

# Who's line is it?

By: Yaniv Bronshtein

# Problem: What can we learn from a movie script?

- Relationships between characters
- Character growth(if we have data from multiple movies/books)
- Who said the line -> NLP -> Multi-category text classification!




# The Datasets

# Lord of The Rings

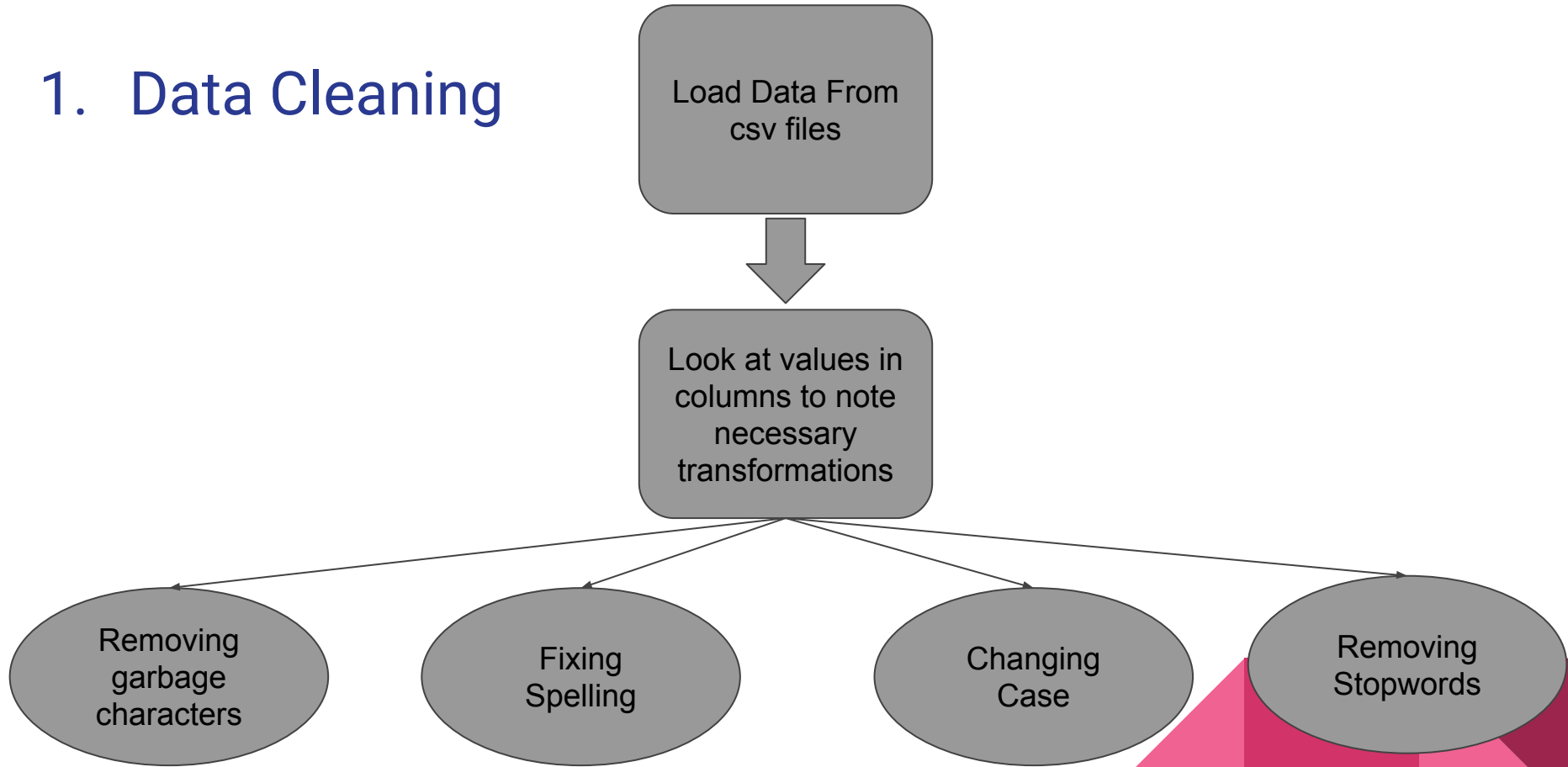
Source:

<https://www.kaggle.com/paultimothymooney/lord-of-the-rings-data>

About:

- Two csv files of which only `lotr_scripts.csv` was used
  - Based on the trilogy containing the Fellowship of the Ring, Two Towers, and The Return of the King
  - `lotr_scripts.csv` contains 3 columns: 'char', 'dialog', and 'movie' where 'char' is the character name, 'dialog' is their line, and Movie is one of three values
- 

# 1. Data Cleaning



## 2. Building a dictionary

Character	Lines	Movie
Frodo	Bloop bleh bloop	Fellowship...
Gollum	My precious precious	Two Towers
Frodo	Ipsum ipsum ipsum	Return of The King

Character	Lines	Movie	Vocab
Frodo	Bloop bleh bloop	Fellowship...	{bloop:2, bleh:1}
Gollum	My precious precious	Two Towers	{my:1, precious:2}
Frodo	Ipsum ipsum ipsum	Return of the King	{ipsum:3}

# THE MEGADICT

```
{  
  Frodo: {  
    bloop:2,  
    bleh:1,  
    Ipsum:3  
  },  
  Gollum: {  
    my:1,  
    precious:2  
  }  
}
```



# 1. Relationships between characters

- For each character in the top 10, use *consolidate\_char\_vocab()* to get their “mega corpus”
- Use depth first search to create a graph using the **networkx** library
- In the *get\_relationships()* function, there is a threshold parameter that allows for choosing the criteria to create in edge
- For example, if thresh=4, an edge(relationship) is created between two characters if a character references another character at least 4 times throughout a given movie

