbeiten. Diese bieten auch die Möglichkeit, interaktive Formulare (Abbildung 6.1) zu erstellen.

Darüber hinaus kann unsere Code-Basis einfach wiederverwendet und erweitert werden. Beispielsweise sind für den Graphen aus Abbildung 5.3 nur wenige Zeilen Code nötig (Code 1).

¹<https://radimrehurek.com/gensim/models/ldamodel.html>

²<https://colab.research.google.com/>

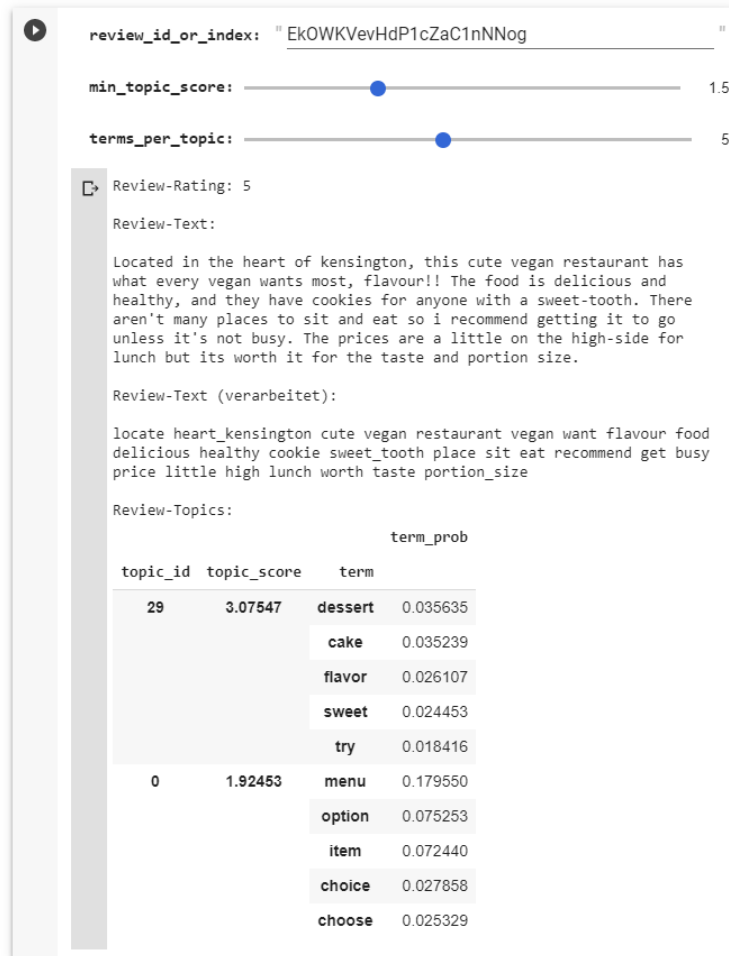


Abbildung 6.1.: Interaktive Code-Zellen in Google Colab: Variablen in Code-Zellen können zu interaktiven Input-Feldern umgewandelt werden.

```
asian_review_ids = max_topic_score_reviews(topic_matrix, 25, 39)
asian_model = TopicAnalyzer.from_review_filter(
    dataframe, review_ids=asian_review_ids, name="asian"
)
asian_model.get_ldavis(num_topics=30, passes=40)
# ... Manuelle Evaluation des Modells mittels LDAvis ...
# Wähle relevante Topics aus.
# 2 = ramen, ...
# 25 = pho, ...
# 26 = udon, ...
asian_model.set_valid_topics([2, 25, 26])
asian_model.plot(2, 25, 26)
```

Code 1: Beispiel-Code zur Detailanalyse ausgewählter Topics

Literaturverzeichnis

- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3 (Jan), 993–1022.
- Boyd-Graber, J., Mimno, D. & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. In E. M. Airoldi, D. Blei, E. A. Erosheva & S. E. Fienberg (Hrsg.), *Handbook of mixed membership models and their applications*. Boca Raton, Florida: CRC Press. Zugriff auf [docs/2014_book_chapter_care_and_feeding.pdf](#)
- Buzalka, M. (2001). Food trends to keep an eye on. *Food Management*, 36 (12), 10.
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (S. 288–296).
- DiPietro, R. B., Roseman, M. & Ashley, R. (2006). A study of consumers' response to quick service restaurants' healthy menu items: attitudes versus behaviors. *Journal of Foodservice Business Research*, 7 (4), 59–77.
- Hoffman, M., Bach, F. R. & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (S. 856–864).
- Honnibal, M. & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. (To appear)
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... et al. (2016). Jupyter notebooks - a publishing format for reproducible computational workflows. In *Elpub*.
- Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.
- Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (S. 3111–3119).
- Mortensen, O. & Konstantinovskiy, L. (2016, Nov). *gensim: Lda model*. Zugriff auf https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html
- Nakatani, S. (2010). *Language detection library for java*. Zugriff auf <https://github.com/shuyo/language-detection>
- Řehůřek, R. & Sojka, P. (2010, 22. Mai). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (S. 45–50). Valletta, Malta: ELRA. (<http://is.muni.cz/publication/884893/en>)

- Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (S. 399–408).
- Roseman, M. G. (2006). Changing times: Consumers choice of ethnic foods when eating at restaurants. *Journal of Hospitality & Leisure Marketing*, 14 (4), 5–32.
- Schofield, A., Magnusson, M. & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (S. 432–436).
- Sievert, C. & Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (S. 63–70).
- Wallach, H. M., Mimno, D. M. & McCallum, A. (2009). Rethinking lda: Why priors matter. In *Advances in neural information processing systems* (S. 1973–1981).
- Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., ... others (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.
- Wieringa, J. E. (2019). A gospel of health and salvation: Modeling the religious culture of seventh-day adventism, 1843 - 1920. *A Gospel of Health and Salvation*. Zugriff auf <http://dissertation.jeriwieringa.com/> (Technische Blogposts auf <https://jeriwieringa.com/portfolio/dissertation/>.)

Anhang A.

Digitale Abgabe

Diese Studienarbeit wird mithilfe eines Ordners auf Panda abgegeben, welcher folgendes enthält:

- **Studienarbeit.pdf** ist dieses Dokument im PDF-Format,
- Der in dieser Studienarbeit verwendete Code wurde in **Python** in Form von Jupyter Notebooks implementiert.
 - `01_Data_Understanding.ipynb` enthält den Code für den Data-Understanding-Abschnitt,
 - `02_Data_Preparation_Modeling.ipynb` enthält den Code für die Data Preparation, das Modeling und der Evaluation.
- Zusätzlich kann unser vollständiges finales Modell online, in einem Google Colab Notebook, ausgeführt werden. Die benötigten Dateien und finalen Modelle werden dabei ohne heruntergeladen zu werden aus einem Google-Drive-Ordner geladen.

[Google Drive Ordner](#),

[Notebook 1: Data Understanding](#),

[Notebook 2: Data Preparation, Modeling & Evaluation](#)