Social Monitor
Topic Extraction von Twitter-Accounts

Jannik Kollmann
Hochschule der Medien
jk148@hdm-stuttgart.de

Einleitung

Das Projekt ”Social Monitor” befasst sich mit der Zuordnung von Themen zu Twitter-Accounts. Gefordert war eine Webanwendung, die ausgewählten Accounts selbstständig Themen zuweist und zudem ähnliche Accounts gruppiert. Zusätzlich sollte der Einfluss, den die jeweiligen Accounts in ihrem Themenbereich haben analysiert werden. Nach Analyse der Daten, sollten die Daten über eine API zur Verfügung gestellt und visuell von einer Frontend-Anwendung dargestellt werden. In einem nächsten Schritt sollten auch Nachbar-Accounts gefunden und dargestellt werden, die ursprünglich nicht in der Liste der angegebenen Accounts vorkamen. Das Projekt sollte die Grundlage für eine weiterführende Analyse im folgenden Semester schaffen.

Ziele

1. Zuordnung von Twitter-Accounts zu Themenbereichen
2. Realisierung als Webservice um eine Integration in eine bestehende Anwendung zu ermöglichen
3. Bereitstellung der Daten über eine API
4. Visualisierung der Daten in einer Frontend-Anwendung
5. Eine Grundlage für eine weiterführende Entwicklung im folgenden Semester

Software Architektur

Das Projekt wurde mit Hilfe von mehreren Webservices und Technologien realisiert:

1. NodeJS Webserver
2. NodeJS Service als Schnittstelle zu den APIs von Twitter
3. Python Service zur Analyse der Daten
4. ReactJS Frontend
5. PostgreSQL Datenbank

Verwendete Algorithmen

1. **Latent Dirichlet allocation (LDA)** ist ein Topic Modell, das verwendet werden kann um aus einer Dokument/Wort Matrix eine bestimmte Anzahl von abstrakten Themen zu bestimmen.
2. **T-Distributed Stochastic Neighbor Embedding (TSNE)** ist ein Algorithmus zur Dimensionsreduktion. Er wird verwendet um hochdimensionale Daten zu visualisieren.
3. **Density-based spatial clustering of applications with noise (DBSCAN)** ist ein Algorithmus zur Clusteranalyse.

Vorgehensweise

Zum Trainieren des **LDA Modells** wird eine Dokument/Term Matrix erstellt, wobei ein Dokument durch alle Statuseinträge eines Accounts repräsentiert wird. Die Wörter, die in der Matrix für den jeweiligen Account verwendet werden, werden in einem Preprocessing-Schritt aus den Statuseinträgen extrahiert. Durch das Training werden 30 Themen gefunden, die den Accounts im nächsten Schritt zugewiesen werden. Jedem Account wird ein Gewichtsvektor mit der Verteilung der Themen zugewiesen. Die 30-dimensionale Account/Gewicht Matrix wird im nächsten Schritt mit Hilfe von **TSNE** auf 2 Dimensionen reduziert. So können die Daten visualisiert werden. Im letzten Schritt wird ein Clustering auf den 2-dimensionalen Daten mit Hilfe des **DBSCAN** Algorithmus durchgeführt.

Ergebnisse

Die Anzahl der Topics bei denen aus den Wortvektoren ein klares Thema hervorgeht ist stark abhängig von den verwendeten Trainingsdaten. Bei der Eingabe von 770 Dokumenten sind ungefähr 13 von 30 Themen klar definierbar.

Topic	Wortvektor	Thema
Topic 0	bigdata, datascience, analytics, data4good, abdsc	Data Science
Topic 10	trump, crime, protest, attack, climate	Internationale Politik
Topic 16	labour, corbyn, jeremy, brexit, smith	Britische Politik

Table 1: Beispiele für gut definierbare Themen

Topic	Wortvektor
Topic 14	poetry, canterbury, brexit, pension, income

Table 2: Beispiel für ein schwer definierbares Thema

Allgemein war die Qualität der Themen stark abhängig von der Menge der Eingabedaten. Je größer die Anzahl der Dokumente war, desto besser waren auch die gefundenen Themen.

