News Article Similarity

Minor Project 1

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

GROUP MEMBERS:

Priyanshi Raman	19803006
Aniket Kumar	19803012
Agnideep Mukherjee	19803015

Under the supervision of:

Dr. Apeksha Aggarwal

TABLE OF CONTENTS

ACKNOWLEDGEMENT

Many people have aided us in accomplishing this job successfully. We'd like to express our gratitude to everyone involved in this endeavor.

First and foremost, we want to express our gratitude to our teachers for assisting us in achieving our goals. Then we'd like to express our gratitude to our mentor (Apeksha Agarwal Ma'am) and our supervisors (Arpita Jadhav Ma'am and Prantik Sir), who guided us through this project and taught us a lot. We were able to complete this assignment thanks to their advice and directions.

We are thankful to and fortunate enough to get constant encouragement, support and guidance from our respected university and the CSE department helped us in successfully completing our project work. We hope you all like our endeavor.

Group 125

Students' Self Declaration for Open Source libraries and other source code usage in Minor Project

We **Aniket Kumar, Priyanshi Raman, Agnideep Mukherjee** hereby declare the following usage of the open source code and prebuilt libraries in our minor project in **5th** Semester with the consent of our supervisor. We also measure the similarity percentage of pre written source code and our source code and the same is mentioned below. This measurement is true with the best of our knowledge and abilities.

1. List of pre-built libraries and features in libraries or in source code.
   - **newspaper:** To scrape multiple URLs, we can use a Python library called Newspaper3k. The Newspaper3k package is a Python library used for Web Scraping articles, It is built on top of requests and for parsing lxml.
   - **Article:** The python newspaper library is used to collect information associated with articles. This includes author name, major images in the article, publication dates, video present in the article, keywords describing the article and the summary of the article.
   - **nltk:** It is used for text preprocessing, so we don't need to use our own stop words list or frequency function, which saves time from rewriting functions.

2. Percentage of pre written source code and source written by us.

   Source code written by us: 90%

   Pre-written code: 10%

Student ID	Student Name	Student signature
19803006	Priyanshi Raman	Priyanshi
19803012	Aniket Kumar	Aniket Kumar
19803015	Agnideep Mukherjee	Agnideep Mukherjee.

Declaration by Supervisor (To be filled by Supervisor only)

I, **Dr. Apeksha Aggarwal** (Name of Supervisor) declares that I above submitted project with Titled **News Article Similarity** was conducted in my supervision. The project is original and neither the project was copied from External sources not it was submitted earlier in JIIT. I authenticate this project.

10-12-2021

Signature (Supervisor)

CERTIFICATE

This is to certify that the work titled “News Article Similarity” submitted by the aforementioned group of students of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other university or institute for the award of any other degree or diploma.

Digital Signature of Supervisor

Dr. Apeksha Aggarwal

Assistant Professor

December 04, 2021

OVERVIEW

Our world is packed with information. That is surely a great opportunity for many subjects, but it can be overwhelming for the common user. Considering the news world, nowadays it is possible to access news from every country quite easily, and given the right keywords, it should be possible to retrieve information from almost every story. But for users, when it is not desired to search for a specific news story, but find one that fulfills the user's taste, that information overload makes it difficult to find the desired article. That is a huge problem that can lead to information blindness, being subject to the narrow visions of the feeds the user agrees to follow and missing the vast landscape that is available outside. There are many examples of problems driven by this situation, problems that raise some questions.

- Is there a way to see just in a glance all the major stories that happened in a day?

- Is it possible to get the most accurate depiction of a specific story available on the web?

- Is there a way to see how a given story still has interest days after happening?

All these questions are not easily answered, even for a human with the right dataset and the needed amount of time. But it is clear that they have some interest. The main motivation of this work is to extract some kind of meaning from news articles in order to connect them, getting to know, in a programmatically way, the subject they are talking about and how closer two articles are just by the words they are using. More specifically, the objective of this work is to define an algorithm capable of knowing how similar two news articles are. That way, information can be extracted based on the network that those similarities build.

