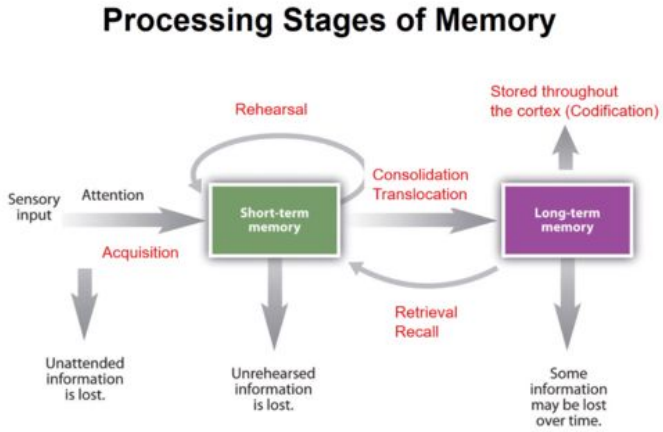
Who’s line is it?

By: Yaniv Bronshtein

Problem

- One of the most ubiquitous problems in the world of Broadway/Hollywood is an actor remembering their line. It actually often takes many cuts for a scene to be just right
- What if we could engineer the mapping to aid in long term mapping?



Solution

- This is an NLP(Natural Language Processing) Problem!
- For the sake of this midterm project, we will focus on all the steps but deliberately stop short of model building
- Goal: Analyze two movie scripts to figure out if given the vocabulary, one can deduce the character

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

Harry Potter

Source:

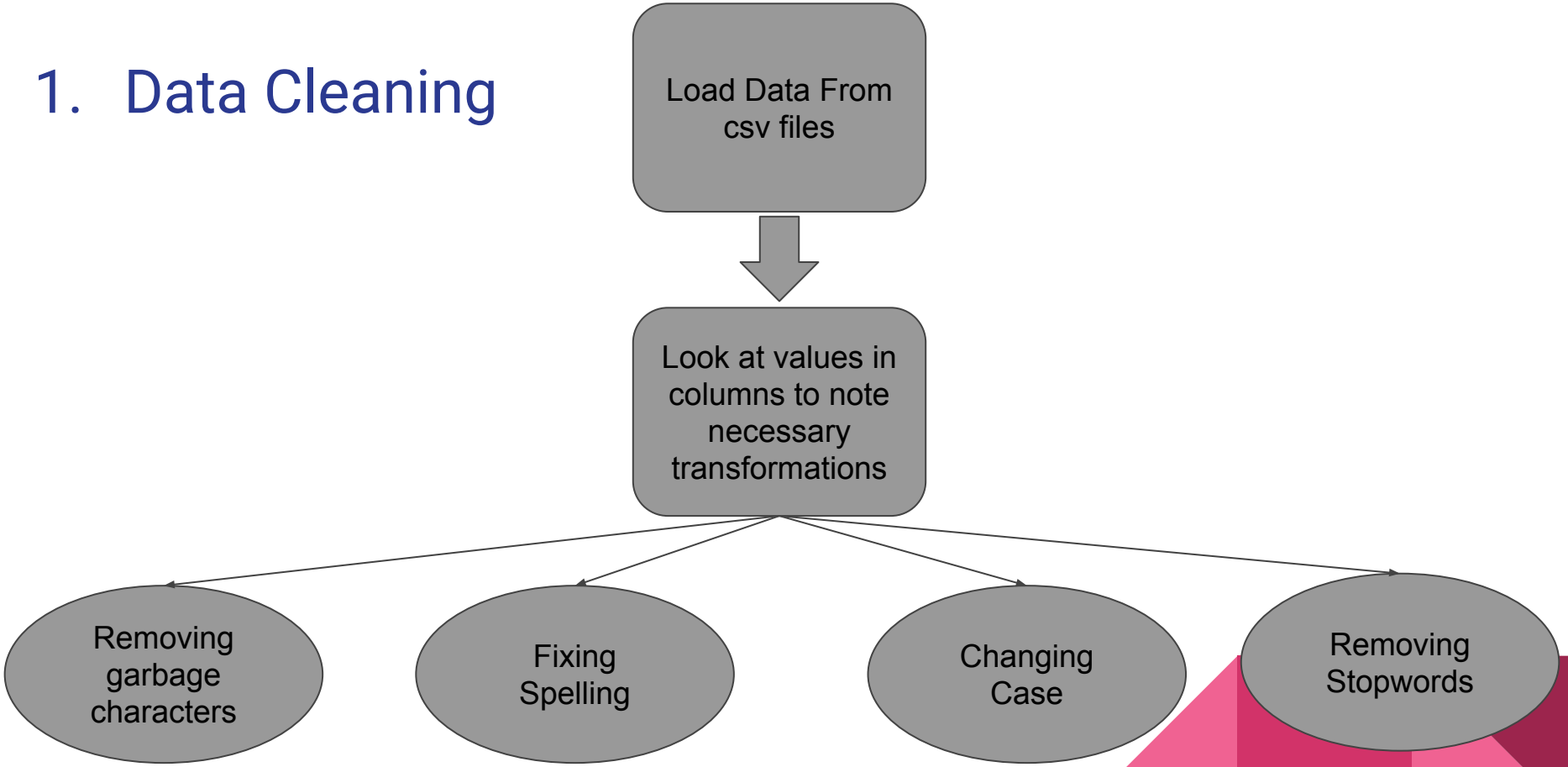
https://www.kaggle.com/gulsahdemiryurek/harry-potter-dataset

About:

- 7 csv files of which only ‘Harry Potter 1.csv’, ‘Harry Potter 2.csv’, and ‘Harry Potter 3.csv’ being used for this analysis
- Based on the first three Harry Potter movies
- Each csv file contains columns: ‘Character’ and ‘Sentence’

