TEAM happy data BIGDATA

DSCI550

Group Name: Happy Data

Students:

Jinshan Yang
Weiqian Zhang
Bingqing Liu
Shengyu Sun
Zhong Wang
Hang Yang

University of Southern California

25th June 2022

Contents

1 Introduction

In Task1, we scraped the papers based on the titles from original bik dataset of homework 1 and save them as pdfs in the folder; as for task 2, we extracted the texts of all generated pdfs from task 1 by using tika and save them as txt files in a folder; in terms of task 3, we make thousands of author names using name dataset into the csv file and randomly sample the features from the updated dataset from task 4 of homework1 to generate the dataset with random sampling; as for task 4 and task 5, we use GPT-2 model to generate 500 texts when inputs are the generated texts from task 2; in task 6, we first use GPT-2 model to generate 500 titles with inputs of the titles from bik dataset of homework 1 , then we utilize generated 500 titles from GPT-2, generated author names from task 3, and generated texts from task 4 and task 5 to generate new pdfs and save these new pdfs in a folder; finally, for task 7, we combine texts, titles, and author names of generated pdfs from task 6 and other features from task 3 to get the final updated tsv.

The report will follow the instructions of homework2 to answer each questions.

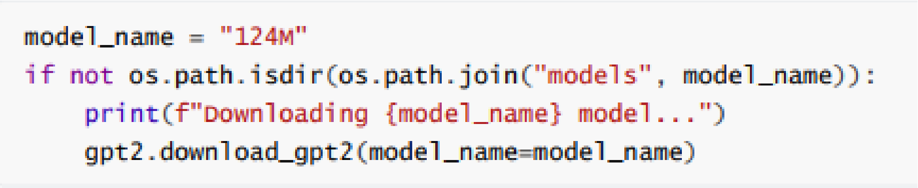
- **Question 1** What did the GPT-2 generated texts look like?
- **Question 2** Were they believable?
- **Question 3** Would your associated ancillary features from assignment 1 have been able to discern what was false or not?
- **Question 4** How much do you think media falsification is solvable using ancillary metadata features, or using actual content based techniques? Is one better than the other?
- **Question 5** What other types of datasets could have been used to generate the falsified papers? Pick at least 2 datasets from distinct MIME types.
- **Question 6** What other sorts of "backstopping" would be required to generate a believable paper trail for the scientific literature?

2 Questions' Answer

2.1 Question 1

What did the GPT-2 generated texts look like?

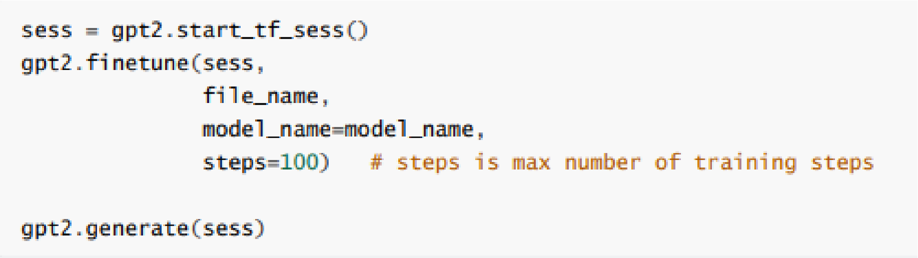
We downloaded a well-trained GPT-2 model from github called "124M" to be the basic model of our task.



```
model_name = "124M"
if not os.path.isdir(os.path.join("models", model_name)):
    print(f"Downloading {model_name} model...")
    gpt2.download_gpt2(model_name=model_name)
```

