




CH5019 GROUP PROJECT REPORT

Prepared by:

Sarath (BS20B033)
Hridith (CH20B049)
Gowtham (CH20B031)
Anuroop (CH20B015)



Problem Statement

An instructor wants to grade answers to descriptive questions automatically. The instructor has a template best answer that she/he has developed and wants to use the same to grade descriptive answers of students. The comparison results between the template and the student provided answer could be categorical (right vs wrong) or continuous (75% similarity and so on). You are expected to develop an AI algorithm that can do this comparison automatically. Please develop an algorithm through any appropriate concept and demonstrate the results on ten test cases. You can use any training approach and test it on any 10 test cases that you feel are appropriate. Your algorithm should take two paragraphs, one correct answer and one student provided answer and return a result. Since this is a rather open-ended problem, the solutions will be largely evaluated based on the creativity associated with problem formulation, solution approach (including feature engineering), and the achieved results. Please remember that there is no single correct method. Please use existing AI/ML libraries as much as possible.

Introduction to NLTK

The Natural Language Toolkit (NLTK) is a Python programming environment for working with human language data in statistical natural language processing (NLP). It includes tokenization, parsing, categorization, stemming, tagging, and semantic reasoning text processing packages.

It also comes with a recipe book and a book that describes the ideas behind the core language processing tasks that NLTK provides, as well as graphical examples and sample data sets.

The NLTK logo is a large, stylized, light blue Python logo (two interlocking snakes) centered on the page. Below it, the letters "NLTK" are written in a light blue, sans-serif font.

NLTK

Link to our code-

<https://colab.research.google.com/drive/17Dw0gd00F6Bm-baUwCJtggoXQfLRa-LU?usp=sharing>

Document Pre-Processing

Many aspects in news reports are not essential (or irrelevant) for text analysis exercises like identifying similarity. As a result, they are pre-processed by changing their words to lower case and removing 'stopwords,' such as 'the,' 'should,' and so on.

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
from nltk.corpus import stopwords

stopwords_en=stopwords.words("english")

def preprocessing(raw):
    wordlist=nltk.word_tokenize(raw)
    text=[w.lower() for w in wordlist if w not in stopwords_en]
    return text

file1=open('t1.txt','r',encoding="utf8")
doc1=preprocessing(file1.read())

file2=open('t2.txt','r',encoding="utf8")
doc2=preprocessing(file2.read())
```

Vectorization using TF-IDF method

Consider each text to be a vector. When compared to one another, each text has some common and rare terms. To account for all eventualities, a word set made up of terms from both sources is created. Words can be vectorised (that is, turned to vectors) in a variety of ways (array of numbers).

TF is a document-specific format. It's a method of determining the value of words (or "terms") in a document based on their frequency of appearance. If a term appears several times in a document, it is considered essential and receives a high score. Although it is simple to calculate, it is confusing.

```

from nltk.probability import FreqDist
word_set=set(doc2).union(set(doc1))

freq_doc1=FreqDist(doc1)
doc1_length=len(doc1)
doc1_tf_dict=dict.fromkeys(word_set,0)
for word in doc1:
    doc1_tf_dict[word]=freq_doc1[word]/doc1_length

freq_doc2=FreqDist(doc2)
doc2_length=len(doc2)
doc2_tf_dict=dict.fromkeys(word_set,0)
for word in doc2:
    doc2_tf_dict[word]=freq_doc2[word]/doc2_length

```

IDF refers to the entire collection. It's a metric for counting how many times a term appears in several texts. If a term appears in a large number of papers, it is not a unique identifier and hence receives a lower score.

```

doc12_idf_dict=dict.fromkeys(word_set,0)
doc12_length=2

for word in doc12_idf_dict.keys():
    if word in doc1:
        doc12_idf_dict[word]+=1
    if word in doc2:
        doc12_idf_dict[word]+=1

import math
for word,val in doc12_idf_dict.items():
    doc12_idf_dict[word]=1+math.log(doc12_length/(float(val)))

```

TFIDF of a word = (TF of the word) * (IDF of the word)

```

doc1_tfidf_dict=dict.fromkeys(word_set,0)
for word in doc1:
    doc1_tfidf_dict[word]=(doc1_tf_dict[word])*(doc12_idf_dict[word])

doc2_tfidf_dict=dict.fromkeys(word_set,0)
for word in doc2:
    doc2_tfidf_dict[word]=(doc2_tf_dict[word])*(doc12_idf_dict[word])