Pipeline/Methodology

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

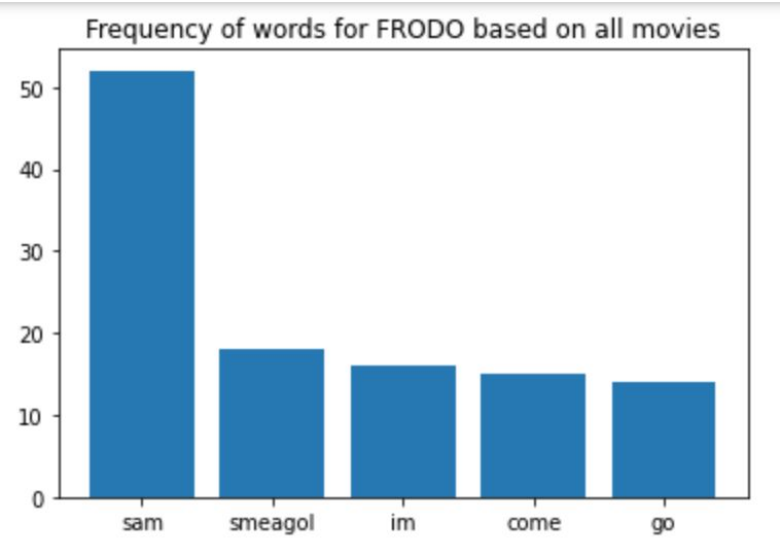
```
{
        Frodo: {
                bloop:2,
                bleh:1,
                lpsum:3
        },
        Gollum: {
        my:1,
        precious:2
        }
}
```

Additional Intermediate Steps

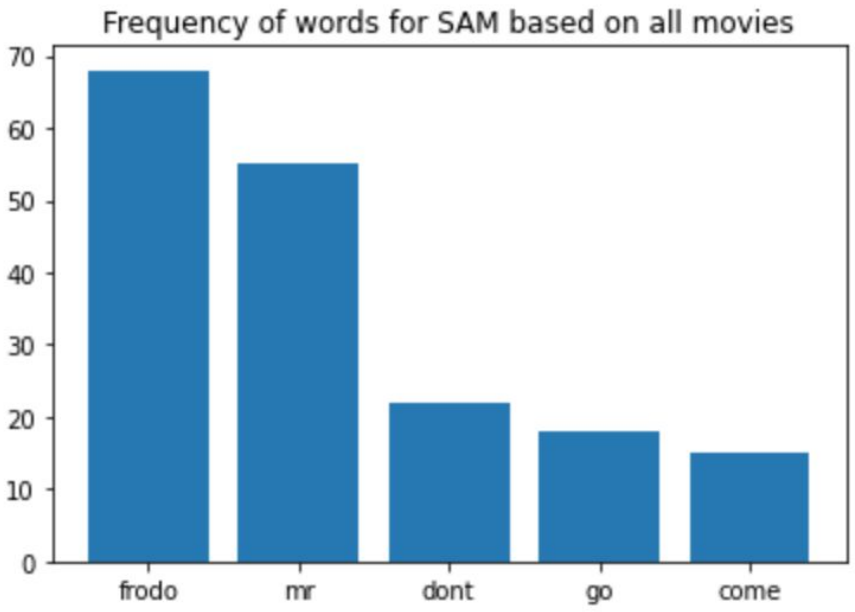
- Used value_counts() to get top 10 characters in each movie by the number of times their name appeared in the dataframe
- Used the line below to get the top 5 words for each character
  - top5_words = sorted(data, key=data.get, reverse=True)[:5]
- Merging the three Harry Potter movie dataframes
- Creating extra Movie column for future analysis in Harry Potter dataframe

Results: Lord of The Rings

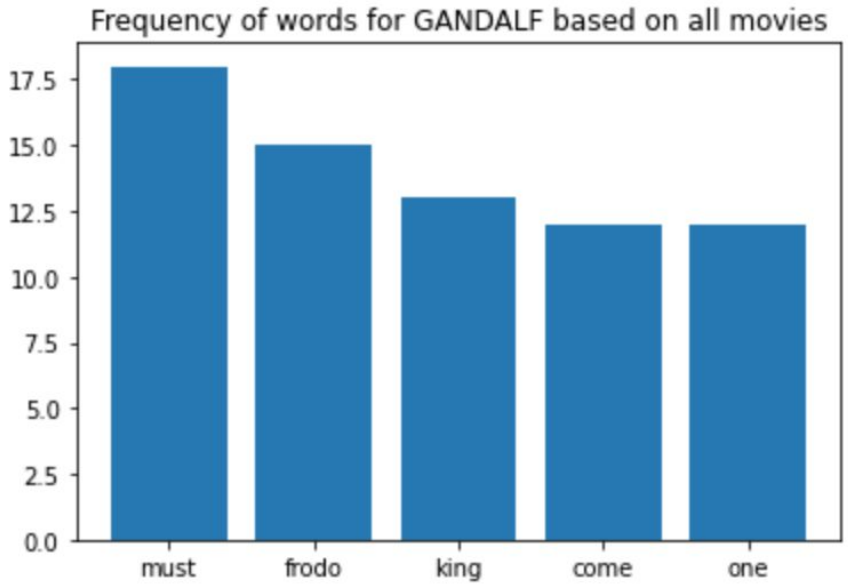
Loves Sam, Scared of Gollum, Crippled with Fear



