Sl.No	Matriculation Nr.	Name	Email address
1	6868664	Akshit Bhatia	abhatia@mail.uni-paderborn.de
2	6896703	Harshith Srinivas	harshith@campus.uni-paderborn.de
3	6852329	Suraj Manjunatha	surajm@mail.uni-paderborn.de
4	6866035	Vinaykumar Budanurmath	budanurm@mail.uni-paderborn.de

Table of contents:

1. Goal:

Given an essay generate a argument prompt.

2. Running instructions:

Environment:

Dependency	Version
Python - 3.7.3	3.7.3
TextBlob	0.15.3
Tensorflow	2.2.0
Tensorflow-estimator	2.2.0
Tensorflow-hub	0.8.0
tensorboard-plugin-wit	1.6.0
Tensorboard	2.2.2

Instructions:

The .zip file consists of two folders. The contents of the folder are described below,

Code folder: This folder contains the python notebook file “assignment4.ipynb” which has the code to execute. This file requires as input, the files “essay_prompt_corpus.json” and “train-test-split.csv” for its execution. These input files are placed in the data folder.

Data folder: The contents of the data folder are tabulated below,

File name	Description
essay_prompt_corpus.json	This file holds the train and test corpus data.
train-test-split.csv	Contains the split of essays marked train and test
test_essay.json train_essay.json	These files include the train and test corpus data after splitting, and are input to the .ipynb file
predictions.json	This file is generated in the data folder after the code in python notebook file is executed

3. Approach:

Research questions:

1. Which of the sentences in the essay contain the highest similarity with the prompt?
    a. How to extract this sentence from the essay?
2. How to rank the sentences in the essay and derive a prompt?

Approach explained:

After reading through the essays, we recognize the following:

- Few of the sentences in the essay have high semantic similarity with the prompt. Some of these sentences include – the major claim and the first sentence of the essay.

We decided to extract these sentences to generate the possible prompt. We named these sentences as "prompt indicator sentences".

Features used to extract prompt indicator sentences:

As mentioned above, we observed that the prompt indicator sentences have high semantic correlation with major claims and the first sentence of the essay. Extracting the first sentence of the essay was a trivial task, but to extract major claim we came up with following approach.

Major claim extraction:

Majority of the major claims in the essay corpus preceded with phrases such as "in conclusion", "prefer", "I think", "I believe", "in my opinion", "to conclude" etc., (Please note that this is not a strict rule)

On extracting the count of these type of phrases, we got the following result:

```
considering all : 4
in the end : 1
i advocate : 2
after analyzing : 2
in a nutshell : 5
i favor : 3
personally : 21
all the above : 1
i support : 2
hence : 6
above reasons : 4
to summarize : 1
to conclude : 20
consequently : 3
in my opinion : 37
agree : 73
it seems to me : 4
to sum up : 23
in conclusion : 99
i would conclude : 4
therefore : 21
in summary : 5
i firmly believe : 6
my view : 11
i believe : 45
i agree : 26
i prefer : 11
i completely agree : 7
i strongly prefer : 1
all in all : 8
from my experience : 1
i think : 36
i suppose : 1
my point of view : 24
however : 51
based on the reasons : 4
although : 39
accordingly : 2
as far as : 13
to me : 9
thus : 11
to sum up : 23
```

We used many of the above phrases to extract prompt indicator sentences. Once we extract these sentences, we clean up these sentences by removing the preceding phrases listed above.

For example, for the essay id 373, the prompt indicator sentences are:

Personally, I think it is no evidence about the reduction of crime rates due to the death penalty because of many reasons

Hence, death penalty neither controls the violent in society nor creates a violent culture.