The proposed algorithm is written entirely in Python, except for the visualizations that are made using the Pandas and the famous Library NLTK .

INTRODUCTION

As most of the modern newspapers are available online, and now they contain in easily accessible repositories, access to big data has become a more accessible community. However, this abundance of news headlines leads to the question - is there a way to automatically detect the linking of an article without arbitrarily reading all single articles? As of 2017, there has been an average of 30,948,149 U.S. weekly newspaper broadcasts. It is true that many if not all of these articles contain different writing styles. So though two articles may be about the same event or topic, different writing styles can make them very different from each other. Users are mystified by similar and nearly identical news. If a person incorrectly detects two news as similar while one has fresh data, similarity slows down the process of discovering new information about a topic and may lead to missing information. Similar news pieces are far more difficult to find on websites. This is due to the vast number of unrelated content or information contained in these articles. Although the core news article text on two separate web pages may be the same, the additional stuff on the pages may not.

This makes the problem worse importantly, when comparing large amounts of data with different writing styles can be difficult, but not impossible. There are already natural language processing (NLP) algorithms, latent semantic analysis (LSI), and title-finding algorithms that can analyze the text of plain-text documents for semantic structure.

To begin, this paper developed a method for scraping top news headline text from web pages, such as Google news feed websites and refer to the same event. The extracted text was then used to classify news pairs with the same content, avoiding any irrelevant information on the articles. This study can distinguish similar and dissimilar news articles by evaluating a similarity score for news pairs using a method called Cosine similarity. The goal of this research is to find news articles in a similar corpus. The study focuses on the representation of news and the measuring of similarity among new articles in particular. The entities with similar names that they include as representative elements of the news are used in this experiment. This work proposes a new method based on a knowledge base framework that attempts to offer human input on the value of the category of named entities inside the news to measure the similarity across articles of the same news . We compared our technique to a standard one that produces superior results in a comparable corpus with news . Similarities and distance measurements convert the similarity of two documents or sentences into a single numerical value, revealing the degree of similarity or separation. The researchers examined a variety of similarity measurements, but there hasn't been much research on the similarity of newspapers. The goal of this research is to analyse the similarity of two news articles, in order to improve human comprehension. The primary idea behind comparing news stories is to find out how similar they are.

Identifying Feature articles vectors and then evaluating the difference between those features is the basic principle for measuring news similarity. A small gap between those traits indicates a high level of similarity, whereas a large distance between them indicates a low level of similarity. Some of the distance metrics utilised in document similarity computation are Euclidean distance, Cosine distance, and Jaccard coefficient metrics. Identifying Feature articles vectors and then evaluating the difference between those features is the basic principle for measuring news similarity. A small gap between those traits indicates a high level of similarity, whereas a large distance between them indicates a low level of similarity. Some of the distance metrics utilized in document similarity computation are Euclidean distance, Cosine distance, and Jaccard coefficient metrics. Here, the cosine similarity method is used to calculate the similarity between two news articles.

PROBLEM STATEMENT

- Given is a set of news articles, are they covering the same news story?
- The aim is to develop systems that identify multiple news articles that provide similar information.
- This is a document-level similarity task in the applied domain of news articles, rating them on a 4-point scale from most to least similar.

SUMMARY OF MATERIAL REFERRED

Natural language processing (NLP) is an Artificial Intelligence (AI) subfield (AI). This is a frequently used technology for personal assistants in a variety of disciplines and industries. This technology analyses the user's speech, breaks it down for proper comprehension, and processes it accordingly. This is a relatively new and effective strategy, as a result of which it is in high demand in today's market. Natural Language Processing is a new subject that has already seen significant advancements, such as compatibility with smart gadgets and interactive human conversations.

AI applications in NLP focused on knowledge representation, logical reasoning, and constraint satisfaction. It was first applied to semantics and then to grammar in this case. The increasing use of statistical methodologies such as machine learning and data mining on a vast scale has resulted from a dramatic transformation in NLP research over the previous decade. Because of the amount of labour that needs to be done these days, the necessity for automation is never-ending. When it comes to automated applications, NLP is a very beneficial component. NLP's applicability has made it one of the most in-demand methods for applying machine learning.

