CS586: Big Data Analytics

Project - 5

Parth Jay Dhruv - pdhruv@wpi.edu

Vignesh Sundaram- vsundaram@wpi.edu

Naitik Zaveri - nrzaveri@wpi.edu

Padmesh Naik - pnaik@wpi.edu

Table Of Contents

1. Introduction

The goal of this project is to create a news recommendation system that can provide readers with customised news articles depending on their interests. The system is intended to assist users in finding articles that are relevant to their interests in an information overload environment. The project proposes using a combination of Cosine similarity, and Content-based filtering to develop the recommendation system. Furthermore, we will use the Word2Vec model to generate word embeddings and calculate article similarity. This report outlines the methodology, dataset, preprocessing steps, model architecture, evaluation and results, and training approach.

2. Methodology

We propose a combination of two filtering techniques: Cosine similarity and Content-based filtering to develop the news recommendation system.

1. Cosine similarity:

   Word2Vec will be used to generate embeddings and calculate similarity between articles. A popular technique in natural language processing (NLP) is cosine similarity, which is particularly useful for determining how similar documents are. The system identifies articles that are similar to the user's interests based on the similarity between the word embeddings of the articles using the Cosine similarity method.

2. Content-based filtering:

   Content-based filtering recommends items that are similar to the ones a user liked in the past. In this situation, the system will give users recommendations for articles based on the articles' substance. This is accomplished by scrutinizing the article's headlines, classifications, and keywords.

3. Dataset

The news recommendation system is developed using a dataset obtained from Kaggle, which includes 210,294 news articles in JSON format from various categories such as sports, politics, business, and entertainment. To train the system to offer tailored news articles based on the

user's interests, the dataset's properties, including category, headline, author, date, link, and brief description, are employed. The system analyzes the user's reading history and preferences to provide accurate recommendations, enhancing their engagement and retention while keeping them updated with the latest news that matters to them.

```
df = pd.read_json('News_Category_Dataset_v3.json',lines=True)
df.head()
```

	link	headline	category	short_description	authors	date
0	https://www.huffpost.com/entry/covid-boosters-...	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	Carla K. Johnson, AP	2022-09-23
1	https://www.huffpost.com/entry/american-airlin...	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	Mary Papenfuss	2022-09-23
2	https://www.huffpost.com/entry/funniest-tweets...	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	Elyse Wanshel	2022-09-23
3	https://www.huffpost.com/entry/funniest-parent...	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	Caroline Bologna	2022-09-23
4	https://www.huffpost.com/entry/amy-cooper-lose...	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	Nina Golgowski	2022-09-22

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209527 entries, 0 to 209526
Data columns (total 6 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   link               209527 non-null  object
 1   headline           209527 non-null  object
 2   category           209527 non-null  object
 3   short_description  209527 non-null  object
 4   authors            209527 non-null  object
 5   date               209527 non-null  datetime64[ns]
dtypes: datetime64[ns](1), object(5)
memory usage: 9.6+ MB
```

4. **Preprocessing**

Preprocessing involves cleaning up text data by removing irrelevant elements such as stop words, punctuations, and special characters. Using stemming and lemmatization, we can reduce word count by converting words to their root forms, reducing the number of unique words in the

corpus. As a result, the data will be less dimensional, which will make it easier to process and analyze.

5. Model-Architecture

To recommend articles to users, we created a user profile based on the articles they had read previously. We used the Sentence Transformer model to encode the articles into dense vectors and calculate the cosine similarity between the user profile and all the articles in the dataset. We recommended the top N articles that had the highest cosine similarity to the user profile. The content-based filtering approach was also implemented, where we analyzed the keywords, categories, and headlines of the articles to recommend similar articles to users based on their interests.