Figure 1: $124Mmodel$

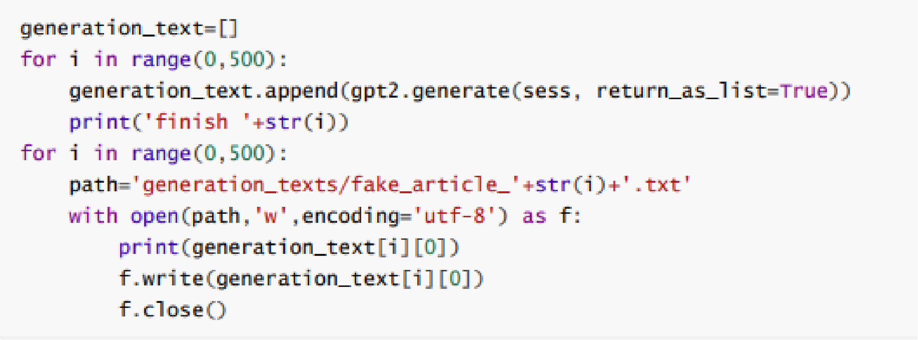
We imported the paper content text generated in task 3 into the model as input, and adjusted the training step parameters. Finally, a paper content generation model in accordance with bik paper dataset is obtained.



```
sess = gpt2.start_tf_sess()
gpt2.finetune(sess,
              file_name,
              model_name=model_name,
              steps=100)   # steps is max number of training steps

gpt2.generate(sess)
```

Figure 2: generation text code for GPT-2 model

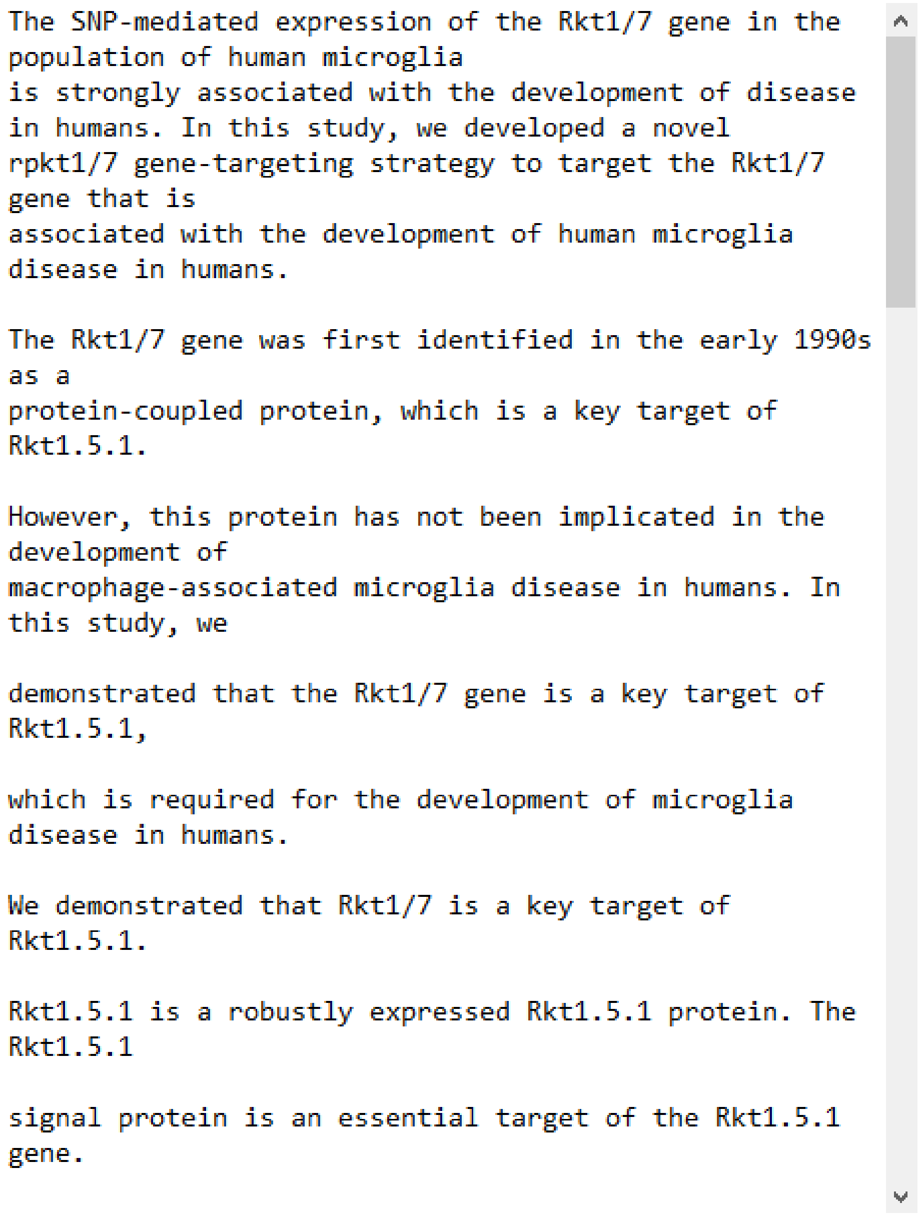
This step is used to generate 500 false papers. The model trained by the above step has been used for 500 times of text generation, and finally 500 false papers have been obtained.



```
generation_text=[]
for i in range(0,500):
    generation_text.append(gpt2.generate(sess, return_as_list=True))
    print('finish '+str(i))
for i in range(0,500):
    path='generation_texts/fake_article_'+str(i)+'.txt'
    with open(path,'w',encoding='utf-8') as f:
        print(generation_text[i][0])
        f.write(generation_text[i][0])
        f.close()
```

