NLP Lab-5

Ali Hassan

March 2023

1 Introduction

We are in the hype of ChatGPT and many people are scared of the future. This includes faculties that are worried about their students doing assignments using this tool and losing the learning experience. However, ChatGPT is simply a language model, trained on the enormous data it has seen. We are far away from having a perfect machine.

2 Problem Statement

Choose 10 questions from Stack Exchange from a topic that you are familiar with and issue them to ChatGPT and save the answers. After that we need to do the manual analysis and automatic analysis.

3 Questions selected from stack exchange

3.1 What type of statistical test should I use for this specific example?

Tags – statistics , regression analysis

3.2 COVID-19 data analysis with Python from Github CSV

Tags – python , pandas

3.3 Get overall covid-19 cases using JavaScript

Tags – javascript

3.4 Folium Heatmap With Time for COVID 19

Tags – python , heatmap

3.5 ”cure for the COVID-19” vs. ”cure for COVID-19”

Tags – articles

3.6 JSON syntax error Covid-19 tracker website

Tags – javascript , htmlTags – javascript , html

3.7 COVID-19 Data visualization with R

Tags – data visualization

3.8 Cannot scrape from Johns Hopkins covid-19 site with scrapy

Tags – scrapy

3.9 Retrieving Covid-19 pandemic statistics per country from DBpedia

Tags – sparql

3.10 Missing Authentication Token for C3.ai COVID-19 Data Lake

Tags – python

4 ChatGPT Results

4.1 What type of statistical test should I use for this specific example?

Tags – statistics, hypothesis testing, statistical test, data analysis

4.2 COVID-19 data analysis with Python from Github CSV

Tags – covid 19 , data analysis

4.3 Get overall covid-19 cases using JavaScript

Tags – javascript , covid19

4.4 Folium Heatmap With Time for COVID 19

Tags – covid19 , folium

4.5 ”cure for the COVID-19” vs. ”cure for COVID-19”

Tags – articles , covid19

4.6 JSON syntax error Covid-19 tracker website

Tags – javascript , json

4.7 COVID-19 Data visualization with R

Tags – data visualization , R

4.8 Cannot scrape from Johns Hopkins covid-19 site with scrapy

Tags – scrapy , python

4.9 Retrieving Covid-19 pandemic statistics per country from DBpedia

Tags – sparql , pandemic

4.10 Missing Authentication Token for C3.ai COVID-19 Data Lake

Tags – API , token

5 Analysis

5.1 Manual Analysis

From checking manually to the correct answers and the answers given by the ChatGPT we can conclude that the we are getting almost the same result from the ChatGPT and the tags correctly explains the question. So we can say that ChatGPT did a really good job in predicting the tags for the questions and which are correct. As we can see that for the first question the original tags are statistics , regression analysis and the tags given by ChatGPT are statistics, hypothesis testing, statistical test, data analysis which highly correlate to the question and to the original tags.

5.2 Automatic Analysis

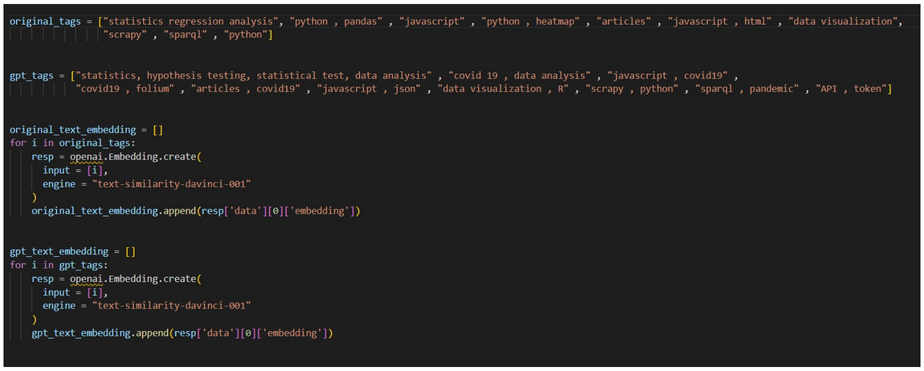
Using the unigram and bigram overlap we have calculated the cosine similarity matrix.

