DHBW Mannheim

Hausarbeit

Trendanalyse mit Hilfe von tf-idf

Projekteinreichung: Neue Konzepte II

Dozent: Dr. Tobias Günther

Vorlesung: Neue Konzepte II

Kursnummer: WWI-18-DSA

Name: Bulkens, Björn;
Köhler, Florian

Matr.-Nr.: 9936663; 4810569

E-Mail: s182146@student.dhbw-mannheim.de;
s18100@student.dhbw-mannheim.de

Fachsemester: 05

Studiengang: BA Wirtschaftsinformatik Data Science

Abgabedatum: 1. Februar 2021

Inhaltsverzeichnis

1 Datenquellen

Die verwendete Datenquelle sind Titel von Forenbeiträgen der communitybasierten Plattform Reddit.com. Reddit hat täglich über 52 Millionen aktive Nutzer und ist innerhalb des letzten Jahres um 30% gewachsen(Reddit.com, 2021). Reddit zeichnet sich dadurch aus, dass sich Menschen für jedes beliebige Thema in Untergruppen, sogenannten Subreddits, zusammenfinden können. Diese Gruppen erstrecken sich von Politik über Nachrichten bis hin zu Gruppen für die Lieblingsserie oder Schauspieler. Falls es den gewünschten Subreddit noch nicht gibt, kann dieser schlichtweg erstellt werden, um für neue Diskussionen zu sorgen und eine Community zu dem Thema zu bilden. Auf einem solchen Subreddit können mehr als 10.000 Beiträge täglich erstellt werden. Dabei kann es viele aktive Nutzer geben oder aber nur ausgewählte Moderatoren, die in der Lage sind Beiträge zu erstellen. Aber so spezifisch diese Subreddits auch auf kleine Nieschenthemen zugeschnitten sein mögen, so lässt sich doch für alles Kategorien finden, denen diese Gruppen zugeordnet werden können. Bei einigen Subreddits besteht jedoch das Problem, dass diese in der Qualität ihrer Beiträge durchwachsen sind. Insgesamt wurden im Jahr 2020 circa 199 Millionen Beiträge veröffentlicht(Reddit.com, 2021), die wiederum von anderen gesehen, bewertet und kommentiert wurden.
Die für diese Arbeit verwendete Datengrundlage ist ein Ausschnitt der Beiträge vierer unterschiedlicher Kategorien: „News“, „Politics“, „Economics“ und „Sports“. Auch diese Kategorien teilen sich wieder in dutzend(e) Subreddits auf. Es besteht jedoch auch die Möglichkeit selbst Themen einzugeben, die für die Zielgruppe von besonderer Releveanz sind. Aus diesen Kategorien wurden Beiträge ab dem 23.01.2021 gesammelt und täglich, um neue Erweitert. Zum Zeitpunkt des Verfassens dieser Arbeit beträgt die Größe der Datengrundlage unter 100.000 Einträgen. Diese wächst jedoch ständig an. Im Anhang befindet sich zusätzlich eine Liste der zum Zeitpunkt der Erstellung verwendeter Subreddits. Diese können sich im Laufe der Zeit ändern, wenn neue hinzukommen.

2 Datentransformation

Die Datentransformation im Rahmen dieses Projektes unterteilt sich in mehrere Schritte. Am Anfang kommt das Sammeln und Abrufen der Daten. Zwecks eins standardisierten und vereinfachten Zugriffs auf die aktuellen Geschehnisse in den unterschiedlichen Subreddits bietet Reddit einen API-Zugriff an. Mit Hilfe eines Package kann über die gewählte Programmiersprachen Python auf die Schnittstelle zugegriffen werden: PRAW - Python Reddit API Wrapper. Dieses bietet Klassen und Funktionen, um vereinfacht über die API-Schnittstelle auf alle Beiträge, die bis zu circa einem Monat zurückliegen (Der Zeitraum wurde empirisch ermittelt, da die Dokumentation hier keine weiteren Information bereithält) zuzugreifen. Hierzu ist jedoch eine OAuth Autorisierung nötig.
Über ein Generatorobjekt können die Beiträge innerhalb eines Subreddits sortiert durchsucht werden. So ist es möglich historische Daten zu archivieren und die Datenbank stets nur um neue zu erweitern. Bei der Durchsuchung der Subreddits werden nun Dataframes mit unter anderem dem Titel, der Beitrags-ID und dem Erstellungsdatum der einzelnen Beiträge erzeugt. Diese werden täglich für die festgelegten Themenbereiche erweitert und als *.pkl* Datei abgelegt.
Bei Aufruf der gewünschten Kategorie wird die Datenbank der Anwendung dann um die Beiträge, die seit der letzten Aktualisierung neu hinzugekommen sind erweitert. Zuletzt werden die Dataframes dann in die aktuellen und die historischen Daten unterteilt und die Titel werden in einzelne Tokens aufgeteilt. Hierzu werden alle Wörter in Kleinbuchstaben umgewandelt und außerdem alle Zahlen oder Sonderzeichen entfernt.
Hiermit ist das Abrufen und Transformieren der Daten abgeschlossen und die Ergebnisse können weiterverarbeitet werden.