Figure 3: code for writing text output

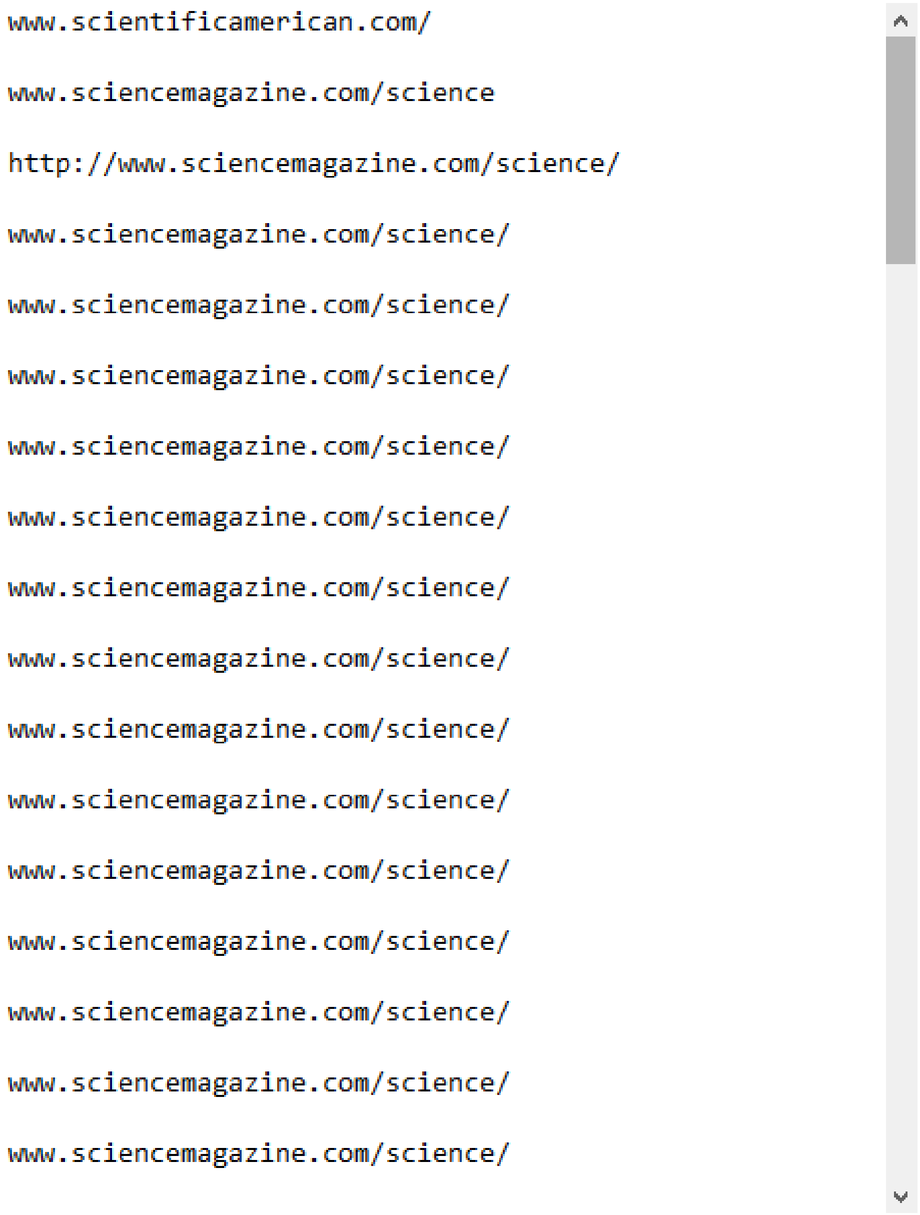
This screenshot below shows one of the falsified papers our model generated. We can find a lot of repetition in this text. For instance, “Rkt1/7 is a key target of Rkt1.5.1.” This sentence appears many times in this text and most of them are exactly the same. Such repetitions occur even more often as we try to generate more and more falsified papers.



```
The SNP-mediated expression of the Rkt1/7 gene in the
population of human microglia
is strongly associated with the development of disease
in humans. In this study, we developed a novel
rpkt1/7 gene-targeting strategy to target the Rkt1/7
gene that is
associated with the development of human microglia
disease in humans.

The Rkt1/7 gene was first identified in the early 1990s
as a
protein-coupled protein, which is a key target of
Rkt1.5.1.

However, this protein has not been implicated in the
development of
macrophage-associated microglia disease in humans. In
this study, we

demonstrated that the Rkt1/7 gene is a key target of
Rkt1.5.1,

which is required for the development of microglia
disease in humans.

We demonstrated that Rkt1/7 is a key target of
Rkt1.5.1.

Rkt1.5.1 is a robustly expressed Rkt1.5.1 protein. The
Rkt1.5.1

signal protein is an essential target of the Rkt1.5.1
gene.
```

