Konzeptioneller Entwurf einer Anwendung zum Thema:

Tweet Analyse von News

Bearbeitet von:

Hamann, Felix

Hartwig, Mattis

Mäder, Hannes

Betreuender Hochschulmitarbeiter:

Viktor Christen

Inhaltsverzeichnis????

**Aufgabenstellung**:

Die Tweets deutscher Nachrichtendienste sollen auf ihre Ähnlichkeit hin überprüft werden. Mit Hilfe des Daten Streaming Frameworks Apache Flink (kurze Erläuterung was das ist?) und eines von Twitter bereitgestellten API, werden die Tweets zunächst gespeichert. Für die Filterung der gespeicherten Tweets wird für jedes Wort dessen inverse Dokumentenhäufigkeit (engl. Inverse Document Frequency (IDF)) berechnet. Der IDF-Wert berechnet sich folgendermaßen:

\text{IDF}_{t}=\log\left(\frac{N_D}{f_t}\right)

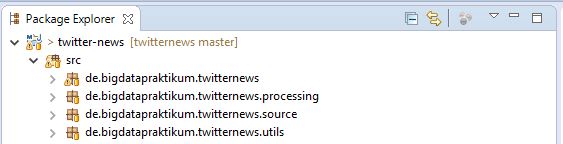
wobei N_D die Anzahl der Dokumente bezeichnet und f_t die Anzahl der Dokumente, die den Term t enthalten. Bei der hier beschriebenen Anwendung stellt jeder Tweet ein Dokument dar und jedes Wort in einem Tweet einen Term.

Basierend auf den gefilterten Termen wird ein Kookkurrenzgraph erstellt. Dieser wird als Gelly-Graph erstellt, der für einen bestimmten Zeitraum konfiguriert werden kann. Auf den erstellen Graphen kann wiederum für ein bestimmtes Zeitintervall ein Clustering-Algorithmus angewendet werden. Zuletzt wird eine Webapplikation zur Verfügung gestellt, um die erarbeiteten Ergebnisse zu visualisieren. Dabei soll es Möglichkeiten zur Eingabe eines Zeitraumes und der Auswahl einer Nachrichtendienstübergreifenden oder spezifischen Darstellung geben.

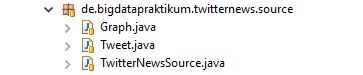
**Konzeptioneller Entwurf**

Als Framework für die Umsetzung wird Apache Flink (hier noch eine genauere Beschreibung?) genutzt.

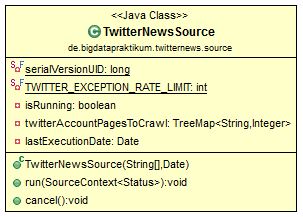
Die Anwendung ist in unterschiedliche Packages gegliedert, welche in nachfolgender Abbildung zu sehen sind.



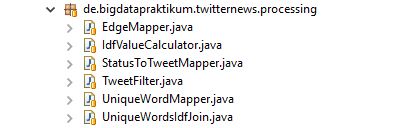
Im Source Package befinden sich die unterschiedlichen Modell-Klassen die die unterschiedlichen für die Anwendung benötigten Objekte enthalten.



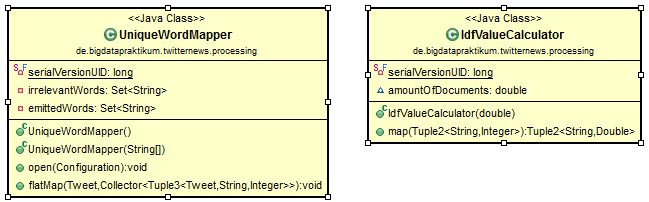
Beispiel Diagramm der Klasse TwitterNewsSoruce:



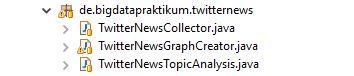
Im Processing Package befinden sich Klassen zum Aufbereiten der Daten, wie etwa der Berechnung der IDF-Werte aller Worte in den Tweets. Außerdem enthalten sind Mapper und Filter Funktionalitäten.



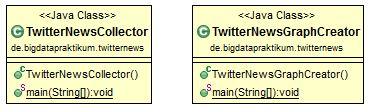
Nachfolgend sind zwei Beispielklassen abgebildet:



Im TweeterNews Package befindet sich der Kern der Anwendung, wie etwa das Speichern der Tweets oder des erstellen des Gelly-Graphen.



Nachfolgend die Klassen TwitterNewsCollector, welche die Tweets aus einer Liste von Twitteraccounts zieht und TwitterNewsGraphCreator welche einen Kookkurrenzgraphen aus den aufbereiteten Daten baut.



Im Utils Package befinden sich verschiedene (eine) Configurations-Dateien.

