Capstone 2 Project Proposal

Priya Sathish

Content based News Recommendation System

Problem

Online platforms support so many of our daily activities that we have become dependent on them in our personal and professional lives. We rely on them to buy and sell goods and services, to find information online and to keep in touch with each other. These platforms could help consumers by recommending items as per their interest and preference by just analyzing your past interaction or behavior with the system. From Amazon to LinkedIn, Uber eats to Spotify, Netflix to Facebook, Recommender systems are most extensively used to suggest "Similar items", "Relevant jobs", "preferred foods", "Movies of interest" etc. to their users. Recommender system with appropriate item suggestions helps in boosting sales, increasing revenue, retaining customers and also adds competitive advantage. Recommender systems use a number of different technologies and can be classified into two broad groups as Content based recommendation and Collaborative filtering.

On a day to day basis, the internet has a lot of sources that generate immense amount of daily news diversified in subject matter. There is continuous demand for new information to be available immediately and with ease by the consumers. So, it is crucial that the news is classified and targets the needs and requirements of the user effectively and efficiently. News services have attempted to identify articles of interest to readers based on the articles that they have read in the past. The similarity might be based on the similarity of important words in the documents or on the articles that are read by people with similar reading tastes. The same principles apply to recommending blogs from among the millions of blogs available or other sites where content is provided regularly.

This project focuses on content-based recommendation using News category dataset. The goal is to recommend news articles which are similar to the already read article by using attributes like article headline, short description, category, author and publishing date.

Client

News readers, blog readers, news agencies, bloggers, retailers and several online platforms.

Data and approach

https://www.kaggle.com/rmisra/news-category-dataset

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. News in this dataset belongs to 41 different categories. Each news record consists of a headline with a short description in our analysis. In addition, we will combine attributes ‘headline’ and ‘short description’ into a single attribute ‘text’ as the input for classification and proceed with developing a deep learning model to build the recommender system.

Citation: "https://rishabhmisra.github.io/publications/"

Deliverables

1. Python code on Jupyter Notebook
    a. Data wrangling
    b. Exploratory Data Analysis
    c. ML/Deep Learning model
2. Prediction model and Report on github.