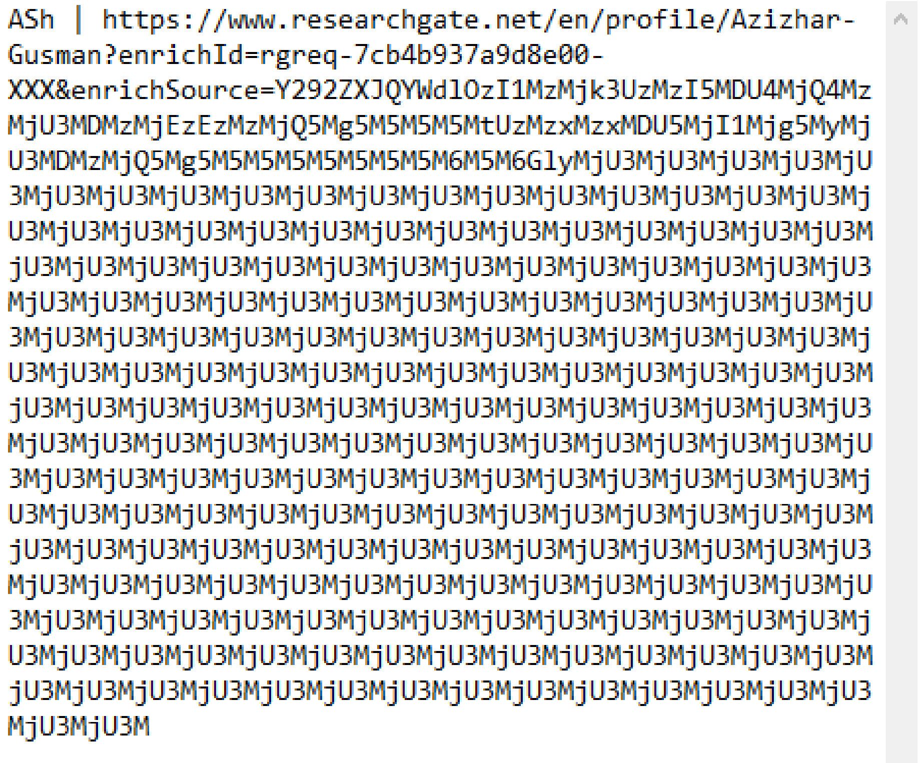
Figure 4: screenshot 1 for generated text

Furthermore, we also noticed that some of the generated text was unreadable, such as repeating urls throughout the text. In those unreadable papers, much of the text is full of repeated or garbled characters.



```
www.scientificamerican.com/

www.sciencemagazine.com/science

http://www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/

www.sciencemagazine.com/science/
```