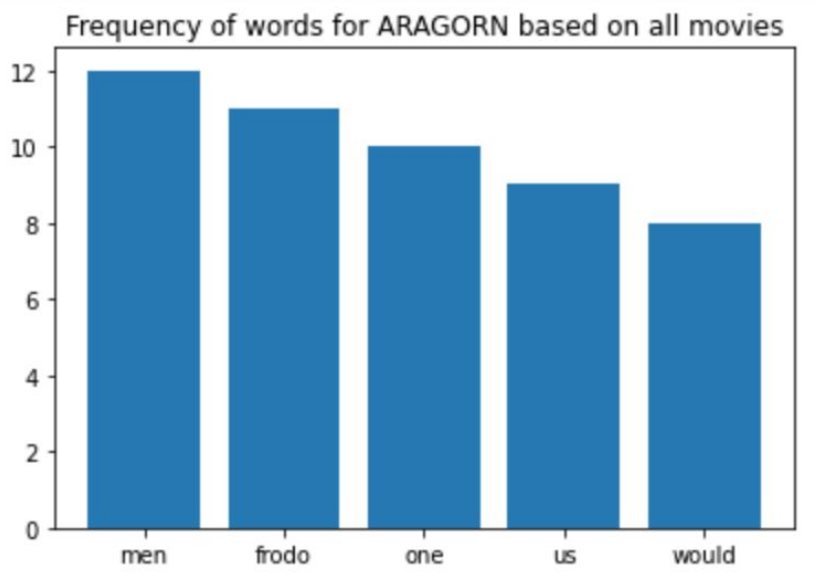
Loves Frodo, Separation Anxiety



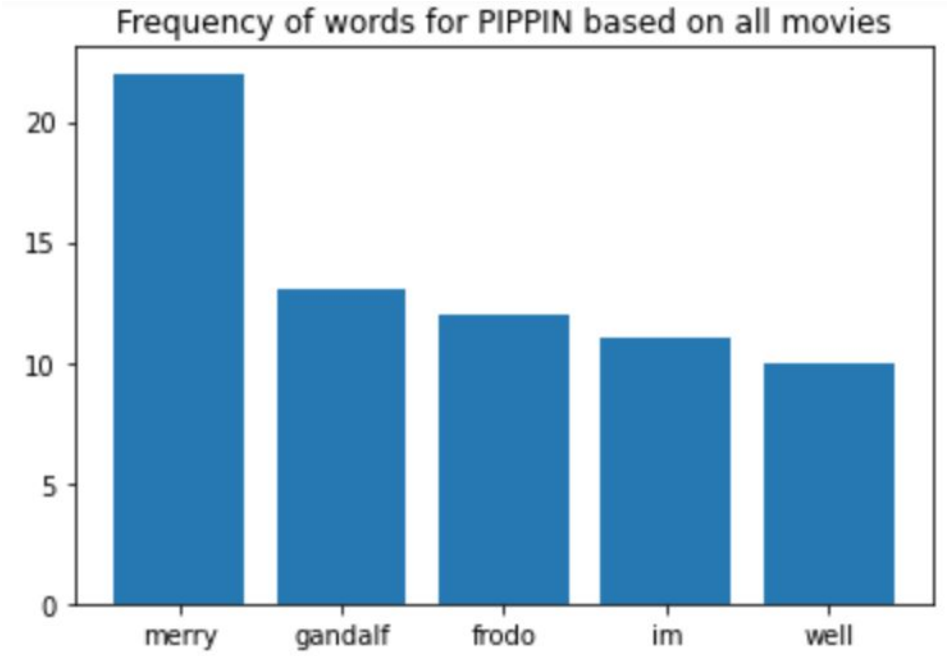
Very Demanding, Loves Frodo, Grandiose thoughts