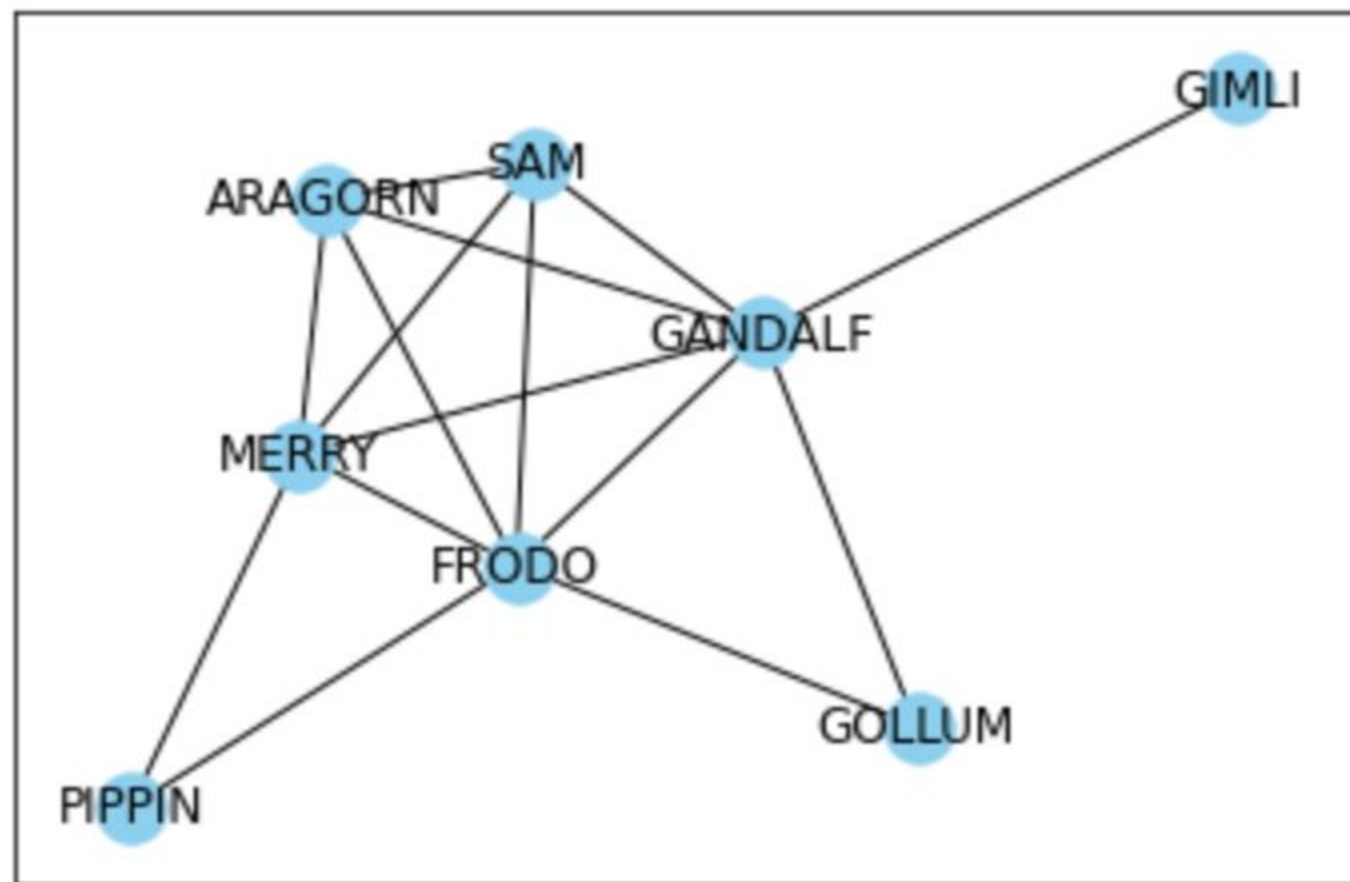




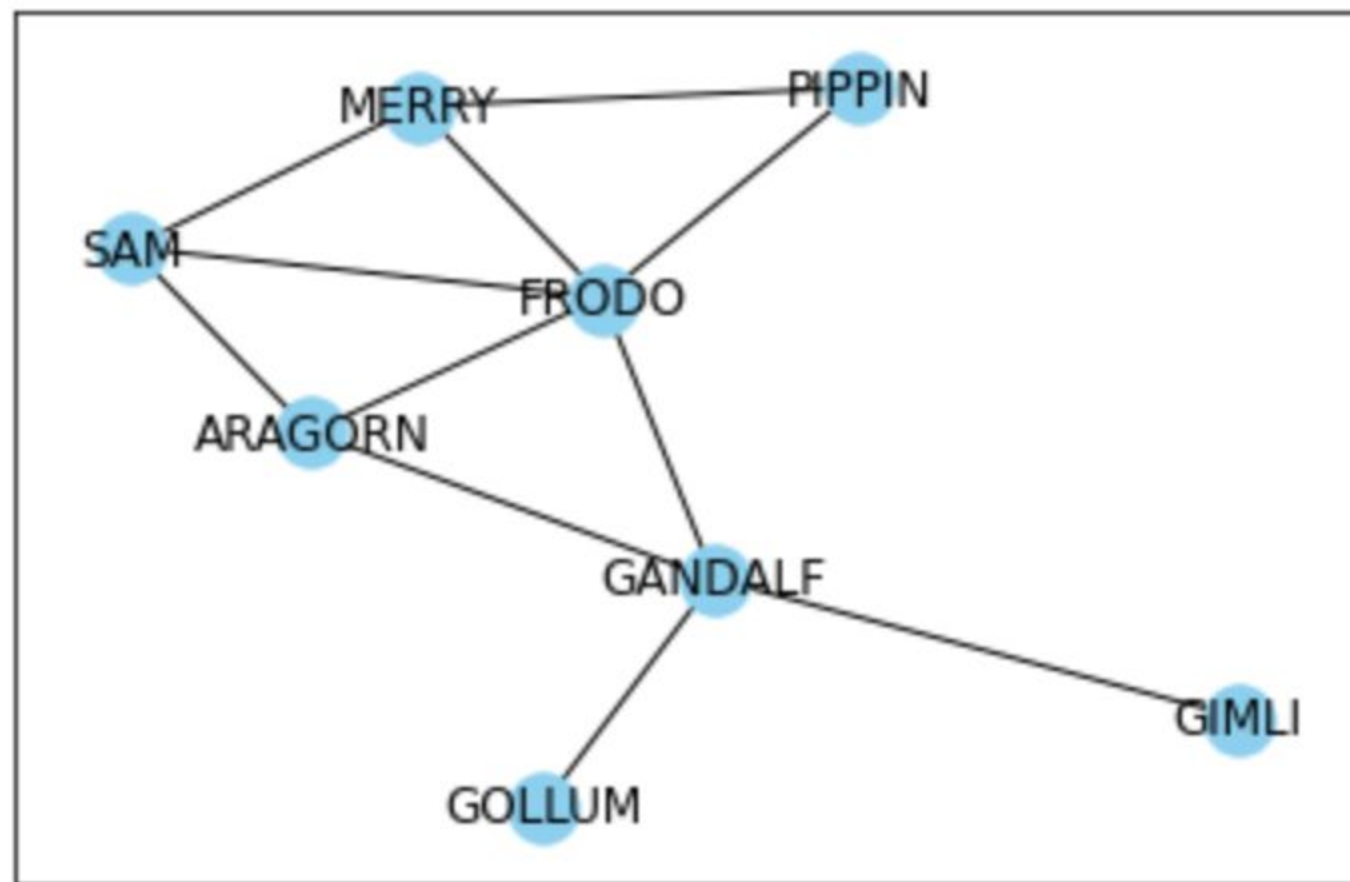


# Relationships: Fellowship of the Ring

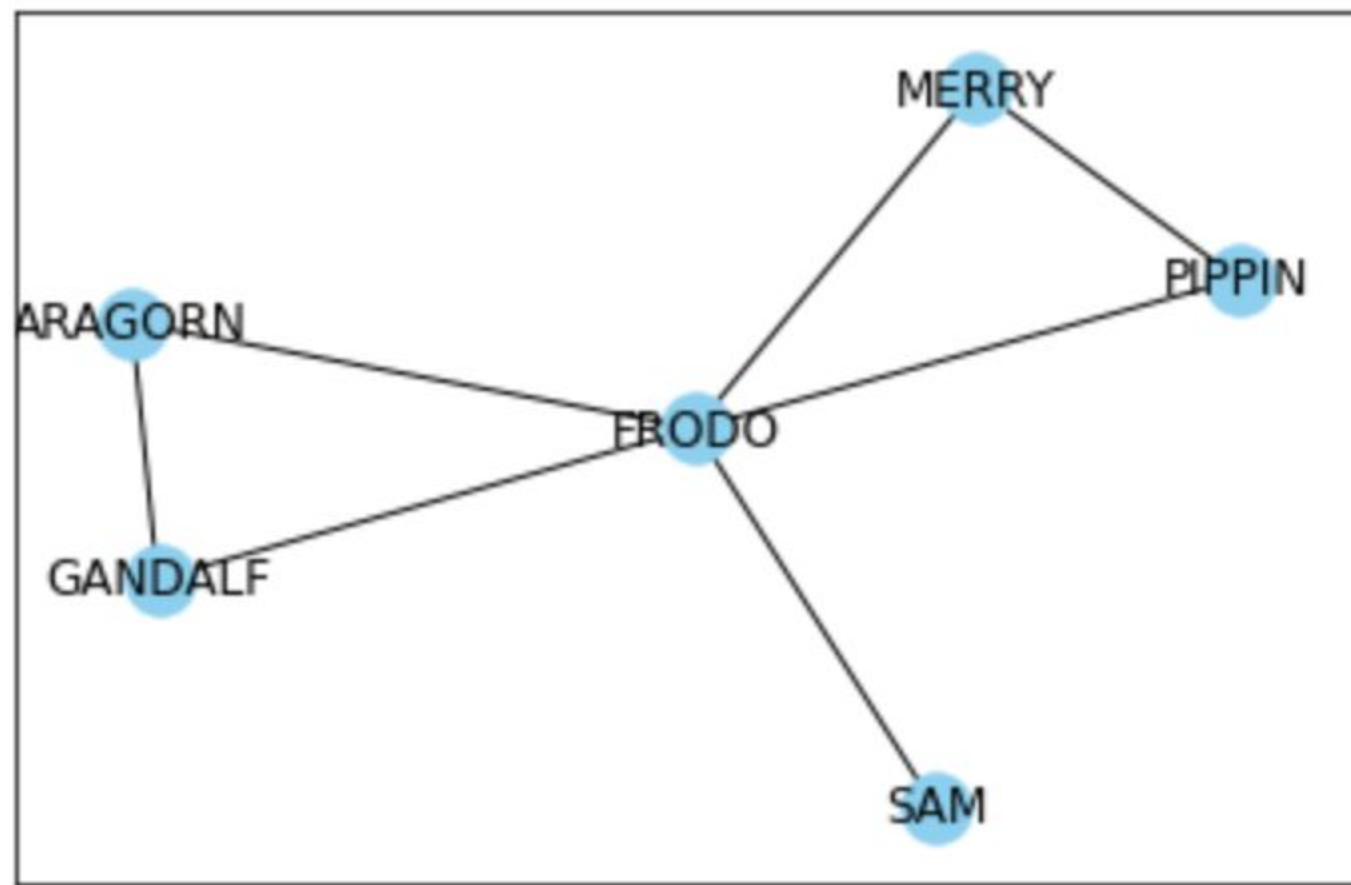
Threshold:0



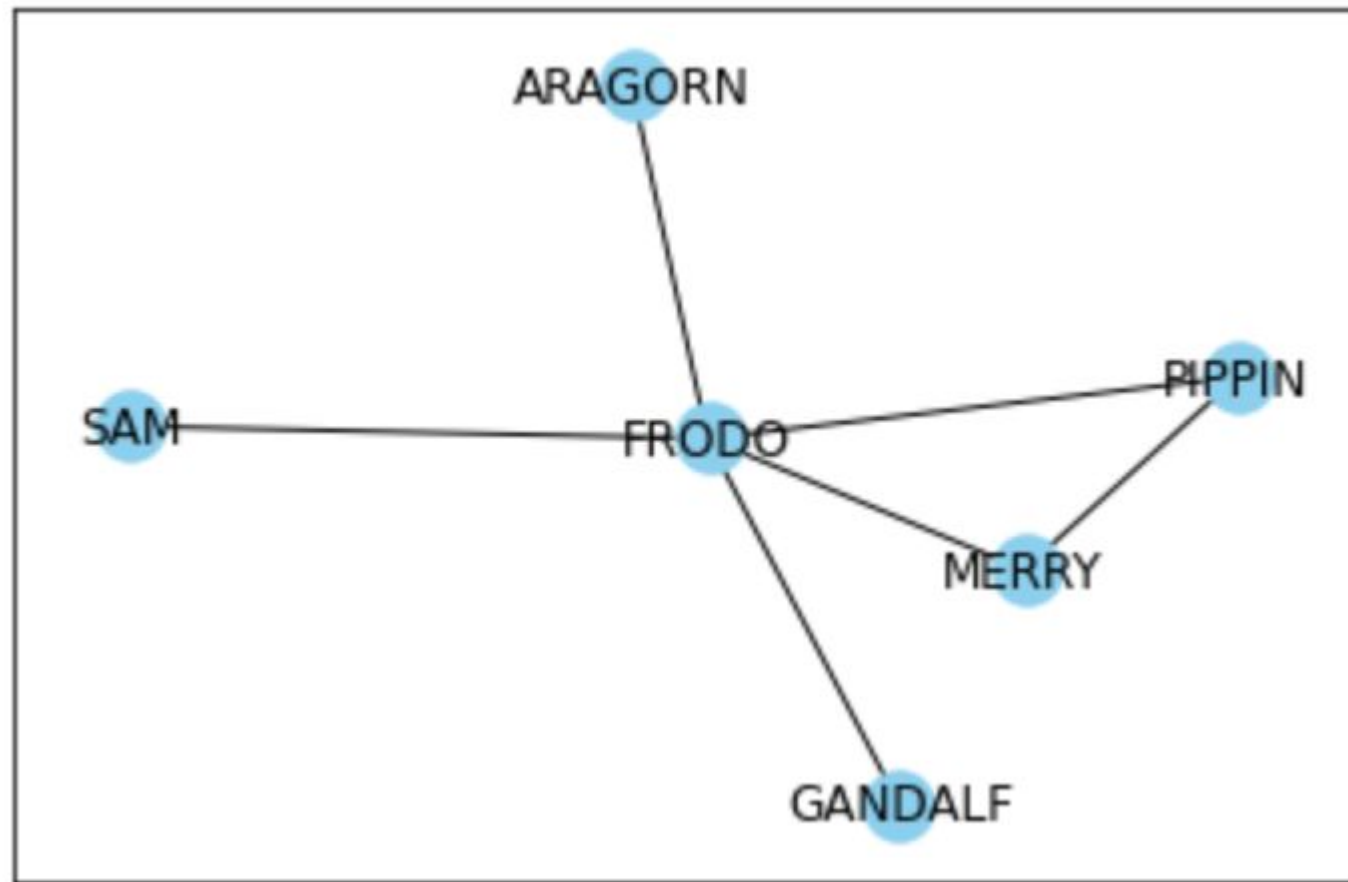
Threshold:1



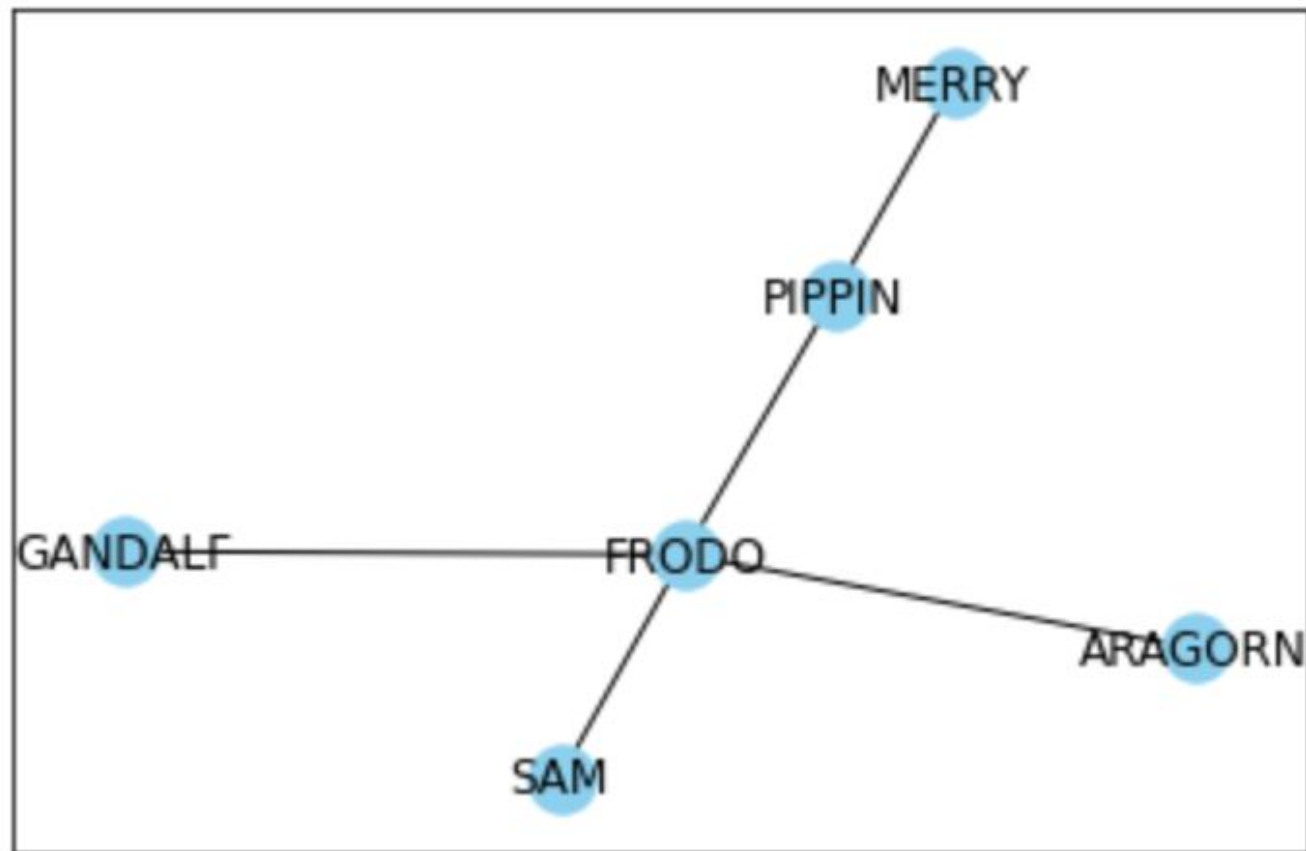
Threshold:2



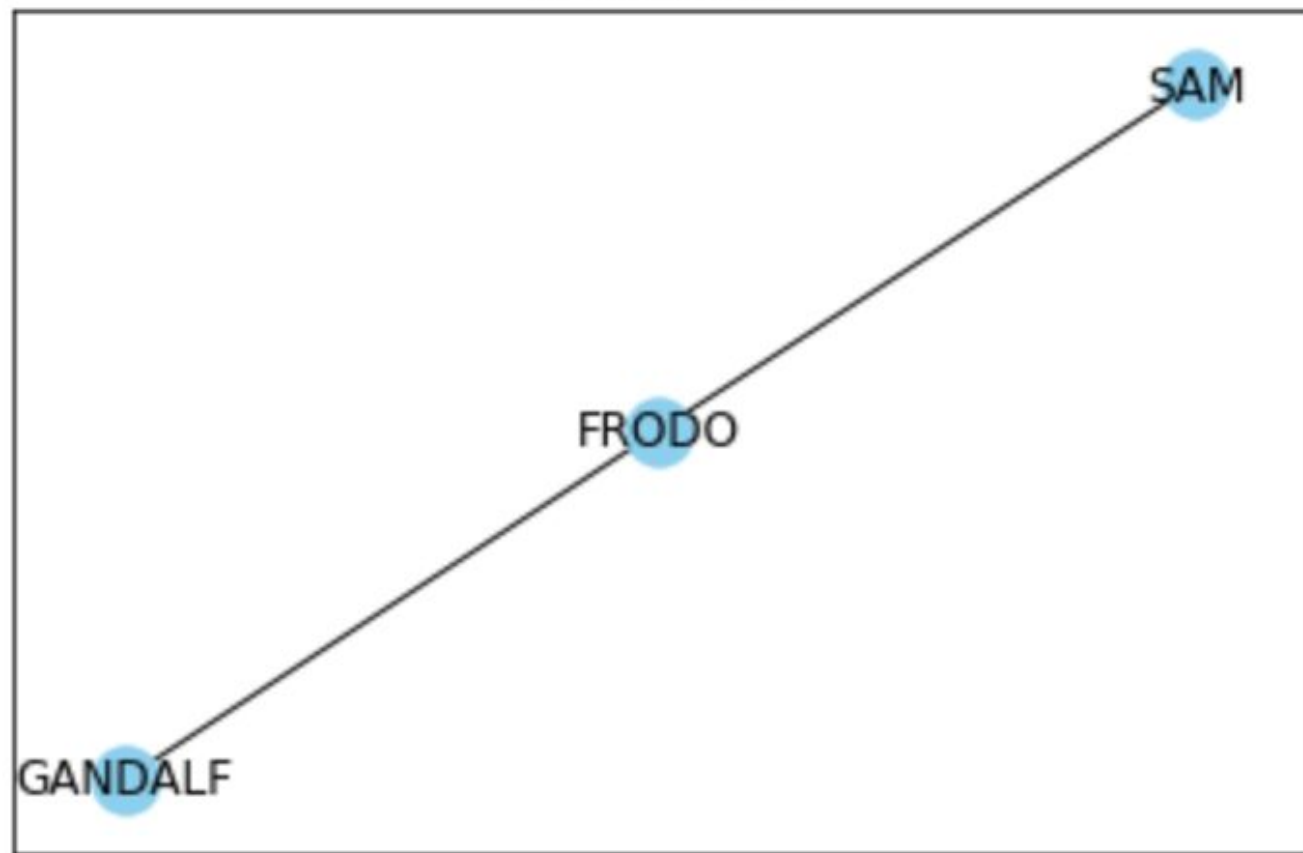
Threshold:3



Threshold:4



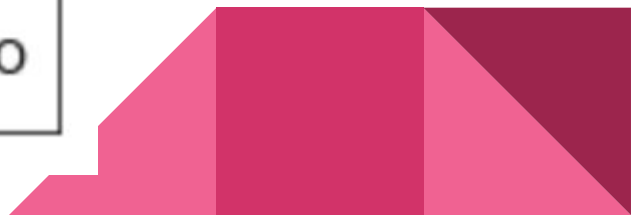
Threshold:5




Threshold:6

SAM

FRODO

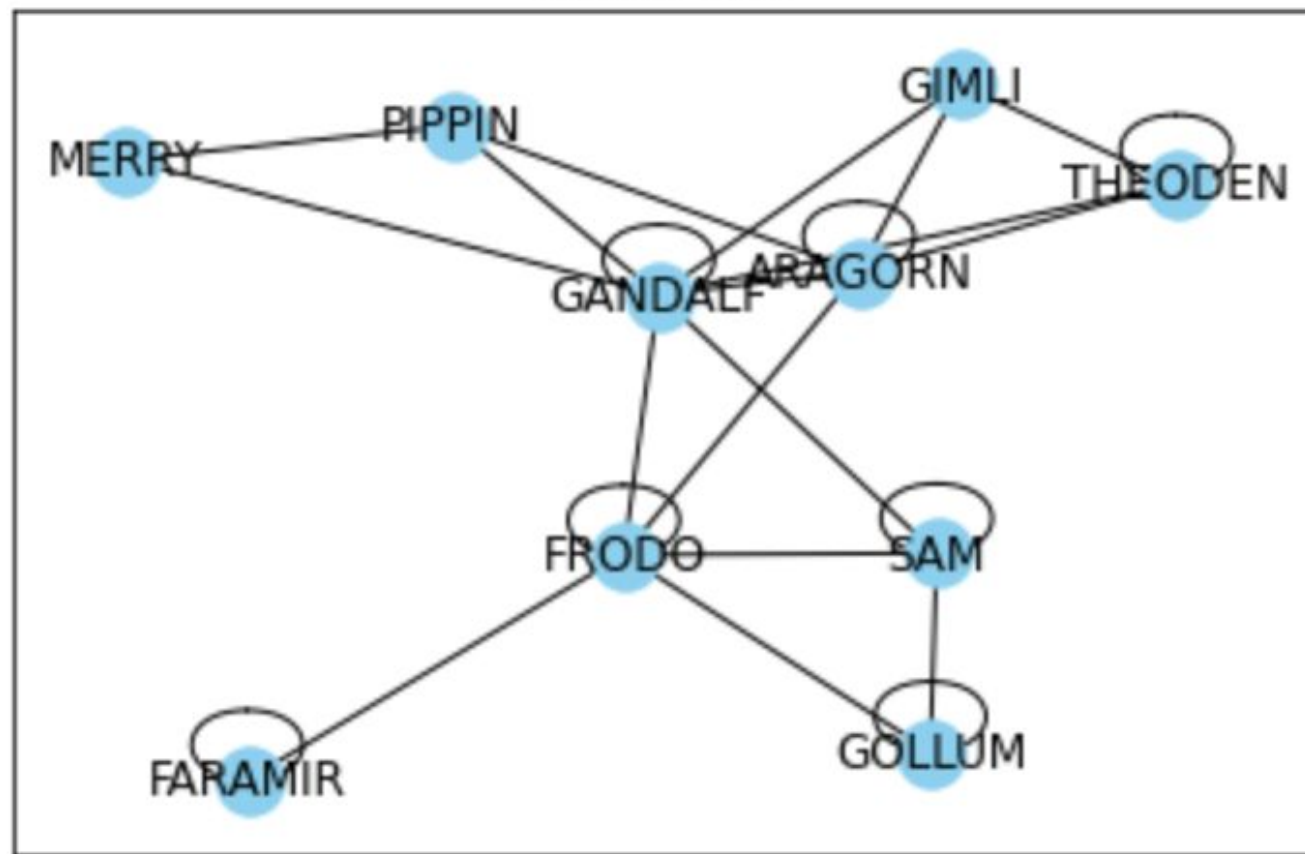