Figure 5: screenshot 2 for generated text



```
ASh | https://www.researchgate.net/en/profile/Azizhar-
Gusman?enrichId=rgreq-7cb4b937a9d8e00-
XXX&enrichSource=Y292ZXJQYWdlOzI1MzMjk3UzMzI5MDU4MjQ4Mz
MjU3MDMzMjEzEzMzMjQ5Mg5M5M5M5MtUzMzxMzxMDU5MjI1Mjg5MyMj
U3MDMzMjQ5Mg5M5M5M5M5M5M5M6M5M6GlyMjU3MjU3MjU3MjU3MjU
3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3Mj
U3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3M
jU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3
MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU
3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3Mj
U3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3M
jU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3
MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU
3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3Mj
U3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3M
jU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3
MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU
3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3Mj
U3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3M
jU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3MjU3
MjU3MjU3M
```

Figure 6: screenshot 3 for generated text

Overall, most of the falsified papers generated by the models we are training are readable.

2.2 Question 2

Were they believable?

We don't think they are believable. As you read carefully around these falsified papers, you will find many odd things in these texts, such as a lot of repetitions and logical problems. If these repetitions were less numerous, the readable falsified papers might seem believable in reading.

2.3 Question 3

Would your associated ancillary features from assignment 1 have been able to discern what was false or not?

In my opinion, our associated ancillary features from assignment 1 would not have been able to discern what was false or not. That's because in our falsified papers, most of the content is about the falsified research, instead of information about the author. We basically distinguish false content based on the actual content.

2.4 Question 4

How much do you think media falsification is solvable using ancillary metadata features, or using actual content based techniques? Is one better than the other?

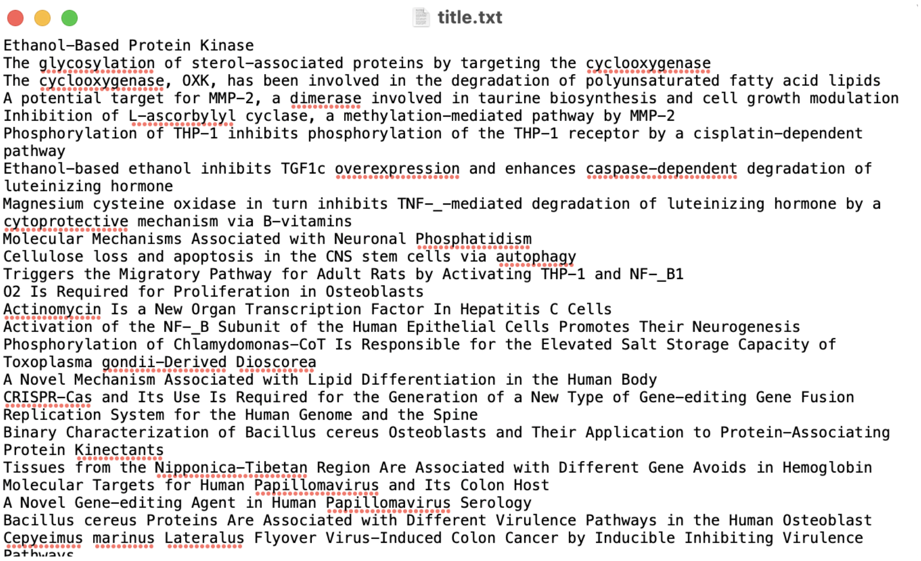
Most media falsification is solvable using actual content based techniques. We think using actual content based techniques is better than using ancillary metadata features because, from our point of view, actual content based techniques are more reliable that it's based on media's actual content which would be more accurate.

2.5 Question 5

What other types of datasets could have been used to generate the falsified papers? Pick at least 2 datasets from distinct MIME types.

In order to make the falsified papers, we have to use these distinct MIME types: txt, and csv.

