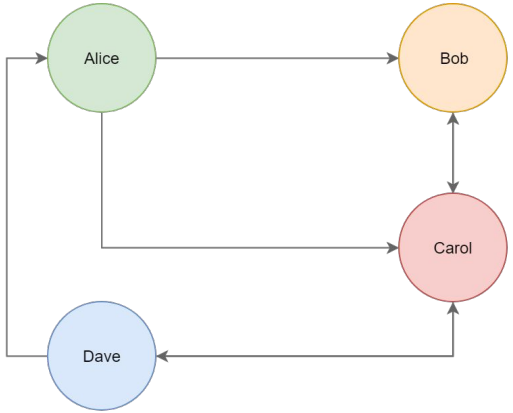
DeepWalk & ParagraphVector on Twitter

Big Data Praktikum 2018

Marvin Hofer, Robert Bielinski
Betreuer: Victor Christen

Twitter Graphen



Alice	#music, #socialtravel, #soundcloud, @Bob, @AppStore, @Apple, @Arclite, @Asana, ...
Bob	#deal, #spring, @Carol, @BillGates, @Microsoft, ...
Carol	#work, #startup, #notreally, @Hotmail, @Hulu, @NewYorker, ...
Dave	...

Unser Datensatz:

https://snap.stanford.edu/data/egonets-Twitter.html

81.306 Knoten, 1.768.149 Kanten, Feature-Listen für jeden Knoten

ParagraphVector Model

Auch Doc2Vec genannt, basiert auf *Word2Vec*

⇒ Continuous Bag-of-Words und Skip-Gram um numerische Feature-Vektoren für Wörter zu erstellen

Baut Feature-Vektoren nicht nur für Wörter, sondern auch für das Dokument selbst.

⇒ Einheitlich lange, numerische Feature-Vektoren, die zum Vergleich von Dokumenten genutzt werden können.

Q. V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. 32, 2014.
https://cs.stanford.edu/~quocle/paragraph_vector.pdf

DeepWalk Model

Basiert auf Word2Vec

Erstellung von Feature-Vektoren aus Graphstrukturen

Generierung von *Random Walks*

1. Start bei einem zufälligen Knoten
2. Laufen zufälliger Pfade entlang der Kanten ausgehend von dem Knoten

Auf diesen Pfaden wird dann ein Modell erstellt.

B. Perozzi and S. Skiena. DeepWalk : Online Learning of Social Representations Categories and Subject Descriptors.
http://www.perozzi.net/publications/14_kdd_deepwalk.pdf

Aufgabenstellung und Ziele

1. Nutzung von Deeplearning4J zur Generierung von ParagraphVector- und DeepWalk-Modellen auf den Testdaten.
2. Darstellung der Daten auf einer Weboberfläche durch D3.js

Potentielle Nutzungsfälle:

Vergleich von Nutzerinteressen und Graphstrukturen zur Generierung von Vorschlägen.

Untersuchung von Sozialen Gruppenstrukturen im Bezug auf Interessengebiete der Nutzer.

DeepLearning4J & D3.js

Deeplearning4J:

- open-source, distributed deeplearning framework
- Implementierung in Java und Scala
- ParagraphVector, DeepWalk, Word2Vec, ...

D3.js:

- Datenbasierte Dokument-Manipulierung mit Hilfe von HTML, SVG und CSS
- JavaScript Framework

https://deeplearning4j.org/, https://d3js.org/

Implementierung

1. Preprocessing:

   Twitter-Daten werden in ein einheitliches Format gebracht und unnötige Daten entfernt
2. Modell-Generierung

   Mit Hilfe von Deeplearning4J werden ParagraphVector- und DeepWalk-Modelle auf den vorbereiteten Daten generiert
3. REST API

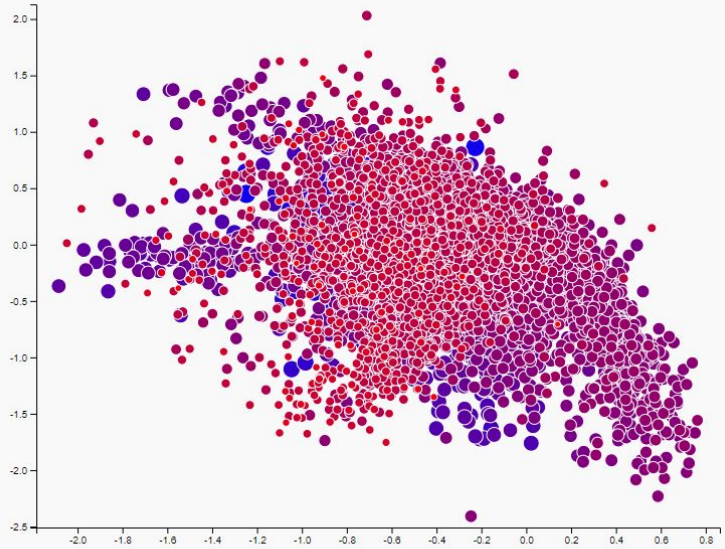
   Algorithmen zur Ähnlichkeitsberechnung, TopK-Berechnung und einfache Feature Anfragen werden über eine REST API nach außen hin zur Verfügung gestellt
4. Javascript & D3 Frontend

   Ein Web-Frontend nutzt die generierten Modelle um eine globale Ansicht der Daten zu erstellen und die REST API um Darstellungen der Beziehungen zwischen spezifischen Knoten zu generieren. Dabei werden die Graphen mit Hilfe von D3 Visualisiert.

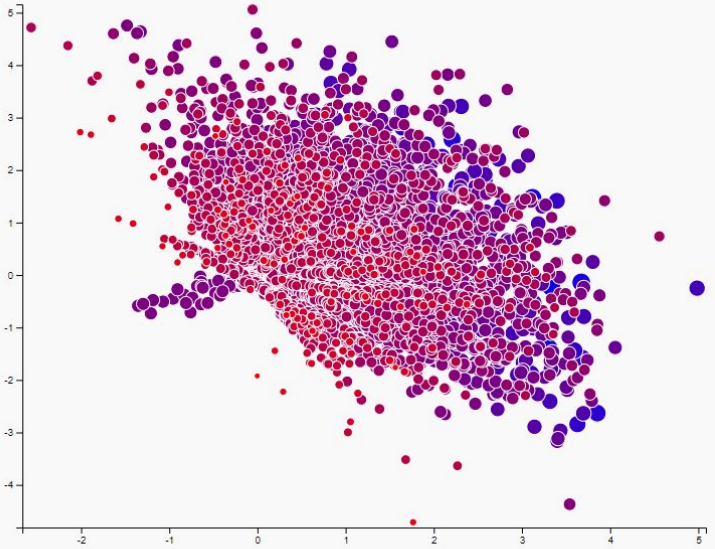
Demo

Node Label: 12831 | Action: GlobalView | Maximum Number of Nodes: 100000

ParagraphVector



DeepWalk



http://localhost:8080/app.html, https://github.com/Termilion/deeplearning-on-twitter