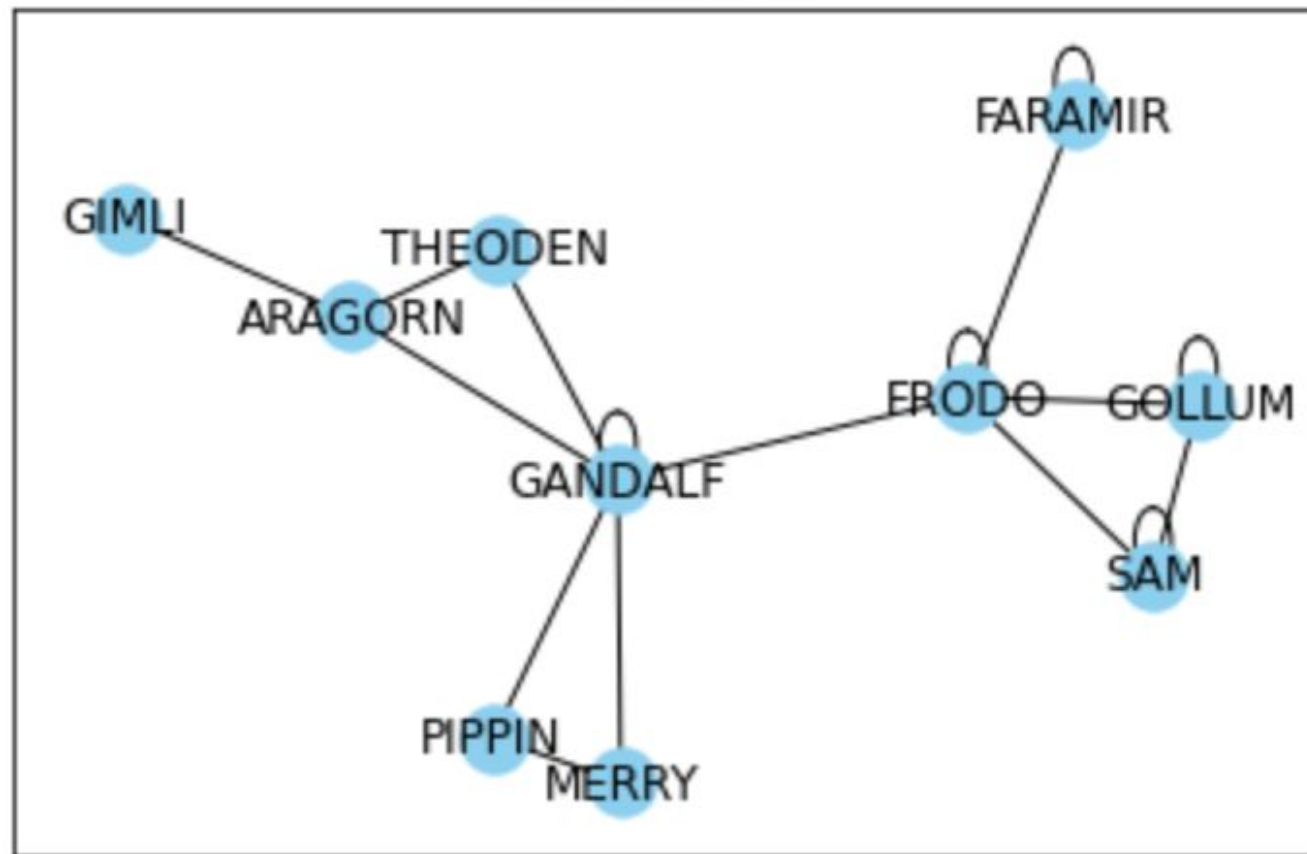


# Relationships: The Two Towers

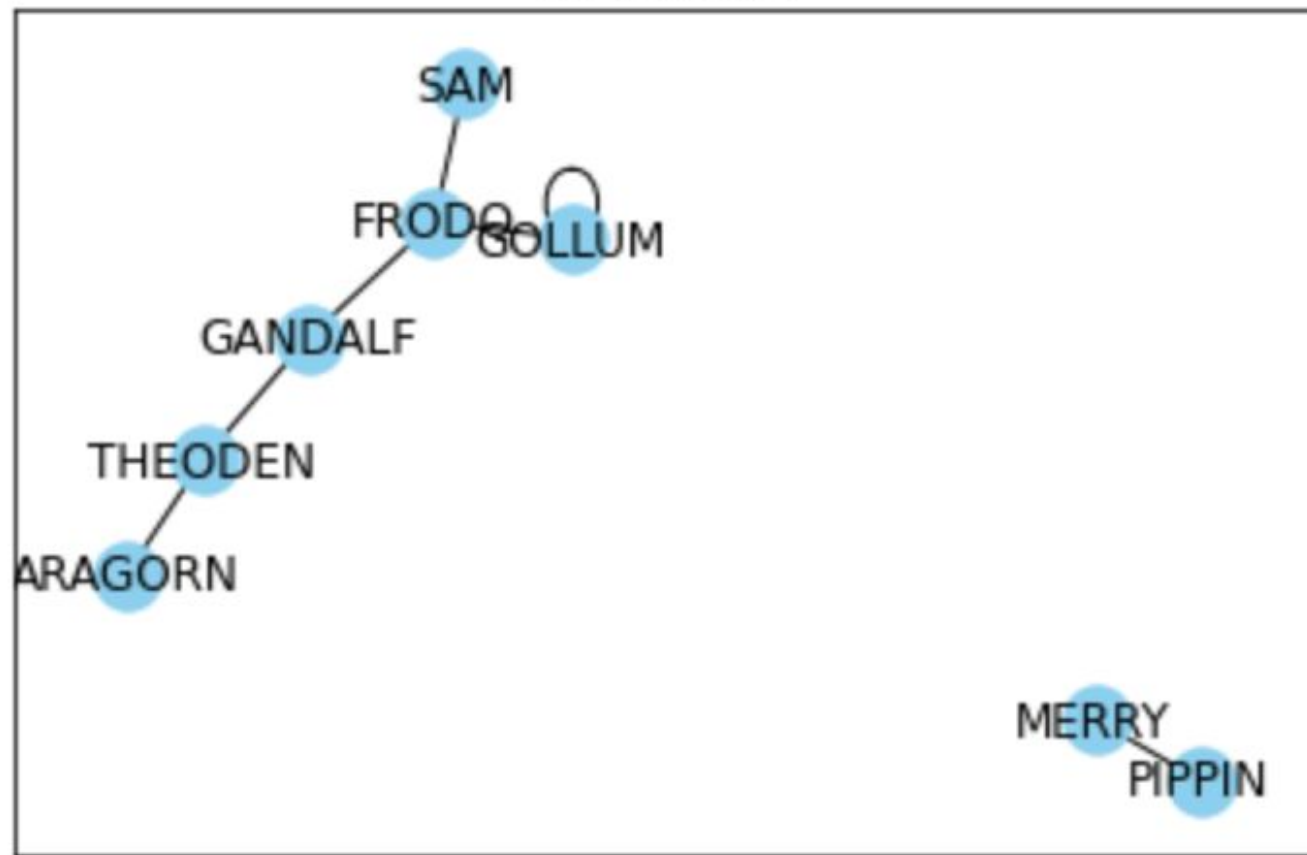
Threshold:0



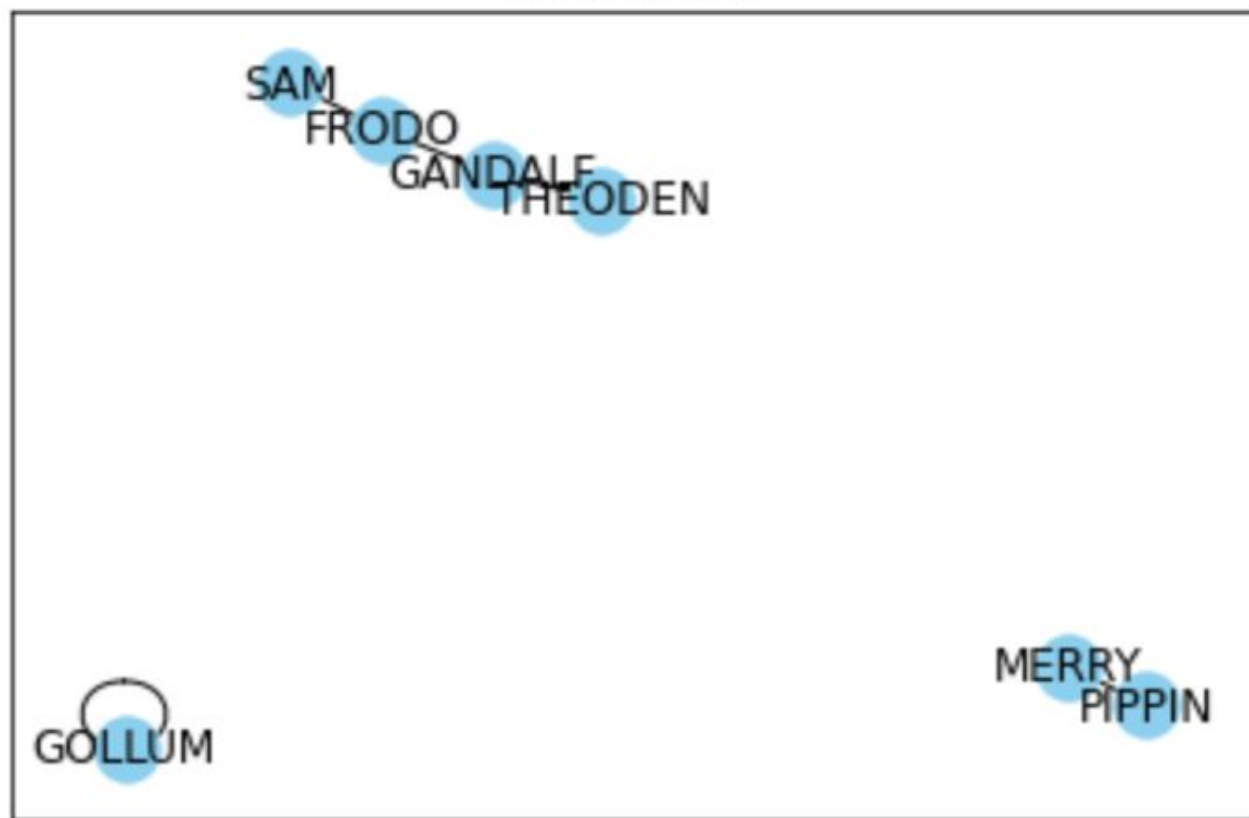
Threshold:1



Threshold:2



Threshold:3



Threshold:4

GOLLUM

THEODEN  
GANDALF

SAM  
FRODO

MERRY  
PIPPIN

Threshold:5

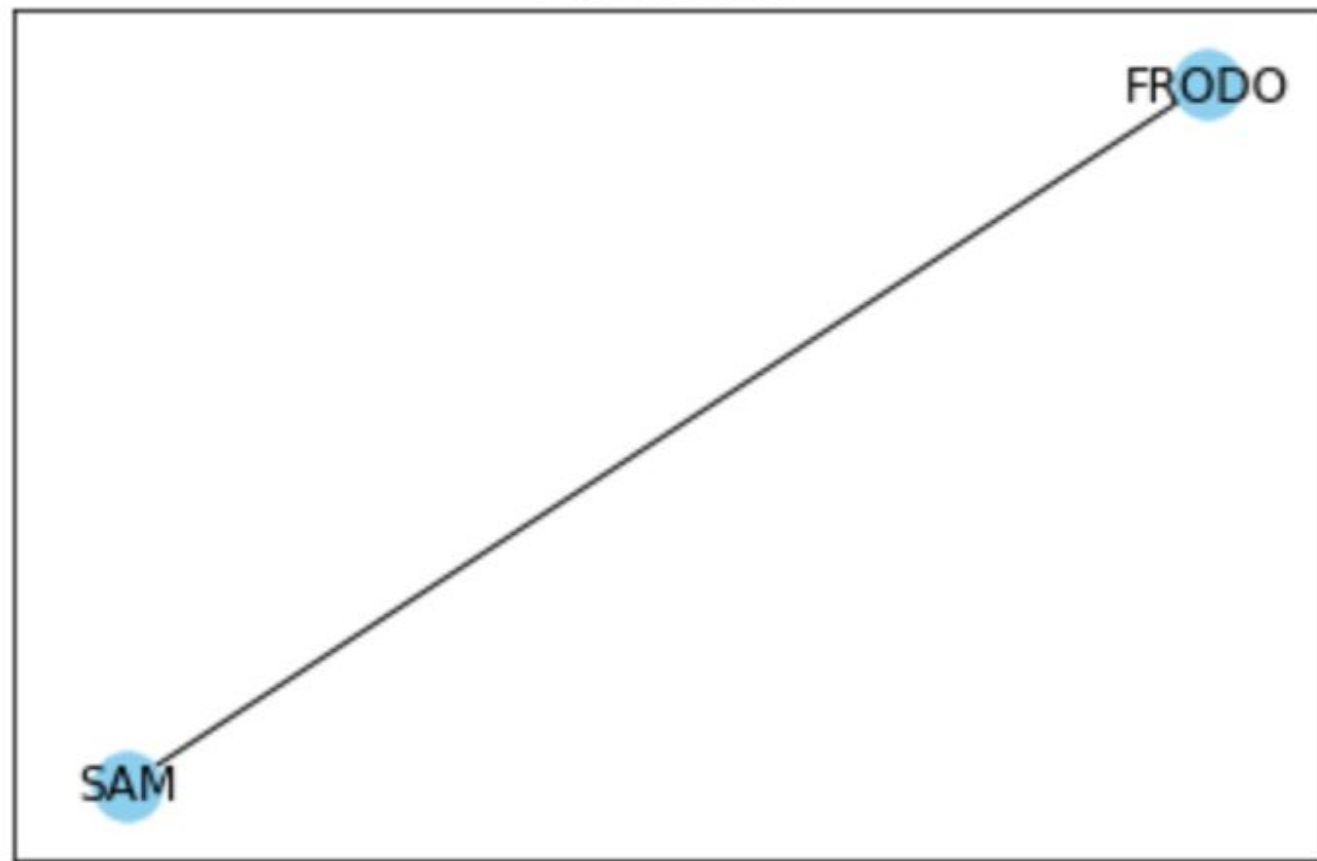


Threshold:6





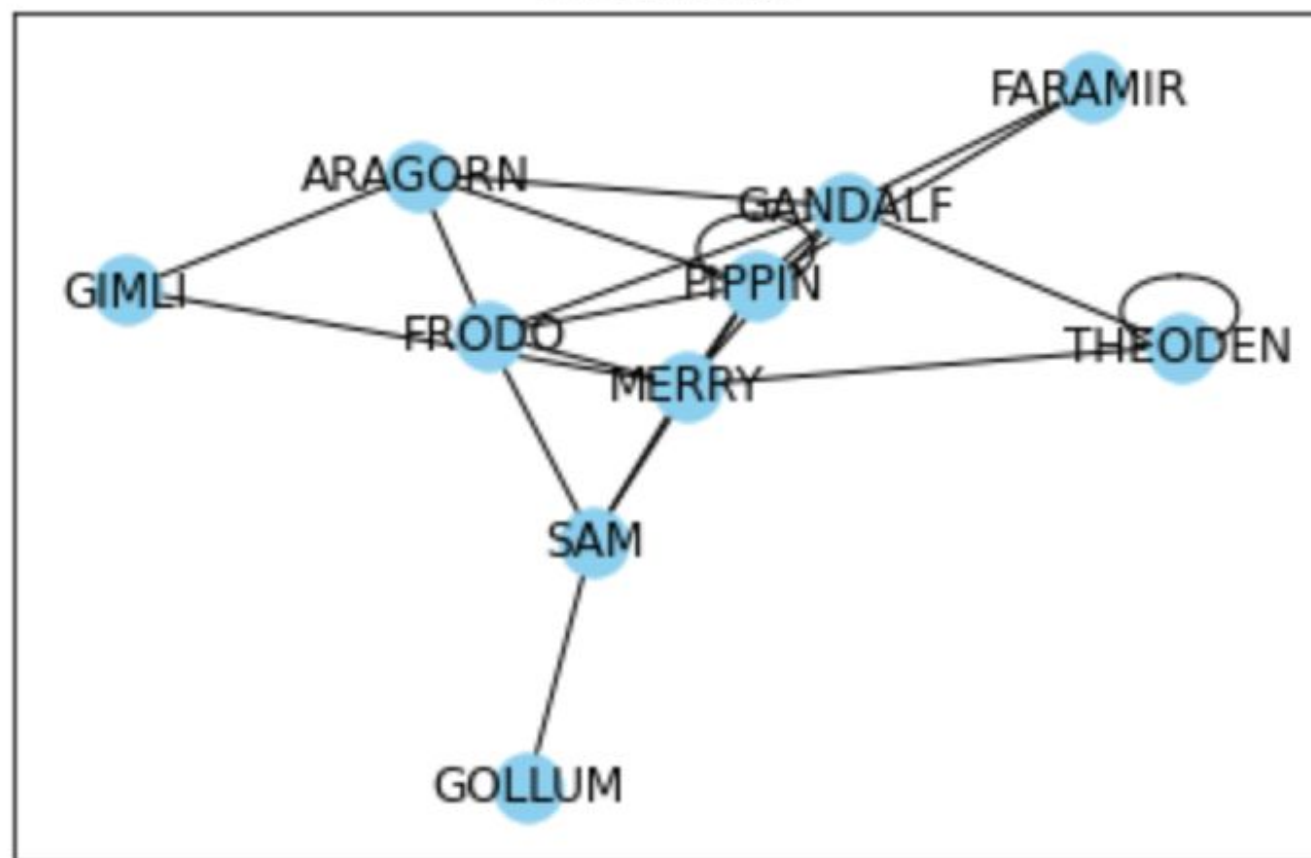
Threshold: 7



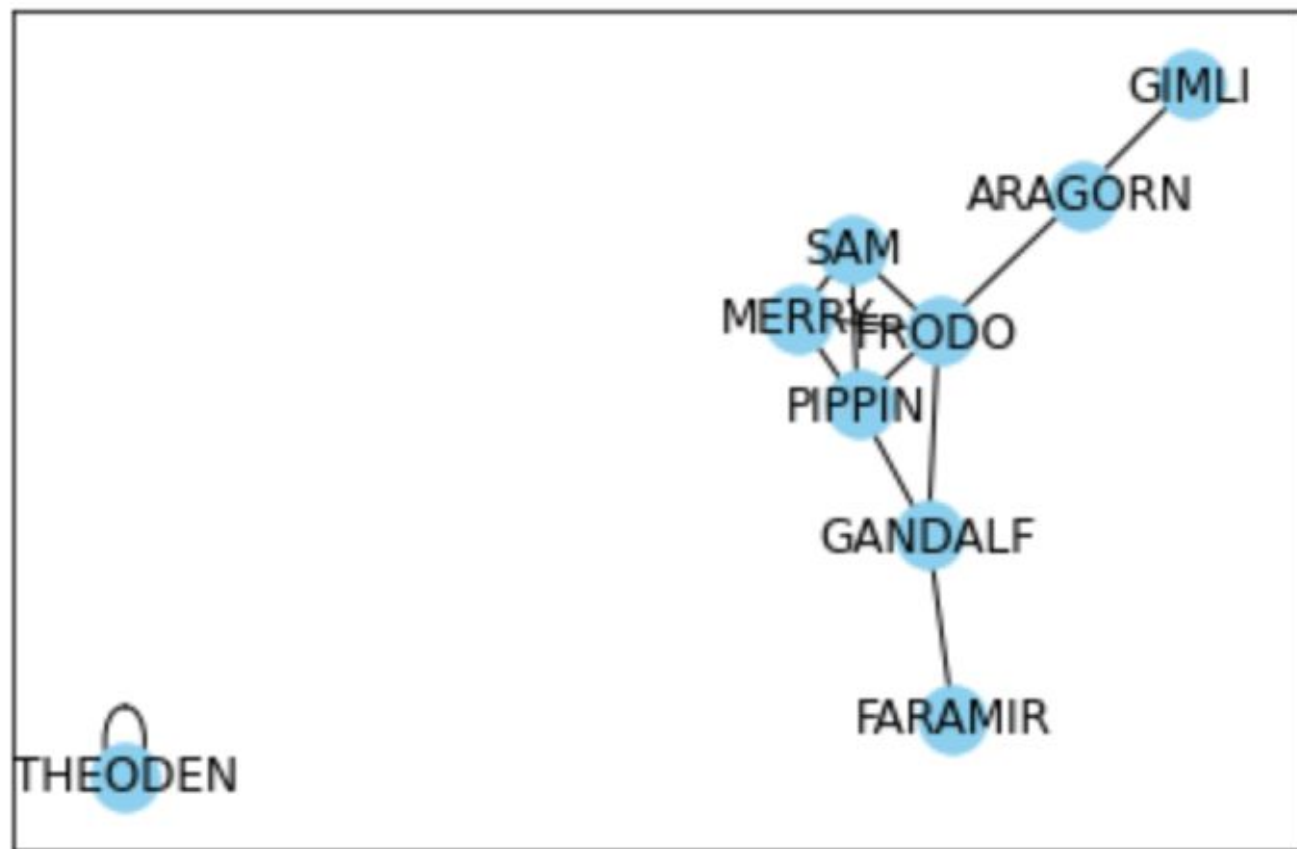


# Relationships: The Return of The King

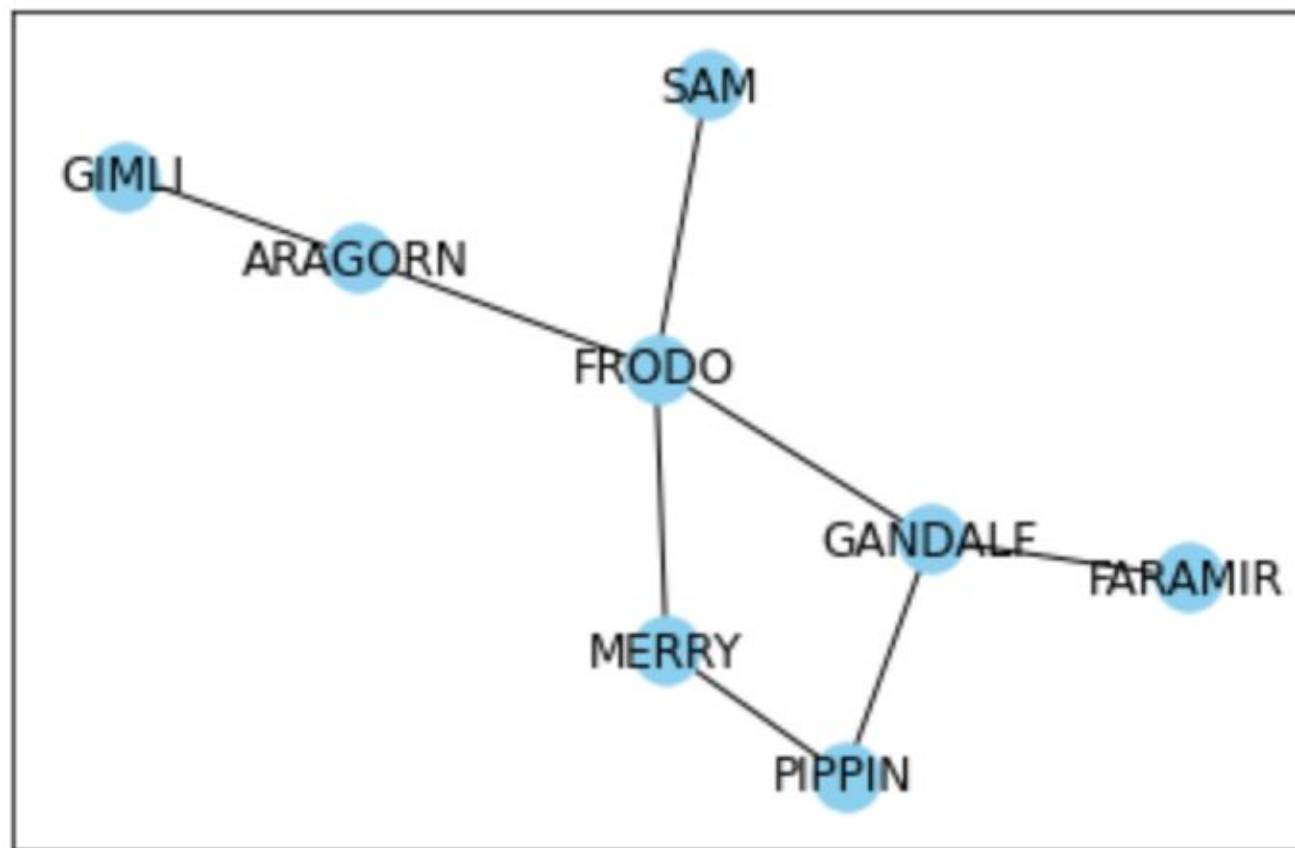