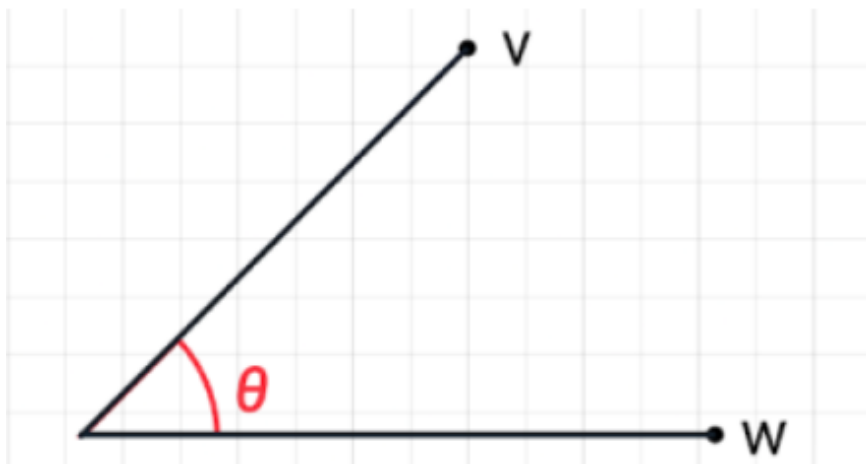
```

Similarity Analysis

Now that we have vectorized the documents, now we have to find out how similar the two vectors are, for that we use cosine similarity analysis.

Cosine Similarity

Cosine Similarity computes the similarity of two vectors as the cosine of the angle between two vectors. It determines whether two vectors are pointing in roughly the same direction. So if the angle between the vectors is 0 degrees, then the cosine similarity is 1.



It is given by-

$$\cos(v, w) = \frac{v \cdot w}{||v|| \times ||w||}$$

Implementation

We implement cosine similarity as follows-

