Source Code zur Bachelorarbeit

Hier befinden sich jeglicher Source Code, der zu der während der Erstellung der Bachelorarbeit verwendet wurde. Dabei ist das Projekt inhaltlich entsprechend der Arbeit gegliedert. Der Source Code steht auch in einem GitHub Repository zur Verfügung (URL: https://github.com/FelixBieswanger/OpinionMining)

Nachfolgend ist eine Beschreibung aller Inhalte dieses Ordners.

Ordner: webscraper

Webscraper (Kapitel 3.2.2.2) Hier sind alle Webscraper enthalten. Zudem auch während dem Scraping erzeugt Logs zum nachvollziehen, ob der Scrape erfolgreich geklappt hat

Ordner: translation

Standardisierung (Kapitel 3.2.3)

- translate.py: Übersetzen des Textes der deutschen Artikel
- translate_headlines.py: Übersetzen der Überschriften der deutschen Artikel (relevant für ein Ansatz der Tonalitätsbestimmtung)

Ordner: data_selection

Zeitliche Selektion (Kapitel 3.2.4.1)

- date_selection.py: Auswahl der Artikel basierend auf den Veröffentlichungsdatum

Inhaltliche Selektion (Kapitel 3.2.4.2)

- topicmodelling_eval.py: Evaluierung der Topic Modelling Modelle
- topicmodelling_train.py: Training verschiedener Topic Modelling Modelle mit Grid Search
- topicmodelling_selection.py: Selektion der Artikel anhand der identifizierten Topics

Zudem befinden sich hier auch alle trainierten Modell im **SubOrdner: lda_models** und eine csv mit den Ergebnissen der Evaulierung.

Ordner: sentiment

Sentiment Analyse (Kapitel 3.2.5)

- evaluation.py: Besimmtung des Median Absoulte Error zwischen manuel versehenen Tonalität und automatisch bestimmten
- labeling_process.py: Manueller Labeling Process (Commandline Interface)
- selection.py: Zufällige Auswahl der Artikel für die vorabdurchgeführte Stichprobe
- calc_sentiment.py: Bestimmung der Tonalität und speichern in der Datenbank
- evaluation_other_approaches.py: Bestimmung der Tonalität der alternativ Ansätze und Erstellung des Plots

Ordner: plots

Ergebnisse (Kapitel 4) zum Teil auch Diskussion (Kapitel 5) Hierin sind alle Scripte für die Erstellung der Plots. Zudem im **Subordner: images** auch die Plots abgelegt als png.

Ordner: resources

Hierin sind alle Resources (Scripte & Files) enthalten, die oft in anderen Scripten verwendet werden.

- database.py: Regelung des Datenflusses mit der Datenbank (mongodb)
- keys.py: Bereitstellung der verwendeten Keys (API und Passwörter), tatsächliche Key-File allerdings nicht in der Abgabe aus Datenschutzgründen.
- logger.py: Implementiertung eines Loggers
- preprocessung.py: verschiedenste Preprocessing funktionenen

Zusätzlich sind hier die Antworten des Survey zur Interpretation der Wordclouds (für besser verständliche Plots)

Ordner: outdated

Hierin sind Ansätze die anganfen wurden zu implementieren aber nicht in die Arbeit geschafft haben. Daruter fällt ein Ansatz zur Datenselektion mit Doc2Vec und der Scrape der Quelle Twitter.

File: requirements.txt

Liste aller verwendeten Pyhton Libraries, die gemäß pip install -r requirements.txt installiert werden können

File: .gitignore

Spezifikation aller Files die nicht von Git versioniert werden sollen