Bachelorarbeit Vorbereitungsfragen

1) Recherchieren Sie einen theoretischen Hintergrund für die Arbeit, den Sie gleichzeitig mit bearbeiten.

- Datensätze:
  - MovieLens : https://grouplens.org/datasets/movielens/25m/
    - https://www.kaggle.com/rounakbanik/the-movies-dataset?select=ratings.csv
  - Reddit Corpus: https://github.com/fau-klue/german-reddit-korpus
    - PDF: https://www.aclweb.org/anthology/2020.lrec-1.774/
- Rekommandation System:
    - https://en.wikipedia.org/wiki/Recommender_system
- Vektorisierung/ Word Embedding
  - Bag of Words/ Count Words
    - https://en.wikipedia.org/wiki/Bag-of-words_model
  - Tfidf
    - https://en.wikipedia.org/wiki/Tf%E2%80%93idf
  - Vec2Word/ Doc2Word
    - http://web2.cs.columbia.edu/~blei/seminar/2016_discrete_data/readings/MikolovSutskeverChenCorradoDean2013.pdf
    - https://cs.stanford.edu/~quocle/paragraph_vector.pdf
  - **Google ELMo**
    - https://arxiv.org/pdf/1802.05365.pdf
  - **Google Bert**
    - https://arxiv.org/pdf/1810.04805.pdf
    - http://jalammar.github.io/illustrated-bert/
- Cosine Similarity
  - https://en.wikipedia.org/wiki/Cosine_similarity
- Frameworks
  - Numpy, Pandas
    - Arbeiten mit Dataensätzen
  - NLTK
    - Hilfe für Tokenisierung
  - SKlearn
    - Vectorizer Implementationen
  - Tensorflow
    - EMLO/ BERT Implementierungen
    - (GPU Beschleunigung)
- Deploment:
  - https://symfony.com/doc/current/components/process.html
  - https://www.sandervanhooft.com/blog/laravel/how-to-use-laravel-with-python-and-the-command-line/

2) Finden Sie einen Titel für die Arbeit

Vergleich von Word Embedding Verfahren für ein Content-Based Recommendation System

3) Schreiben Sie eine kurze Zusammenfassung / Abstract (0,5 Seiten) für die Arbeit und den geplanten Inhalt.

Das Ziel der Bachelorarbeit ist ein wertender Vergleich verschiedener Word Embedding Verfahren für die Nutzung in einem Recommendation System. Dieses soll in einem sozialen Netzwerk eingebunden werden, wo Benutzer Gruppen beitreten und Posts veröffentlichen können.

Um den „Cold Start“, d.h. einen Start ohne Daten für ein funktionierendes System, zu verhindern, wird auf ein inhaltsbasiertes (Content based) Recommendation System gesetzt. Das System ist so aufgebaut, dass zuerst die Rohtexte vorbereitet/gereinigt werden (Filtern der Stopwörter und Tokenisierung). Die so bearbeiteten Daten werden dann im Vektorisierungsverfahren auf ihre Eigenschaften(Features) komprimiert und in einer Matrix zusammengefasst. Dann können die vektorisierten Posts mit Hilfe der Kosinus Ähnlichkeit (Cosine-Similarity) verglichen und eingestuft werden.

Beim Vergleich der verschiedenen Word Embedding Verfahren soll der Unterschied und Fortschritt zwischen den Modellen analysiert werden. Dabei liegt der Fokus besonders auf dem BERT Verfahren von Google. BERT steht für “Bidirectional Encoder Representations from Transformers” und ist das erste Transformer Modell seiner Art. Transformer gehören zu den Deep-Learning-Architekturen und werden mit einer großen Menge von Daten vortrainiert. Mit dem Prinzip von Transfer-Learning kann das Modell nun mit Finetuning auf ihre Aufgabe angepasst werden.

Dazu werden alle Verfahren in einem Test Durchlauf mit dem MovieLens Datensatz eingebunden und untersucht (siehe unter 4. Evaluation). Für Implementierung wird dann das am besten funktionierende Modell genutzt.

Das Recommendation System kann auf verschiedene Arten im Netzwerk eingesetzt werden. (Wie die genaue Implementierung sein soll, muss noch besprochen werden.)

Beispiele:

1. Jedem Benutzer wird auf Basis seiner gefolgten Gruppen und interagierenden Posts neue Post aus anderen Gruppen vorgeschlagen.
2. Gruppen werden untereinander verglichen und dem Benutzer als mögliche interessante Gruppe vorgeschlagen.
3. Posts können in mehreren ähnlichen Gruppen angezeigt werden

(Möglichkeit User-Bewertungen der vorgeschlagenen Posts, Like/Dislike, mit einzubeziehen)

Für die Integration in das System sind Python Skripts, die mit Hilfe von Symfony Process ausgeführt werden, vorgesehen. Außerdem soll ein Verfahren für die zyklische Einbindung der neusten Daten erstellt werden.

4) Wie möchten Sie die Ergebnisse Ihrer Arbeit evaluieren?

Um ein Rekommandation System komplett zu evaluieren, muss man es in der Praxis beobachten und mit den vorhandenen Daten im Laufe der Zeit verbessern. Grundsätzlich kann man sagen, dass die Effizienz des Systems mit der Menge an verfügbaren Daten ansteigt.

Für die Evaluation liegt nun das Problem vor, dies ohne reale Daten zu testen. Somit wäre ein Datensatz nötig der dem späteren, realen Datensatz möglichst gleich kommt**.** Dieser ist aber zurzeit nicht zur finden. Besonders Datensätze in deutscher Sprache sind rar. Der am besten passende Datensatz wäre der Reddit Corpus, eine Sammlung aller Reddit Posts, dessen Daten mit einem Skript für deutsche Texte gefiltert werden können.

Mein Vorschlag wäre für die Evaluation der unterschiedlichen Word Embedding Verfahren, den oft im Bereich der Rekommandation Systemen verwendeten Datensatz von MovieLens zu benutzen. Dieser enthält eine Vielzahl von Filmbeschreibungen und User Bewertungen. Mit einer Train/Test Teilung kann dann die Genauigkeit des Systems mit den Messungsverfahren „Mean Average Percision“ (MAP) und „Normalized Discounted Cumulative Gain“ (NDCG) berechnet und verglichen werden. (http://fastml.com/evaluating-recommender-systems/)

Für die letztendliche Implementierung könnte man dann mit dem Reddit Corpus als Test Datensatz arbeiten und stichprobenartige Überprüfungen durchführen.

5) Erstellen Sie eine Projektskizze bzw. einen Projektplan für Ihr Praxisprojekt.