3 Theorie des Lösungsansatzes

Die Theorie des Lösungsansatzes ist, dass potentielle Trends vorliegen, sobald eine Gruppe an Personen überdurchschnittlich häufig über ein bestimmtes Thema schreibt. Hierzu sollen historische Daten herangezogen werden, mithilfe deren festgestellt werden kann, wie häufig welches Wort in der Vergangenheit verwendet wurde. Diese historischen Werte sollen dann mit aktuellen Daten entsprechend verglichen werden. Dies bedeutet, dass ein potentieller Trend nicht darin zu erkennen ist, wie häufig ein Wort verwendet wird, sondern darin wie stark es in seiner Verwendung steigt oder, ob es gar neu aufkommt. Wurden Wörter in der Vergangenheit bereits häufig verwendet, handelt es sich hierbei also um keinen potentiellen neuen Trend. Hinzu kommt der Punkt, dass potentielle Trendwörter in ihrer Relevanz bemessen werden und der Algorithmus möglichst nicht bei der zufälligen besonders häufigen Verwendung kleinerer Füllwörter ausschlägt. Stattdessen soll der Algorithmus aussagekräftige Begriffe liefern, welche in einer weiteren Recherche über das Thema verwendet werden können. Hierzu muss ein Gütemaß entwickelt werden, welches all diese Anforderungen vereinen und entsprechend bewerten kann, welchen Informationsgehalt das Wort insgesamt hat. Dabei soll sich der Algorithmus sehr sensitiv verhalten, da nur auf diese Art Trends bzw. Ereignisse frühzeitig erkannt werden können, ohne dass es sich bereits um allgemein bekanntes Wissen handelt.
Da der angestrebte Nutzen der Applikation im Bereich des *Information Retrieval* einzuordenen ist wurde sich für die Umsetzung der Hauptfunktionalität an einem der meist verwendeten Techniken für diese Art von System orientiert: „Term frequency–inverse document frequency“ (TF-IDF) Aizawa (2003).
TF-IDF baut auf zwei grundlegenden Maßstäben auf: Dem Beliebtheitsmaß, in diesem Fall repräsentiert durch die „term frequency“ (TF), und dem Spezifitätsmaß, in diesem Fall repräsentiert durch die „inverse document frequency“ (IDF) (Aizawa, 2003) . Die TF gib es in zwei grundsätzlichen Verwendungsarten. Dabei gilt es nach Robertson, 1990 zwischen den Möglichkeiten der Term-Findung und -Gewichtung zu unterscheiden. Denn Techniken, welche für die eine Anwendung geschickt sind, sind es nicht unbedingt auch für die andere (Robertson, 1990, S. 364) . Da es bei dem vorliegenden Projekt darum geht Trends zu selektieren und die

Größe der einzelnen Dokumente, also den Titeln der Beiträge, relativ klein ist gegenüber üblicher Anwendungsfälle des TF, wurde sich lediglich auf das Maß der „Term Selection“, abzulesen aus der Tabelle: Übersicht der Maßstabskategorien, fokussiert, um die Ergebnisse zu kategorisieren. „Total TF“ beschreibt die absolute

	M. of popularity	M. of specificity	M. of discrimination	M. of representation
M. for term selection	Total TF	IDF, signal-to-noise ratio	IG relevance weighting	(Total TF)×IDF
M. for term weighting	Within-df	Pairwise mutual information	Relevance weighting	(Within-df)×IDF

