Quotes Recommender System, Report 2

Semester Project, Practical Machine Learning and Deep Learning, Fall 2023, Innopolis University

Team

Name	E-mail	Responsible for
Vladimir Makharev	v.makharev@innopolis.university	Data, method, demo, presentation
Artem Batalov	a.batalov@innopolis.university	Method, demo
Georgii Budnik	g.budnik@innopolis.university	Method, organization, presentation

Responsibilities can be combined between us

GitHub repository

https://github.com/kilimanj4r0/quotes-recsys

Description

"If I quote others, it is only to better express my own thought"
— Michel de Montaigne is a French philosopher of the Renaissance

Inspirational quotes for everyday help a person achieve goals and believe in himself, do not give up and move on. The goal of our project is to instantly select a quote based on the answer to the question "How was your day?" to lift the mood and improve the next day.

The goal of our project is to develop a system of recommendations for quotations based on the answer to the question to lift your mood and improve your mood for the next day. The recommendation engine will be based on NLP mechanisms and extraction of valuable features for deep learning.

Tags

RecSys, NLP, PML, DL, Feature Extraction

Progress

The past weeks have been devoted to discussing and shaping the construction of a common work process. Several meetings were organized, discussions and brainstorming of all ideas for the implementation of the project task were held.

As a result of the meetings, the main direction to solving the problem was chosen, and experiments were conducted. The code of the experiments can be found in the repository and it will be developing in further weeks.

Brainstorm

The main difficulty of the logical implementation, which we have in the project, is the classification of Quotations (quotes theirselves) and Posts (response from users). In the process of generating ideas for solving this problem, several possible solutions were proposed: creating tags for quotes and responses from users. This approach boils down to the task of multi-tabling, which was considered in different variations. Our main points (hypotheses) were:
- Get tags from Post
  - Predict sentiment of Post
  - Predict multi-label tags of Post
- Get tags from Quote
- Generate matching tags from Post & Quote
  - Get mean embeddings of matching tags
  - Find the closest embedding to each other, return most suitable quote
  - And other experiments with different embeddings generating process
- Build index of quotes (maybe full search will be slow)

Code development

To evaluate the list of hypotheses, the following procedural steps were executed:
- Acquisition of datasets from the following sources: https://huggingface.co/datasets/go_emotions, https://huggingface.co/datasets/sem_eval_2018_task_1
- The decision was made to utilize pre-existing models based on RoBERTa, in order to avoid redundancy. These models can be found at the following URLs: https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion-multilabel-latest, https://huggingface.co/SamLowe/roberta-base-go_emotions
- An initial assessment was conducted by applying the 'roberta-base-go_emotions' model without the classification head, focusing solely on cosine similarity between embeddings. The outcome demonstrated

improved performance compared to the base RoBERTa model when applied to data closely resembling our dataset.

- An investigation led to the discovery of the 'torch.nn.CosineEmbeddingLoss' function (as documented [here](#)). This discovery suggests the possibility of creating a model composed of two embedding models, one for the post and another for the quote, with a target value of 1 or -1. This approach is utilized for measuring the similarity or dissimilarity between two inputs and presents another proposal for how the recommender system may function.

Additionally, it's worth noting the information found in the document titled 'How to Build a Text-Based Recommender System with NLP,' available at https://mobidev.biz/blog/how-to-build-text-based-recommender-system-with-nlp. We plan to leverage options 1 and 3 from this source as part of our recommender system.

Further work

In the following weeks, the application part will be disassembled, so that it was laid down in the past weeks. Various principles of data processing and primary training of what is already available will be considered. Several hypotheses will be tested.