Threshold:0



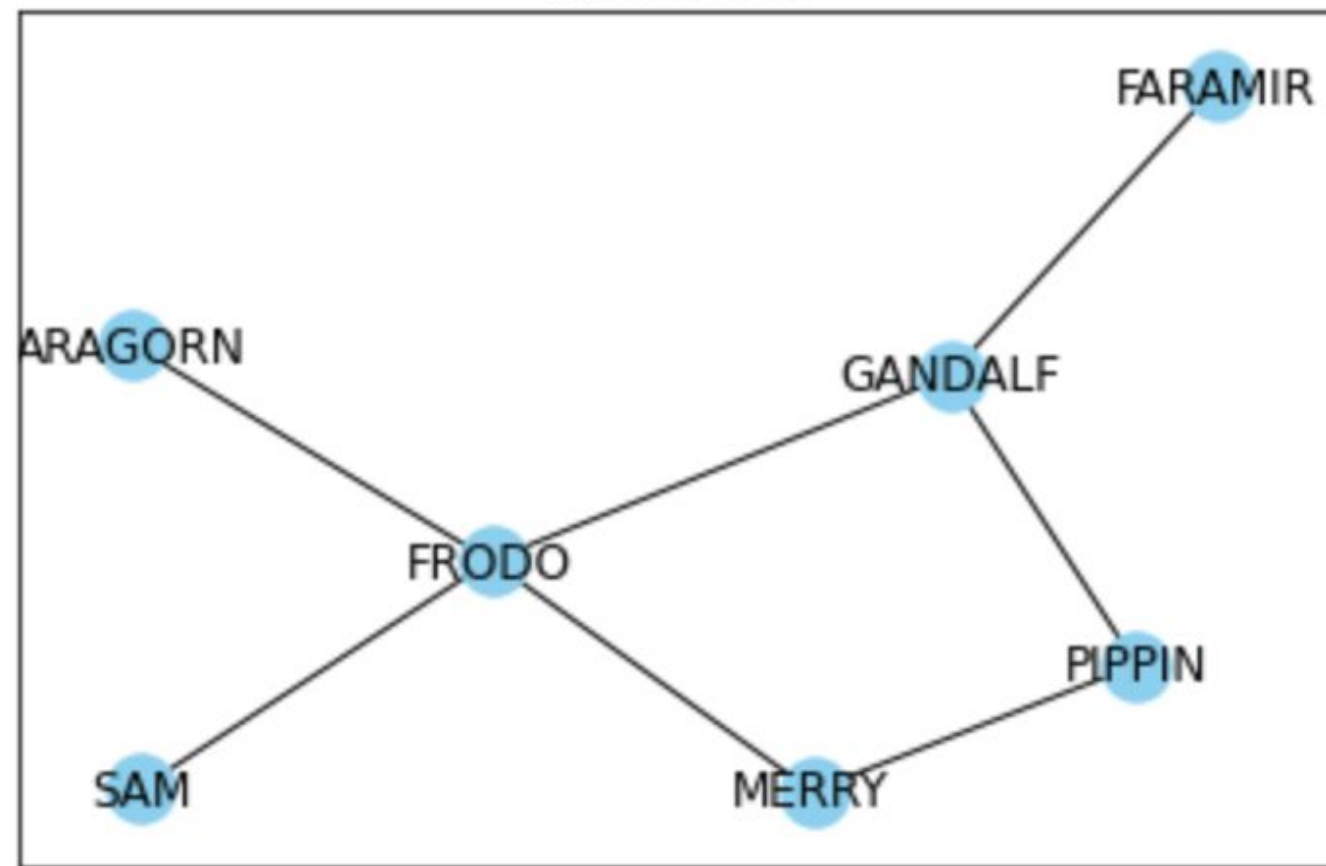
Threshold:1



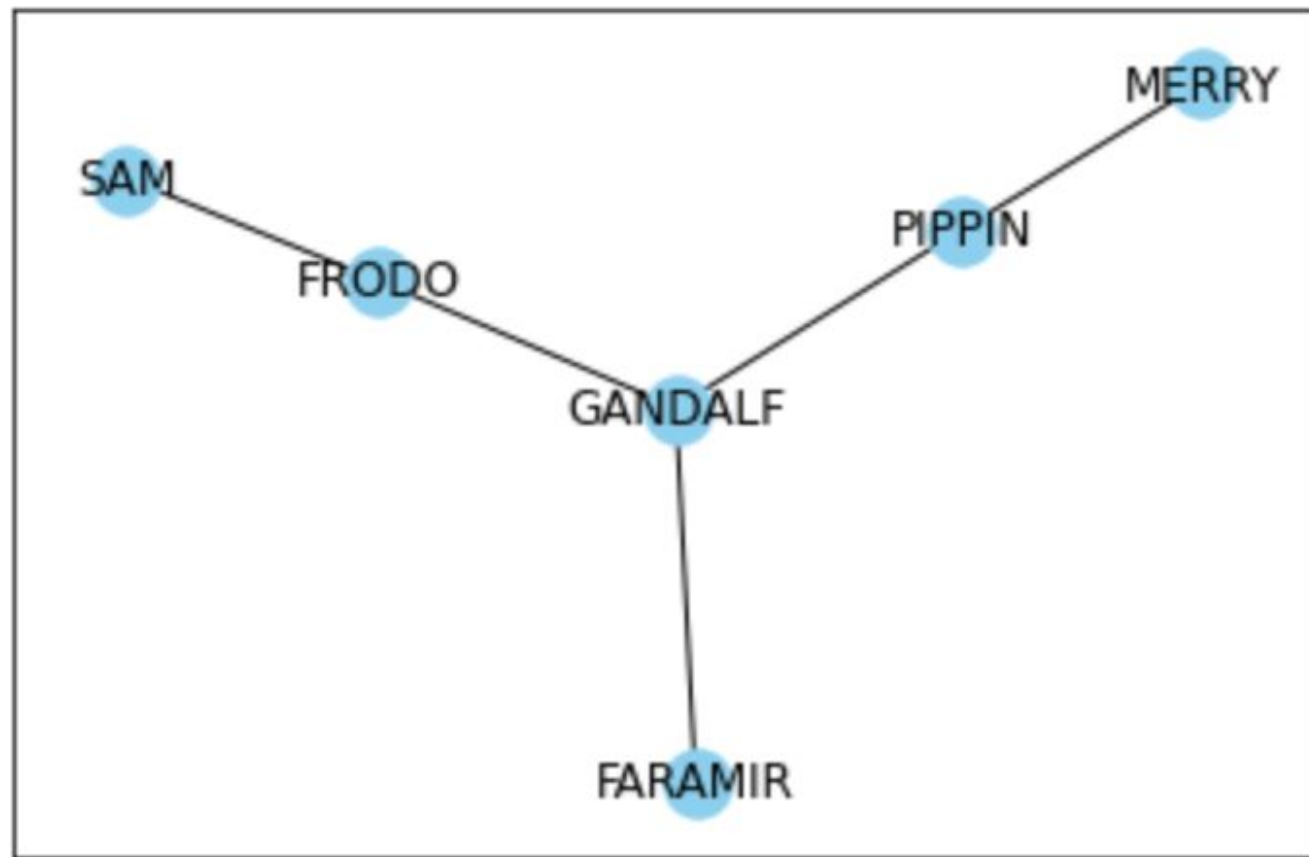
Threshold:2



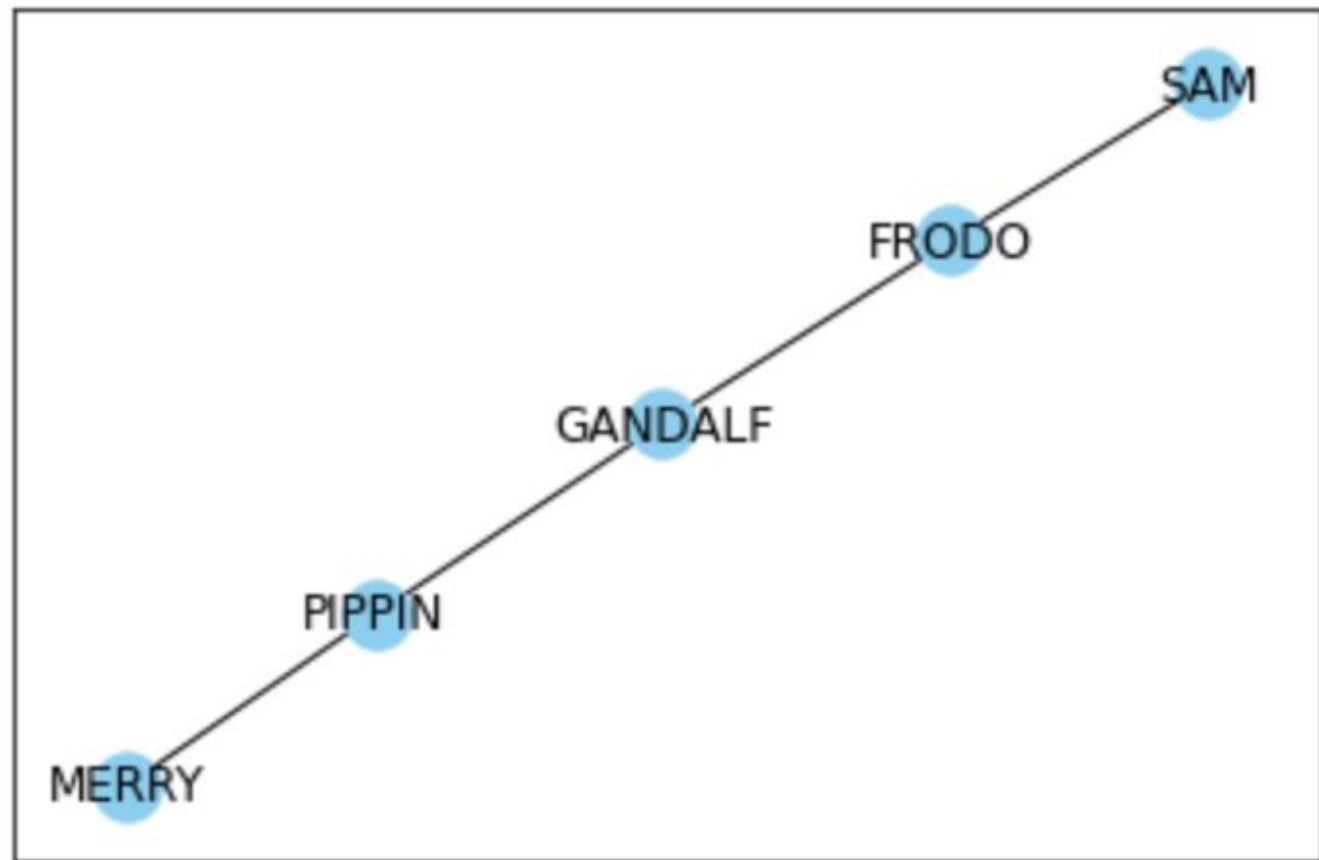
Threshold:3



Threshold:4

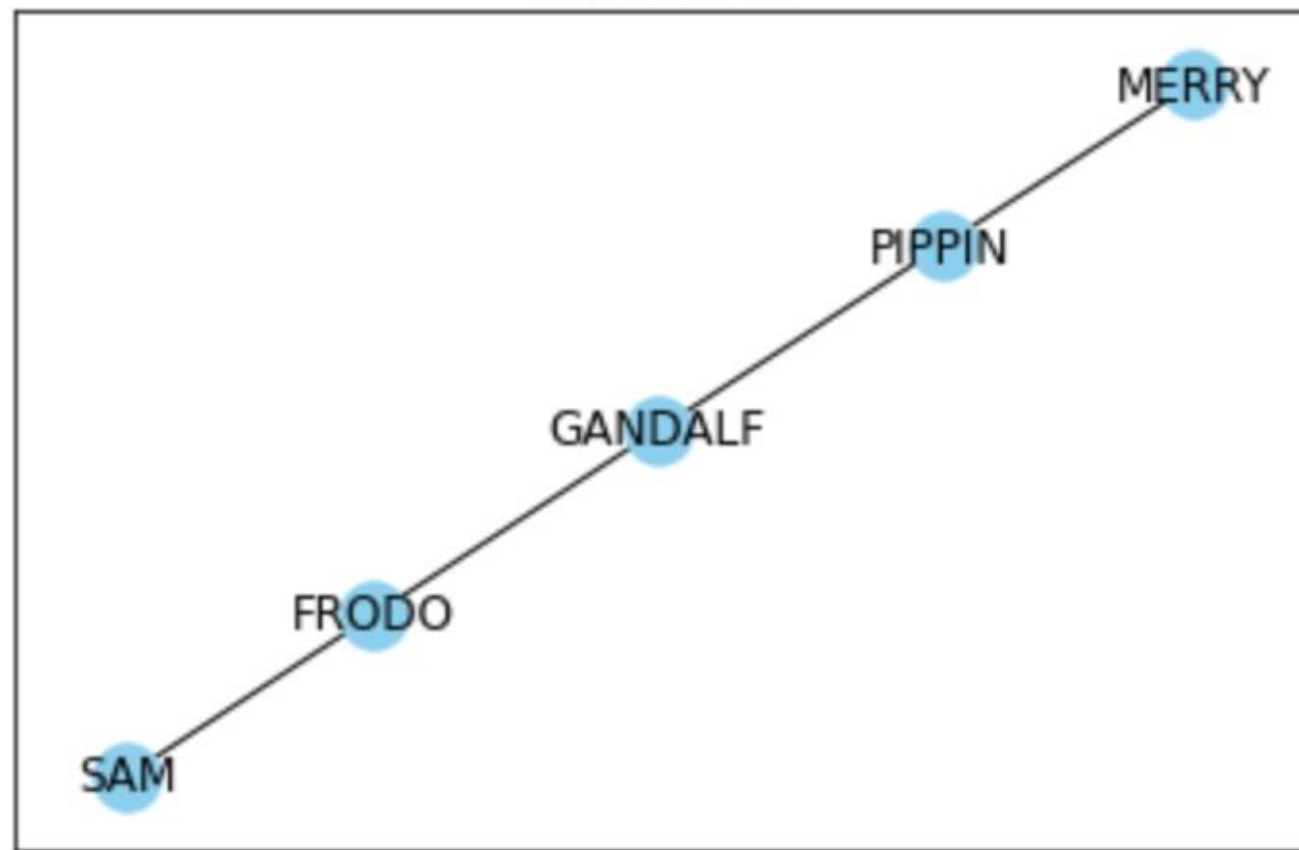


Threshold:5





Threshold:7

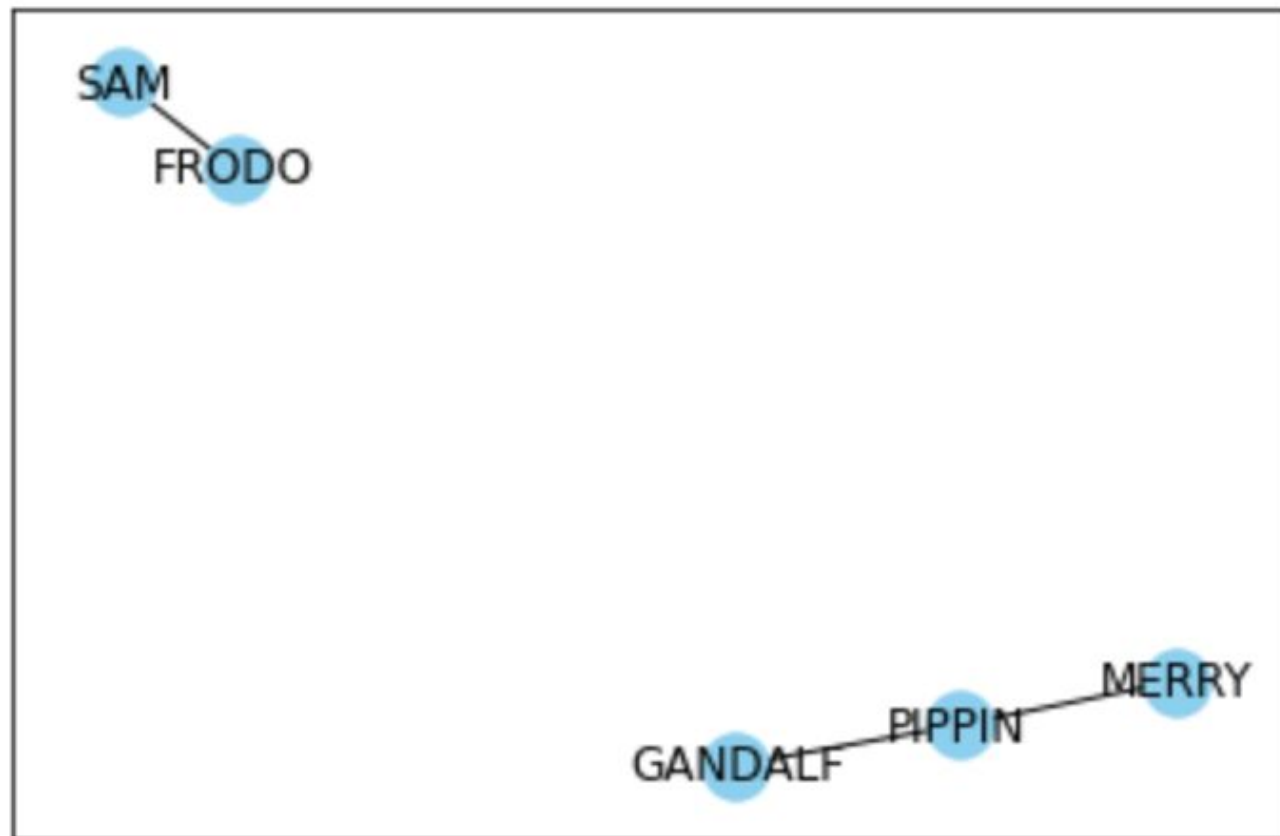


Threshold:8

FRODO — SAM

GANDALF — PIPPIN — MERRY

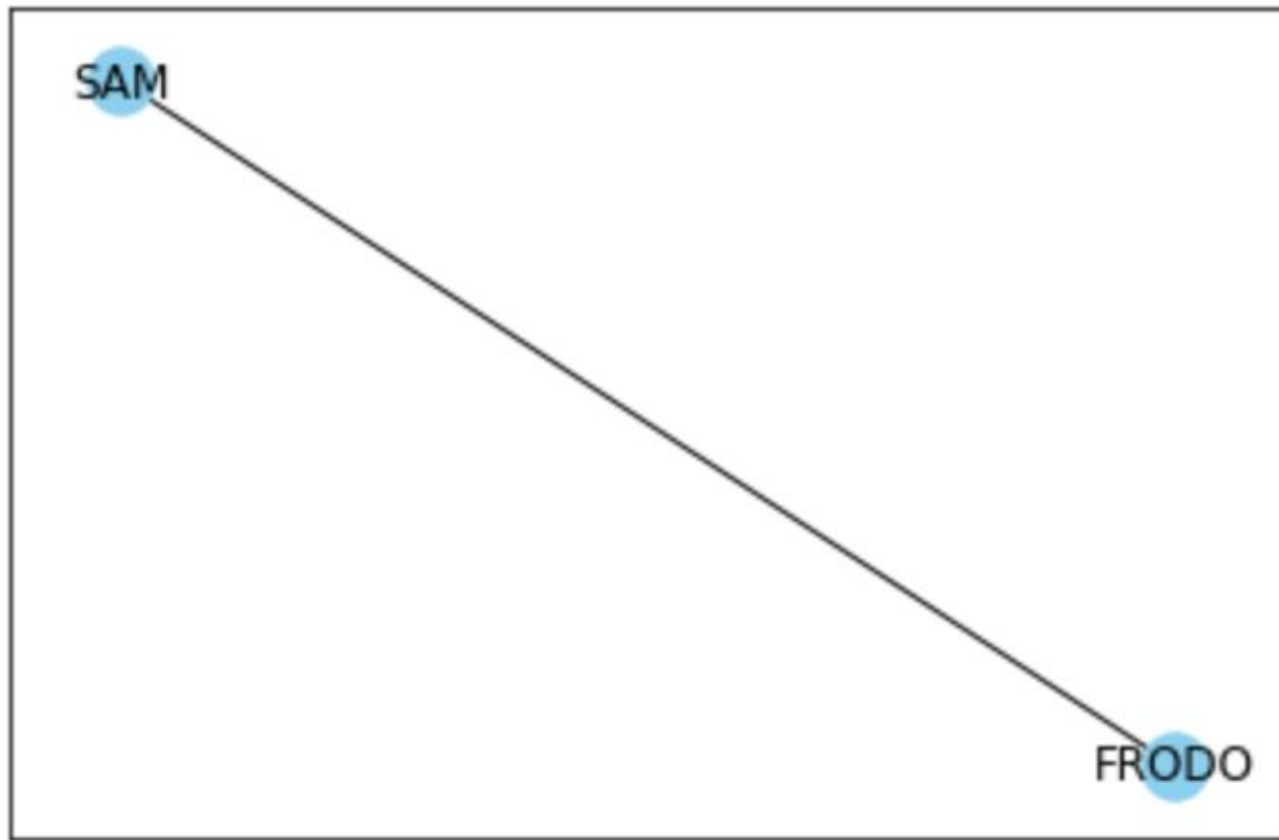
Threshold:9



Threshold:10



Threshold:13



