Artificial Intelligence

TERM PROJECT

CLASSIFICATION WITH NLTK

Gabrielle Good

FALL 2020

LIBRARIES

- Natural Language Toolkit (NLTK)
- Matplotlib
- Native Python library

DATASET

- Dataset consisted of 5 text files. These files contained 15-20 movie plot descriptions each.
- The files:
  - Action.txt: 959 words
  - Comedy.txt: 1,041 words
  - Drama.txt: 935 words
  - Horror.txt: 961 words
  - Romance.txt: 900 words

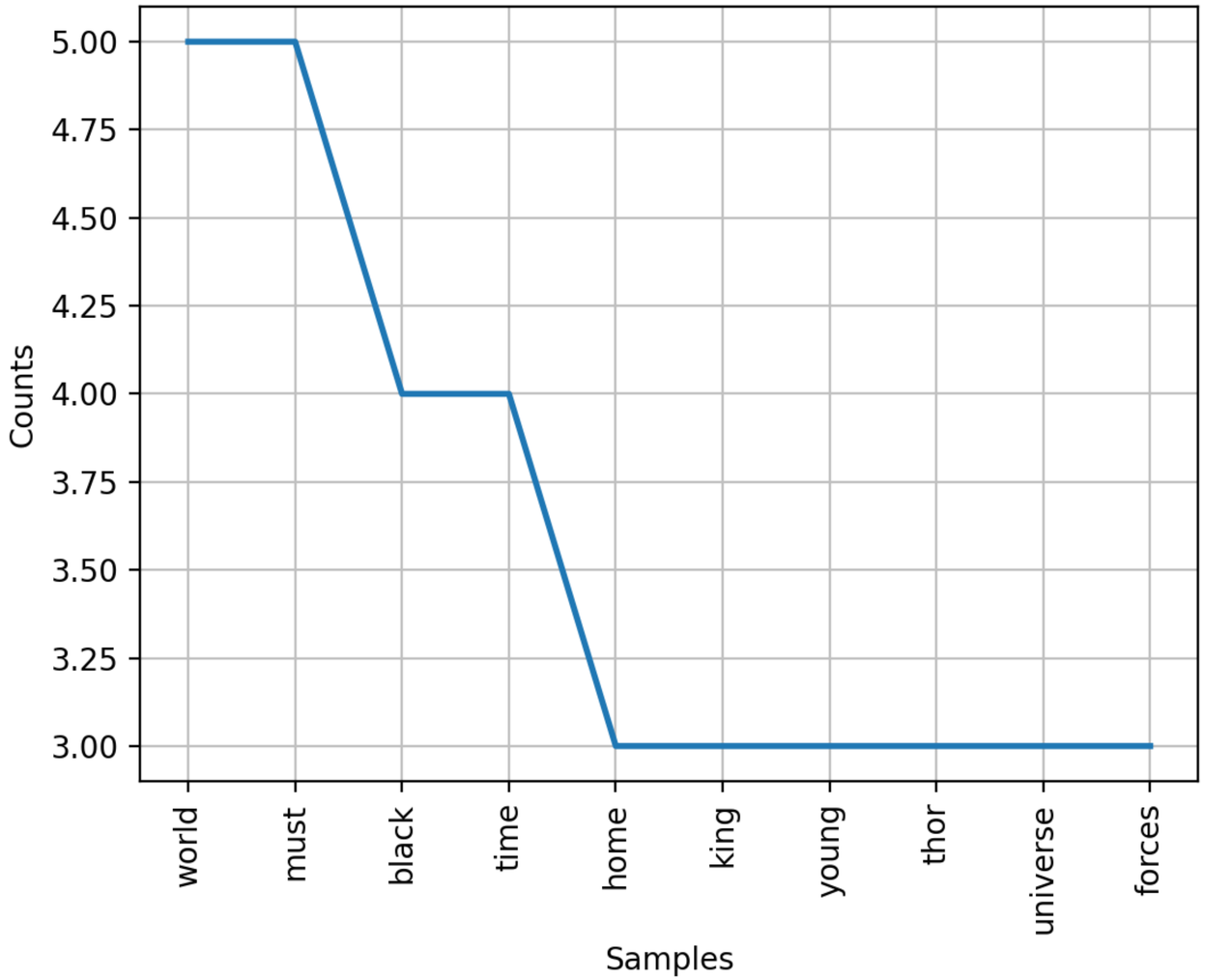
PROCESS

- Each data file was read into the Python file and cleaned.
- This process included converting each word to lowercase, removing punctuation, stripping each word to their root word, and removing stopwords.
- The datasets were combined after a label was added to each word in the individual genre set. Then the final dataset was randomized.
- The classifier would predict what category the word in the set belonged to.
- The feature set was split between training data and test data. The first thousand words were training data and the next thousand were used for testing