```
v1=list(doc1_tfidf_dict.values())
v2=list(doc2_tfidf_dict.values())
similarity=1-nltk.cluster.cosine_distance(v1,v2)
print("Similarity : {:.2f} %".format(similarity*100))
```

Results on our test cases-

Template-

Here are some of the problems caused by plastic bags. Plastic bags are non-biodegradable. Thus, disposing of plastics is the biggest challenge. They are destroying nature due to their harmful effect. Plastic bags have become the main cause of land pollution today. The plastic bags entering into the water bodies are a major cause of water pollution. Hence we can conclude that these are deteriorating our environment in every possible way. Animals and marine creatures unknowingly consume plastic particles along with their food. Research shows that waste plastic bags have been a major reason for untimely animal deaths. The production of plastic bags releases toxic chemicals. These are the leading cause of serious illness

Student provided answer-

Here are some of the issues that plastic bags cause. Plastic bags do not decompose. As a result, disposing of plastics is the most difficult task. They are destroying nature as a result of their negative impact. Today, plastic bags are the leading cause of land pollution. Plastic bags that end up in aquatic bodies are a major source of pollution. As a result, we might conclude that these are wreaking havoc on our ecosystem in every manner possible. Plastic particles are inadvertently consumed by animals and aquatic species. According to research, waste plastic bags are a major cause of untimely animal fatalities. Toxic substances are released during the manufacture of plastic bags. These are the most common causes of death and serious disease. The dirty environment is a major contributor to a variety of ailments.

Result-

Similarity : 69.61%

Template-

Travelling is an amazing way to learn a lot of things in life. A lot of people around the world travel every year to many places. Moreover, it is important to travel to humans. Some travel to learn more while some travel to take a break from their life. No matter the reason, travelling opens a big door for us to explore the world beyond our imagination and indulge in many things. Therefore, through this Essay on Travel, we will go through everything that makes travelling great.

There are numerous benefits to travelling if we think about it. The first one being, we get to meet new people. When you meet new people, you get the opportunity to make new friends.

Student provided answer-

Traveling is a fantastic method to pick up new skills and discover new things. Every year, a large number of people travel to various locations around the world. In addition, travelling to humans is critical. Some people travel to expand their knowledge, while others travel to unwind. Whatever the cause, travel allows us to discover the world beyond our wildest dreams and partake in a wide range of activities. As a result, we'll go through everything that makes travelling so enjoyable in this Essay on Travel.

If we think about it, there are several advantages to travel. We get to meet new individuals, for starters. It is possible to make new friends when you meet new people.

Result-

Similarity : 60.30%

Template-

Books are referred to as a man's best friend. They are very beneficial for mankind and have helped it evolve. There is a powerhouse of information and knowledge. Books offer us so many things without asking for anything in return. Books leave a deep impact on us and are responsible for uplifting our mood.

There are different genres of books available for book readers. Every day, thousands of books are released in the market ranging from travel books to fictional books. We can pick any book of our interest to expand our knowledge and enjoy the reading experience.

Student provided answer-

A man's best buddy is said to be books. They are extremely valuable to humanity and have aided in its evolution. There is a wealth of information and knowledge available. Books give us so much without expecting anything in return. Books have a profound effect on us and are responsible for elevating our spirits.

For book lovers, there are a variety of genres to choose from. Every day, thousands of books, ranging from travel to fiction, are released on the market. We might choose any book that piques our attention in order to broaden our horizons and enjoy the reading experience.

Result-

Similarity : 56.94%

Template-

A food chain is a linear network of links in a food web starting from producer organisms such as green plants or grass which use radiation from the Sun to make their food via photosynthesis and ending at an apex predator species like lion or killer whales, detritivores like earthworms or woodlice, or decomposer species insuch as fungi or bacteria. A food chain also shows how organisms are related to each other by the food they eat. Each level of a food chain represents a different trophic level.

Student provided answer-

A food chain is a network of animals that begins with producer organisms like plants, trees that use solar energy to make their food via photosynthesis and ends with apex predators like lion, tiger, detritivores such as earthworms , or decomposer species like fungi or bacteria. A food chain also demonstrates linkage between organisms by food. Each trophic level in a food chain is represented by a distinct level of the food chain.

Result-

Similarity : 68.03%

Template-

Social media is a tool that is becoming quite popular these days because of its user-friendly features. Social media platforms like Facebook, Instagram, Twitter and more are giving people a chance to connect with each other across distances. In other words, the whole world is at our fingertips all thanks to social media. The youth is especially one of the most dominant users of social media.

Like how there are always two sides to a coin, the same goes for social media.

When we look at the positive aspect of social media, we find numerous advantages. The most important being a great device for education. All the information one requires is just a click away. Students can educate themselves on various topics using social media. Despite having such unique advantages, social media is considered to be one of the most harmful elements of society. If the use of social media is not monitored, it can lead to grave consequences.

It is harmful because it invades your privacy like never before. The oversharing happening on social media makes children a target for predators and hackers. It also leads to cyberbullying which affects any person significantly.

Student provided answer-

Social media is becoming increasingly popular these days. People can connect with each other across distances through social media platforms such as Facebook, Instagram, Twitter, and others. To put it another way, social media has brought the entire world to our fingertips. The youth, in particular, are among the most active users of social media.

The same way there are two sides to a coin, there are two sides to social networking.

When we consider the positive aspects of social media, we can see that there are numerous benefits. The most crucial being that it is an excellent educational tool. All of the information needed is only a mouse click away. Students can use social media to educate themselves on a variety of topics.

Despite its numerous positives, social media is widely regarded as one of society's most dangerous elements. If social media usage is not closely managed, it can have serious effects.

It is damaging because it intrudes on your privacy in ways you have never experienced before. Children are becoming a target for paedophiles and hackers as a result of their oversharing on social media. It also leads to cyberbullying, which has a huge impact on anyone.

Result-

Similarity : 77.65%

Template-

