Title: Designing a Content-Based News Article Recommendation System.

Abstract: In this report, we present the design and implementation of a content-based news article recommendation system that suggests personalized articles to users based on their reading history and interests. We utilized the News Category Dataset from Kaggle, which contains around 210k news headlines from HuffPost. We preprocess the data by joining the Headline and Short Description columns, replacing missing values, and dropping null values. We also perform stopword removal and lemmatization using the nltk package and vectorization using Google's pre-trained Word2Vec model from gensim package. Our recommendation system is based on content-based filtering, which recommends articles based on their similarity to a given article. We build three models, one based on the 'Text' column, one based on the 'Text' and 'Category' columns, and one based on the 'Text', 'Category', and 'Authors' columns. Our models calculate similarity between articles based on the average of their Word2Vec vectors and use weighted-couple distance to combine similarities across different columns. Future work includes building a simple UI/Website for our recommendation system, recording user interactions, and implementing collaborative filtering.

Keywords: recommendation system, content-based filtering, natural language processing, Word2Vec, News Category Dataset, Kaggle

I. Introduction

The abundance of information available online has made it increasingly difficult for users to find relevant content. Recommendation systems have become an essential tool to provide personalized content suggestions to users. In this report, we present the design and implementation of a content-based news article recommendation system that suggests personalized articles to users based on their reading history and interests.

II. Dataset and Preprocessing

We utilized the News Category Dataset from Kaggle, which contains around 210k news headlines from HuffPost. The dataset includes columns for category, headline, authors, link, short_description, and date. To preprocess the data, we joined the Headline and Short Description columns and made one 'Text' column out of it. We replaced missing values in the 'authors' column with 'Unknown' and dropped the rows with null values in the 'text' column. We also performed stopword removal and lemmatization using the nltk package and vectorization using Google's pre-trained Word2Vec model from gensim package.

III. Model Building

In order to build a recommendation system that recommends news articles based on their similarity to a given news article. We first use Word2Vec, a natural language processing technique, to create vector representations of words in article headlines. Then, we calculate the similarity between different headlines based on the average of the Word2Vec vectors of their words.

For recommendations based on single feature ie using only ‘Text’ Column - we write a function that returns the "num_similar_items" most similar articles to a given article, based

on the Euclidean distance between their average Word2Vec vectors. The recommended articles are displayed in a DataFrame along with their publishing dates and similarities to the given article.

For recommendations based on multiple features ie ‘Text’ column and ‘Category’ column. We first do one hot encoding on ‘Category’ column and then based on formula: weighted_couple_dist = (w1 * w2v_dist + w2 * category_dist)/float(w1 + w2), we get recommended articles that are from the same category as the queried article category due to high w2 weight.

For recommendations based on multiple features ie ‘Text’ column, ‘Category’ column and ‘Authors’ column. We do one hot encoding on ‘Authors’ column and then based on formula: weighted_couple_dist = (w1 * w2v_dist + w2 * category_dist + w3 * authors_dist)/float(w1 + w2 + w3), we get recommended articles that are from the same author as the queried article author due to high w3 weight.

IV. Conclusion

In this report, we presented the design and implementation of a content-based news article recommendation system that suggests personalized articles to users based on their reading history and interests. Our models are based on content-based filtering and utilize natural language processing techniques such as stopword removal, lemmatization, and Word2Vec vectorization.

V. Future Work

In the future, we plan to build a user interface or website for our content-based recommendation system. This website will allow users to interact with our system and provide feedback on the recommended articles. By recording user interactions with the system, we can start building up user data on the news articles, such as ratings, time spent on reading the article, search history, etc. This user data will be valuable in implementing collaborative filtering, a technique that recommends articles based not only on article features but also on similar users. Collaborative filtering can significantly enhance the recommendation system's robustness, and we plan to integrate it with our content-based filtering method to create a more powerful and personalized recommendation system.

References:

[1] News Category Dataset, Kaggle. https://www.kaggle.com/rmisra/news-category-dataset