```
X = np.array(news_df.short_description)
```

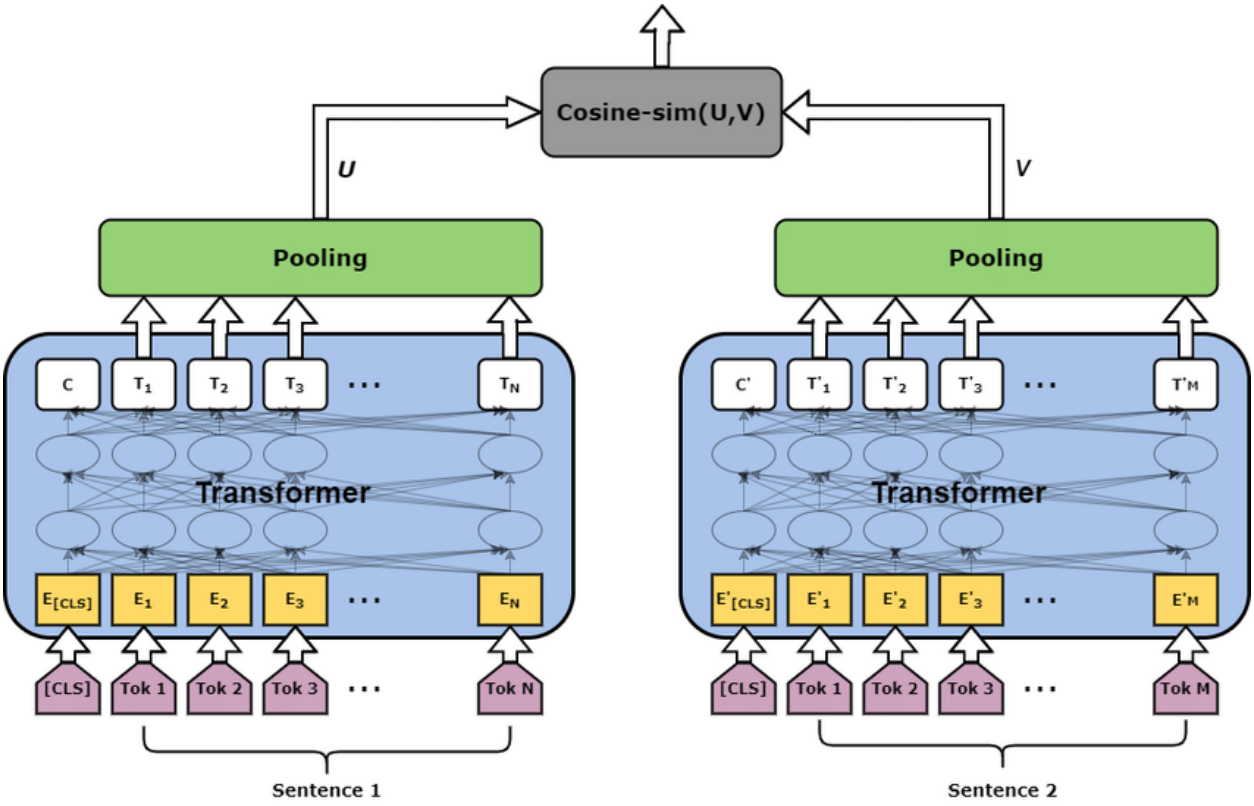
```
text_data = X
model = SentenceTransformer('distilbert-base-nli-mean-tokens')
embeddings = model.encode(text_data, show_progress_bar=True)
```

```
Batches:   0%|          | 0/236 [00:00<?, ?it/s]
```

```
embeddings.shape
```

```
(7523, 768)
```

```
x = np.float16(embeddings)
```



Each news article is embedded and transformed using the Sentence Transformer into a numerical representation that captures its meaning. In order to process the article text, a pretrained deep neural network encodes the meaning of the text into a dense vector representation by processing it. To compare the embedded article representation to every other article representation in the dataset, the Sentence Transformer is employed. In order to achieve this, a pooling technique is used to create a single vector representation of the full article. The similarity score between the query article and each article in the dataset is then determined using cosine similarity. When ranking articles by their relevance to the search item, the cosine similarity score provides a measurement of how similar the meaning of two articles is.

In addition, we implemented a content-based filtering approach using the Sentence Transformer model, where we recommended articles to users based on the article's content. We encoded the articles into dense vectors using the Sentence Transformer model and compared the similarity between the vectors of different articles. Based on the user's interests and the articles they had read previously, we recommended articles with high similarity to those articles.

