AUTOMATIC DESCRIPTIVE GRADING REPORT, TEAM 69:

Project group and Project members:

The project was done by team 69 which have

- Landa Hema EE19B036
- Anvith Pabba EE19B070
- Kandimalla Pravallika EE19B030

Aim:

This project aims for automatic descriptive grading which compares the student's answer to the original or reference template. We have done the evaluation based on the "**most important words**" the documents and their related corpus contain.

Procedure or what we have done:

- Importing data
- Cleaning the data
- Creating word2vector instances
- Finding the percentage of similarity between the answer and reference template

Importing Data:

- We have used google collab for the project and had our original template and test cases as txt files in the drive. The used data can be accessed here.
- We have written a function that accesses the drive and reads all the text files and stores each test case in a string format in a list known as 'text'
- So the first element in the list ''text'' is the reference template.

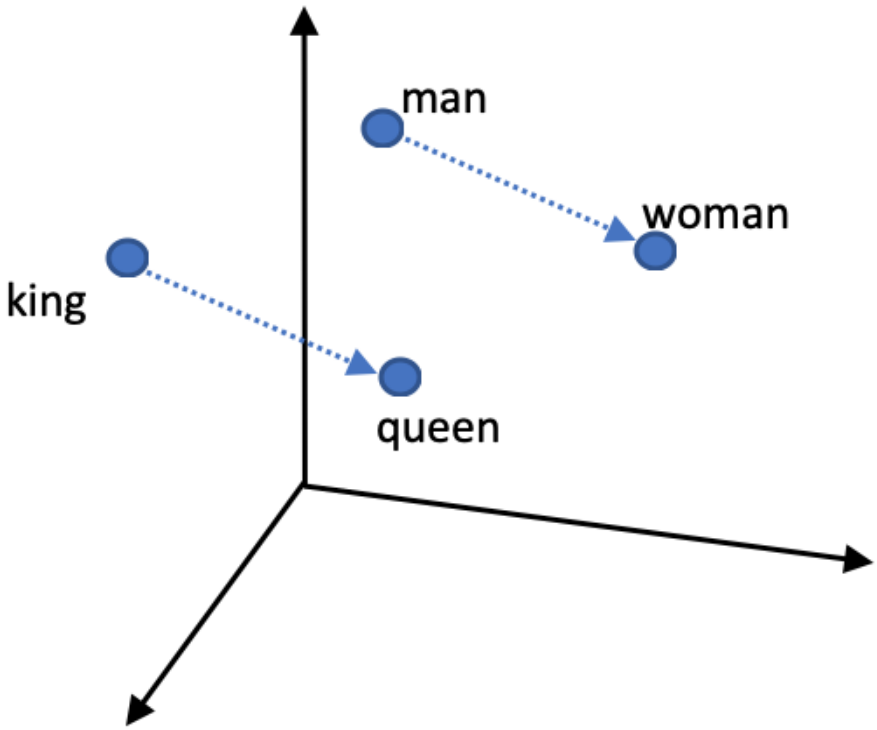
Cleaning the data:

- We have created a data frame with the testcases in one column and cleaned paragraphs in another column.
- We have imported the''stopwords'' library from 'nltk' and used inbuilt word function to get all the stop words in english.

- We have parsed the original sentence and removed all the stopwords if they are in the above list, removed special characters and converted everything to lower case letters and made a new sentence with the remaining important or special words.
- We have kept these in ‘’cleaned_text’’ column.

Creating Word2Vector instances:

Embeddings: Embeddings are the vector representations of text where word or sentences with similar meaning or context have similar representations.



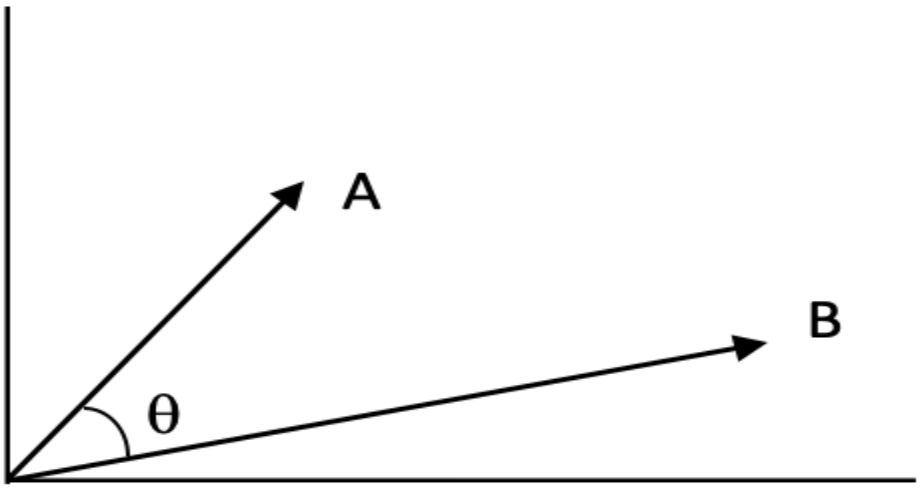
- We have used the basic idea of inverse text frequency popularly known as “**tf-idf**”**(term frequency and inverse document frequency)** algorithm.
- It is a common method to evaluate how important a single word is to a corpus.
- Let us take we are talking about movie reviews, here the term “movie” can be repeated several times in each review but the term doesn’t carry much weightage.
- We have considered the following three steps to determine the weightage of a particular term in the answer script or document.