# Summary Stats

	Max Thresh	Chars at Max Thresh	Comments
<b>Movie #1</b>	6	Sam, Frodo	Has the smallest # of chars
<b>Movie #2</b>	7	Sam, Frodo	Has cycles
<b>Movie #3</b>	13	Sam, Frodo	Has highest thresh



## 2. Character Growth

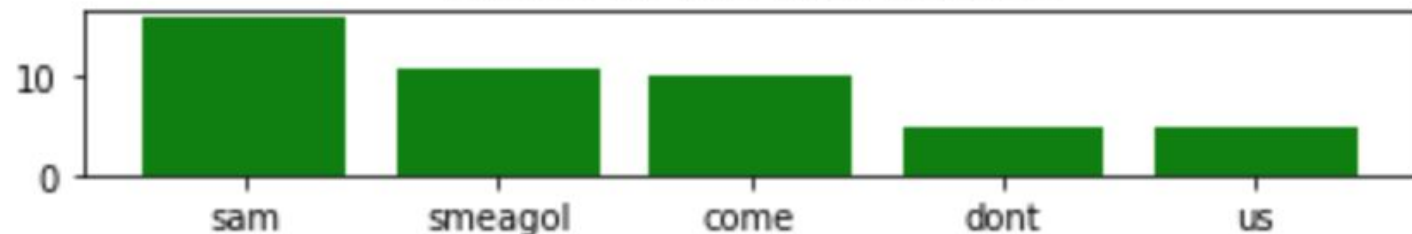
- Previously, bar plots were used to analyze the top 5 words for a character for the entire trilogy
- **Problem?** Some characters do not have lines in all the movies
- Character growth can mean change of vocabulary or even just more speaking lines