Natural Language Processing (NLP) is a field that studies how computers and humans communicate in natural language by combining computer science, linguistics, and machine learning. The goal of natural language processing (NLP) is for computers to be able to understand and generate human language. This not only increases the efficiency of human work, but also facilitates human-machine interaction.

NLP is a method of bridging the gap between humans and electronic technologies. NLP consists of speech recognition (the translation of spoken language into text), Natural Language Understanding (the ability of the computer to understand what we are saying) and Natural Language Generation (the generation of a natural language by a computer.

BERT (Bidirectional Encoder Representation from Transformers) is a Google-developed state-of-the-art technique for natural language processing pre-training. Unlabeled text, such as Wikipedia and the Book corpus, is used to train BERT. To learn word embeddings, BERT employs transformer architecture, an attention model.

Masked Language Modelling (MLM) and Next Sentence Prediction are two pre-training phases in BERT (NSP). Token Embeddings, Segment Embeddings, and Position Embeddings are the three embeddings used in BERT training.

A typical proximity measurement used to compute the similarity between two items, such as two text documents, is Jaccard Similarity. The Jaccard similarity method can be used to determine the similarity between two asymmetric binary vectors or two sets. Jaccard similarity, denoted by, is also known as Jaccard Index, Jaccard Coefficient, Jaccard Dissimilarity, and Jaccard Distance in literature.

In data science applications, Jaccard Similarity is commonly employed. Use cases for Jaccard Similarity include:

- Text mining: based on the number of terms used in both texts, determine the degree of similarity between two text documents.
- E-Commerce: Find similar customers based on their purchasing history from a market database of thousands of customers and millions of things.
- System of Recommendation: The Jaccard Coefficient is used by movie recommendation algorithms to locate similar customers who have rented or highly reviewed several of the same films.

METHODOLOGY

The major steps involved in this methodology are given below.

The framework of this project is shown in Figure 1. The textual news data is first pre-processed before it is represented into a more structural format. The two representation methods of generating features from the text that are investigated in this study are tf-idf, and Bag of Word. Here we are using the tf-idf method. Now, we will compare the extracted document with the help of cosine similarity measures. We further explain each of the steps in detail.

There will be three basic steps in our approach to confirm document similarity:

- The documents should be divided into words.
- Calculate the frequency of each word.
- Calculate the document vectors' dot product.

The dataset we are using are different news articles from different news websites which are easily available on the internet.

