CIND820 – Final Capstone Project

Supervisor: TBD

By: Daria Yip (500721106)

Submission Date: January 23, 2023

In the past few years, with the advent of the Internet and the rise of instant gratification, companies who have been able to provide services on-demand have been the greatest winners. One of the companies that have led the transition to digital media consumption is Netflix. During the pandemic, they achieved even greater success through a significant increase in subscriptions. One of the technologies that streaming platforms like Netflix and its peers (i.e. Disney, Crave, Hulu) share is recommendation systems.

This paper looks at the logic behind recommendation systems and applying it to the Yelp dataset to provide personalized recommendations to the user based on data such as their reviews, ratings, location, and more. The dataset provides the information in .json form, and provides six files with information on users, businesses, reviews, check-ins, tips, and photos. Two common candidate generation approaches include content-based filtering, which is item-based filtering, and collaborative filtering, which uses user-based filtering (Google, 2022). The research questions that I will attempt to answer are: “For restaurant recommendations, do content-based filtering models or collaborative filtering models provide better suggestions?” and “What level of accuracy/performance can I achieve given these data points?”

To evaluate my models, I propose using the metric recall. As we are looking to create a Top-N recommendation list, recall is known to be effective in evaluating the ability of the model in finding relevant results (Cremonesi, P., & Turrin, R. (n.d.)). Currently, I am contemplating using the leave-one-out cross validation method (LOOCV), where it involves leaving a known rating to the side and later trying to predict it. Alternatives would be to use the k-fold cross-validation or the train-test split. As the Yelp dataset is very large, it may be computationally expensive to use k-fold cross-validation or the LOOCV methods. However, I will attempt to use the k-fold cross validation first, and then use the LOOCV, then the train-test split respectively.

I will be using Python to code this project. Tools that I am proposing to use include Python’s Pandas, Scikit-Learn, Numpy, Weka, and MatPlotLib libraries. With these tools, I will perform dimensionality reduction, data cleaning, exploratory data analysis, create training and test sets, and develop the content-based and collaborative. Then, I will evaluate this model with the aforementioned metric of recall.

# Dataset Link

Dataset Link: https://www.yelp.com/dataset

# GitHub Link

<https://github.com/dariayip/Final-Capstone-Project> (currently private)

# References

Google. (2022, July 18). *Candidate Generation Overview  |  machine learning  |  google developers*. Google. Retrieved January 23, 2023, from https://developers.google.com/machine-learning/recommendation/overview/candidate-generation

Cremonesi, P., & Turrin, R. (n.d.). *An evaluation methodology for Collaborative Recommender Systems - polimi.it*. Retrieved January 24, 2023, from https://chrome.deib.polimi.it/images/8/8e/Cremonesi\_2008\_CROSSMEDIA.pdf