Klassifizierung der Accounts

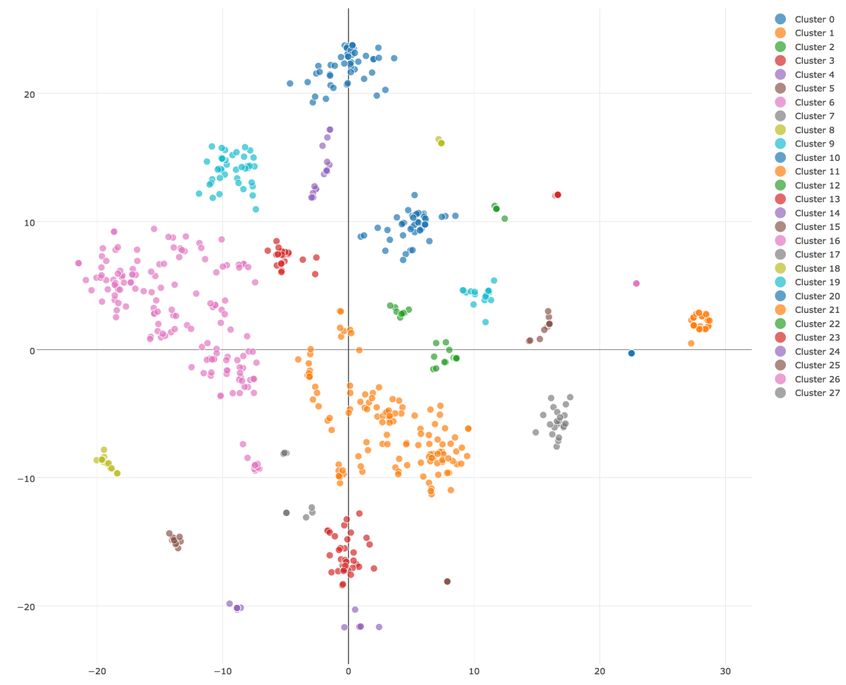

Figure 1: Plot der kategorisierten Twitter-Accounts

In der Abbildung sind die Twitter-Accounts in einem zweidimensionalen Plot abgebildet. Jeder Punkt repräsentiert einen Account. Accounts die im gleichem Cluster vorkommen, besitzen eine ähnliche Themenverteilung. Die Cluster sind durch die unterschiedlichen Farben gekennzeichnet. In dem dargestellten Beispiel wurden 28 Cluster gefunden, Ausreißer wurden nicht dargestellt. Innerhalb eines Clusters existiert meist ein Hauptthema und mehrere Unterthemen. Beispielsweise ist das Hauptthema für Cluster 10 Britische Politik. Zusätzlich gibt es Accounts im gleichen Cluster deren Topic eine Kombination aus Britischer und Internationaler Politik ist.

Fazit

- Die Wahl des Modells war wichtig um gute Themen zu finden, allerdings war es wichtiger eine große Anzahl von Daten zu sammeln. Gerade bei Twitter ist ein großer Anteil der Daten unbrauchbar. Deshalb hat die Vorverarbeitung der Daten eine entscheidende Rolle gespielt.
- Der Einfluss der jeweiligen Accounts wurde noch nicht analysiert, allerdings wurden bereits Accounts analysiert, die nicht im vorgegebenen Datensatz enthalten waren.
- Da das Themengebiet im Voraus nicht bekannt war, hätte es sich angeboten zunächst einige Prototypen zu erstellen um stärken und schwächen unterschiedlicher Herangehensweisen schneller abwägen zu können.
- Das Aufteilen der Anwendung im Voraus ging mit vielen Nachteilen einher. Ohne Automatisierung verlangsamt sich der Entwicklungsprozess enorm.

Ausblick

Das Projekt sollte von Anfang an als Grundlage für eine weitere Arbeit in diesem Bereich dienen. Im kommenden Semester soll die Anwendung verbessert und erweitert werden. Zusätzlich soll sie in eine bereits existierende Webanwendung integriert werden.