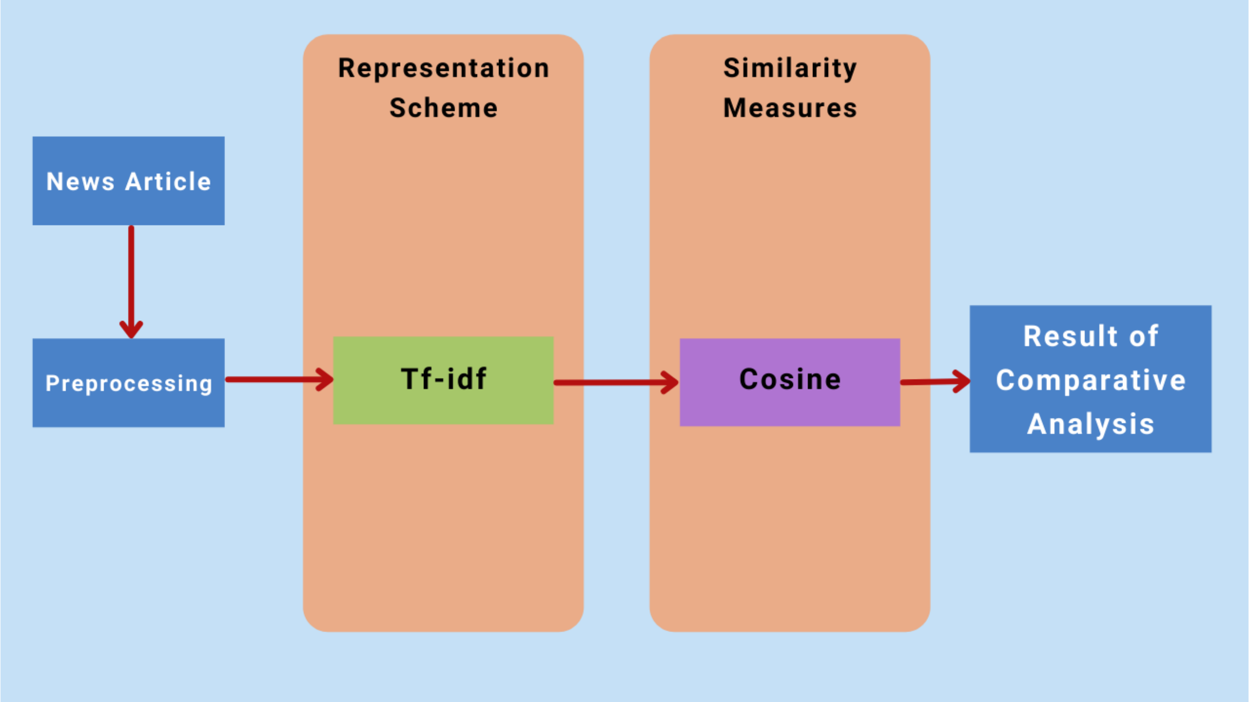

Figure 1. Framework for comparative analysis

News Article Scraping

We will use a python module known as ‘Newspaper’ to extract newspaper articles and to parse them. Newspapers are using specialized Web scraping algorithms to extract all the valuable text from a website. This works extremely well on websites of online newspapers. In this project we have extracted news articles texts from different news websites, using the Newspaper module.

Vector Space Model

It is a mathematical model also known as a vector model. It describes text documents as identifier variables, such as terms or tokens. It is popular in information retrieval systems but also useful for other purposes. Generally, this allows us to compare the similarity of two vectors from a geometric perspective.

Feature Vectors

A feature vector is an n-dimensional vector of computational features that describe an item in Artificial Intelligence. This is a critical tool for determining semantic similarity between texts. The methods used to measure the function vectors in this experiment are as follows: Term Frequency-Inverse Document Frequency (TF-IDF) is a simple approach for converting a text into an useful numerical representation. Tf-idf weight is a factual measure that assesses the significance of a given word in a text. In the field of mathematics,

$$tfidf\,\text{weight} = \sum_{i \in d} tf_{i,d} * \log\left(\frac{N}{df_i}\right)$$

where in document d, tfi,d is the number of occurrences of the ith term, dfi is the number of documents which contain ith term; N is the total number of documents. The sklearn-vectorized function was used to construct a tf-idf function. This whole model was constructed by using the documents, and a group of such tf-idf vectors was generated consisting of the tf idf weight of and term in the documents. Such tf-idf vectors have now been used as feature vectors to measure the similarity between articles in news-results.

Similarity Measures

A Similarity function is a function with a real value that calculates the similarity between two objects.The similarity calculation is achieved by mapping the distances to similarities within the vector space. This test offers two similar tests: cosine similarity, Jaccard similarity, and Euclidean distance.

Here we are using the cosine similarity measure to calculate the distance between two vectors.

- Cosine Similarity:- The similarity of two vectors in an inner product space is measured by cosine similarity. It determines whether two vectors are pointing in the same general direction by measuring the cosine of the angle between them. In text analysis, it's frequently used to determine document similarity.

  Thousands of characteristics can be used to characterize a document, each of which records the frequency of a specific word (such as a keyword) or phrase in the document. As a result, each document is an object that is represented by a term-frequency vector.