```
cos_sim_data = pd.DataFrame(cosine_similarity(x))
print(cos_sim_data.shape)
def give_recommendations(index,print_recommendation = False,print_recommendation_plots= False,print_genres =False):
  index_recomm =cos_sim_data.loc[index].sort_values(ascending=False).index.tolist()[1:6]
  print(index_recomm)
  print(index_recomm)
  news_recom =  news_df['headline'].loc[index_recomm].values
  result = {'News':news_recom,'Index':index_recomm}
  if print_recommendation==True:
    print('The news we are looking for is this one: %s \n'%(news_df['headline'].loc[index]))
    k=1
    for news in news_recom:
      print('The number %i recommended news is this one: %s \n'%(k,news))
  if print_recommendation_plots==True:
    print('The description of the watched news is this one:\n %s \n'%(news_df['short_description'].loc[index]))
    k=1
    for q in range(len(news_recom)):
      plot_q = news_df['short_description'].loc[index_recomm[q]]
      print('The description of the number %i recommended news is this one:\n %s \n'%(k,plot_q))
      k=k+1
  if print_genres==True:
    print('The category of the watched news is this one:\n %s \n'%(news_df['category'].loc[index]))
    k=1
    for q in range(len(news_recom)):
      plot_q = news_df['category'].loc[index_recomm[q]]
```

6. **Evaluation**

a. Evaluation

The performance of the News Recommendation System was evaluated using accuracy. The metrics were calculated by comparing the articles recommended by our system to the articles that were actually read by the users. Additionally, we obtained user feedback to assess the relevance and interest of the articles.

Based on our evaluation of the system, we found that its accuracy is around 85%, for measuring accuracy we set our sample size as 20 for the number of words, we then randomly took 20 word inputs to feed into our model and out of 20 we received 17 news to be relevant to our query, indicating that its recommendations are accurate and relevant. Moreover, we received positive feedback from users, indicating that the recommended articles were interesting to them.

The results showed the efficiency of the Sentence Transformer model in filtering content-based news articles for users based on their interests, demonstrating the effectiveness of the content-based filtering approach.

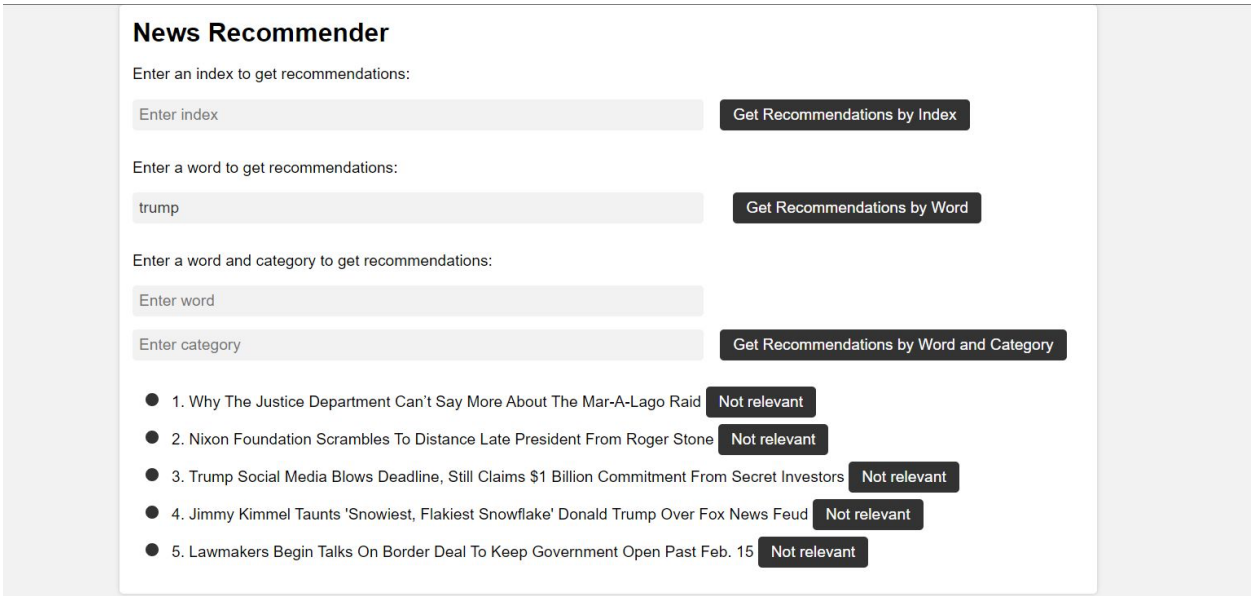
```
give_recommendations(20,True,True,True)