Tabelle 1: Übersicht der Maßstabskategorien
Beispiele für Arten von Maßstäben zur Repräsentation der Signifikanz eines Terms übernommen aus Aizawa, 2003. Legende: M. = Measure, TF = term frequency, DF = document frequency

Häufigkeit des Terms im gesamten Datensatz und basiert auf der Annahme, dass auch häufig vorkommende Wörter Relevanz besitzen können(Luhn, 1957). Jedoch ist die Häufigkeit eines Wortes allein nicht genug, um seiner Signifikanz Ausdruck verleihen zu können, da es noch von vielen Füllwörtern umgeben ist, welche es herauszufiltern gilt. Hier kommt dann das Spezifitätsmaß IDF ins Spiel. Diese besteht aus der Anzahl aller Dokumente (N) geteilt durch die Menge an Dokumenten in denen das jeweilige Wort vorkommt. Da die berechnete Zahl durch ihre inhärente Natur, ausgehend von vielen tausenden Dokumenten die durchsucht werden, numerisch sehr klein wird, wird mithilfe des Logarithmus skaliert. Dies hat zum Vorteil vor allem die kleinsten Werte numerisch besser differenzieren zu können. Die Alternative des Information Gains ist an dieser Stelle nicht nützlich, da dieser Werte die schlicht in einem einzigen Titel vorkommen stark bevorzugen würde(Aizawa, 2003). Dies wurde in diesem Projekt als nicht zielführend erachtet, da Trends durch ein vermehrtes Aufkommen charakterisiert werden. TF-IDF ist letztlich das Produkt des verwendeten TF und des IDF und stellt so das Repräsentationsmaß des Terms dar.

4 Eingesetzter Algorithmus

Diese Berechnungen wurde zwecks Anpassung an den verwendeten Anwendungsfall wie folgt verwendet: Sei N die Menge aller Elemente und T eine echte Teilmenge von N, welche die Dokumente, die in den letzten 24 Stunden erstellt wurden, beinhaltet. tf_{j_N}, idf_{j_N} und tf-idf_{j_N} seien der jeweilige TF, IDF bzw. TF-IDF bezogen auf die Menge N bzw. T und das untersuchte Wort j. Um diese zu berechnen wird df_j benötigt, die Anzahl der Dokumente in denen der Term j vorkommt.

$$\text{tf-idf}_{j,N} = \underbrace{\text{tf}_{j,N}}_{\text{Anzahl aller } j \text{ in } N} \cdot \underbrace{\log \frac{N}{\text{df}_j}}_{\text{idf}_{j_N}}, (Aizawa, 2003)$$

Die erzielte Trendwertung (TW) ergibt sich für jedes Wort j aus:

$$\text{TW}_j = \text{tf}_{j_T} * \frac{1}{\text{tf-idf}_{j_T}} * \text{idf}_{j_N}$$

Die durchschnittliche Länge eines verwendeten Dokuments beträgt nur circa 10 bis 15 Wörter, die das angesprochene Thema knapp zusammenfassen sollen. Die Menge T besteht momentan durchschnittlich noch aus weniger als 1/10 der Dokumente in N. Dieser Wert wird mit der Zeit linear kleiner werden, da sich die Datenbasis mit jeder Aktualisierung vergrößert. Dies hat zur Folge, dass der tf-idf_{j_T} hier für sich alleine nicht dazu geschickt ist auszudrücken, wie wichtig ein Wort tatsächlich ist. Der Ansatz ist es deshalb die TF_{j_T}, als die aktuelle Popularität des Terms j, mit seinem Informationsgehalt über alle Dokumente idf_{j_N} zu verrechnen und diesen Wert schließlich über das Inverse der tf-idf_{j_T} zu regulieren. Letzteres hat den Sinn, dass empirische Untersuchungen zeigten, dass idf_{j_T} stark in Richtung lediglich einmal vorkommender Worte tendiert und somit nicht zielführend, während tf-idf_{j_T} überproportional in Richtung von Füllwörtern, wie „the“ und „to“, tendiert. So werden populäre Terme zuverlässig um Füllwörter bereinigt, während potentielle Trends erhalten bleiben.

5 Implementierungsdetails

Zu Beginn werden die Beiträge der letzten 24 Stunden (nachdem die Titel in Tokens umgewandelt wurden) wie in Kapitel 2 beschrieben herausgefiltert. Im hier dargestellten Code wird diese Aufteilung mithilfe einer Mask realisiert. Der Code zeigt die Erstellung eines Dataframes.

