2. Testat

Konzeptioneller Entwurf einer Anwendung zum Thema:

Tweet Analyse von News

Bearbeitet von:

Hammann, Felix

Hartwig, Mattis

Mäder, Hannes

Betreuender Hochschulmitarbeiter:

Victor Christen

Konzeptioneller Entwurf

Generelles

Dieses Testat beschäftigt sich hauptsächlich mit dem Sichtbarmachen der Analysen auf den Twitter-Daten. Damit wir unser Backend, das die Analysen durchführt elegant ansprechen können, haben wir uns dazu entschieden das Spring-Framework zu nutzen. So können wir unsere Backend-Services ganz einfach über eine REST-Schnittstelle ansprechen. Sobald die Backend-Services ein Ergebnis zurückliefern, kann dieses im Frontend dann angezeigt werden. Die Ergebnisse der Analyse werden dabei nicht jedes Mal übertragen, sondern an einem gemeinsamen Speicherort gespeichert. So kann das Frontend auch nach einem Neustart noch auf die Ergebnisse der letzten Analyse zurückgreifen.

Zur Umsetzung des Frontends wurde die Bibliotheken D3 und Cytoscape genutzt.

Insgesamt haben wir seit dem 07.06.2016 eine lückenlose Abdeckung der Tweets von den geforderten vier Nachrichtenagenturen. Ein Teil der Daten geht noch weiter in die Vergangenheit, da z.B. die neue Welt weniger twittert und immer die letzten 10.000 Tweets geladen werden können. Insgesamt haben wir 28.000 Tweets gespeichert die analysiert werden können.

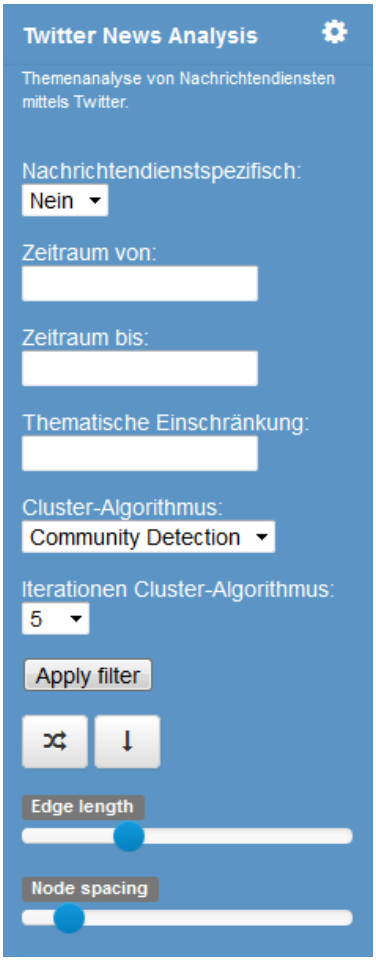
Im Folgenden beschreiben wir die Umsetzung der einzelnen Bausteine.

Clustering

Wir führen ein Clustering auf dem im vorherigen Testat generierten Kookkurrenzgraphen durch. Wir nutzen dazu zwei verschiedene Cluster-Algorithmen. Zum einen verwenden wir den in Flink bereits implementierten Community Detection Algorithmus. Dieser clustert dieser ist mit dem Scatter-Gather-Pattern implementiert und liefert mit der richtigen Anzahl an Iterationen und dem richtigen Delta relativ vernünftige Ergebnisse.

Zusätzlich haben wir noch den Chinese-Whisper-Algorithmus implementiert. Problematisch ist, dass der Algorithmus wie er von Biemann (2006) beschrieben wurde ein sequentieller Algorithmus ist, da die Reihenfolge der Abarbeitung entscheidend ist. Ignoriert man die Reihenfolge einfach, so kann es unter Umständen dazu führen, dass einzelne Knoten ihre Cluster alterieren, anstatt, dass der Algorithmus eine feste Grenzverteilung anstrebt. Trotzdem bekommt man mit dem Algorithmus einigermaßen gute Ergebnisse.

Die Wahl des Algorithmus und die Werte für die Parameter lassen sich bequem über unser Frontend steuern.