Postulates of Rutherford's model of an atom: The mass of an atom is concentrated in a small space called the nucleus. Atoms majorly consist of positively charged particles. Negatively charged electrons revolve around atoms in circular paths called orbits at very high speed. An atom is electrically neutral that is it has no net charge.

Student provided answer-

According to Rutherford's model, the mass of an atom is concentrated in a small region termed the nucleus.

Positively charged particles make up the majority of atoms and are orbited by negatively charged electrons.

An atom has no net charge and is electrically neutral.

Result-

Similarity : 46.01%

Template-

Atomic radius is the distance from the atom's nucleus to the outer edge of the electron cloud. In general, atomic radius decreases across a period and increases down a group. Across a period, effective nuclear charge increases as electron shielding remains constant. A higher effective nuclear charge causes greater attractions to the electrons, pulling the electron cloud closer to the nucleus which results in a smaller atomic radius. Down a group, the number of energy levels increases, so there is a greater distance between the nucleus and the outermost orbital. This results in a larger atomic radius.

Student provided answer-

The atomic radius is defined as the distance between the nucleus and the outermost region of electron cloud

The atomic radius reduces as you progress through a period and increases as you progress through a group.

The effective nuclear charge decreases across a period and this attracts electrons more strongly, drawing them closer to the nucleus.

Down a group the number of energy levels increases and they are more separated. This leads to increase in radii down the group.

Result-

Similarity : 47.93%

Template-

Language is often described as the sine-qua-non or the most important and distinguishing characteristic of a culture or civilisation. There has been a the consistent relationship between the level of advancement of a society and the complexity and development of its language. In fact, one may say that civilisation or for that matter the very idea of knowledge is closely intertwined with language.

Scientists interested in the study of the evolution of behaviour of societies point out that there are four distinct features which have made the human organism distinctly superior to the highest evolved sub-human organisms like the chimpanzee. These are, attainment of an erect posture, the growth of the cerebral cortex and its complexity, the prolonged period of socialisation, and finally the acquisition of advanced and complex linguistic capacities and abilities.

Student provided answer-

Language is frequently referred to as a culture's sine qua non, or its most significant and distinctive feature. The complexity and growth of a society's language have always had a continuous relationship. Indeed, one may argue that language is inextricably linked to civilisation, or even the concept of knowledge itself.

Scientists studying the evolution of social behaviour have identified four key characteristics that distinguish the human organism from the most evolved subhuman creatures, such as the chimp. These include the development of an upright posture, the expansion and complexity of the cerebral cortex, a protracted period of socialisation, and finally the acquisition of advanced and complicated linguistic talents and abilities.

Result-

Similarity : 46.33%

Template-

The Indian Entertainment and Media Industry has performed better in the Indian economy and is considered one among fastest growing sectors of India. It is increasing on the base of economic growth and accelerating income levels that our country has been witnessing in the last few years. This is indeed proving useful to the entertainment and media industry in India as this is a rather sensitive one and it prospers faster when it is an expanding economy. An added advantage to the entertainment and media industry in India is from the point of view of demographics where the spending of the consumer is increasing as a result of increase in disposable incomes due to sustained growth in income levels and reduced income tax in the last ten years. The present size of the industry is estimated at US\$ 7 billion in 2004 and is estimated to grow at a CAGR of 14 per cent to US\$ 13bn by the year 2009. The Filmed Entertainment and Television segment rules the industry succeeding to Print, Radio and the Music segments.

Student provided answer-

The Indian entertainment and media industry has outperformed the Indian economy and is one of the country's fastest-growing industries. It is growing since our country has experienced economic expansion and rising income levels in recent years. This is proven to be beneficial to India's entertainment and media industry, as it is a sensitive one that thrives faster in a growing economy. The entertainment and media sector in India has a demographic advantage in that consumer spending is increasing as a consequence of increased disposable incomes as a result of steady growth in income levels and lower income taxes over the previous 10 years. The industry's current size is anticipated to be US\$ 7 billion in 2004, with a CAGR of 14% expected to reach US\$ 13 billion by 2009. The Filmed Entertainment and Television section dominates the industry, with the Print, Radio, and Music divisions trailing behind.

Result-

Similarity : 53.92%

Template-

Global warming refers to the gradual rise in the overall temperature of the atmosphere of the Earth. There are various activities taking place which have been increasing the temperature gradually. Global warming is melting our ice glaciers rapidly. This is extremely harmful to the earth as well as humans. It is quite challenging to control global warming; however, it is not unmanageable. The first step in solving any problem is identifying the cause of the problem. Therefore, we need to first understand the causes of global warming that will help us proceed further in solving it.

Global warming has become a grave problem which needs undivided attention. It is not happening because of a single cause but several causes. These causes are both natural as well as manmade. The natural causes include the release of greenhouse gases which are not able to escape from earth, causing the temperature to increase.

Further, volcanic eruptions are also responsible for global warming. That is to say, these eruptions release tons of carbon dioxide which contributes to global warming. Similarly, methane is also one big issue responsible for global warming.

Student provided answer-

The progressive rise in the overall temperature of the Earth's atmosphere is referred to as global warming. There are a variety of things going on that are steadily raising the temperature. Our ice glaciers are fast disappearing due to global warming. This is incredibly detrimental to both the environment and people. Controlling global warming is difficult; yet, it is not impossible. Identifying the problem's cause is the first step in solving any problem. As a result, we must first comprehend the causes of global warming before moving on with solutions.

Global warming has evolved into a serious issue that requires our full attention. It is occurring due to a combination of factors rather than a single cause. These factors are both natural and man-made. The emission of greenhouse gases that are unable to escape from the planet, causing the temperature to rise, is one of the natural causes.

Furthermore, volcanic eruptions contribute to global warming. To put it another way, these eruptions spew masses of carbon dioxide into the atmosphere, contributing to global warming. Methane, on the other hand, is also a major contributor to global warming.

Result-

Similarity : 72.22%

Individual Contributions-

- Document preprocessing- Sarath and Anuroop
- Vectorization- Hridith
- Similarity analysis- Gowtham
- Test case generation- Sarath & Anuroop