To conclude, capital punishment is a form of legalized revenge, it is an easy way for serious crimes, and nobody has rights to take others life; thus, it neither demines crimes of violence nor be essential to control violence in society

please see below the screen shot of indicator sentences after removal of “claim indicator” phrases for the essay id 373

```
Out[10]: [',  it is no evidence about the reduction of crime rates due to the death penalty because of many reason
         s.',
          ', death penalty neither controls the violent in society nor creates a violent culture.',
          '"Capital punishment or the death penalty is a legal process whereby a person is put to death by the sta
         te as a punishment for a crime."',
          ', capital punishment is a form of legalized revenge, it is an easy way for serious crimes, and nobody h
         as rights to take others life; , it neither demines crimes of violence nor be essential to control violen
         ce in society.']
```

Each essay yields two to four prompt indicator sentences as shown above. So, we used text ranking mechanism to rank these sentences and get one top ranked sentence.

Text ranking is done by first finding the top ten most commonly occurring words in the tokenized "prompt indicator sentence list" and ranking all the sentences in the list based on occurrence of these common tokens. The highest ranked sentence is then selected. We have named this as the “top prompt indicator sentence”.

Below image shows a Dataframe output of prompt indicator sentences before ranking.

	Essay_id	text	promptIndicatorSentenceList	Actual_prompt
0	373	["Capital punishment or the death penalty is a...	[, it is no evidence about the reduction of c...	Capital punishment; 51% countries have polishe...
1	61	[Computer-a device which has given a whole new...	[Computer-a device which has given a whole new...	Computers - use, future prospects and over-dep...
2	180	[During our life, it is inevitable that we may...	[during our life, it is inevitable that we may...	Why are groups or organizations important to p...
3	211	[Students have become more and more stressed d...	[however, that it is not a good idea because ...	Non academic subjects should be removed from s...
4	229	[There is an argument regrading weather lettin...	[that a mistake can not brake friendship if i...	Friendship is more important than mistake by a...

Below image shows the dataframe output after ranking the indicator sentences. The “prompt” column of a data frame is the possible prompt of the essay

Top ranked indicator sentence ⇩

	id	promptIndicatorSentenceList	prompt	Actual_prompt
0	373	[, it is no evidence about the reduction of c...	, it is no evidence about the reduction of cr...	Capital punishment; 51% countries have polishe...
1	61	[Computer-a device which has given a whole new...	Computer-a device which has given a whole new ...	Computers - use, future prospects and over-dep...
2	180	[during our life, it is inevitable that we may...	during our life, it is inevitable that we may ...	Why are groups or organizations important to p...
3	211	[however, that it is not a good idea because ...	, school education should not only focus on th...	Non academic subjects should be removed from s...
4	229	[that a mistake can not brake friendship if i...	however, that friendship is more important th...	Friendship is more important than mistake by a...

Some of the sentences that are ranked, though have highest similarity with the prompt, but we believe that they cannot be directly termed as “prompt”. So, we need a mechanism to generate a prompt.

Prompt generation:

To generate a prompt, we decided to write a prompt generator module. We used the prompts from the training data to train the module.

Prompt generator module:

The base idea of prompt generation is to use the prompts from the train data and generate prompts of the test data. In the context of a prompt, we believe that “a noun” in a prompt and “three words” post that noun are atomic in nature.

For example, consider the following prompts:

Way to reduce the amount of traffic?

People are now able to overcome long distances in a short amount of time

Large amount of violence in television programs

Spending equal amount of money on libraries and sports?

Amount of control on media information

If the noun extracted is “*amount*”, then we get the following phrases

“amount of traffic”, “amount of time”, “amount of violence”, “amount of money”, “amount of control”

Using the train data, we create a memory of “noun” to “prompt phrase”

Steps of prompt memory generation:

Step 1:

1. Collect all the prompts in train data and store it in one superset text

2. Collect all the nouns from the prompt superset.

3. For all the nouns in the prompt superset, collect phrases that follow post these nouns

The atomic phrases for the noun “amount” would be the first two to three words after the noun. Below is the screenshot of the output from the code.

	noun	prompt
0	amount	of traffic?
1	amount	of time
2	amount	of violence in television programs
3	amount	of money on libraries and sports?
4	amount	of control on media information

This memory is used in the prompt generation process of the test data – ie., when a noun is found in the recognized “top prompt indicator sentence”, we get all the possible prompt phrases of that noun and try to fit the noun’s prompt phrase in the “top prompt indicator sentence”. We fit the noun’s prompt phrase in the indicator sentence only if the semantic similarity of the “prompt phrase” and the “top prompt indicator sentence” is above a threshold.