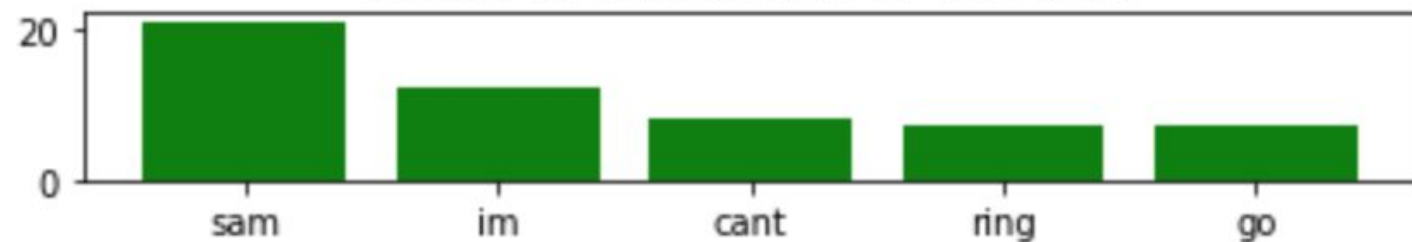
FRODO in: THE FELLOWSHIP OF THE RING



FRODO in: THE TWO TOWERS

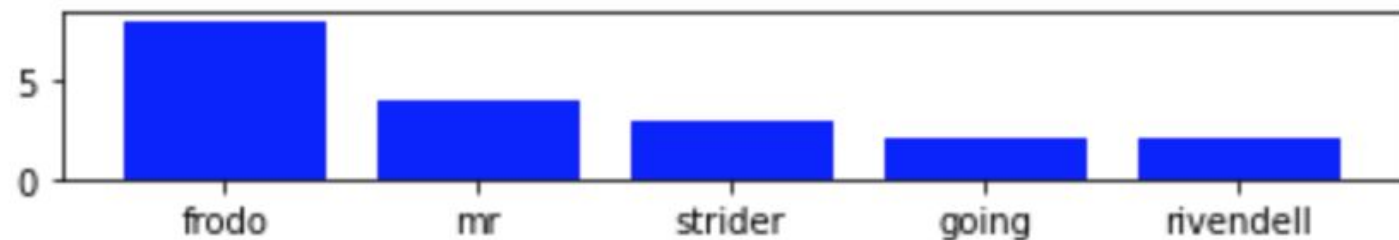


FRODO in: THE RETURN OF THE KING

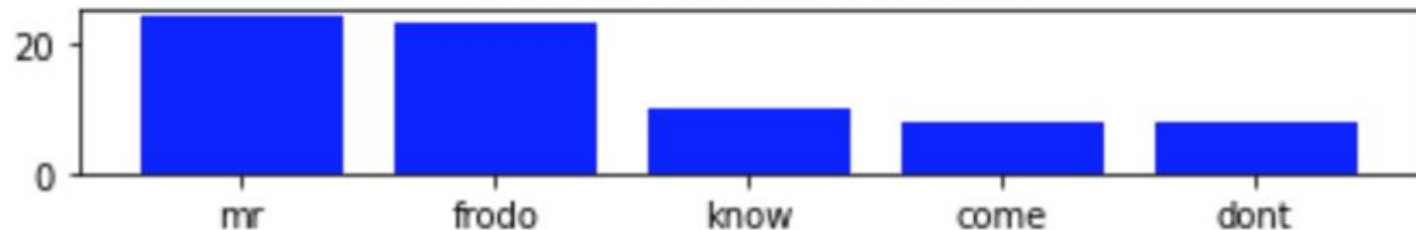




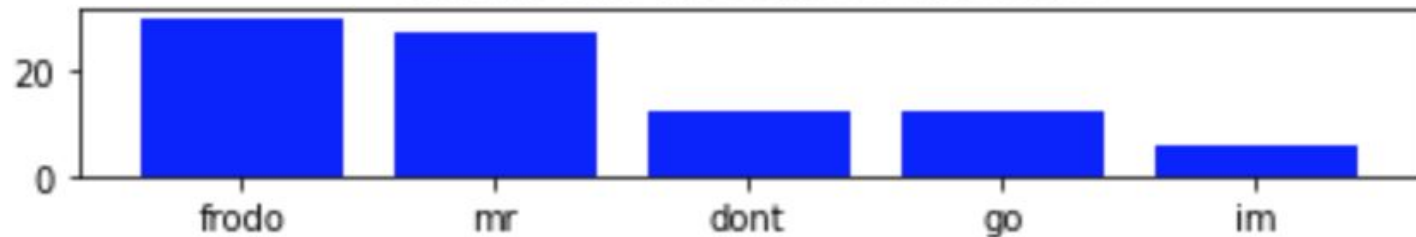
SAM in: THE FELLOWSHIP OF THE RING



SAM in: THE TWO TOWERS



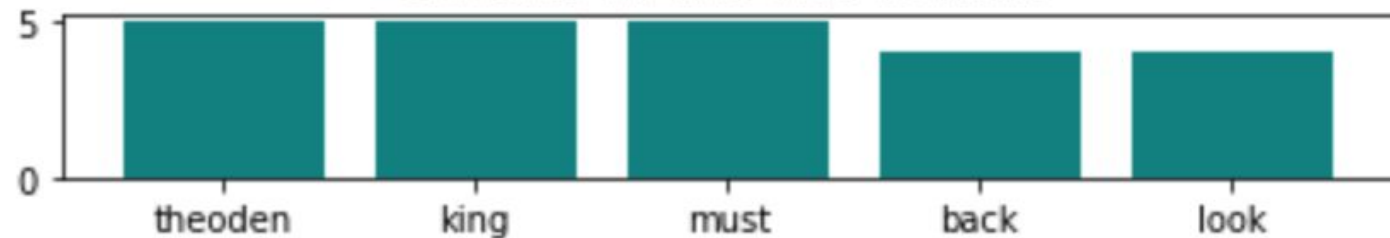
SAM in: THE RETURN OF THE KING



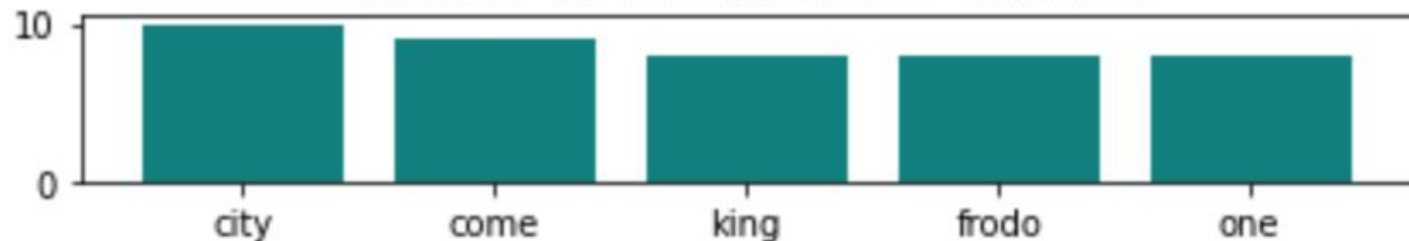
GANDALF in: THE FELLOWSHIP OF THE RING



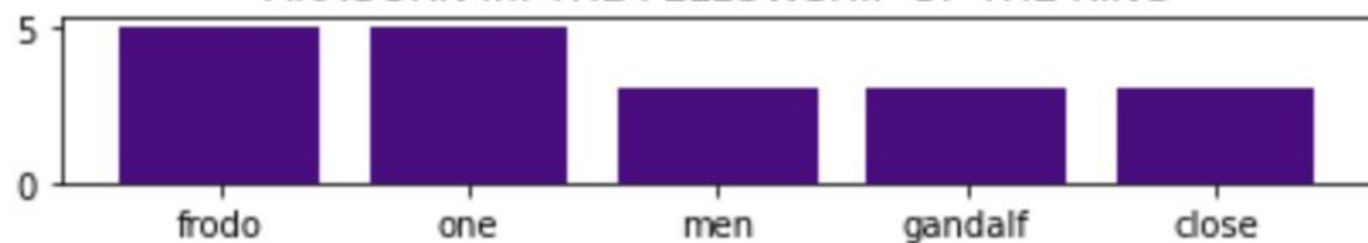
GANDALF in: THE TWO TOWERS



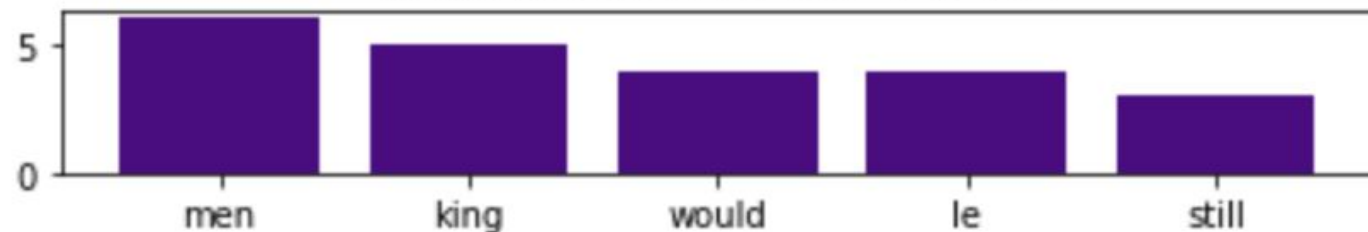
GANDALF in: THE RETURN OF THE KING



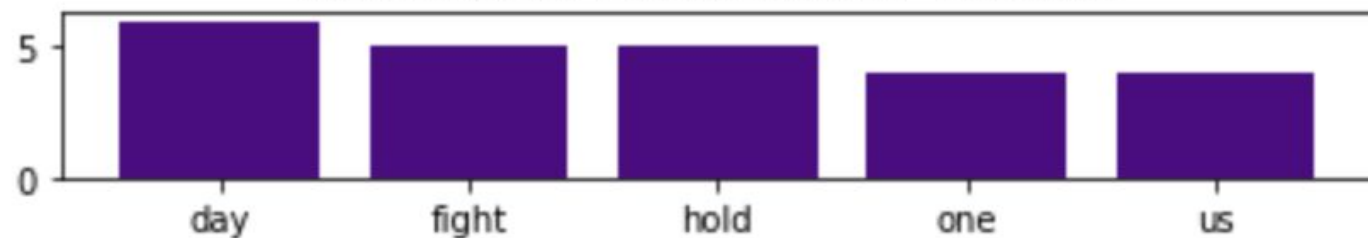
ARAGORN in: THE FELLOWSHIP OF THE RING



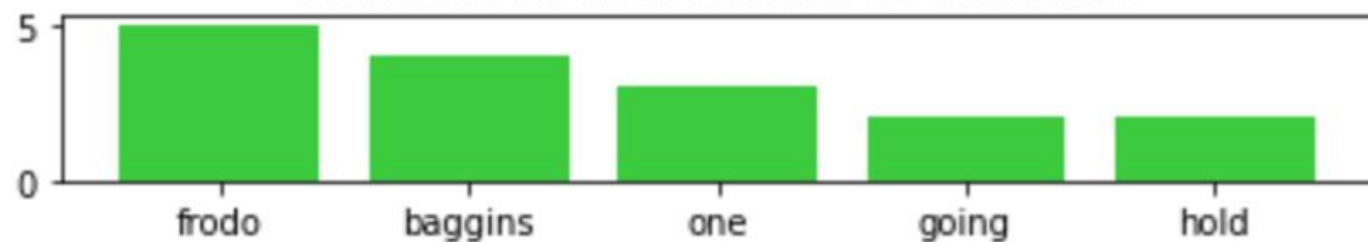
ARAGORN in: THE TWO TOWERS



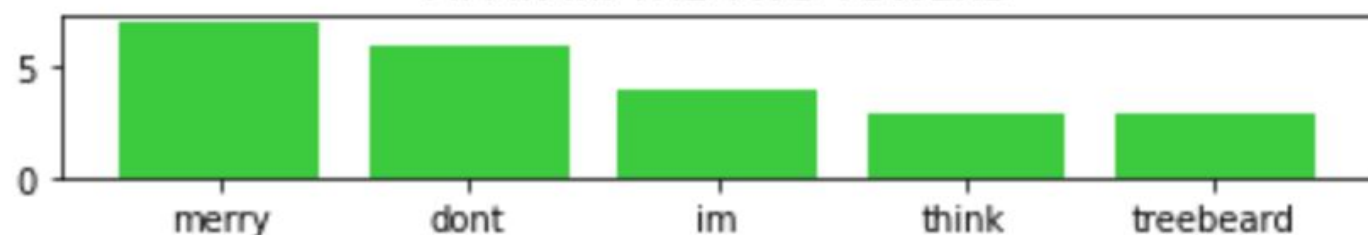
ARAGORN in: THE RETURN OF THE KING



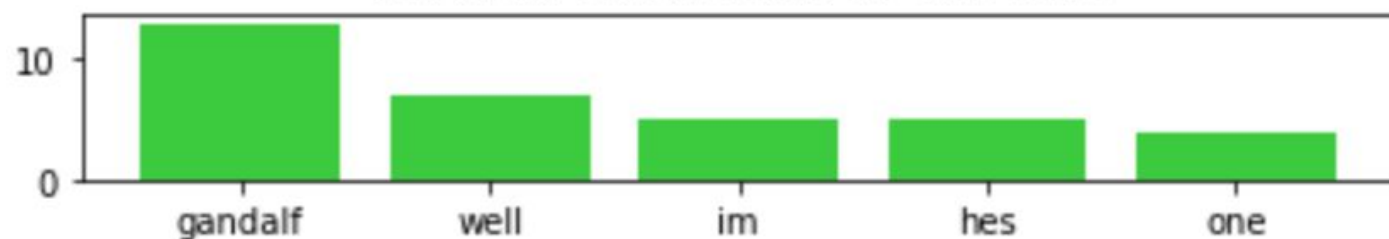
PIPPIN in: THE FELLOWSHIP OF THE RING



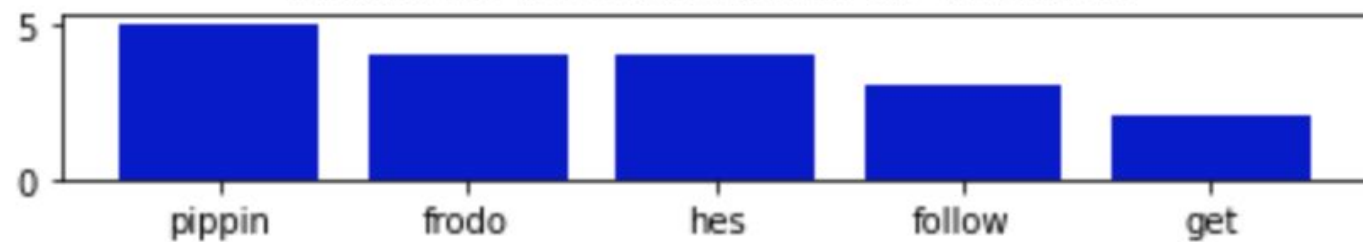
PIPPIN in: THE TWO TOWERS



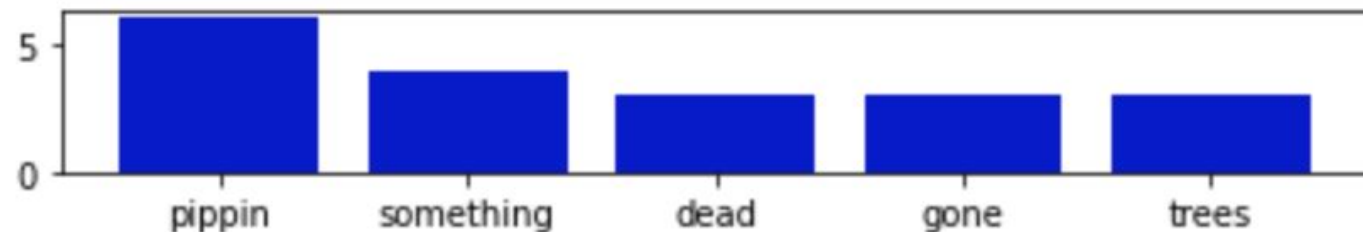
PIPPIN in: THE RETURN OF THE KING



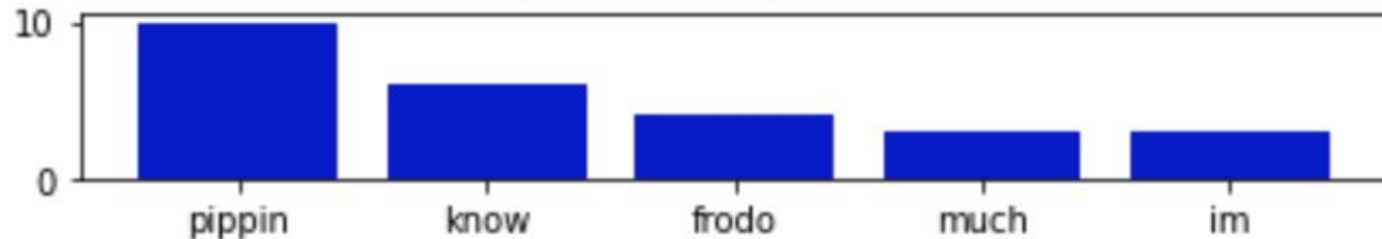
MERRY in: THE FELLOWSHIP OF THE RING



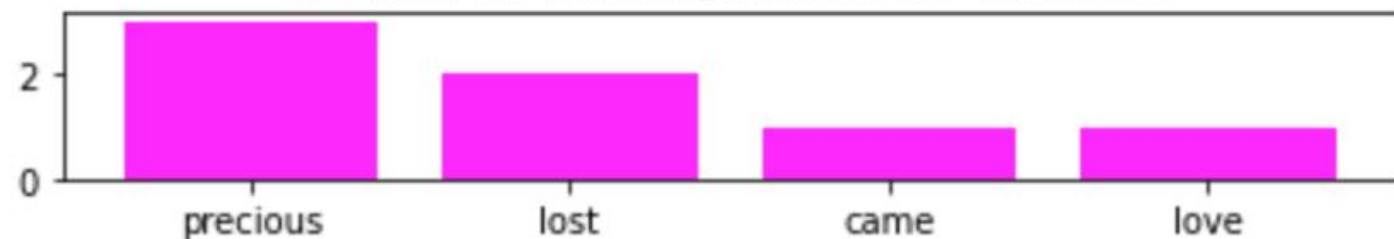
MERRY in: THE TWO TOWERS



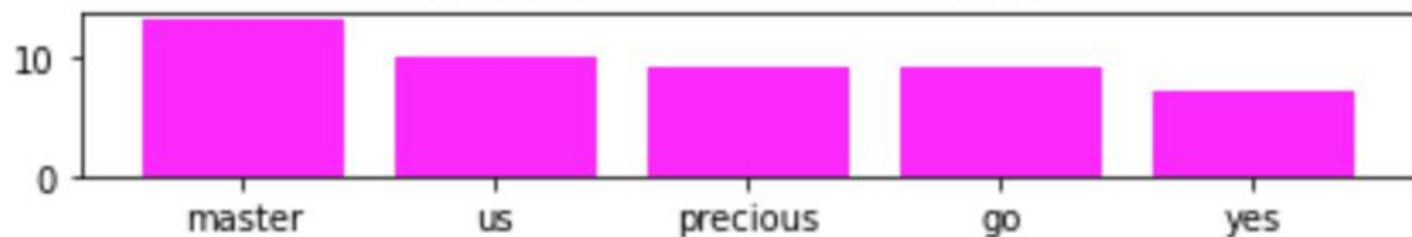
MERRY in: THE RETURN OF THE KING