Algorithm 1: Filtern auf Beiträge der letzten 24 Stunden

```
yesterday = today() - timedelta(days=1)
mask_now = (Vektor.df["created"] > yesterday )
self.now = Vektor.df.loc[mask_now]
self.now.reset_index(inplace=True, drop=True)
```

Um die Berechnung der Trendwertung, wie in Kapitel 4 beschrieben durchführen zu können, wird der $\mathrm{tf}_{j,T}$ aller j in T benötigt. Dieser wird mithilfe einer Dictionary Comprehension in einer Update-Funktion realisiert. Den Input für die Berechnung stellt das Dataframe mit den Beiträgen der letzten 24 Stunden aus Algorithmus 1 dar.

Algorithm 2: Berechnung $\mathrm{tf}_{j,T}$ aller j in T

```
tf = dict()
for tokens in df["title"]:
    token_count = dict(Counter(tokens))
    tf.update({key:(tf.get(key,0)+value) for
       key, value in token_count.items()})
```

Außerdem wird der idf_N und der idf_T benötigt. Hier wird zuerst die Anzahl der Beiträge des Datensatzes gezählt. Daraufhin wird mithilfe einer Dictionary Comprehension in Verbindung mit einer Update Funktion gezählt, in wie vielen Beiträgen jeder Token mind. 1x vorkommt. Zuletzt wird daraufhin in einer Weiteren Dictionary Comprehension der idf, wie in Kapitel 4 beschrieben berechnet.

Algorithm 3: Berechnung idf_N und idf_T

```
# total number of posts
N = len(df)
# number of posts containing a term
tD = dict()
for tokens in df['title']:
    tD.update({token:( tD.get(token, 0)+1) for
        token in set(tokens)})
# calc idf out of the above calcs
idf = {key:math.log10(N / value) for key, value
    in tD.items()}
```

Des Weiteren wird der tf-idf_T zur Berechnung der Trendwertung benötigt. Hier wird, wie in Kapitel 4 beschrieben, der TF mit dem IDF der Daten der letzten 24 Stunden verrechnet. Diese Berechnung wurde auch wieder mit einer Dictionary Comprehension gelöst.

Algorithm 4: Berechnung tf-idf_T

```
{key:(tf[key]*idf[key]) for key in tf.keys()}
```

Nun werden die Informationen zusammengeführt, um den Trendwert aus Kapitel 4 für jeden Token zu berechnen. Die Schlüssel des erstellten Dictionary mit dem höchsten Wert stellen dabei die Tokens dar, die am ehesten mit einem Trend in Verbindung stehen.

Algorithm 5: Berechnung Trendwerte

```
{key:(value*(1/trend["tf_idf_total"][key])*
    whole["idf"][key]) for key, value
    in trend["total_tf"].items()}
```

Zuletzt wurde eine Nebenfunktionalität umgesetzt, welche den ermittelten Wörtern im Kontext stehende Begriffe, sortiert nach ihrem IDF, zuordnet. Dies dient der Einordnung des Trends, falls der ermittelte Trendbegriff nicht eindeutig sein sollte. Die Darstellung dieser Funktionalität würde jedoch den Rahmen dieser Arbeit übersteigen und kann im Quellcode unter der Funktion „Vektor.context“ gefunden werden.

