CS 505 Text-Location project

Names:

1- Ghalia Alshanbari

2- Jena Jordahl

3- Bhaskar Abbireddy

Link to github with code: https://github.com/jenajjedu/NLP_Named_Entities

We thought about how the location of a story isn't really explicitly given in the text. So we decided a new transformer neural network might be able to tell us where a scene takes place. Our main goal was to use the text classification and graph theory we used in class to determine the approximate location of where a scene takes place in a fiction or a nonfiction text. As part of our process we generated a summary text that is based on location specific sources. We then trained an LSTM and a transformer as well to generate location summary text. So, our approach was to train the neural network with the Game of Thrones script and the Gettysburg battle texts from the gutenberg press. We leveraged SCC and colab as well as local machines to run our code because of GPU issues. We visualized the significance of characters in a script/scene via word clouds and then we graphed using networkx and pyvis a forced directed graph to break the related texts into scenes.

Participation Statement: We completed the project definition, presentation and write-up together. We meet weekly in-person to help each other on tasks. Certain people focused on special tasks, see below. We all located research resources. We worked together to overcome obstacles listed in the presentation. Each of us independently completed portions of the homework assignments so that we could use that knowledge to complete a research task.

We met on average 3 hours daily for the past week. We all worked on every deliverable due to our joint sessions. Point Person for specific deliverables:

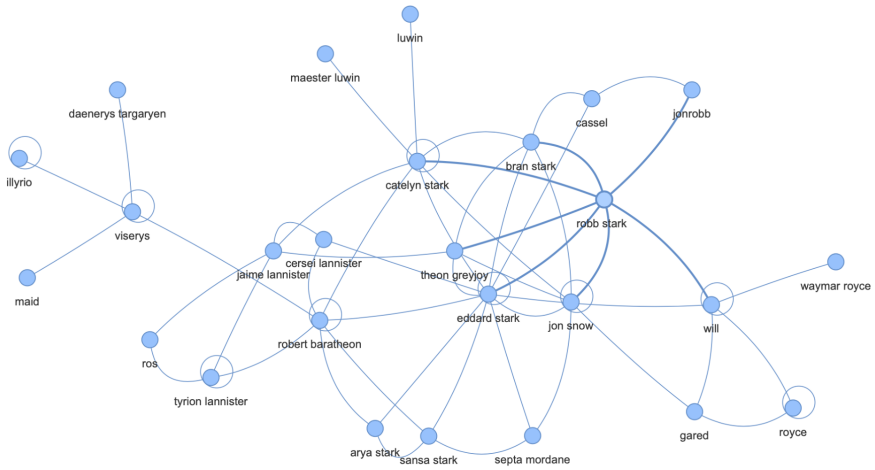
- General Architecture for Location Classification: Jena
- Text generation via a trained LSTM on Gettysburg & Game of Throne scripts: Jena
- Screen scraping of gutenberg press text and location information websites: Jena
- Researched networkx as a tool to breakout the scene boundaries: Jena
- Located Game of Thrones script and played all the episodes in the background while working to pick up the scene boundaries and understand location changes: Jena
- Read a significant portion of The Star of Gettysburg book, Gettysburg script (adapted from Killer Angels book) and location websites: Jena
- Generated commands from text to build the graph model: Jena
- Built the tutorial for movie plot summarization and fine-tuning sentence comparison: Jena
- As part of the research work for our project, solved homework 6 I, 7 II, and 7 III: Jena
- Reviewed Coursera NN videos: Jena
- Bert sentence comparison of scene text from Gettysburg & Game of Throne scripts: Jena
- Created prompts to gauge subjective accuracy of encoded location information: Jena
- Preprocessed the Game of Thrones text and generated the word clouds of who spoke the most for the whole series and specific scenes as well as used the BERT model to train the data: Ghalia

- Worked on HW-6 Part 1 and part 2 training using an LSTM model to help with our project in order to implement it in our preprocessing for our data and to help with research:Ghalia
- Worked on PA7 II and PA7 GPT to help with the BERT implementation (linked in github): Ghalia
- Completed the presentation and the paper with help from Jena and Bhaskar as well as research:Ghalia
- Reviewed videos on BERT and compared with other models as well ashow to implement it to help with our project: Ghalia
- Worked on training BERT and breaking down the GOT script based on next sentence prediction into different scenes (list of scenes for entire GOT series): Bhaskar
- Used **graph theory** (with the help of jena) to plot the graph with nodes(characters) and edges(conversation) for the scenes broken down from BERT; eventually grouping the scenes on the graph: Bhaskar
- Researched work on different blogs and papers to get information aligned to our requirement for different libraries like **Pyvis** (medium, towards data science, springer): Bhaskar
- Scrape the data from web (gettysburg) (helped jena with it) researching over beautifulSoup: Bhaskar
- As part of the research work for our project, solved homework 7 I and homework 7 II which helped me give a better understanding of Transformers and BERT to later implement that in the project: Bhaskar
- Watched couple of videos on Datacamp to get a better understanding of deep learning topics: Bhaskar