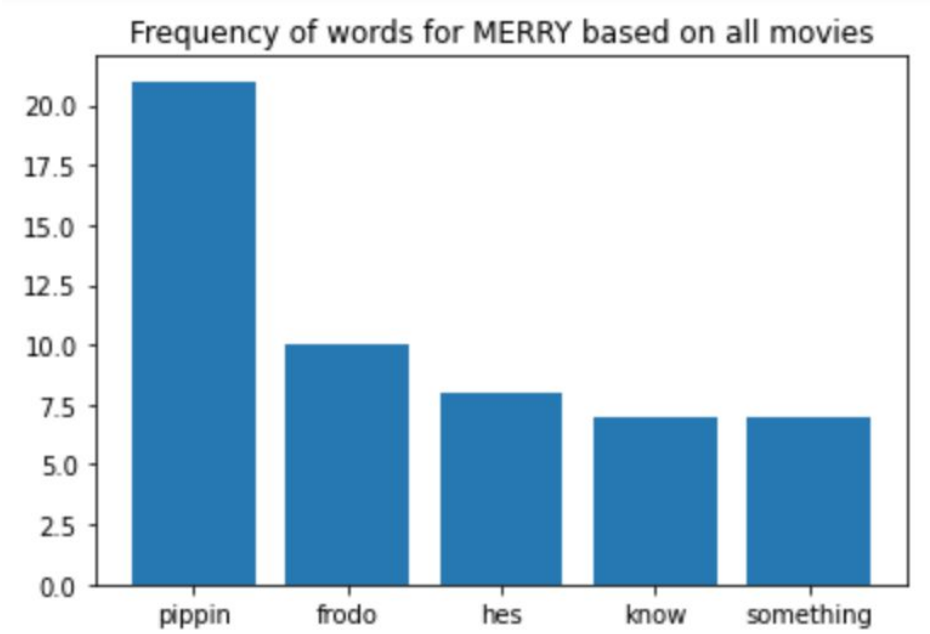
Unifying figure, Always saving Frodo



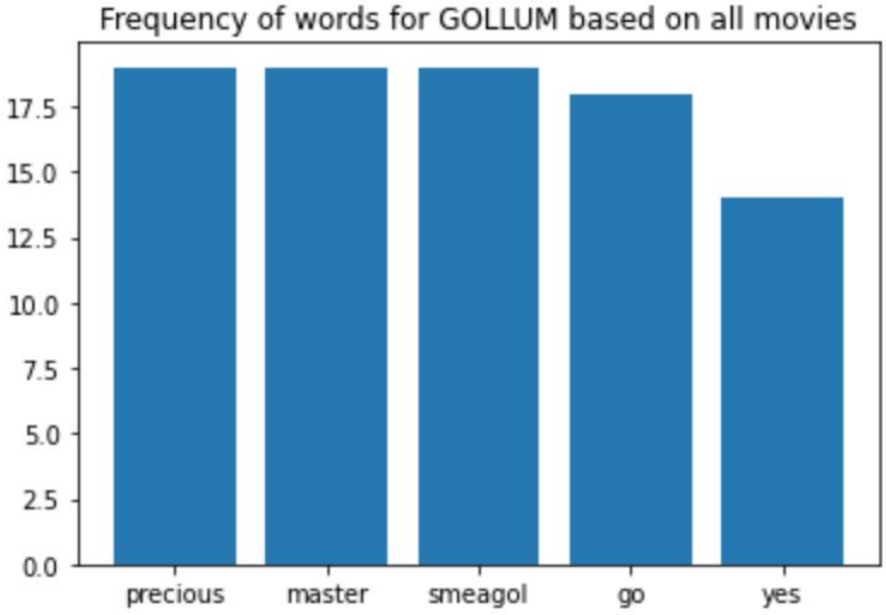
Loves Merry alot, Optimistic



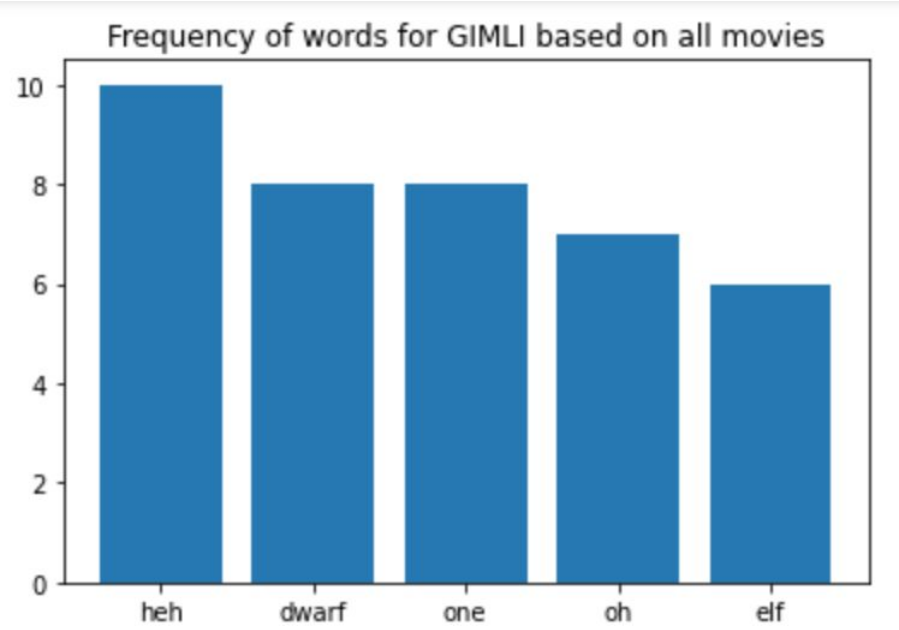
Loves Pippin, Very paranoid



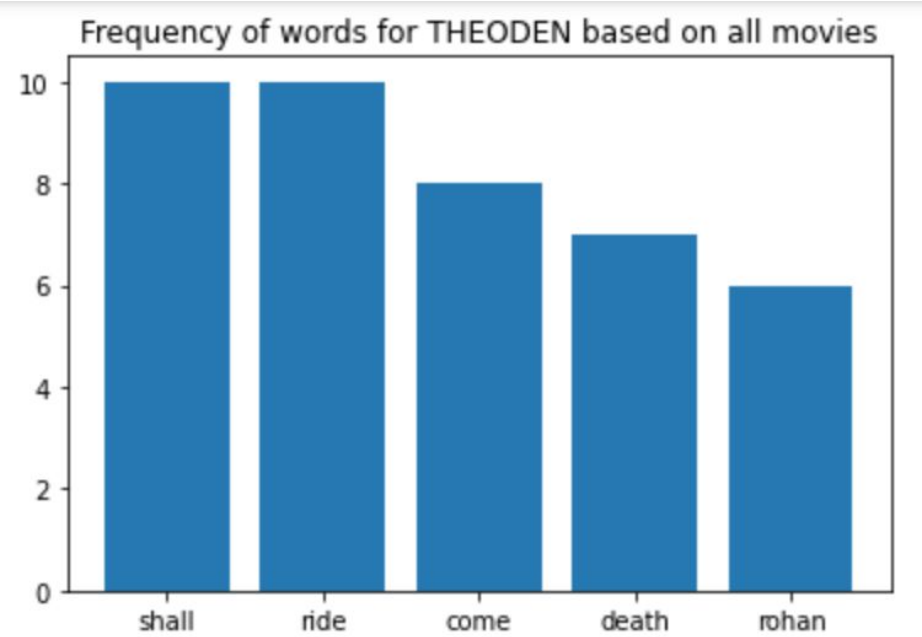
Loves his precious



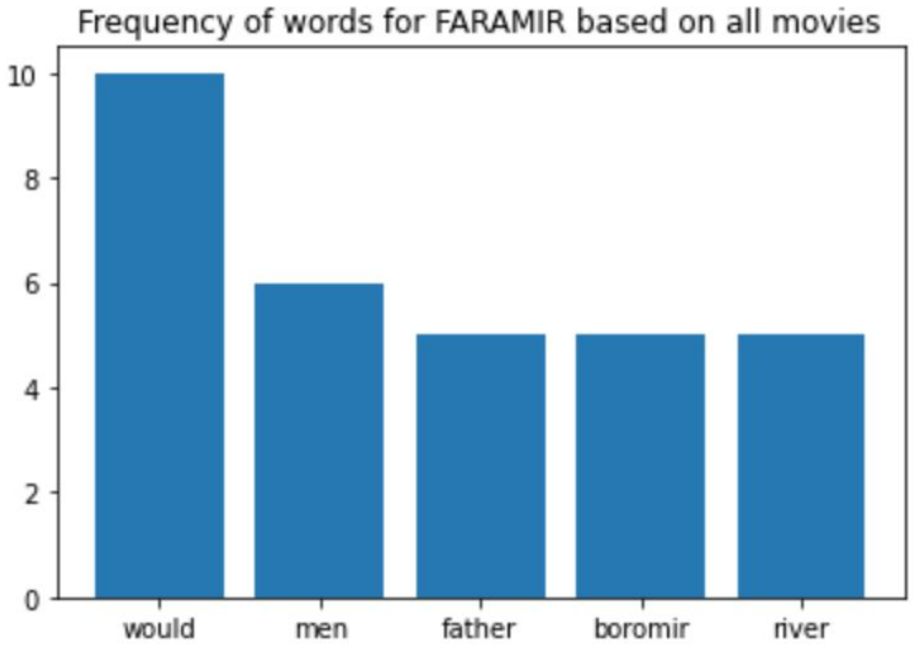