Concrete example from test data:

Essay id: 211

Current possible prompt: *, school education should not only focus on the academic development of a student as it is much more crucial to teach them how to be independent and live a good life.*

Recognized Noun: *school*

The possible prompt phrases from memory for the noun school:

	noun	prompt
515	school	make music lessons compulsory
516	school	or get a job?
517	school	
518	school	for students
519	school	day
520	school	subjects/focus on one subject
521	school	students should be taught how to manage money
522	school	in person
524	school	children should study art and music
525	school	for parents who have children studying in pri...

Selected prompt phrase from prompt phrase memory which is above similarity threshold is:

students should be taught how

New prompt after replacing noun with generated prompt:

, school students should be taught how education should not only focus on the academic development of a student as it is much more crucial to teach them how to be independent and live a good life.

Recognized Noun: *education*

The possible prompt phrases from memory for the noun education:

	noun	prompt
132	education	should be available only to good students or not
133	education	is needed for success
134	education	and prevention issues
135	education	plays an important role in the socioeconomic ...
138	education	
139	education	needs
140	education	and health care or not?
141	education	is a good idea
142	education	in some aspects
144	education	restriction
145	education	and trade
146	education	be free for everyone interested?
148	education	should be only available for good students

Generated prompt above similarity threshold:

should be only available for

New prompt after replacing noun with generated prompt:

, school students should be taught how education should be only available for should not only focus on the academic development of a student as it is much more crucial to teach them how to be independent and live a good life.

4. Other approaches tried but not considered:

Approach 1:

Recognize the top 10 sentences of an essay using the text ranking mechanism. Use these top sentences summarize the content using summarizer library of genism.

Advantages:

1. Gensim summarizer was coming up with meaningful sentences.

Disadvantages:

1. Biggest disadvantage – We believe that summarizing approach is wrong for the “prompt generation”, because a summary is quite different than a prompt. A summary is a meaningful and sometimes nondebatable gist of an essay. Whereas a prompt is a debatable sentence.

2. Very low score during evaluation.

Approach 2:

We tried to group essays with similar contents and give a possible prompt. To this end, we created embeddings of essays. Then we group the essays which are similar above a threshold value. After grouping the essays, we created a memory of “essay to prompt”. With this memory, when we get a new essay, we compare the essay with the existing essays and list all the possible prompts. Finally choose one prompt which is having highest similarity with the new essay.

Disadvantages:

1. Very low score during evaluation.

5. Evaluation:

We have observed that the prompt generated with this approach may not lead to meaningful sentences. But the generated prompts tend to be as close to the claim of the essay. So, we conclude that it is a naïve attempt to generate possible prompt phrases given a noun.

Evaluation score:

During evaluation we have observed the following strange behavior: Every time we run our solution, and generate a new predictions file, and then run the evaluation python code, the evaluation score was different on different runs, though we had not changed solution code between any two evaluations.

Some of the scores generated between two runs

{'rouge-l': {'f': 0.17713941660161786, 'p': 0.12752130497301187, 'r': 0.34283903596403587}}

{'rouge-l': {'f': 0.172792892106078, 'p': 0.12426878340708361, 'r': 0.3351944236319234}}

{'rouge-l': {'f': 0.163223667718758, 'p': 0.11705541504124684, 'r': 0.3178092740592739}}

{'rouge-l': {'f': 0.19184530118316967, 'p': 0.13896472634490492, 'r': 0.3679485445110443}}

6. Conclusion:

The approach that we have implemented may or may not generate meaningful sentences for this given corpus. But we believe that our approach will perform well in the following scenario:

Consider a given a list of prompts and consider that, essays are written for some of the prompts out of the given prompt list. When we feed only the essays to our implemented solution, our approach can guess for which prompt an essay has been written. We also agree that the accuracy would not be high.