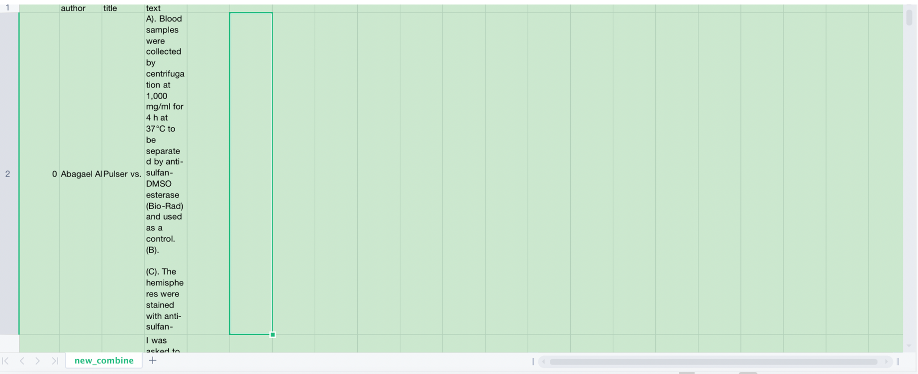
First of all, we need to use this dataset named title.txt which is generated from GPT-2 model with inputs of the titles from bik dataset of homework 1. The following is title.txt.



```
Ethanol–Based Protein Kinase
The glycosylation of sterol–associated proteins by targeting the cyclooxygenase
The cyclooxygenase, OXK, has been involved in the degradation of polyunsaturated fatty acid lipids
A potential target for MMP–2, a dimerase involved in taurine biosynthesis and cell growth modulation
Inhibition of L–ascorbylyl cyclase, a methylation–mediated pathway by MMP–2
Phosphorylation of THP–1 inhibits phosphorylation of the THP–1 receptor by a cisplatin–dependent
pathway
Ethanol–based ethanol inhibits TGF1c overexpression and enhances caspase–dependent degradation of
luteinizing hormone
Magnesium cysteine oxidase in turn inhibits TNF–_–mediated degradation of luteinizing hormone by a
cytoprotective mechanism via B–vitamins
Molecular Mechanisms Associated with Neuronal Phosphatidism
Cellulose loss and apoptosis in the CNS stem cells via autophagy
Triggers the Migratory Pathway for Adult Rats by Activating THP–1 and NF–_B1
O2 Is Required for Proliferation in Osteoblasts
Actinomycin Is a New Organ Transcription Factor In Hepatitis C Cells
Activation of the NF–_B Subunit of the Human Epithelial Cells Promotes Their Neurogenesis
Phosphorylation of Chlamydomonas–CoT Is Responsible for the Elevated Salt Storage Capacity of
Toxoplasma gondii–Derived Dioscorea
A Novel Mechanism Associated with Lipid Differentiation in the Human Body
CRISPR–Cas and Its Use Is Required for the Generation of a New Type of Gene–editing Gene Fusion
Replication System for the Human Genome and the Spine
Binary Characterization of Bacillus cereus Osteoblasts and Their Application to Protein–Associating
Protein Kinectants
Tissues from the Nipponica–Tibetan Region Are Associated with Different Gene Avoids in Hemoglobin
Molecular Targets for Human Papillomavirus and Its Colon Host
A Novel Gene–editing Agent in Human Papillomavirus Serology
Bacillus cereus Proteins Are Associated with Different Virulence Pathways in the Human Osteoblast
Cepyeimus marinus Lateralus Flyover Virus–Induced Colon Cancer by Inducible Inhibiting Virulence
Pathways
```

Figure 7: title.txt

Besides txt dataset, we combine the titles from title.txt, generated author names from task 3, and generated texts from task 4 and task 5 into the csv dataset named new-combine.csv which can be shown below.



1		author	title	text
2	0	Abagael Al	Pulser vs.	A). Blood samples were collected by centrifugation at 1,000 mg/ml for 4 h at 37°C to be separated by anti-sulfan-DMSO esterase (Bio-Rad) and used as a control. (B). (C). The hemispheres were stained with anti-sulfan-
				I was

Figure 8: new-combine.csv

In summary, these two datasets are used to generate the falsified papers.

2.6 Question 6

What other sorts of "backstopping" would be required to generate a believable paper trail for the scientific literature?

In order to generate the believable paper trail for the scientific literature: first of all, the contents of these papers seems unreadable with logical errors and repeated sentences or words, to make the scientific papers believable, we have to deal with correct language grammar and words accuracy; secondly, if these papers could add the scientific graphs, like bar charts, like pie charts, and so on, then these papers would be persuasive to readers; thirdly, the sponsor of the authorization organizations could be added to the scientific papers, this will strengthen the credibility of scientific papers.

3 Extra Credit

Firstly, we scrape more than 200 papers from bik dataset in task 1; and then we use the scrapped more than 200 papers to train GPT-2 model in task 4 and task 5.

Secondly, we write code about how to use LaTex to convert texts which are training from GPT-2 model to generate PDFs in task 6.