CS410 Text Info System - Project Progress Report

Fall 2023

Qinxi Wang | qinxiw2@illinois.edu

qinxiw2

The progress writeup reports on how the project implementation has been coming along since the proposal. We will answer 3 main questions:

1) Which tasks have been completed?

We have completed the initial data preparation task, including data cleaning and explory on the dataset selected. We identified the issues with the raw dataset, including ambiguous identifier and data representation, missing data, ill-formats, outliers, duplications, and data skewness. We used basic NLP skills to clean those up, and saved the post-progressed dataset for later steps.

We also started on project development layout, and conducted research on data transformation and aggregation for performative schema storage and search index tasks. We also read up on literature on how to perform sentiment analysis and opinion mining on similar datasets but in different applications and domains.

2) Which tasks are pending?

We still have the pending tasks as the following that are either still work in progress, or not started yet:

1. Build search index and allow for searching wine by description using free text across columns
2. Categorize reviews by positive, neutral, or critica sentiment using a fine-tuned sentiment analysis classifier
3. Retrieve relevant wine search by variety, country, designation, etc and build similar wine recommendations using content based and collaborative filtering for the wines retrieved from the search
4. Implement functionality where a user can search, select, and retrieve similar / recommended wines
5. Recommend similar wines via content based and collaborative filtering, and tree-based models to learn bias / weights on various discriminative features; Run model eval for offline experiments, and add instrumentation of online metrics
6. Documentation of code, produce a demo and write up for the final submission

3) Are you facing any challenges?

We have not hit any major blockers yet, but so far we have identified the following areas that imposes challenges or are harder than we originally anticipated:

- Indexing and retrieval on raw and long text is rather slow, and our system has not been as performative as we thought an interactive program should be
- It’s difficult to find a performative and reliable pretrained sentiment analysis model, might need to apply fine tuning but our dataset size is quite limited for that to be useful
- As we plan for the further implementation, integrating a UI with python/jupyter environment which we have been utilizing for exploration and visualization seems hard
- Another non-technical challenge is that we simply lack the time to work on it due to other MP or shifting deadlines, and other courses’ deadlines also. We hope to wrap those up then return to this project with more focus post fall break.