Loves himself and Legolas



Definition of true grit

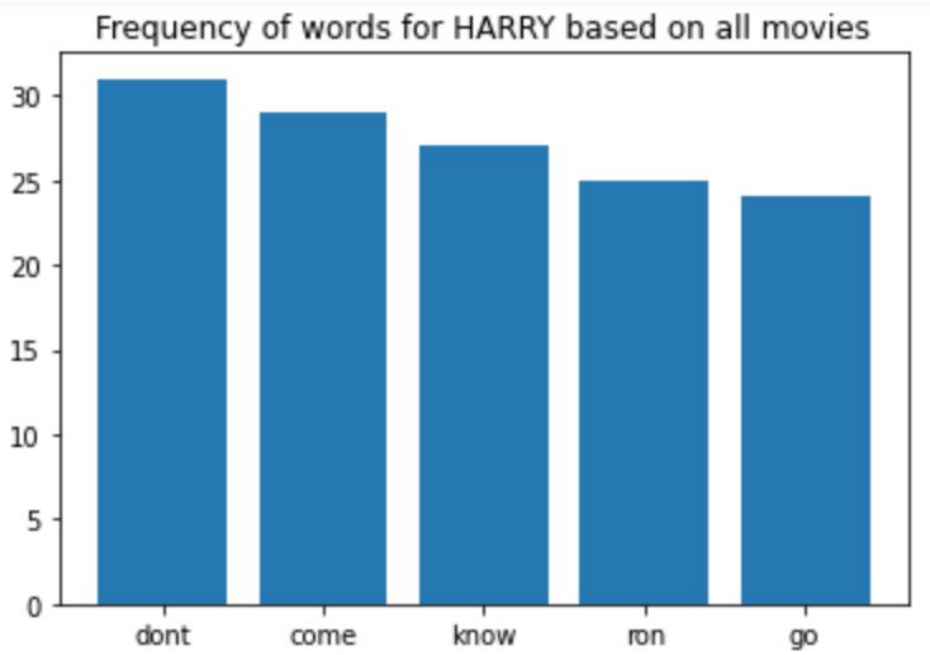


Very masculine

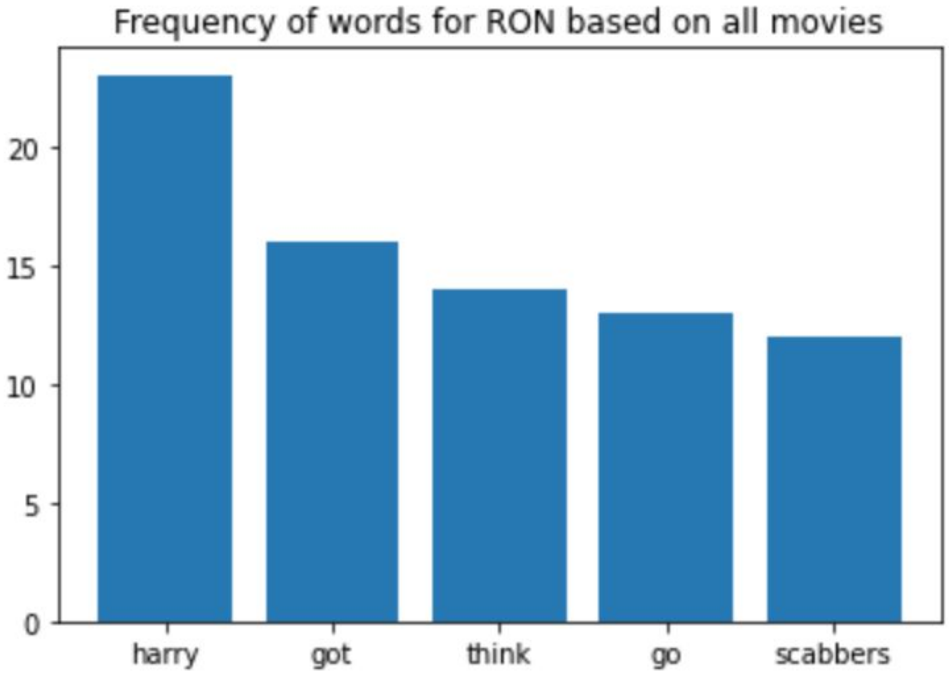


Results: Harry Potter

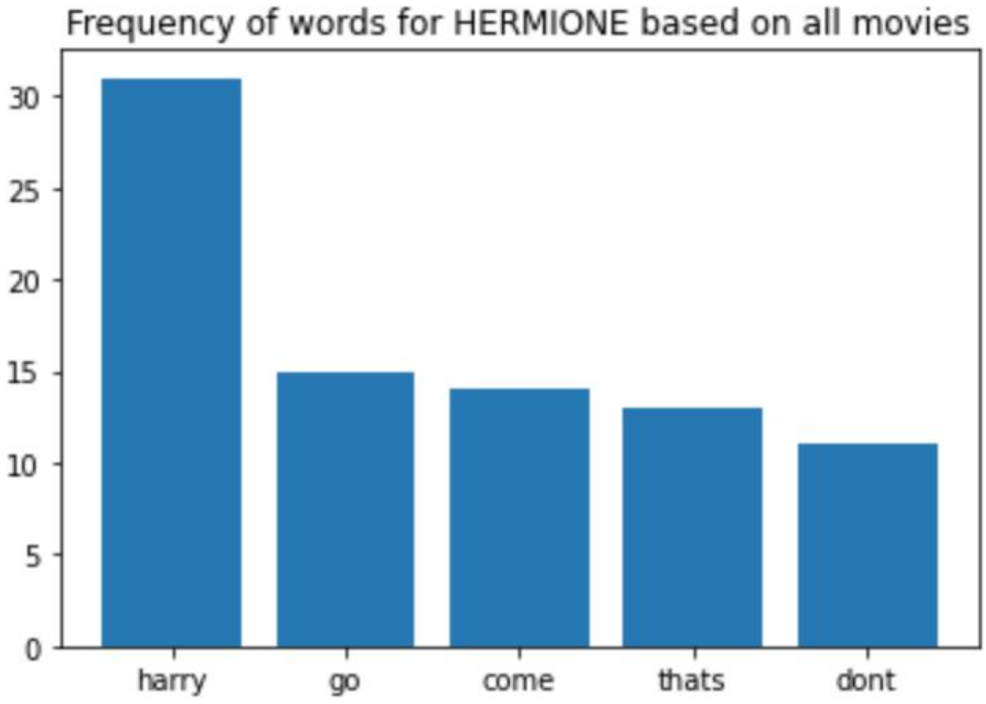
Always says no before saying yes



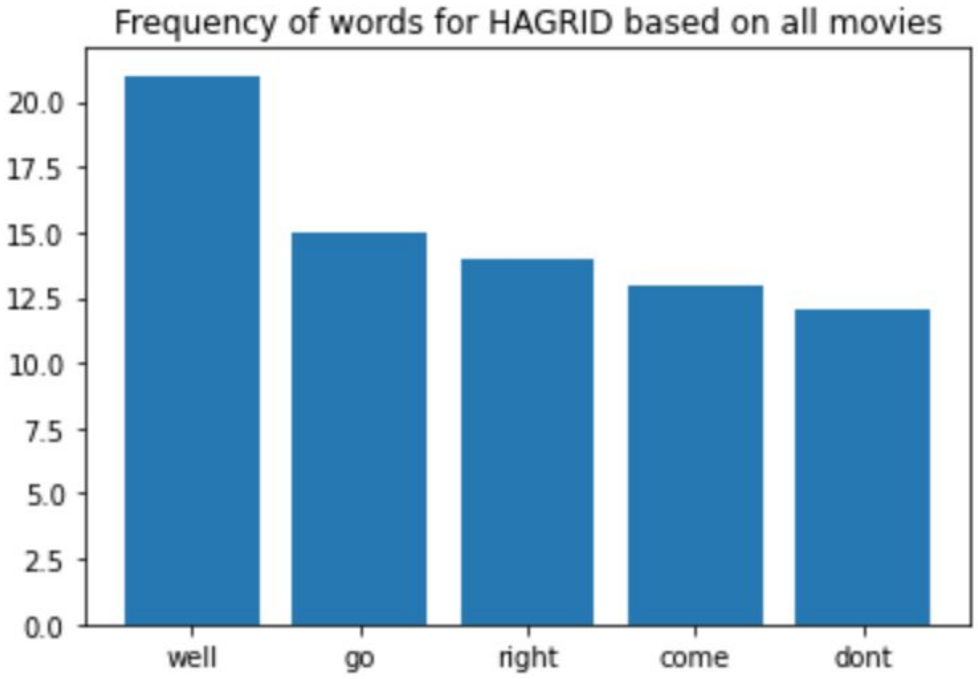
Follows Harry around like a lapdog