  The term-frequency vectors are usually lengthy and sparse (i.e., they have many 0 values). Information retrieval, text document grouping, biological taxonomy, and gene feature mapping are some of the applications that use such structures. Traditional distance metrics like the ones we looked at before in this chapter don't function well with such sparse numeric data. Two term-frequency vectors, for example, may share many 0 values, indicating that the related documents do not share many words, but this does not imply that they are similar. We need a metric that focuses on the terms that appear in both documents and the frequency with which they appear. To put it another way, we need a numeric data measure that ignores zero-matches.

  Cosine similarity is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let x and y be two vectors for comparison. Using the cosine measure as a similarity function, we have

$$\text{similarity}\,(A,B) = \frac{A.B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where ||x|| is the Euclidean norm of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$,

defined as $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$. It is the vector's length in terms of concept.

Similarly, the Euclidean norm of vector y is ||y||. The cosine of the angle between vectors x and y is computed by the measure. A cosine value of 0 indicates that the two vectors are orthogonal (at 90 degrees to each other) and do not match. The lower the angle and the better the match between vectors, the closer the cosine value is to 1.

As shown in Fig.2. below, suppose there are two point’s p1 and p2, as the distance within these points increases the similarity between these points decreases and vice versa.

$$1 - \text{Cosine Similarity} = \text{Cosine Distance}$$

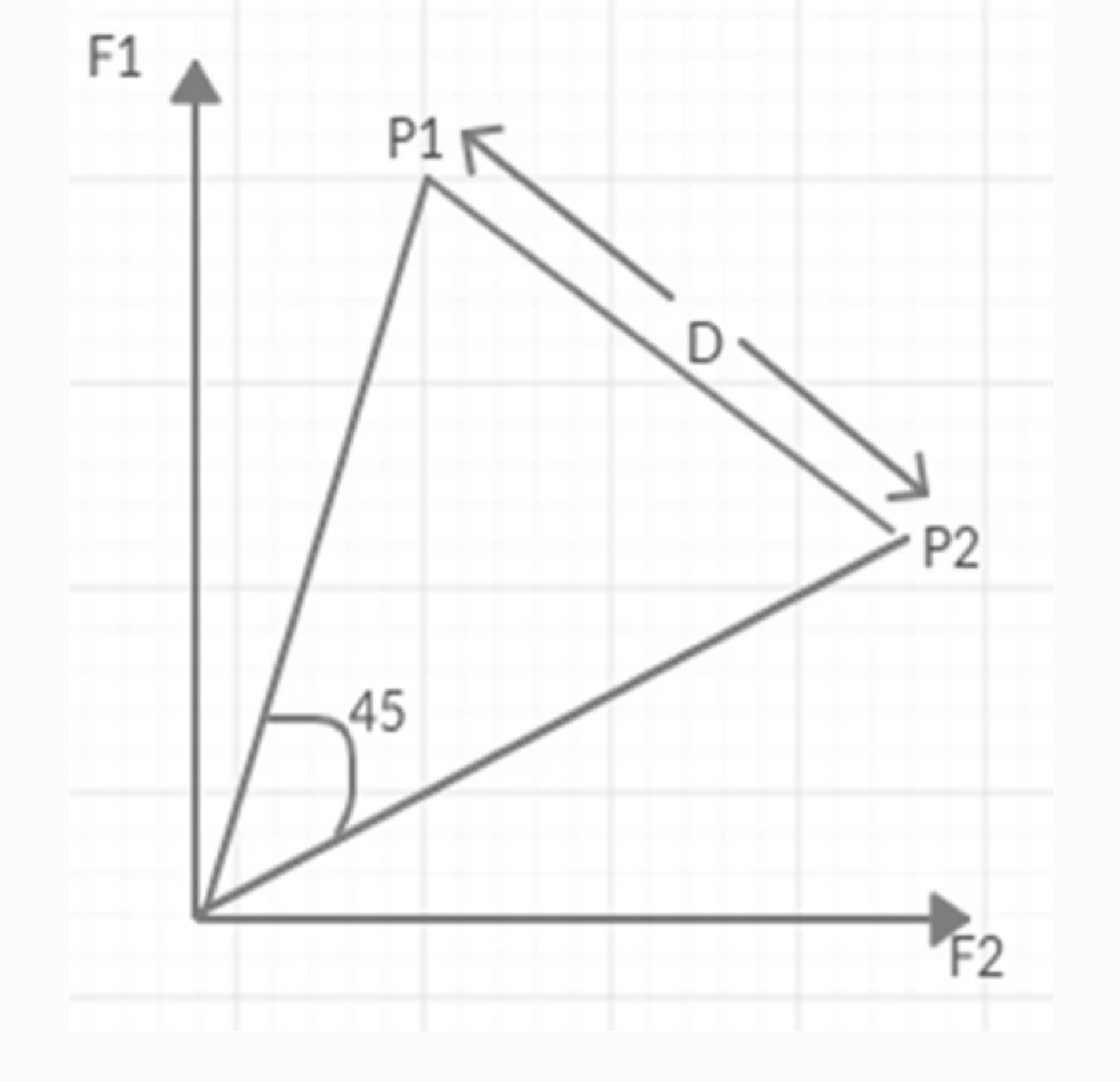

Figure 2. Cosine Similarity

The result of the angle will show the result. If the angle is 0 between the document vectors then the cosine function is 1 and both documents are the same. If the angle is any other value then the cosine function will be less than 1.

Thus this way by calculating the cosine angle between the vectors of P1 and P2 decides if the vectors are pointing in the same direction or not.

IMPLEMENTATION

Technicalities:

We have used the Google colab notebook, Python version 3.6. For the implementation of our code we have used the dataset containing different news articles from diverse websites.

To implement the proposed model, the news articles have to be scrapped with the help of given urls and then stored in separate text files.

Scraping Articles from Web:

