Quotes Recommender System, Report 3

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

We revised our RecSys problem and documented it all from the beginning.

Our Initial Thoughts

1. Find labeled datasets of diary-style texts and quotes, then somehow play with their labels.
2. Fine just datasets of texts that are close to the diary-style domain and quotes datasets (maybe parse from somewhere), then label all of these manually using one multi-label classifier.

Found datasets

1. go_emotions
2. sem_eval_2018_task_1
3. journal-entries-with-labelled-emotions
4. Diary-Entry-To-Rap
5. Quotes Dataset

Found classifiers from the top HuggingFace Trending Models

1. roberta-base-go_emotions — 27 labels + Neutral label
2. twitter-roberta-base-emotion-multilabel-latest — 11 labels (or 4 labels from tweetnlp [1])
3. bert-base-uncased-emotion — 6 labels (Ekman emotions)
4. emotion_text_classifier — 6 labels (Ekman emotions)
5. EmoRoBERTa — 27 labels + Neutral label

We will experiment with the **first two classifiers**. There are reasons to decline 3, 4, and 5:

- roberta-base-go_emotions outperforms EmoRoBERTa on the same dataset
- bert-base-uncased-emotion and emotion_text_classifier are trained to predict only 6 Ekman emotions, so we consider it as not enough for our task. Also, according to Demszky *et al.* [2], the 6 emotion categories proposed by Ekman in 1992 are very basic and recent findings in psychology offer a more complex "semantic space" of emotion.

Goal

The goal is to build a content-based *recommendation system*. By choosing this approach, we will recommend to user x similar to previous items rated highly by x.

We need to construct a dataset of user interactions with items. Our setting is:

- **user**: represented by unique id and text (diary-style) + features of the text (emotions)
- **item**: represented by unique id and quote + features of the quote (emotions)
- **interaction**: weight that denotes whether user liked the item or not, i.e., it can be either -1 or 1

Therefore, by combining all things we will get our utility matrix (dataset).

Preliminary Solution Steps

1. Solve Gathering Ratings problem
    1. Collect samples that represent **user** and **item** (filtering suitable data)
    2. Construct feature vector for each **user** and **item** by using multi-label classifier (choosing somehow appropriate classifier)
    3. Utilize some model that will create a dataset of "reasonable" pairs (diary-style text, quote) so the we can annotate it.
    4. Build a dataset manually (or using external tools) by creating interaction examples based on the dataset of pairs (diary-style text, quote).
    5. Split the dataset into train and test
2. Solve Extrapolating Utilities problem
    1. Make utility matrix dense (probably it will be dense since we have only numerical features and binary interaction weight)
    2. Train different RecSys models
3. Evaluate extrapolation methods
    1. Evaluate (compare) trained models by well-known RecSys metrics

Problems to solve

- How to build a user profile? Most probably, it will be a collection of quotes user like and dislike.
- How to avoid overspecialization? User might have multiple quotes that they like

Footnotes

RecSys models and metrics can be taken from the most recent and new open-source library RecTools released by MTS [3]. These findings were inspired by Habr article about RecTools [4] and lecture about Recommendation Systems (on the PMLDL course).

Further work

Finish the planned pipelined (solution steps), visualize results, create a demo (add author to quotes) and presentation.