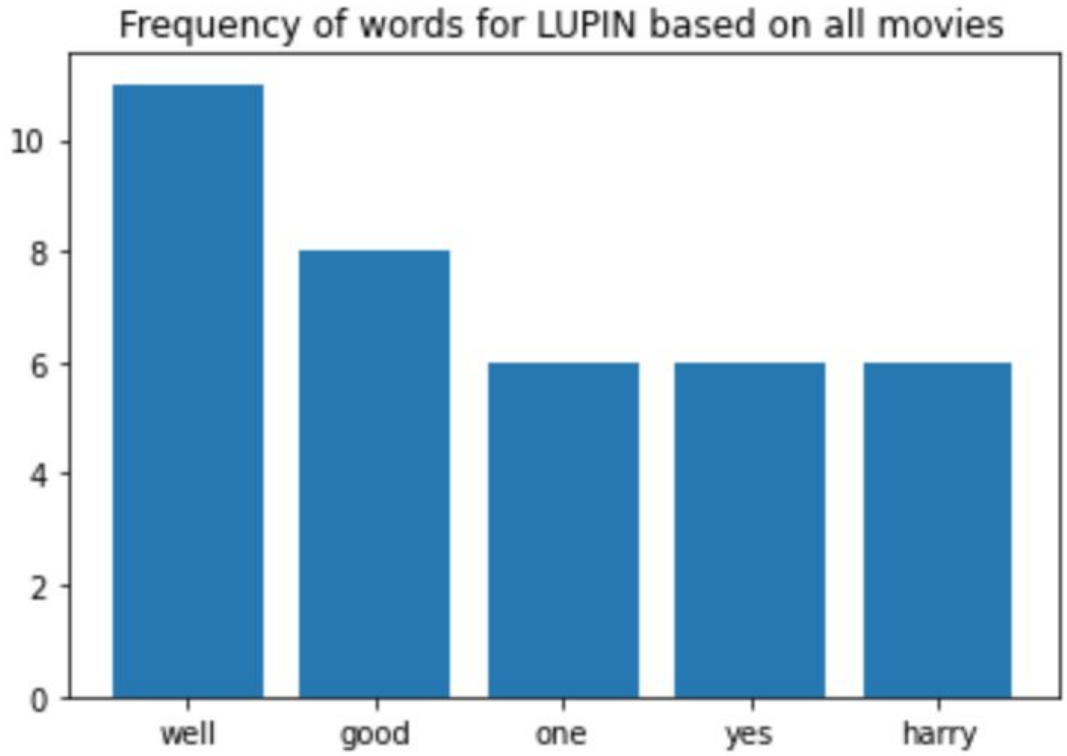
Married the wrong guy



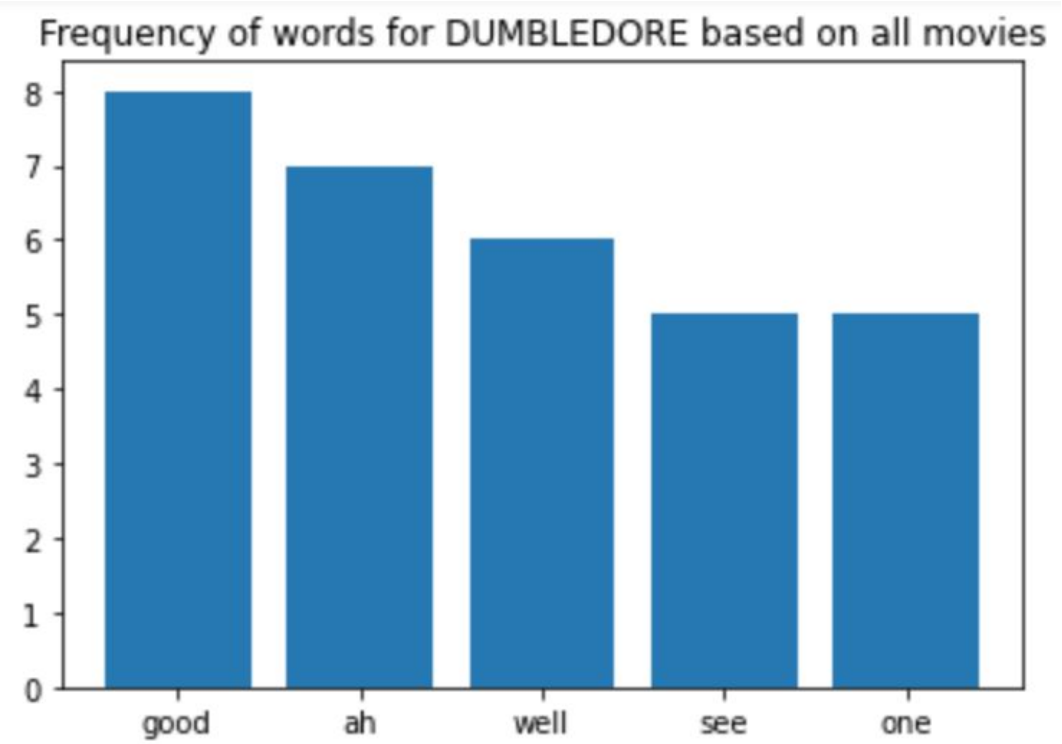
Good old Hagrid



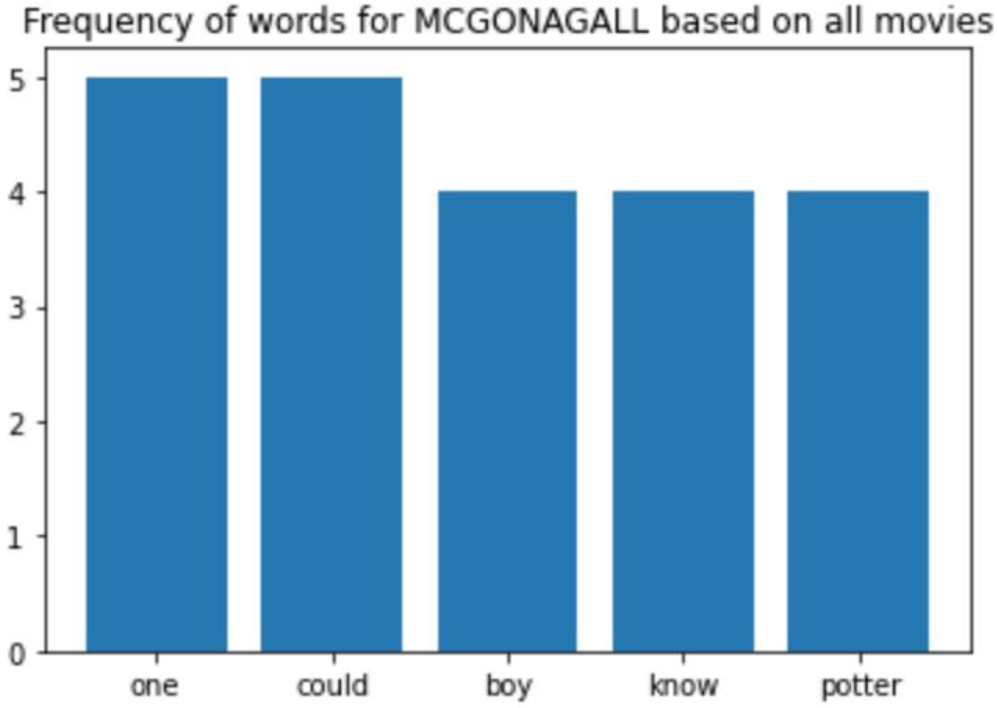
Uses many of the same words as Hagrid



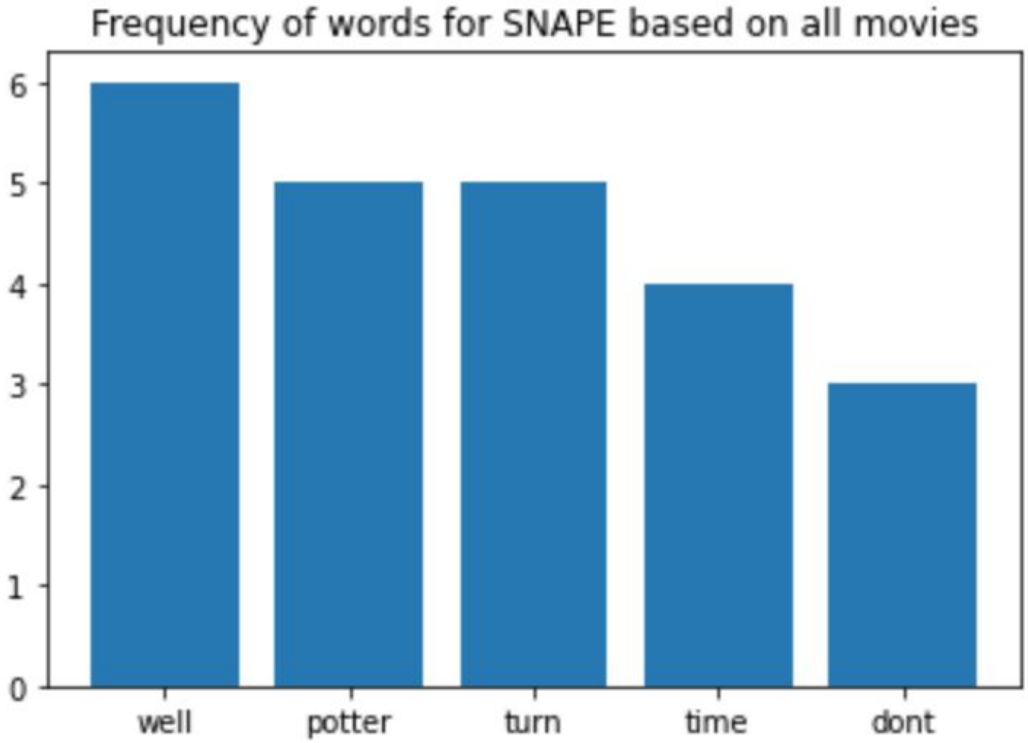
Uses similar words to Hagrid and Lupin



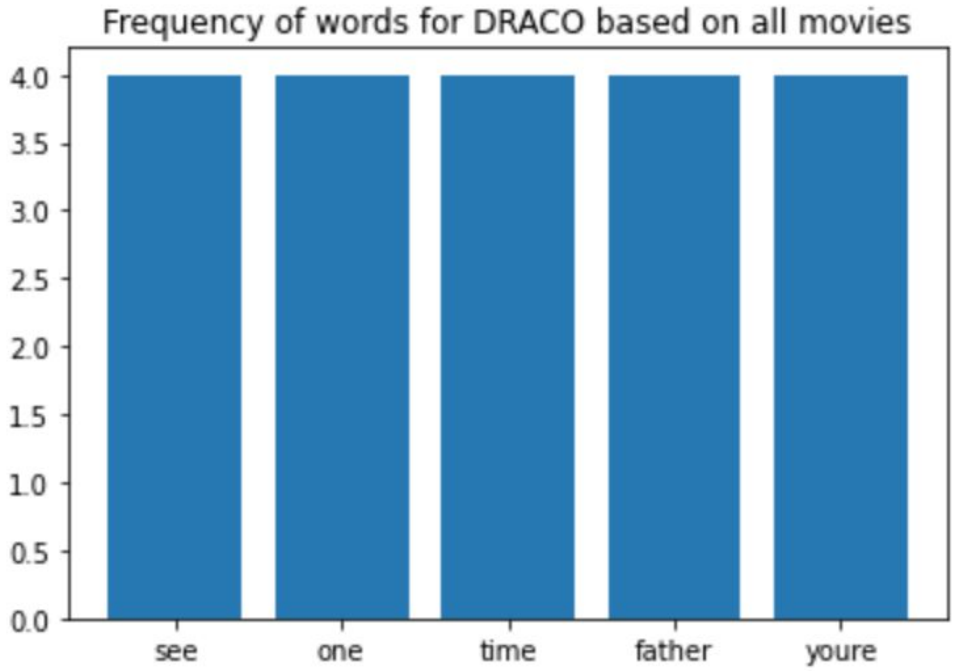