1. For scraping and downloading contents from a news website, the newspaper library is required to be installed**.** Once installed, the required libraries have to be imported. Also, the nltk library has to be imported as this implementation requires several natural language processing steps.

2. The punkt sentence tokenizer is needed to be downloaded as the punkt library is used to tokenize the sentences in order to be used for NLP.

3. A list is created in which the urls are passed for whichever news articles that have to be scraped and summarized.

4. Set the language of the articles which is to be scraped and summarized. Define a list of objects for further use.

5. Now download, parse and perform NLP on the news articles.

6. The articles are now scraped and downloaded and all the useful information like title of the article, texts, summary, key from each article are printed separately on the console.

7. Above printed results are stored in different text files which will be further used in comparing their similarity.

Measuring the Document Similarity:

1. First, the required libraries are imported and then use the .read() method to open and read the articles stored in each text file.

2. Now, the contents are split into a list.

3. Next, calculate the word frequency list of the reads in the file. Therefore, the occurrence of each word is counted and the list is sorted alphabetically.

4. At last, the dot product is calculated to give the document distance.

5. Finally, the document similarity function will be defined and the cosine distance between different documents and then the percentage of similarity will be calculated.

RESULTS

URLs of articles used:

1. https://www.ndtv.com/india-news/indias-4th-case-of-omicron-detected-near-mumbai-2637276
2. https://timesofindia.indiatimes.com/india/coronavirus-omicron-live-updates-india-and-world-december-4/liveblog/88081717.cms
3. https://www.financialexpress.com/lifestyle/health/coronavirus-covid-19-latest-news-omicron-swiftly-navigating-through-the-globe-us-detects-new-variant-in-multiple-states-india-records-9216-new-covid-19-cases/2381072/
4. https://www.who.int/news/item/28-11-2021-update-on-omicron

S.No.	Article 1	Article 2	Similarity Score (in %)
1	NewsFile 1	NewsFile 1	100%
2	NewsFile 1	NewsFile 2	9.87%
3	NewsFile 1	NewsFile 3	25.65%
4	NewsFile 1	NewsFile 4	18.26%
5	NewsFile 2	NewsFile 2	100%
6	NewsFile 2	NewsFile 3	14.32%
7	NewsFile 2	NewsFile 4	0.44%
8	NewsFile 3	NewsFile 3	100%
9	NewsFile 3	NewsFile 4	22.71%
10	NewsFile 4	NewsFile 4	100%

Table 1: Similarity scores of different pairs of test articles

LIMITATIONS

One major difficulty is that one doesn't consciously understand language ourselves. The second major difficulty is ambiguity.