1. How often can we find term **t** in document **D**? As long documents may lead to distortions of this question (as terms may occur more often), we consider the **maximum frequency** of a term t' in order to **normalize #t in D**.
2. In how many documents of the overall corpus do we find the term **t**? (Terms that do occur in all considered documents are less important if the occur many times in those considered documents)
3. In the last step, the weight of the term t in the given document D is calculated as the product of the two above calculations
4. We will use the ft-id case as described before. For this approach, we will consider **all available documents** instead of just counting words within as single document

- So by considering the top n terms which have highest scores as calculated by using above algorithm, we can calculate their corresponding vectors for each example.
- For this we have used the inbuilt library **"TfidfVectorizer"** from **"sklearn.feature_extraction.text"**

Finding the similarities:

- For this we have used **cosine similarity algorithm** which calculates the dot product between two normalized vectors to get the angle between them.
- This angle refers to how close the two vectors are.



- cos theta of 1 represents they are perfectly similar, 0 represents they are entirely different.
- We have calculated this using inbuilt module "**cosine similarity**" and converted to percentages.

We can see the above implementation in this snippet.

```
#converting text to data frame
text_df=pd.DataFrame(text,columns=['text'])

# removing special characters and stop words from the text
stopwords_list=stopwords.words('english')
text_df['text_cleaned']=text_df.text.apply(lambda x: " ".join(re.sub(r'[^a-zA-Z]',' ',w).lower() fo

#forming the vectors
tfidfvectoriser=TfidfVectorizer()
tfidfvectoriser.fit(text_df.text_cleaned)
tfidf_vectors=tfidfvectoriser.transform(text_df.text_cleaned)

#finding the similarity using cosine angle
cosine_similarities=np.dot(tfidf_vectors[0],tfidf_vectors.T).toarray()

#printing the similarities of given test cases with the original template
print(cosine_similarities[0][1:])
```

```
[1.         0.91246258 0.91715989 0.87389989 0.82892847 0.88022668
 0.77152106 0.35874342 0.2010791  0.0085588 ]
```

Note: Kindly see the submitted code for full lines of code, which we could not fit in the picture.

Test Cases:

We have taken the **reference template** as follows:

"*We all know that health is wealth. With its intricate network of bones, muscles, and organs, a well-functioning human body is much like an orchestrated symphony. To keep this orchestra playing well, we need physical exercise. It may take the form of sports, yoga, or even regular walking. It is well-known that people who engage in physical exercise stay happier and live longer. Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need physical exercise. Exercise also sharpens our intellect. It keeps a balance between our body and mind. With the help of regular physical training, we will stay healthier, happier, and more alert.*"

1. Firstly we have taken exact paragraph as TC1 for which we got 100% similarity.

2. The second and third test cases, we have taken by changing just one line. Here we can see a change in the first line

“We are all aware that one's health is one's most valuable asset. With its intricate network of bones, muscles, and organs, a well-functioning human body is much like an orchestrated symphony. To keep this orchestra playing well, we need physical exercise. It may take the form of sports, yoga, or even regular walking. It is well-known that people who engage in physical exercise stay happier and live longer. Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need physical exercise. Exercise also sharpens our intellect. It keeps a balance between our body and mind. With the help of regular physical training, we will stay healthier, happier, and more alert. ”

For this we got similarity of 91.24 percent.

3. Here we have changed the line ‘It is well-known that people who engage in physical exercise stay happier and live longer’.

“We all know that health is wealth. With its intricate network of bones, muscles, and organs, a well-functioning human body is much like an orchestrated symphony. To keep this orchestra playing well, we need physical exercise. It may take the form of sports, yoga, or even regular walking. Physical activity is well known for making people happier and extending their lives. Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need physical exercise. Exercise also sharpens our intellect. It keeps a balance between our body and mind. With the help of regular physical training, we will stay healthier, happier, and more alert. ”

For this we got similarity of 91.72 percent.