Tired of Harry, Ron, and Hermione's shenanigans



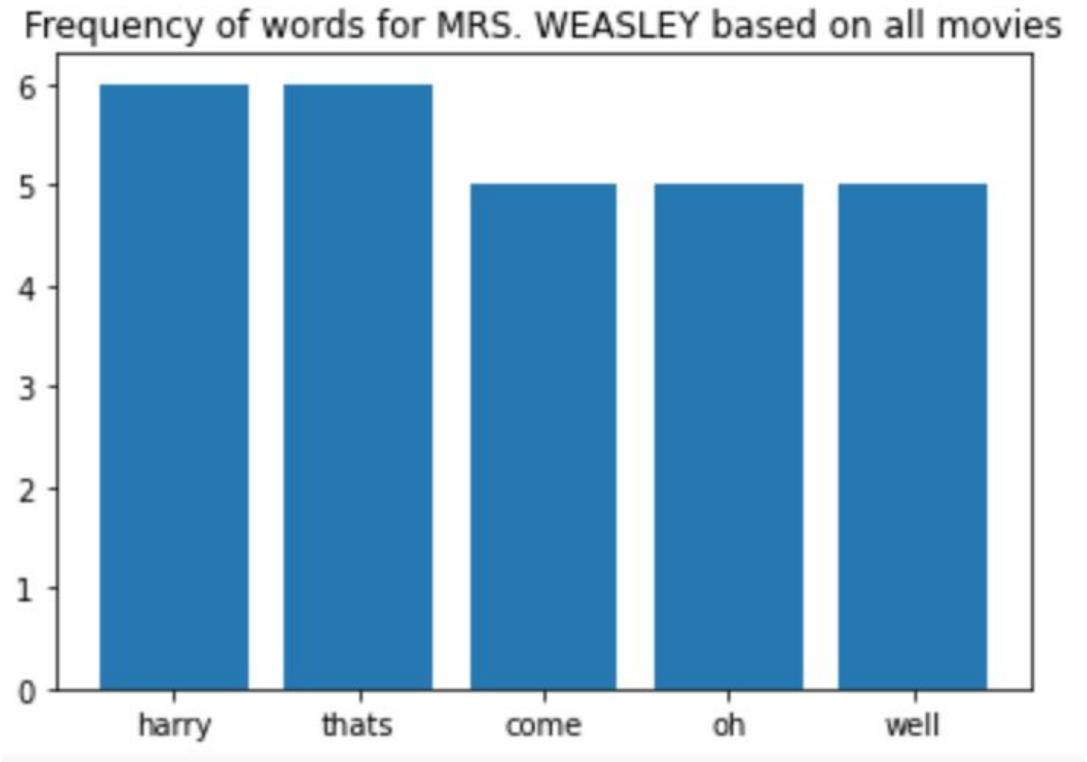
Has bad blood with Harry that goes deep



Relies on his dad to bail him out of sticky situations



Wishes Harry was her son



Conclusion

- Basic Data Analysis provided useful but not rigorously quantifiable data-> Need ML
- Should try observing changes over time as part of sentiment analysis
- Merging the lotr_scripts.csv data with lotr_characters.csv data would create a more complicated ML problem of determining whether environmental factors contributed to characters utilizing certain words(e.g Elven language)
- The same can be said for the Harry Potter dataset

That’s all folks!