6 Evaluierung der Umsetzung

Die Evaluierung der Umsetzung ist aufgrund der Art der Anwendung nicht anhand üblicher Gütemaße möglich, sondern beruht rein auf den empirischen Beobachtungen der Autoren und deren Bewertung.
Zum einen ist die Datenbasis noch recht klein, was die Eingrenzung relevanter Terme erschwert. Diese Eingrenzung wird dabei im Laufe der Zeit vermutlich stets besser werden, durch die ständige Erweiterung aller Datensätze. Es bleibt offen, ob die Datengrundlage ab einem gewissen Grad ihre optimale Größe erreicht hat und alte Daten wieder gelöscht werden sollten.
Der zweite Kritikpunkt ist, dass es bisher kein Einbeziehen der aufkommenden Kontroverse des Trend-Themas gibt. Dies könnte in einem weiteren Schritt anhand der Anzahl der Kommentare und der Bewertung im Verhältnis zur Größe der Community und der Zeit, die die dazugehörigen Beiträge bereits online sind stattfinden. Auch wenn momentan noch nicht beachtet werden kann, wie „bedeutend“ der ermittelte Trend ist, so zeigen empirische Analysen doch, dass die ermittelten Themengebiete oft sehr kurzweilige Berichte sind. Dies ist zumindest für uns im deutschsprachigen Raum der Fall, da rund 50% aller Nutzer aus den USA kommen (Reddit.com, 2021) und Analyseergebnisse entsprechend eher in Richtung dort relevanter Trends tendieren.
Ein letzter Punkt ist, dass die Umsetzung einer Nebenfunktionalität, welche im Kontext relevante Wörter, basierend auf deren IDF, dem Trendwort zuordnet wird als sehr hilfreich empfunden. Denn oft ist es schwierig ein einzelnes Wort einem Sachverhalt zuzuordnen. Die ermittelten potentiellen Trends sind trotzdem häufig zielführend und decken bis dato, zumindest im deutschsprachigen Raum, oft übersehene Umstände auf, die die Welt beschäftigen. Diese Themen beschäftigen die Menschen manchmal so stark, dass deren Potential über einen kurzweiligen Bericht hinausgeht und das Thema ein länger anhaltendes Diskussionsthema werden kann. Somit liefert die Analyse die Grundlage, um Spezialisten die Möglichkeit zu geben, diese potentiellen Trends selbst evaluieren zu können. Die Ergebnisse können dann für ihr Unternehmen oder ihren Kontext eingeordnet werden. So kann der Informationsfluss besser gefiltert werden und relevante Themen frühzeitig erkannt werden.

Literatur

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, *39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), 309–317.

Reddit.com. (2021). Reddit's 2020 Year in Review [Online; accessed 30 January 2021]. https://redditblog.com/2020/12/08/reddits-2020-year-in-review/

Robertson, S. E. (1990). On term selection for query expansion. *Journal of documentation*, *46*, 359–364.

Anhang

GitHub zum Projekt: https://github.com/FatManWalking/Reddit_Trend

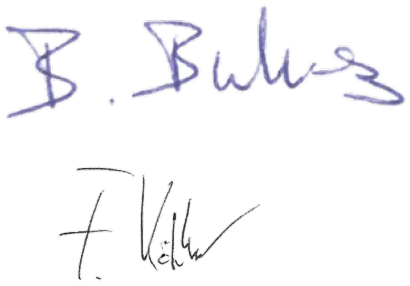
Die genannten Subreddits sind zu finden unter reddit.com/r/*Name*:

Algorithm 6: Liste verwendeter Subreddits

```
sport =   [ "sports", "running", "bicycling", "golf", "fishing",
            "skiing", "sportsarefun", "tennis",
            "rugbyunion","discgolf","cricket","sailing","nfl",
            "CFB","fantasyfootball", "baseball","mlb",
            "fantasybaseball", "nba","collegebasketball",
            "fantasybball", "skateboarding","snowboarding",
            "longboarding", "formula1","MMA","squaredcircle",
            "ufc","boxing","wwe", "MMAStreams","hockey",
            "nhl", "olympics","apocalympics2016",
            "soccer","worldcup","Bundesliga"]
politics = ['Politics','worldpolitics','anarchism','socialism',
            'conservative','politicalhumor','Libertarian',
            'neutralpolitics','politicaldiscussion','ukpolitics',
            'geopolitics','communism','completeanarchy',
            'politicalcompassmemes']
economics = ["Economics","business","entrepreneur","marketing",
             "BasicIncome","business","smallbusiness",
             "stocks","wallstreetbets","stockmarket"]
news = [ "worldnews", "news", "nottheonion", "UpliftingNews"
        "offbeat", "gamernews", "floridaman", "energy",
        "syriancivilwar", "truecrime"]
```

Eigenständigkeitserklärung

Hiermit versicheren wir, dass wir die Hausarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, alle Ausführungen, die anderen Schriften wörtlich oder sinngemäß entnommen wurden, kenntlich gemacht sind und die Arbeit in gleicher oder ähnlicher Fassung noch nicht Bestandteil einer Studien- oder Prüfungsleistung war.



Mannheim, den 1. Februar 2021

Björn Bulkens, Florian Köhler