4. We have taken test case 4 by removing one sentence completely and without changing the other lines.

“We all know that health is wealth. With its intricate network of bones, muscles, and organs, a well-functioning human body is much like an orchestrated symphony. To keep this orchestra playing well, we need physical exercise. Physical activity is well known for making people happier and extending their lives. Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need

physical exercise. Exercise also sharpens our intellect. It keeps a balance between our body and mind. With the help of regular physical training, we will stay healthier, happier, and more alert. ”

For this we got a similarity of 87.39 percent.

5. We have taken test cases 5,6,7 by changing two lines or removing a sentence and changing a line.

“We all understand that one's health is the most valuable asset one may own. With its intricate network of bones, muscles, and organs, a well-functioning human body is much like an orchestrated symphony. To keep in condition, this symphony necessitates physical exertion. It may take the form of sports, yoga, or even regular walking. It is well-known that people who engage in physical exercise stay happier and live longer. Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need physical exercise. Exercise also sharpens our intellect. It keeps a balance between our body and mind. With the help of regular physical training, we will stay healthier, happier, and more alert. ”

For this we got a similarity of 82.89 percent.

6.

“We all know that health is waste. With its intricate network of bends and cracks. To keep this orchestra playing well, we need physical exercise. It may take the form of sports, yoga, or even regular walking. It is well-known that people who engage in physical exercise stay happier and live longer. Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need physical exercise. Exercise also sharpens our intellect. It keeps a balance between our body and mind. With the help of regular physical training, we will stay healthier, happier, and more alert. ”

For this we got a similarity of 88 percent.

7.

“We all understand that one's health is the most valuable asset one may own. With its intricate network of bones, muscles, and organs, a well-functioning human body is much like an orchestrated symphony. To keep in condition, this symphony necessitates physical exertion. It may take the form of sports, yoga, or even regular walking. It is well-known that people who engage in physical exercise stay happier and live longer.

Our society is turning towards more and more technical sophistication and automation. The machine has replaced our physical labor. To compensate for this change in lifestyle we need physical exercise. With the help of regular physical training, we will stay healthier, happier, and more alert."

For this we got a similarity of 77.15 percent.

8. We have taken test cases 8 and 9 as paragraphs with similar meaning but not the exact words.

"We are all aware that one's health is the most valuable thing one may own. With its complex network of bones, muscles, and organs, a well-functioning human body is analogous to an orchestrated symphony. This symphony requires physical exertion to stay in good shape. Sports, yoga, and even daily walking can all be beneficial. People who indulge in physical activity are believed to be happier and live longer. Our society is becoming more and more automated and technologically advanced. Machines have taken the place of physical labour. Physical activity is required to compensate for this shift in lifestyle. Physical activity sharpens our intellect just as much as it sharpens our bodies. It makes sure that our bodies and minds are in sync. If we eat healthier, happier, and more alert foods, we will stay healthier, happier, and more alert."

For this we got a similarity of 35.87 percent.

9.

"We all know that good health is a waste of time. With its complex web of twists and fissures. Emotions have an important part as well. Our society is becoming increasingly technologically oriented. Our hard, physical effort has been substituted by the machine. Physical activity is required to compensate for this shift in lifestyle. Exercise improves our intelligence and endurance as well. It maintains a delicate and delicate equilibrium between our body and mind. We will stay healthy, sweet, and dark with the help of regular fun."

For this we got a similarity of 20.11 percent.

10. This we have taken a completely different paragraph which doesn't match with the reference template at all.

"The animal I like most is the dog. The dog is a pet animal. It is a four-footed animal. It has two bright eyes. It has two ears, sharp teeth and a small tail. Dogs are of many kinds. Some dogs have fur on their bodies. The dogs are of different colours. They are of different sizes. The dog is a very useful and a faithful animal. The dog can swim in the water. It is found everywhere in the world. It loves its master very much. It guards

the house from thieves with care. The policemen use dogs in tracing out thieves and criminals."

For this we got a similarity of 0.86 percent.

We can see the code snippet of printing all these percentage similarities.

```
#printing the similarities of given test cases with the original template
for i in range(1,11):
  print('The answer {} is {:.4f}% matching'.format(i,cosine_similarities[0][i]*100))
```

```
The answer 1 is 100.0000% matching
The answer 2 is 91.2463% matching
The answer 3 is 91.7160% matching
The answer 4 is 87.3900% matching
The answer 5 is 82.8928% matching
The answer 6 is 88.0227% matching
The answer 7 is 77.1521% matching
The answer 8 is 35.8743% matching
The answer 9 is 20.1079% matching
The answer 10 is 0.8559% matching
```

Thus we have implemented the algorithm successfully.

Contributions:

- Hema Landa: have written the code.
- Hema and Anvith: have written the report
- Pravallika: have written the test cases.