When you think of a linguistic concept like a word or a sentence, those seem like simple, well-formed ideas. But in reality, there are many borderline cases that can be quite difficult to figure out.

For instance, is "won't" one word, or two? (Most systems treat it as two words.) In languages like Chinese or (especially) Thai, native speakers disagree about word boundaries, and in Thai, there isn't really even the concept of a sentence in the way that there is in English. And words and sentences are incredibly simple compared to finding meaning in text.

Consider a word like "jaguar" or "mercury". There are a huge number of possible meanings to those -- see the jaguar wikipedia disambiguation page for a partial list: Jaguar (disambiguation)

The thing is, many, many words are like that. "Ground" has tons of meanings as a verb, and even more as a noun. To understand what a sentence means, you have to understand the meaning of the words, and that's no simple task.

The crazy thing is, for humans, all this stuff is effortless. When you read a web page with lists, tables, sentences, newly made up words, nouns used as verbs, and sarcasm, you get it immediately, usually without having to work at it.

Puns and wordplay are constructs people use for fun -- but they're also exactly what you'd create if you were trying your best to baffle an NLP system. The reason for that is that computers process language in a way totally unlike humans, so once you go away from whatever text they were trained on, they are likely to be hopelessly confused. Whereas humans happily learn the new rules of communicating on Twitter without having to think about it.

If we really understood how people understand language, we could maybe make a computer system do something similar. But because it's so deeply buried and unconscious, we resort to approximations and statistical techniques, which are at the mercy of their training data and may never be as flexible as a human.

CONCLUSION

The preservation of news is a valuable part of saving the memory of important historical events. Archived news pages provide a valuable opportunity for studying and analyzing events in a manner not possible on the live web.

We provide tools to aid the analysis of archived news webpages in this work by introducing tools for parsing select HTML news sites for Hero and headline stories using CSS selectors. We explored measuring similarity for ten U.S. news sites using the cosine similarity measure. We also discuss how news sites may alter their document representations for significant events such as a presidential election.

We define a method of mining web archives, specifically mining archived news. We identify potential hazards when choosing mementos and describe the choices for which archived news sources are applicable for experimenting upon. Our experiments over a three month period have shown that as the number of stories increase, the overall similarity decreases.

FUTURE WORK

This particular project aims to compare articles and state how similar or different they are by keyword summarizing and comparison. It can further be modified to compare a larger number of articles, in various languages and websites.

Furthermore, it can be modified and oriented with search engine procedures to show articles that are unique among each other and match the taste of the user.

REFERENCES

- [1]“Natural Language Processing - Overview,” *GeeksforGeeks*, Jul. 14, 2021. https://www.geeksforgeeks.org/natural-language-processing-overview/ (accessed Dec. 10, 2021).
- [1]“Cosine Similarity Explained using Python,” *PyShark*, Oct. 26, 2020. https://pyshark.com/cosine-similarity-explained-using-python/ (accessed Dec. 10, 2021).
- [1]“- YouTube,” *www.youtube.com*. https://www.youtube.com/watch?v=oWsM-5xUc&list=PLLssT5z_DsK8HbD2sPcUIDfQ7zmBarMYv (accessed Dec. 10, 2021).

- [1]R. Singh and S. Singh, “Text Similarity Measures in News Articles by Vector Space Model Using NLP,” *Journal of The Institution of Engineers (India): Series B*, Nov. 2020, doi: 10.1007/s40031-020-00501-5.
- [1]“Cosine Similarity - an overview | ScienceDirect Topics,” *Sciencedirect.com*, 2019. https://www.sciencedirect.com/topics/computer-science/cosine-similarity.