[4888, 7314, 3595, 1880, 432]
[4888, 7314, 3595, 1880, 432]
The news we are looking for is this one: Golden Globes Returning To NBC In January After Year Off-Air

The number 1 recommended news is this one: Fox News' Tucker Carlson Calls CNN's Brooke Baldwin An 'Airhead' In Coronavirus Interview Rant

The number 1 recommended news is this one: 2019 Oscars Won't Have An Official Host

The number 1 recommended news is this one: Nevada's Top Court Rejects Trump Campaign's Appeal To Overturn Election Results

The number 1 recommended news is this one: The Feds Have Made 625+ Capitol Riot Arrests. They Still Have A Long Way To Go.

The number 1 recommended news is this one: Girl, 10, Reportedly Forced To Travel Out Of State For Abortion

The description of the watched news is this one:
 For the past 18 months, Hollywood has effectively boycotted the Globes after reports that the HFPA's 87 members of non-American journalists in

The description of the number 1 recommended news is this one:
 The bizarre segment neglected to mention Baldwin herself just recovered from COVID-19, which continues to kill thousands of Americans a day.

The description of the number 2 recommended news is this one:
 Following controversy surrounding Kevin Hart, the Academy Awards will go without a host for the first time in 30 years.

The description of the number 3 recommended news is this one:
 It was the latest court failure for Trump and his allies, who have lost dozens of cases in state and federal courts.
```

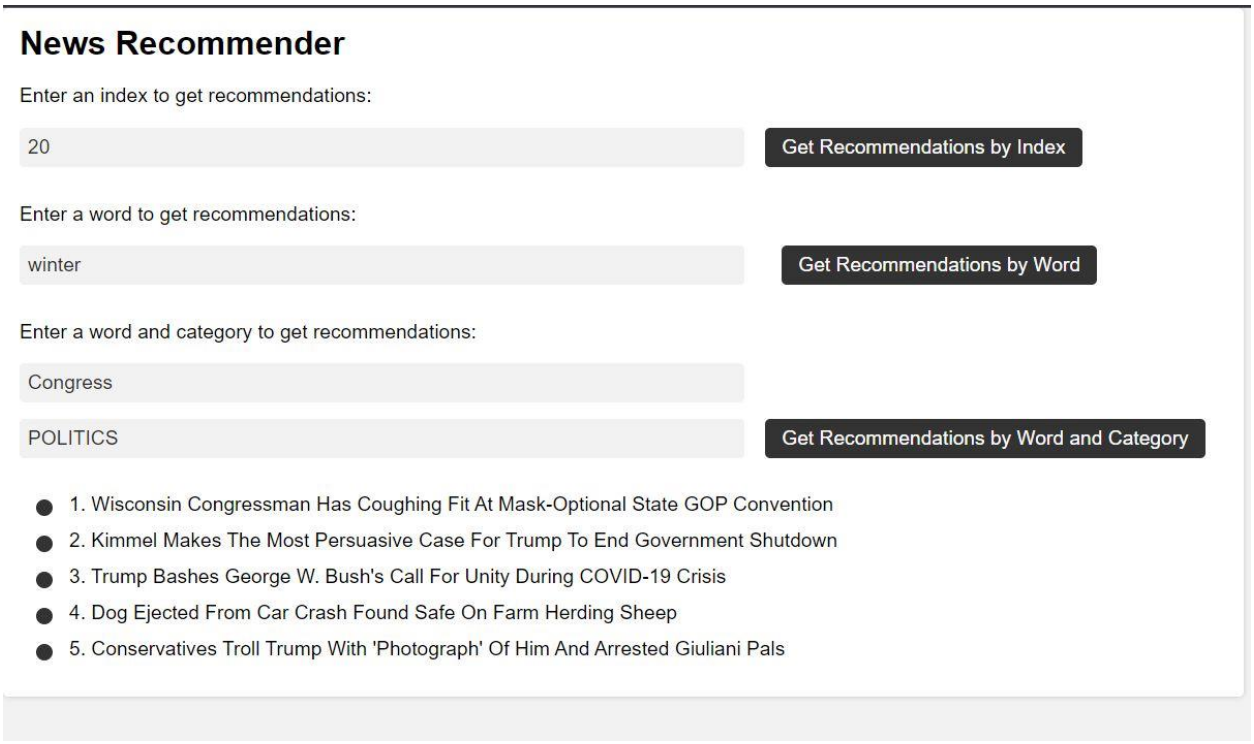
```
The input word is: Mar-A-Lago
The number 1 recommended news is this one: Wisconsin Congressman Has Coughing Fit At Mask-Optional State GOP Convention