Filter

In unserem Frontend lassen sich folgende Einstellungen vornehmen:

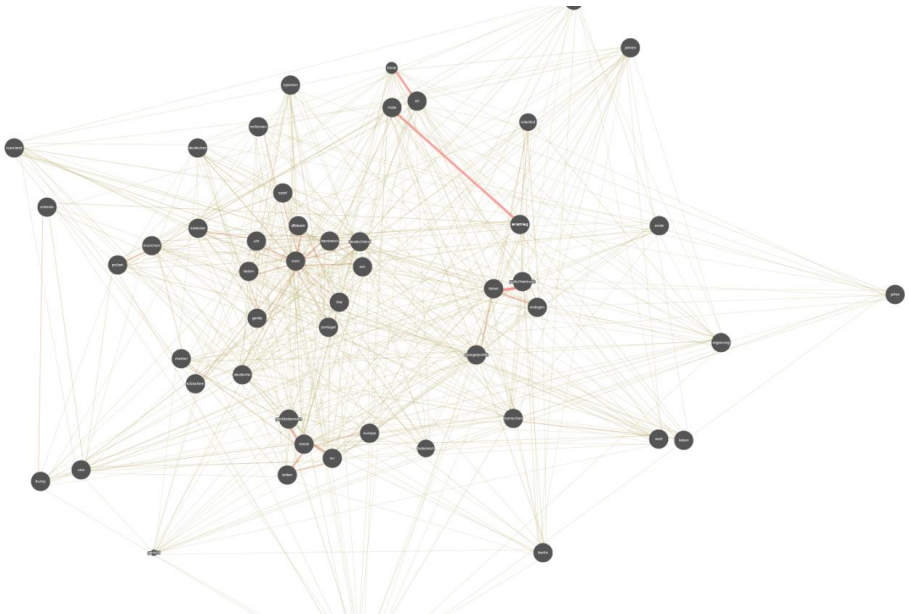
1. Betrachtungszeitraum
2. Einschränkung auf Nachrichtendienste
3. Cluster-Algorithmus
4. Parameter für Algorithmus
5. Thematische Einschränkung (besonders relevant für historische Daten)

Verändert man die Einstellungen kann man das Backend anfordern die Ergebnisse neu zu berechnen. Es ist leider nicht möglich, dass die gleichen Ergebnisse genutzt werden, da die Bewertungen der wichtigen Terme, die Graphen usw. von den Filtereinstellungen abhängig sind.

Dargestellte Informationen

Von unserem Frontend werden Informationen in drei Tabs dargestellt. Ein Tab für den Kookkurrenzgraph, ein Tab für die Tag-Clouds inkl. Detailinformationen und ein Tab für eine historische Darstellung der Tweets.

Die Darstellung des Kookkurrenzgraphen wurde bereits im letzten Testat besprochen. Prinzipiell bestimmt die Größe der Knoten die Häufigkeit mit der ein Begriff genannt wurde und die Dicke der Kanten wie häufig zwei Begriffe gemeinsam auftauchten.



Im Tab mit den Tag-Clouds ist für jedes Cluster dargestellt welche Begriffe zu diesem Cluster gehören. Je größer die Wörter desto relevanter ist das Thema. Klickt man auf eine Tag-Cloud bekommt man zusätzlich Detailinformationen zu dem Thema angezeigt. Zum einen wird angezeigt welcher Nachrichtendienst zu diesem Thema wie oft getwittert hat und zum anderen noch wie viel Prozent der Nachrichten jedes Nachrichtendienstes zu diesem Thema waren. So lassen sich die Unterschiede an zwei Kennzahlen messen, wovon eine auch die Größe des Nachrichtendienstes (bzw. die Gesamtanzahl an Tweets) berücksichtigt.

Im Tab mit den historischen Daten wird angezeigt wie viele Tweets bei den gewählten Filter-Einstellungen historisch verfasst wurden. Besonders interessant ist diese Information, wenn eine Thematische Einschränkung vorgenommen wird. So lässt sich erkennen, ob bestimmte Themen nur eine bestimmte Zeit in den Nachrichten waren. Unten ein Beispiel mit dem Begriff „Nizza“