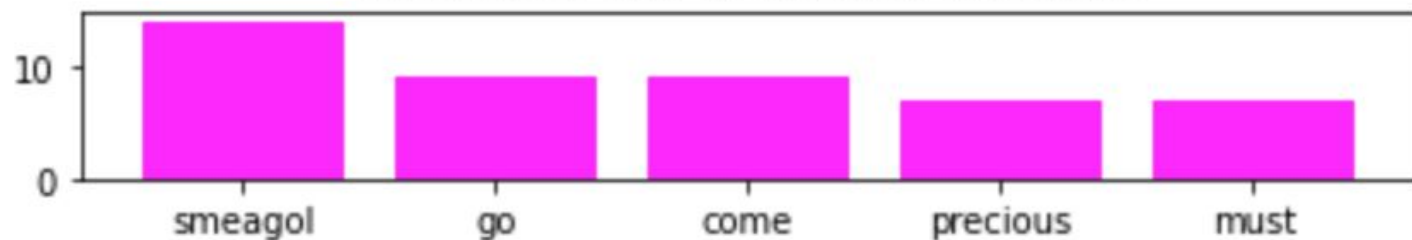
GOLLUM in: THE FELLOWSHIP OF THE RING



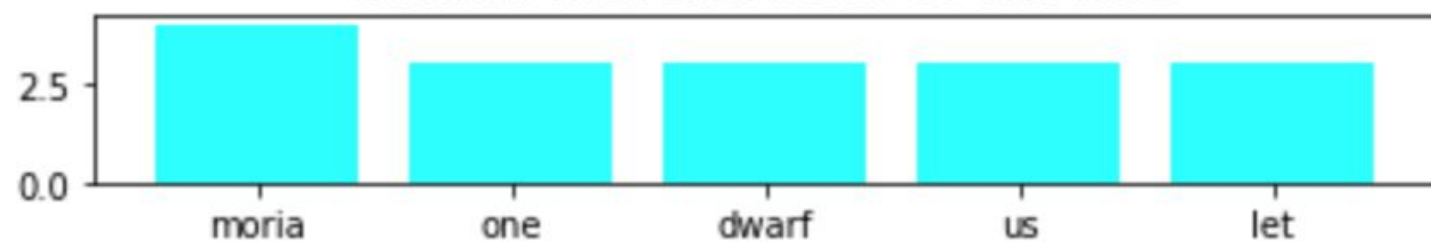
GOLLUM in: THE TWO TOWERS



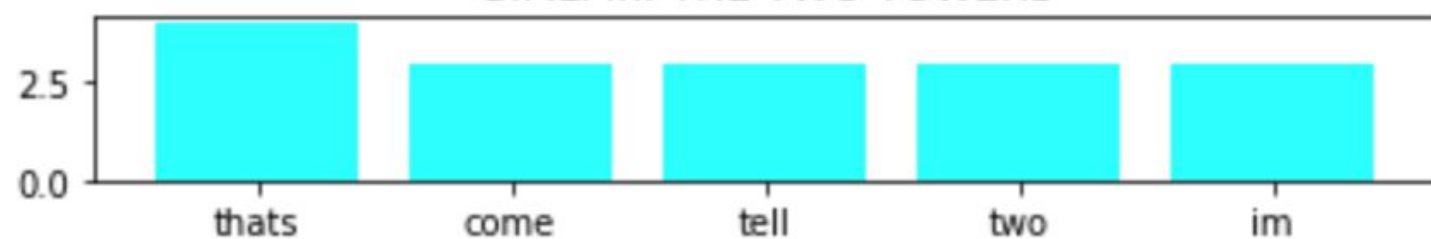
GOLLUM in: THE RETURN OF THE KING



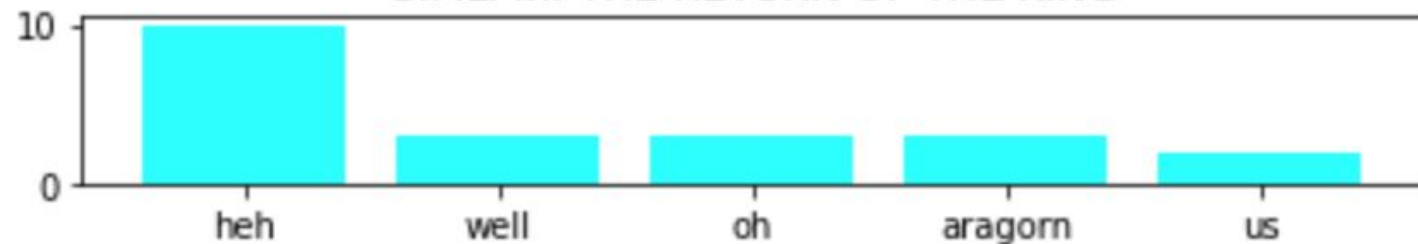
GIMLI in: THE FELLOWSHIP OF THE RING



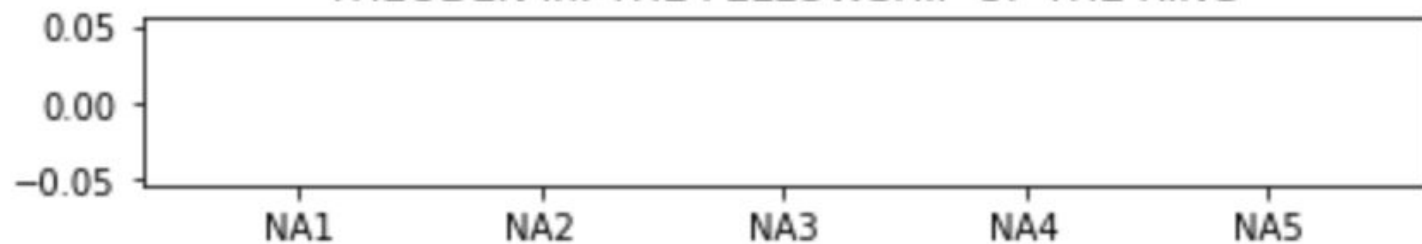
GIMLI in: THE TWO TOWERS



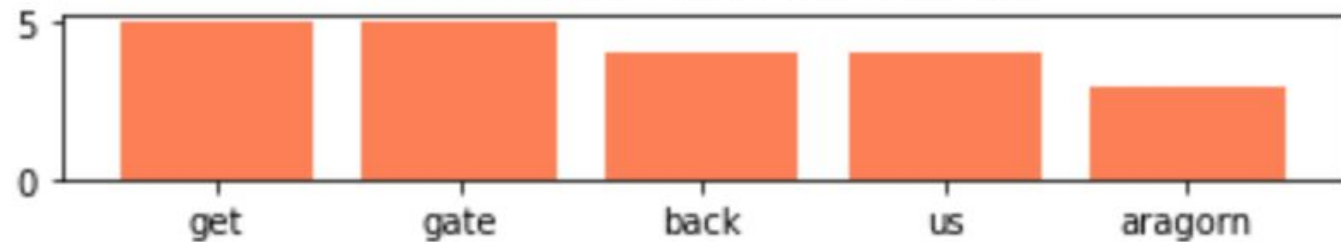
GIMLI in: THE RETURN OF THE KING



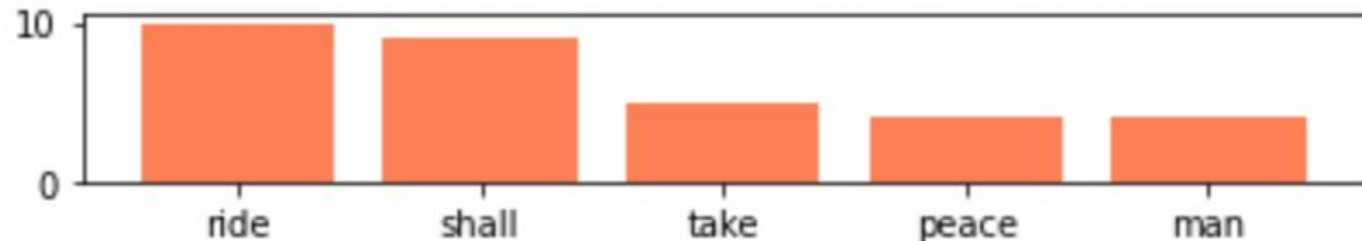
THEODEN in: THE FELLOWSHIP OF THE RING



THEODEN in: THE TWO TOWERS

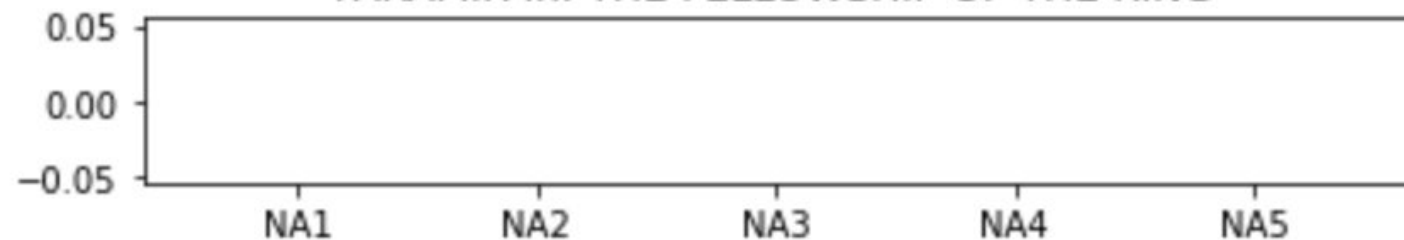


THEODEN in: THE RETURN OF THE KING

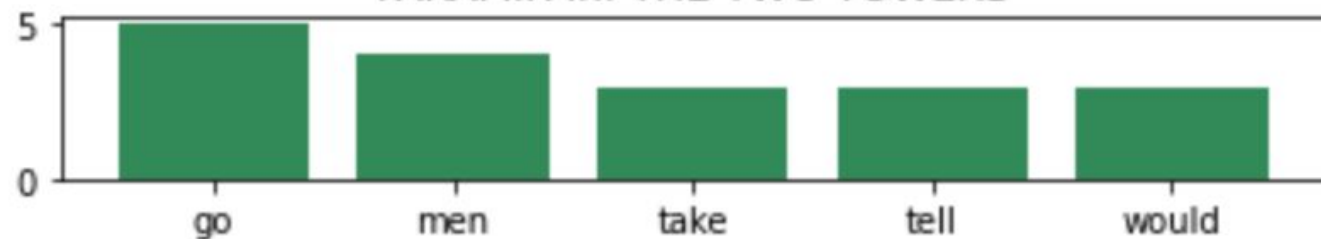




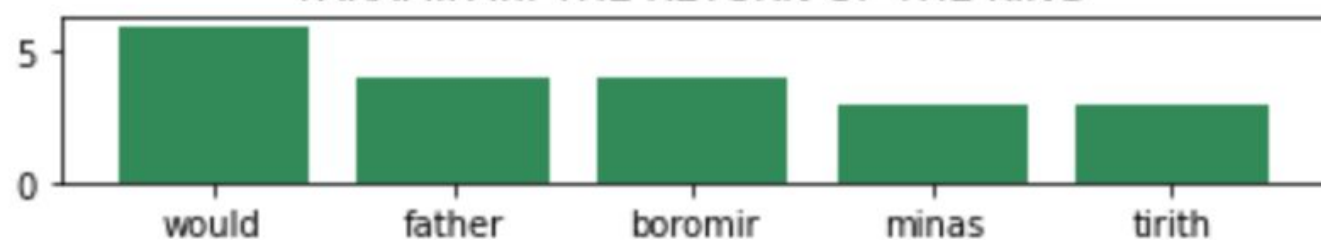
FARAMIR in: THE FELLOWSHIP OF THE RING



FARAMIR in: THE TWO TOWERS



FARAMIR in: THE RETURN OF THE KING

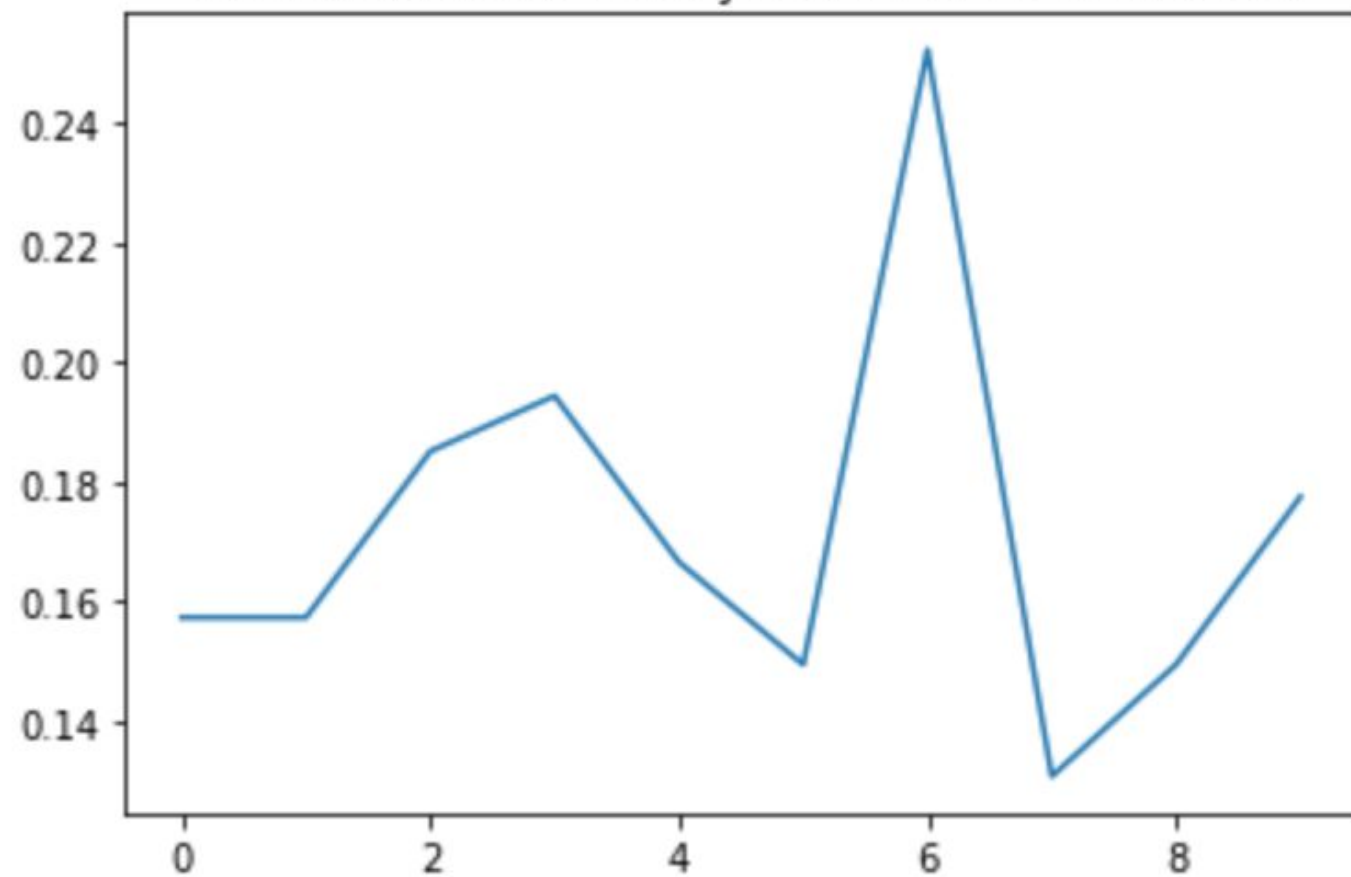


### 3. Who said it? ML

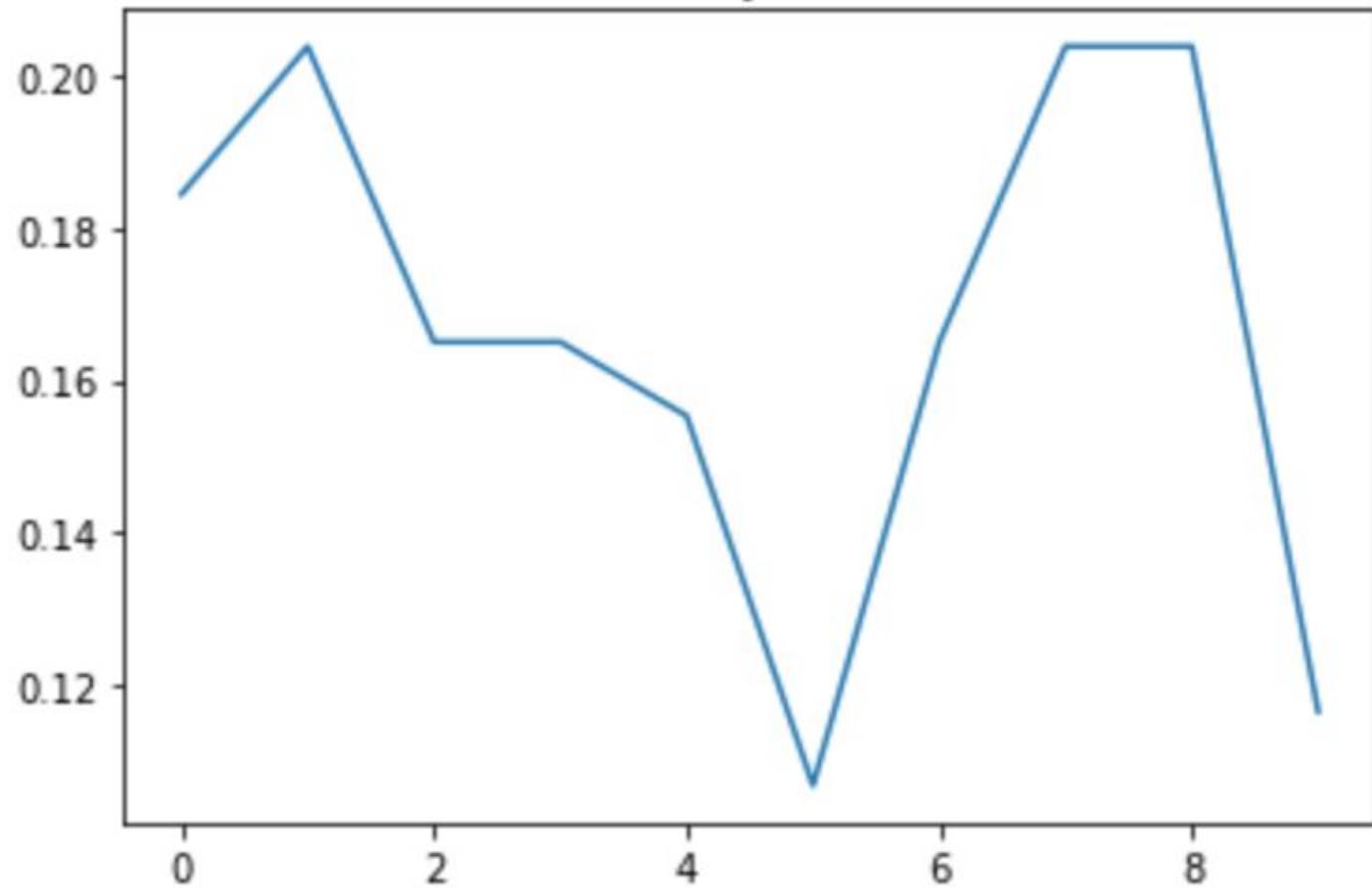
- Used Random Forest Classifier trained on optimal inputs derived using RandomizedSearchCV
- Used accuracy as scoring metric



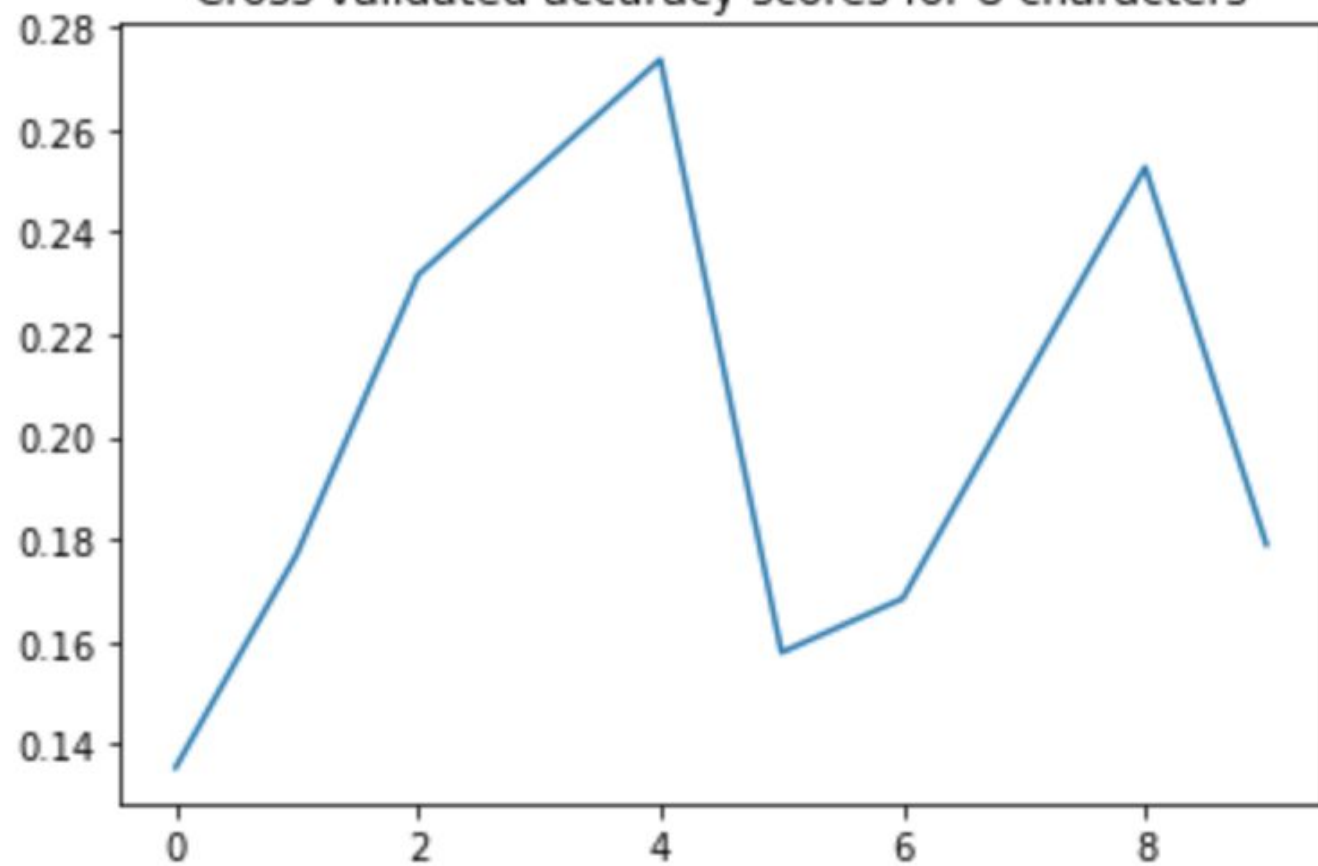
Cross validated accuracy scores for 10 characters



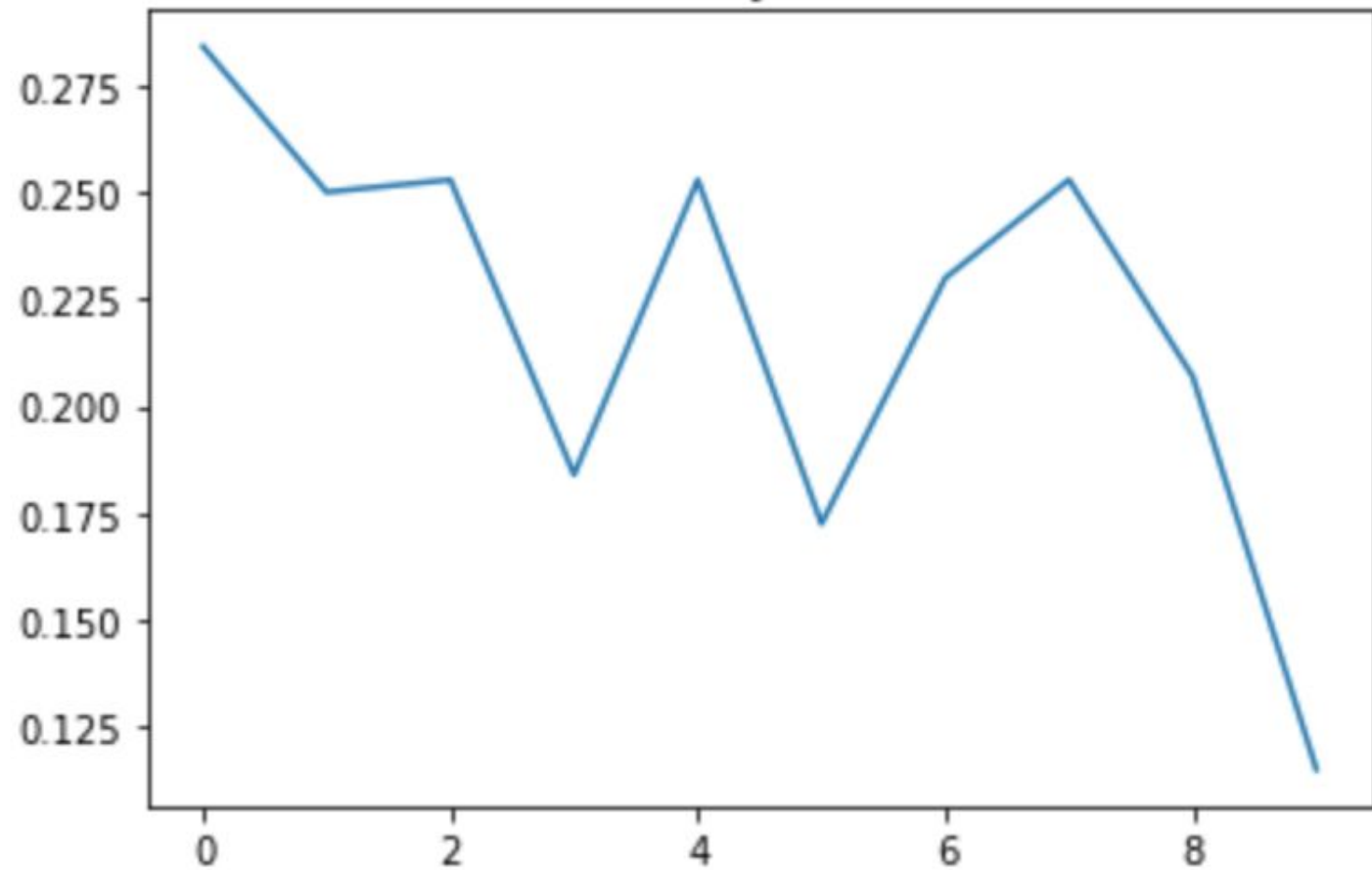
Cross validated accuracy scores for 9 characters



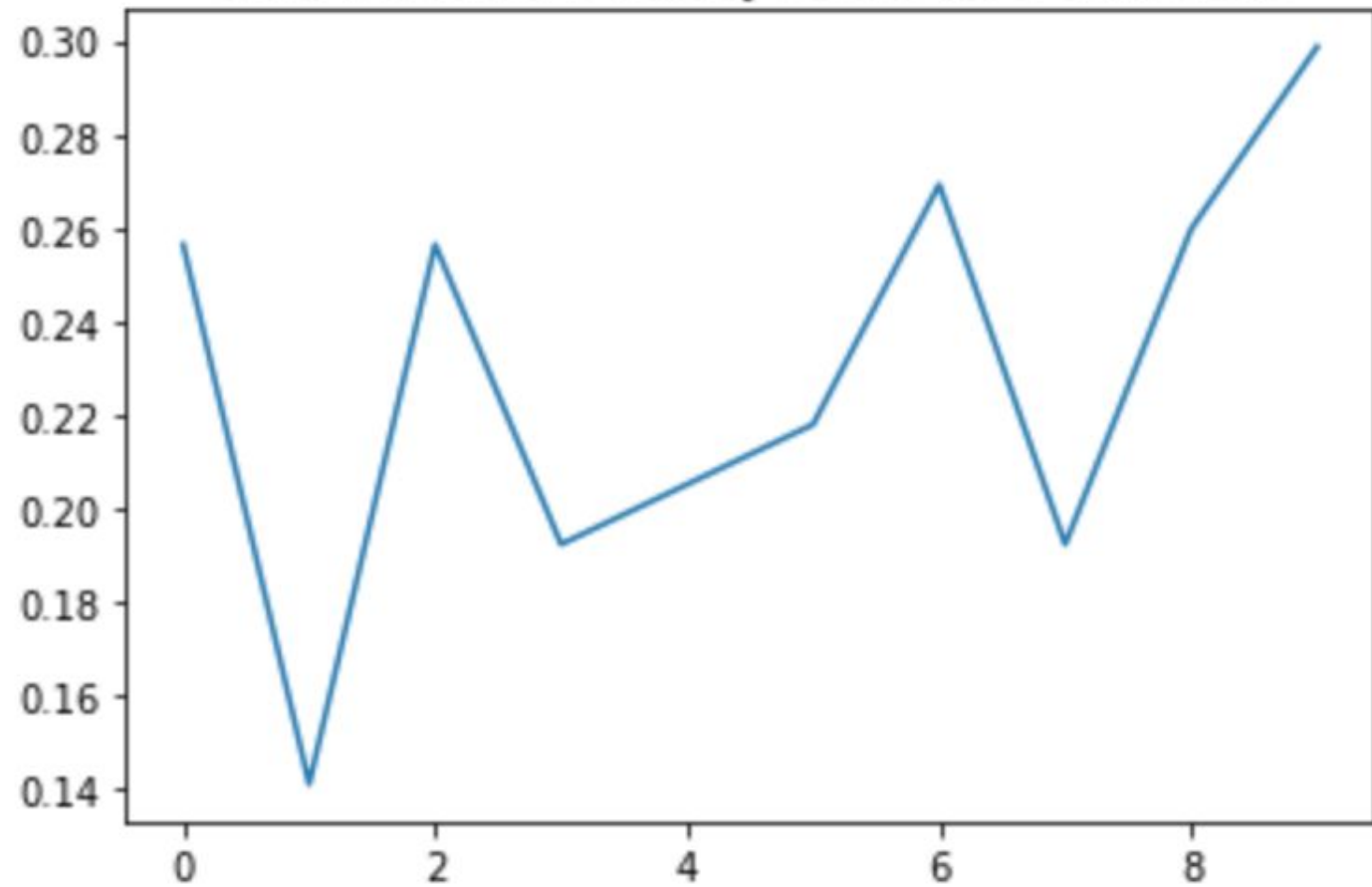
Cross validated accuracy scores for 8 characters