Tags	0	1	2	3	4	5	6	7	8	9
0	0.8									
1		0.79								
2			0.86							
3				0.77						
4					0.85					
5						0.91				
6							0.91			
7								0.91		
8									0.87	
9										0.81

So we have calculated the cosine similarity between the ith tag of original tag and the ith tag of the gpt created tag. So this depicts the cosine similarity between the original tags and the gpt tags.

6 Word Embeddings using Openai

Source Code



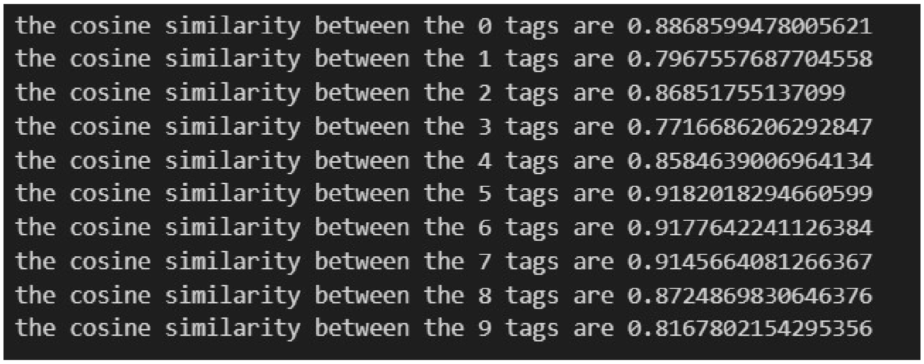
```
original_tags = ["statistics regression analysis", "python , pandas" , "javascript" , "python , heatmap" , "articles" , "javascript , html" , "data visualization",
                 "scrapy" , "sparql" , "python"]


gpt_tags = ["statistics, hypothesis testing, statistical test, data analysis" , "covid 19 , data analysis" , "javascript , covid19" ,
            "covid19 , folium" , "articles , covid19" , "javascript , json" , "data visualization , R" , "scrapy , python" , "sparql , pandemic" , "API , token"]


original_text_embedding = []
for i in original_tags:
    resp = openai.Embedding.create(
      input = [i],
      engine = "text-similarity-davinci-001"
    )
    original_text_embedding.append(resp['data'][0]['embedding'])


gpt_text_embedding = []
for i in gpt_tags:
    resp = openai.Embedding.create(
      input = [i],
      engine = "text-similarity-davinci-001"
    )
    gpt_text_embedding.append(resp['data'][0]['embedding'])
```

Output



```
the cosine similarity between the 0 tags are 0.8868599478005621
the cosine similarity between the 1 tags are 0.7967557687704558
the cosine similarity between the 2 tags are 0.86851755137099
the cosine similarity between the 3 tags are 0.7716686206292847
the cosine similarity between the 4 tags are 0.8584639006964134
the cosine similarity between the 5 tags are 0.9182018294660599
the cosine similarity between the 6 tags are 0.9177642241126384
the cosine similarity between the 7 tags are 0.9145664081266367
the cosine similarity between the 8 tags are 0.8724869830646376
the cosine similarity between the 9 tags are 0.8167802154295356
```

7 Conclusion

After going through the whole process we can clearly see that chatgpt provided a really high similarity tags to the original ones. Chatgpt performed very well as we can see that the cosine similarity score is almost above 80 percent in all the cases which is amazing. After doing the manual analysis also we can conclude that chatgpt performed really well as the tags were matching the question and provided a good extract of the question.

Original Tags	Chatgpt Tags
statistics regression analysis	statistics, hypothesis testi
python , pandas	covid 19 , data analysis
javascript	javascript , covid19
python , heatmap	covid19 , folium
articles	articles , covid19
javascript , html	javascript , json
data visualization	data visualization , R
scrapy	scrapy , python
sparql	sparql , pandemic
python	API , token