Themenanalyse von Nachrichtendiensten mittels Twitter

18. April 2016

1 Ziele

Es soll eine Webapplikation zur Darstellung einer zeitabhängigen Themenanalyse bzgl. der Nachrichtendienste realisiert werden. Mithilfe der Webapplikation sollen z.B. ähnliche Nachrichtendienste für einen bestimmten Zeitraum, der Themenverlauf für einen Zeitraum, überschneidende Themen, etc. dargestellt werden.

2 Aufgabenspezifikation

1. *1. Testat*

   - *Datenaquisition*
     Mithilfe der Twitter API sollen Sie für jeden Nachrichtendienst die Tweets ermitteln, die in der vergangenen Zeit veröffentlicht worden. Dabei sollen mindestens die Nachrichtendienste FAZ, Spiegel, Süddeutsche Zeitung, und die Junge Welt betrachtet werden. Bei der Abfrage ist die Twitter4j Bibliothek hilfreich `http://twitter4j.org`. Des Weiteren benötigen Sie für die Abfrage einen Authentifizierungsschlüssel `https://apps.twitter.com/`. Ein Beispiel für eine Anfrage finden Sie z.B. auf Stackoverflow `http://stackoverflow.com/questions/2943161/get-tweets-of-a-public-twitter-profile`. Dabei speichern Sie folgende Daten zu einem Tweet ab:
     - Datum
     - Quelle
     - Inhalt
     - Retweet Anzahl

     Die Tweets sollen im HDFS in einem Format gespeichert werden, die das Laden in ein DataSet erlauben(JSON oder CSV). Da es nur

möglich ist eine begrenzte Anzahl von vergangenen Tweets zu beschaffen, überlegen Sie sich eine Lösung um eine automatische bzw. semi-automatische Datenakquisition zu gewährleisten.

- *Preprocessing*
  Ein Thema eines Tweets wird bei dieser Aufgabe durch eine Menge von Termen repräsentiert, die bzgl. ihrer IDF Werte ein bestimmtes Kriterium erfüllen. Diesbezüglich sollen sie sich Gedanken machen wie dieses Kriterium definiert werden kann, z.B. *Top k* IDF Terme umfasst die höchsten k Terme bzgl. der IDF Sortierung, Threshold alle Terme, deren IDF Wert über einen Threshold liegt,etc.

  Um Terme dementsprechend zu filtern, muss für jedes Wort der IDF(inverse document frequency)-Wert bestimmt werden. Der IDF- Wert für einen Term t bzgl. einer Menge von Dokumenten D berechnet sich wie folgt:
  $idf(t) = log\frac{|D|}{|D_k|}$

  Implementieren Sie eine Funktion in Flink, um die IDF- Werte für die Terme von Tweets zu berechnen. Basierend auf dem Resultat, führen Sie eine Filterung der Tweets durch.

- *Kookkurrenzgraph erstellung*
  Die Spezifikation eines Themas mittels der gefilterten Wörter ist für eine Analyse nicht hinreichend. Um eine Verallgemeinerung des Themas zu realisieren, soll ein Kookkurrenzgraph basierend auf den Termen erstellt werden. Der Graph $G_{tweet} = (V, E, w)$ ist wie folgt definiert:
  $V := \{t | \exists tweet : t \in tweet\}$
  $E := \{(t_1, t_2) | \exists tweet : t_1 \in tweet \wedge t_2 \in tweet\}$
  $w : E \Rightarrow \mathbb{N}$, $w((t_1, t_2)) =$ number of co-occurrences

  Erstellen Sie den Graphen als Gelly-Graph für einen konfigurierbaren Zeitraum. Erstellen Sie diesbezüglich eigene Klassen, um die notwendigen Werte für einen Knoten bzw. einer Kante zu speichern.

2. *2. Testat* Führen Sie auf den resultierenden Graphen für ein bestimmtes Zeitintervall einen Clustering- Algorithmus durch. Ein Cluster umfasst dabei die Terme, die den Knoten des Clusters entsprechen. Dementsprechend ist ein gefilterter Tweet in einem Cluster und befasst sich eine Quelle mit diesem Thema, wenn alle Terme im Cluster enthalten sind.

   Implementieren Sie den Chinese-Whisper Algorithmus auf Flink.

   Entwerfen Sie eine Webapplikation unter Verwendung der D3.js [1] Bibliothek, um die gewählten Hypothesen darzustellen. Die Mindestanforderung für die Darstellung sind folgende Aspekte:

   (a) *Eingabe*

   - Zeitraum

[1] `https://d3js.org/`

- Nachrichtendienstübergreifend vs spezifischer Nachrichtendienst
- Wenn spezifischer Nachrichtendienst , Zeitfenster

(b) *Darstellung*

(c) Nachrichtendienst übergreifend
Es sollen alle Themen(=Cluster) dargstellt werden, die von den Nachrichtendiensten für diesen Zeitraum angeboten werden, z.B. als Tag-Cloud(Gewichtung mithilfe der Anzahl der Tweets, die Term enthalten) sowie die Anzahl der Tweets pro Nachrichtendienst für jedes Cluster.

(d) spezifischer Nachrichtendienst
Für jedes Zeitfenster soll der Kookkurrenz-Graph sowie die Cluster dargestellt werden. Eine Möglichkeit für die Darstellung, ist die Präsentation mehrerer Graphen mit dem dazugehörigen Zeitintervall oder eine Art Simulation bei der man mithilfe eines Sliders das aktuelle Zeitfenster konfigurieren kann und der aktuelle Graph dargestellt wird. Es können auch andere Ideen realisiert werden.