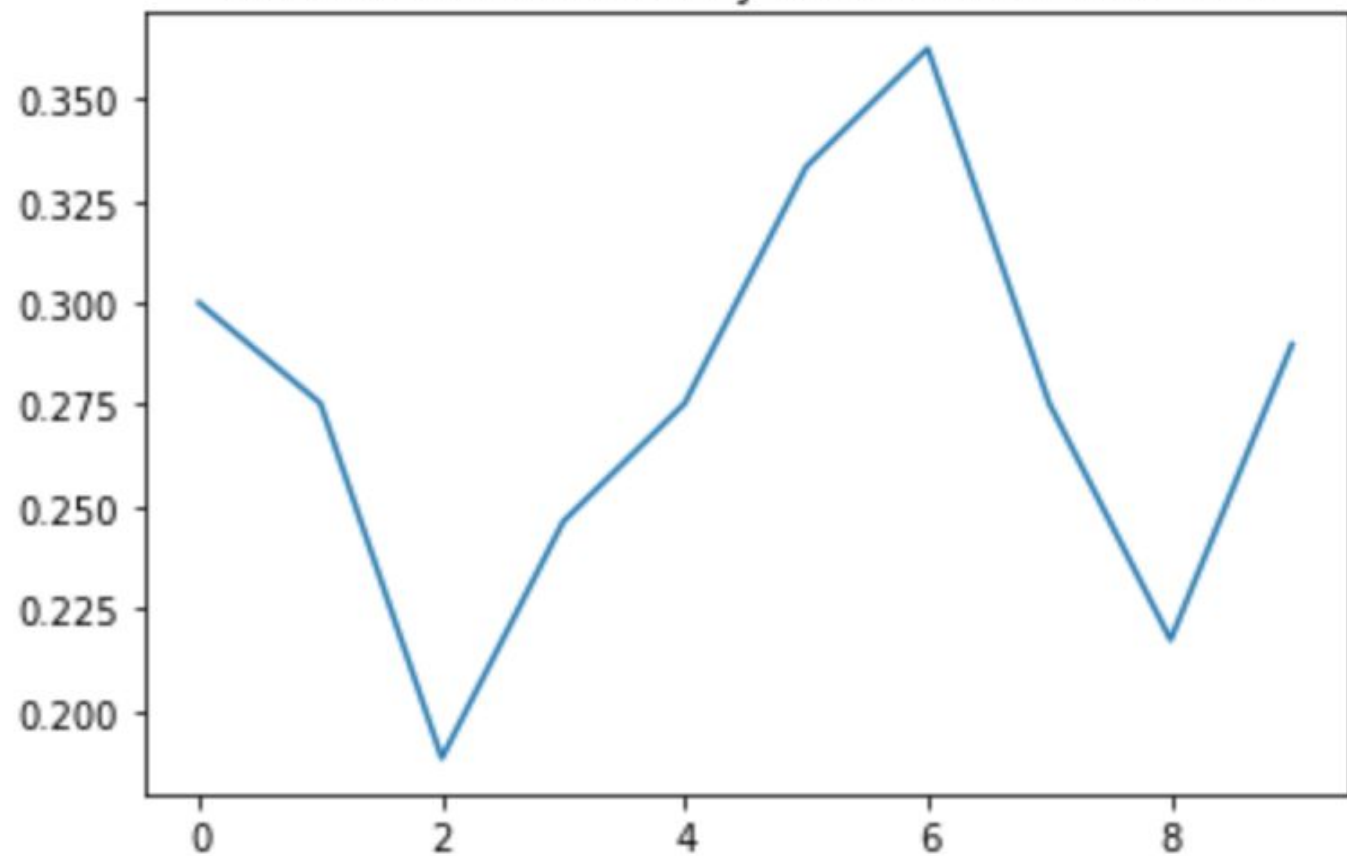
Cross validated accuracy scores for 7 characters



Cross validated accuracy scores for 6 characters

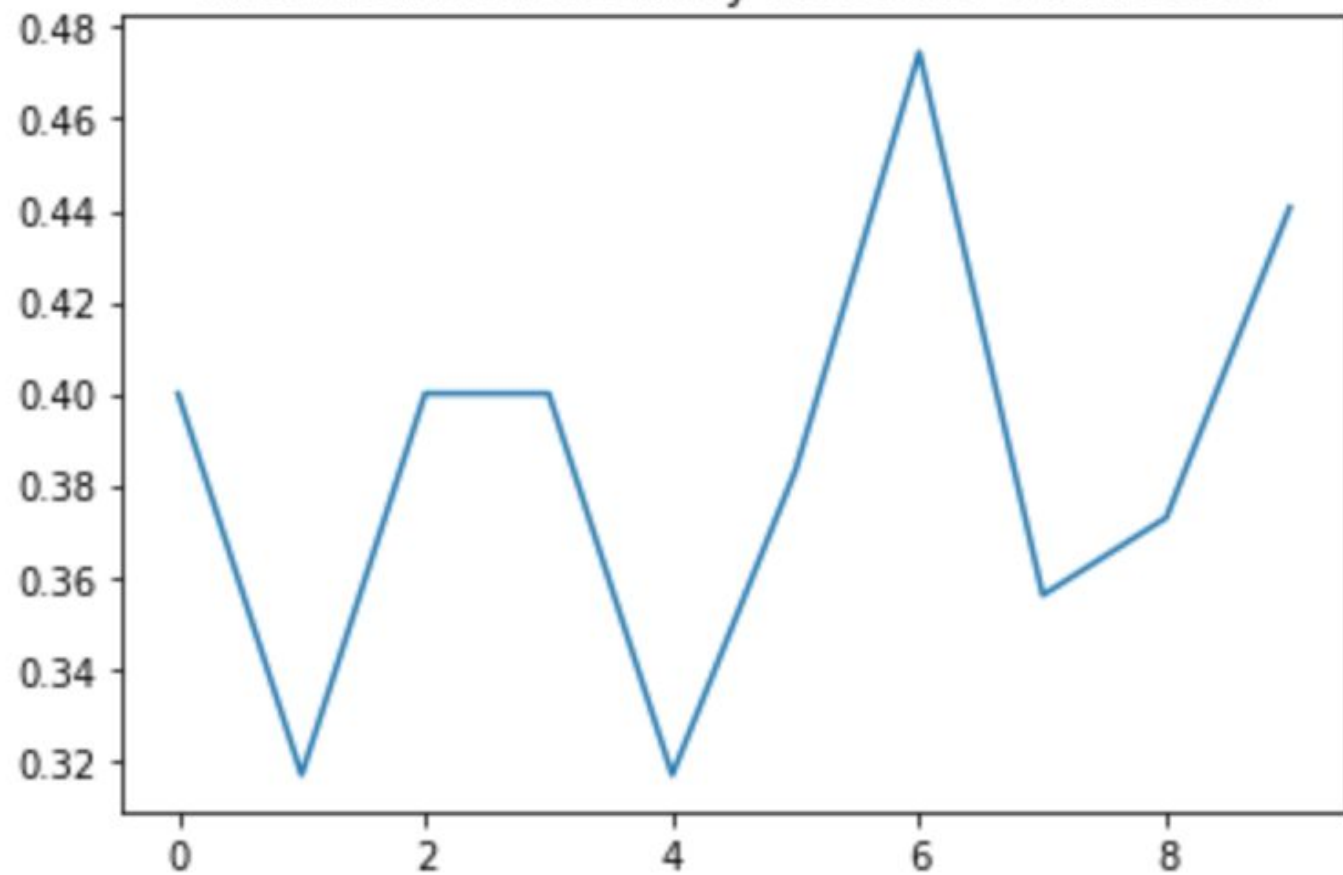


Cross validated accuracy scores for 5 characters

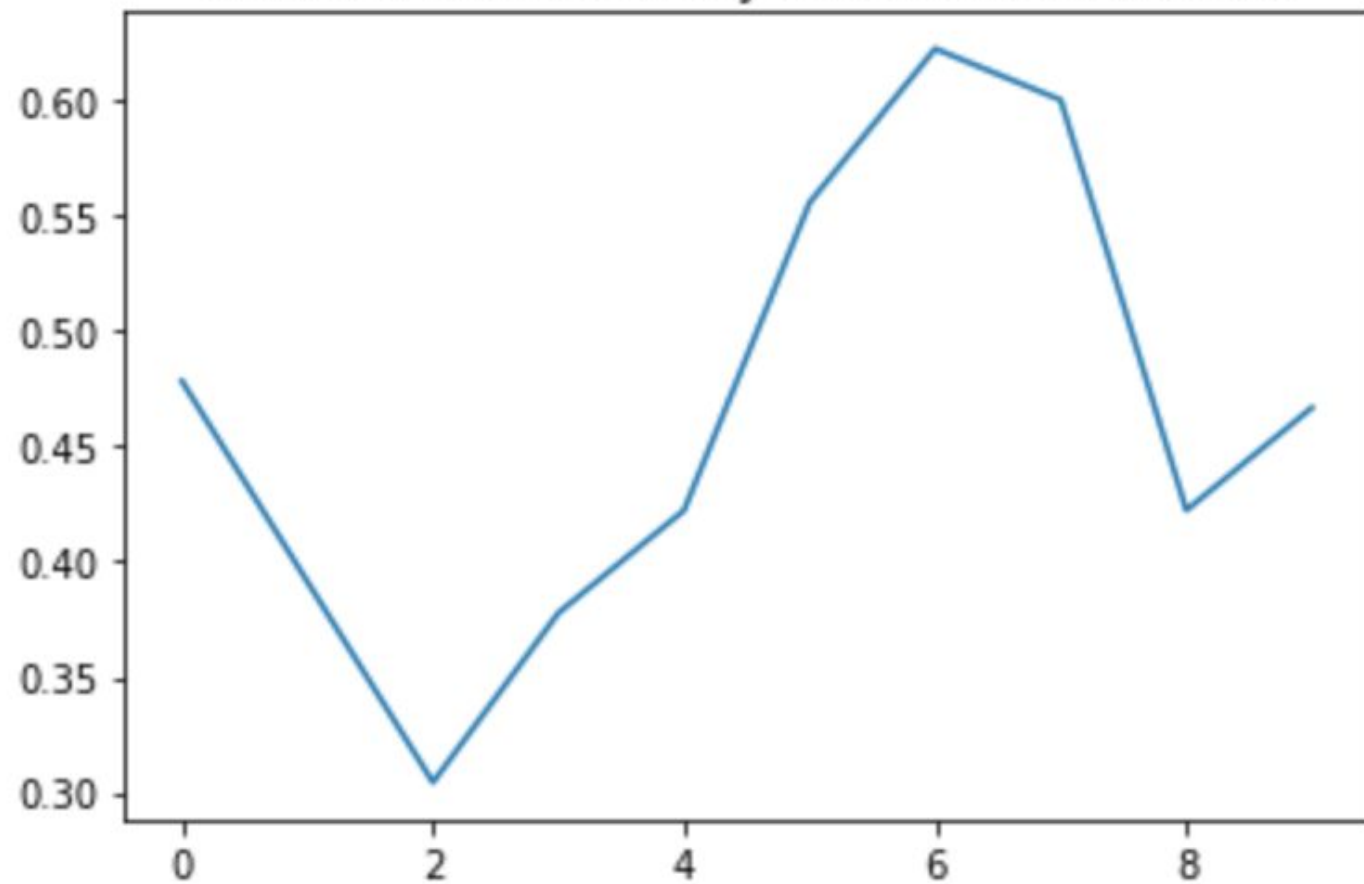




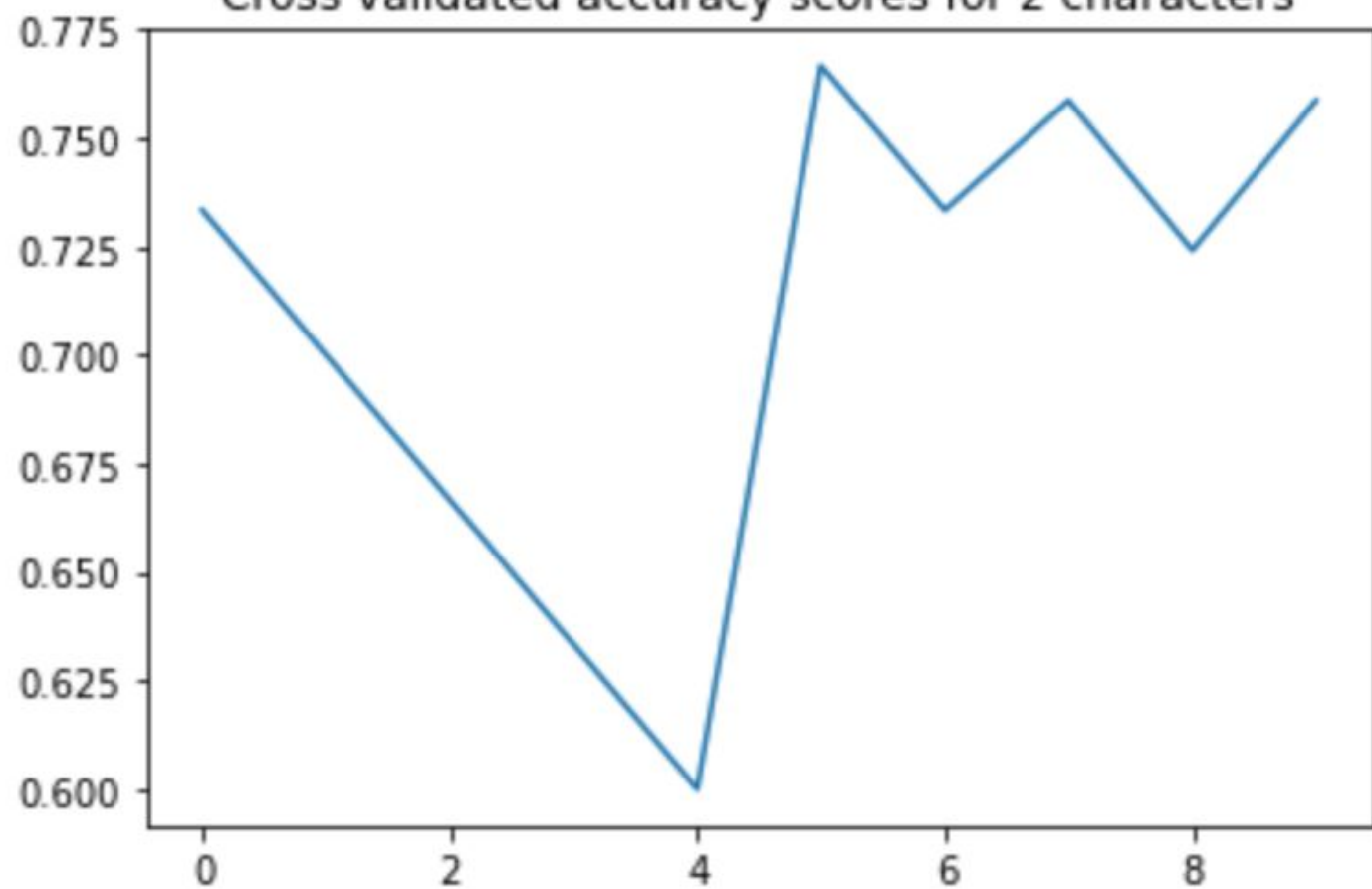
Cross validated accuracy scores for 4 characters



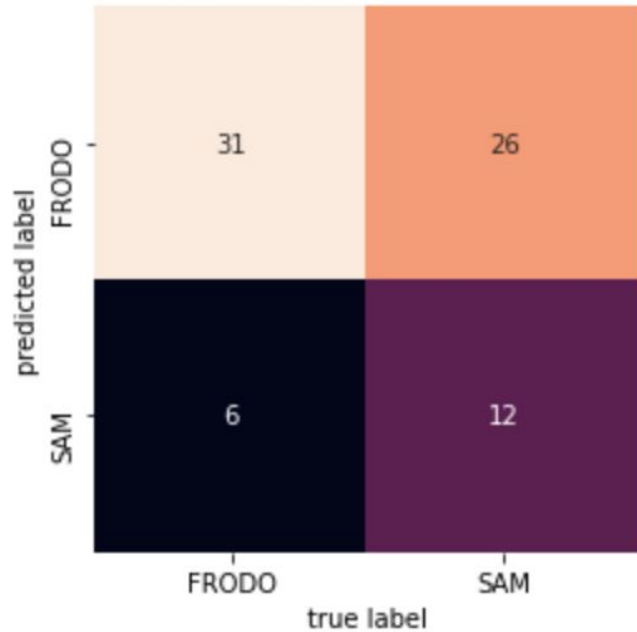
Cross validated accuracy scores for 3 characters



Cross validated accuracy scores for 2 characters



# Confusion matrix for 2 class model



Accuracy score: 0.5733333333333334

Execution time: 3.235245943069458 seconds

# ML Summary

# Classes	10	9	8	7	6	5	4	3	2
Max Accuracy	.25	.21	.27	.30	.30	.37	.47	.65	.76



# Conclusion

- With datasets that are not specifically designed for text classification or where there are multiple languages involved, text classification using standard techniques does not perform as well as one would like
- Even after hyperparameter tuning, every model had erratic CV scores
- For two classes, the best performance was under 80%





That's all folks!