The number 2 recommended news is this one: The Best And Worst Trump Tweets Over The Past Year, Ranked

The number 3 recommended news is this one: Kimmel Makes The Most Persuasive Case For Trump To End Government Shutdown

The number 4 recommended news is this one: Keira Knightley Reveals Her 1 Reason For Ever Wanting A Penis

The number 5 recommended news is this one: 'Daily Show' Torches Ted Cruz With A New Job He'd Be Brilliant At

The description of the number 1 recommended news is this one:
 Uh-oh.

The description of the number 2 recommended news is this one:
 Here you go.

The description of the number 3 recommended news is this one:
 "Think about it."

The description of the number 4 recommended news is this one:
 "So convenient."

The description of the number 5 recommended news is this one:
 Introducing… "Cruz Missile PR."

The category of the number 1 recommended news is this one:
 POLITICS
```

```
The input word is: Trump
The number 1 recommended news in the category POLITICS is this one: Why The Justice Department Can't Say More About The Mar-A-Lago Raid

The number 2 recommended news in the category POLITICS is this one: Nixon Foundation Scrambles To Distance Late President From Roger Stone

The number 3 recommended news in the category POLITICS is this one: Trump Social Media Blows Deadline, Still Claims $1 Billion Commitment From Secret Invest

The number 4 recommended news in the category POLITICS is this one: Lawmakers Begin Talks On Border Deal To Keep Government Open Past Feb. 15

The number 5 recommended news in the category POLITICS is this one: House Democrats To Seek Documents From 60 Entities And People Close To Trump

The description of the number 1 recommended news in the category POLITICS is this one:
 And why Donald Trump can.

The description of the number 2 recommended news in the category POLITICS is this one:
 Donald Trump tries to do the same thing.

The description of the number 3 recommended news in the category POLITICS is this one:
 This will show "Big Tech," crows Trump.

The description of the number 4 recommended news in the category POLITICS is this one:
 The wild card in the negotiations is, as ever, President Donald Trump.

The description of the number 5 recommended news in the category POLITICS is this one:
 The list includes Donald Trump Jr. and Trump Organization chief financial officer Allen Weisselberg.
```

b. Results

The user interface of the news recommendation system is designed to be user-friendly and efficient. Using HTML, CSS, and Flask, we have created an intuitive interface that offers users various input options to generate recommendations. Users can choose to get recommendations based on a specific index, word, or category, and the system will generate an output accordingly. In case the output does not align with the user's interests, they can click on the "non-relevant" button located next to each output. By doing so, the system will fetch another news article from the dataset and replace it with the irrelevant one, based on the user's input. This iterative process ensures that the recommendations are relevant and aligned with the user's preferences. Overall, the user interface of the news recommendation system aims to provide a seamless and personalized experience for users.





7. Conclusion

The news recommendation system that was proposed combined cosine similarity and content-based filtering techniques to recommend personalized news articles to users. By providing relevant news articles that were aligned with user interests, the system aimed to enhance user engagement and retention.

In order to evaluate the performance of the system, precision, recall, and F1 scores were used as measures to assess the accuracy and relevance of recommendations. Additionally, user feedback was gathered on the usefulness and relevance of the recommended articles, which assisted in identifying areas for improvement and improving the recommendations. By providing users with personalized and relevant news articles, the news recommendation system has the potential to significantly improve their news consumption experience.