Results:

Word Cloud Gives Weights to the Edges on the Graph:

Graph Demonstrates which Characters Speaks one after the Other in an



Episode:

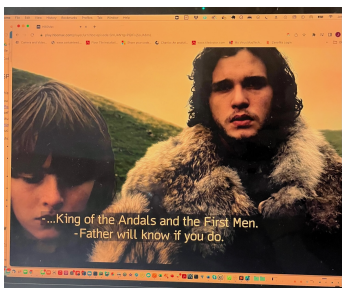
Found Scene Boundaries and Correlated Them With Scenes:

```
    result = torch.argmax(outputs.logits)

    if result.item() == 0:
      inter_result.append(sentences[i+1])
    else:
      result_set.append(inter_result)
      print(i,inter_result)
      inter_result = []
  else:
    result_set.append(inter_result)

print(result_set)
```

```
3 ['What d'you expect? They're savages One lot steals a goat from another lot and before y
36 ['We should head back to the wall', 'Do the dead frighten you?', 'Our orders were to tr
37 ['Don't look away']
38 ['King of the Andals and the First Men']
46 ['Father will know if you do', 'Lord of the Seven Kingdoms and protector of the realm,
48 ['Is it true he saw the White Walkers?', 'The White Walkers have been gone for thousand
49 ['So he was lying?']
56 ['A madman sees what he sees', 'What is it?', 'Mountain lion?', 'There are no mountain
59 ['There are no direwolves south of the Wall', 'Now there are five', 'You want to hold i
72 ['Where will they go? Their mother's dead', 'They don't belong down here', 'Better a qu
```



Scenes take place in a location:

Use Spacey Parser to Identify the Named Entity Type: Location

```
Robert E 162 170 PERSON
Lee 173 176 PERSON
Virginia 771 779 GPE
Maryland 869 877 GPE
Pennsylvania 887 899 GPE
Lee 938 941 PERSON
Abraham Lincoln 1131 1146 PERSON
Lee 1219 1222 PERSON
Washington 1315 1325 GPE
Lee 1583 1586 PERSON
Harrison 2184 2192 PERSON
Harrison 2297 2305 PERSON
Jeb Stuart 4062 4072 PERSON
mule 4509 4513 PERSON
Jeb Stuart's 4711 4723 PERSON
Sorrel 5054 5060 GPE
Lee 5128 5131 PERSON
Eleventh 5715 5723 PERSON
Harrison 5952 5960 PERSON
Mississippi 6042 6053 GPE
George Meade's 6425 6439 PERSON
Harrison 6461 6469 PERSON
George Meade 6529 6541 PERSON
Pennsylvania 6543 6555 GPE
Chamberlain 7359 7370 PERSON
bucko 7433 7438 GPE
Maine 8882 8887 GPE
Maine 9158 9163 GPE
```

Generate Location Related Text using Prompts:

Sentence Generation

<s> i 'm sorry i 'm not going back to king in the south ! ” it 's the only way to die . " i 'll have a baby , you 're the only person left in the city , and we have a choice , i 'll be the first of my own father , the queen of the iron islands . ” and i 'll be sure . . </s>

Prompt Completion

```
King's Landing is east of the dead and we 're not going home , but we 'll
  be the first man in Westeros ? ” it is the only one i ever thought to
 confess , but you 'll have a choice , you know . ! ” it is the best who
knows what i 've seen the people of Westeros and their lives are defeated
                              and _UNK . </s>
```

Prompt Completion ACCURACY, determined subjectively:

Gettysburg: 20 correct; 52 incorrect; correct percentage 27%
Gettysburg: 13 correct; 59 incorrect; correct percentage 18%
Conclusion: LSTMS are not good at predicting location.

Table of Subjective Results:

Fiction Game of Thrones

Saved LSTM model is 19 M

72 * 2 prompt completions and X number of accurate completions

Nonfiction Gettysburg Script

Saved LSTM model is 19 M

72 * 2 prompt completions and X number of accurate completions

Bert Model for Sentence Completion

	sentence1	sentence2	label
1951	According to Naeye, the newfangled telescopes ...	The Big Bang is the primordial explosion from ...	1
1312	We present a novel approach to enhance avalanc...	Robots are used to find avalanche victims.	1
694	RJ Reynolds Tobacco announced yesterday that i...	RJR built factories in Turkey.	1
2188	Robert Szabo concludes that circumcision, as p...	It is incorrect to assert that circumcision pr...	1
1505	During a IRS raid in Spring 2006, the governme...	Hovind was found guilty on 59 federal counts.	1