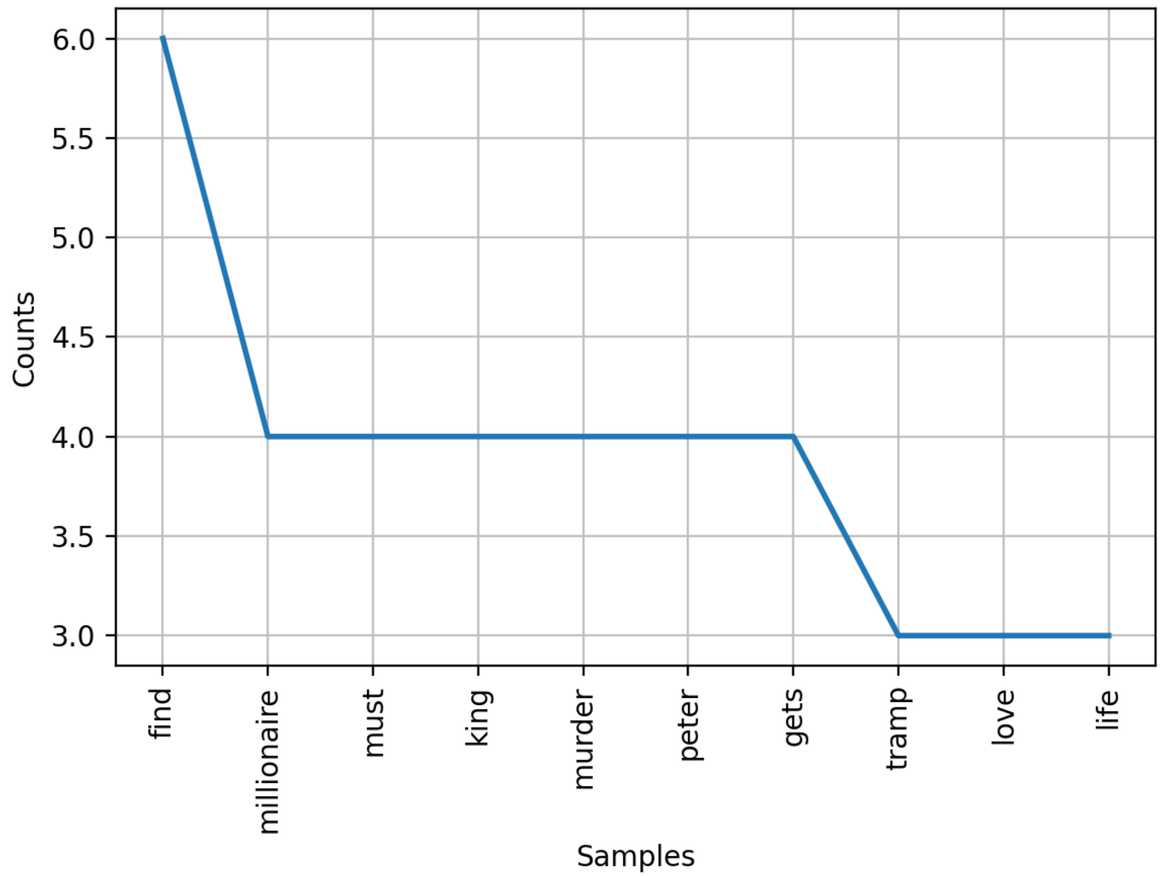
DIFFICULTIES

- The issue with this is that there are common words between each categories of Action, Comedy, Drama, Horror, and Romance.
- These commonalities can affect the prediction.
- Also, there were a lot of names present. Names of characters and actors that also skewed the categories.

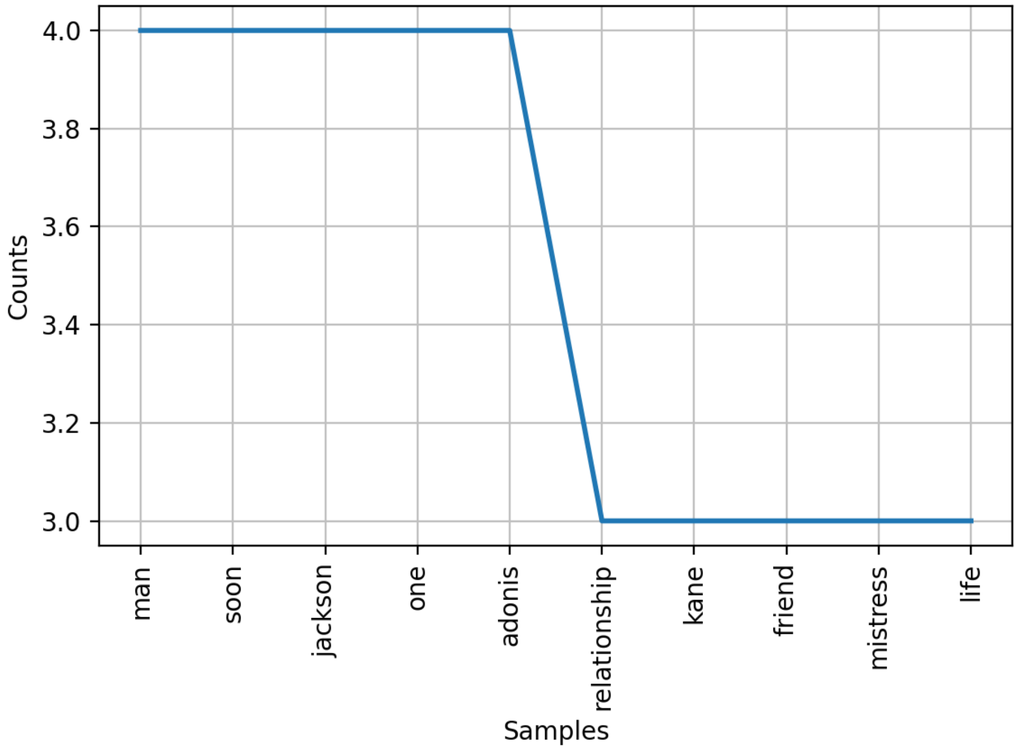
ACTION FILE


5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
Counts
world
must
black
time
home
king
young
thor
universe
forces
Samples

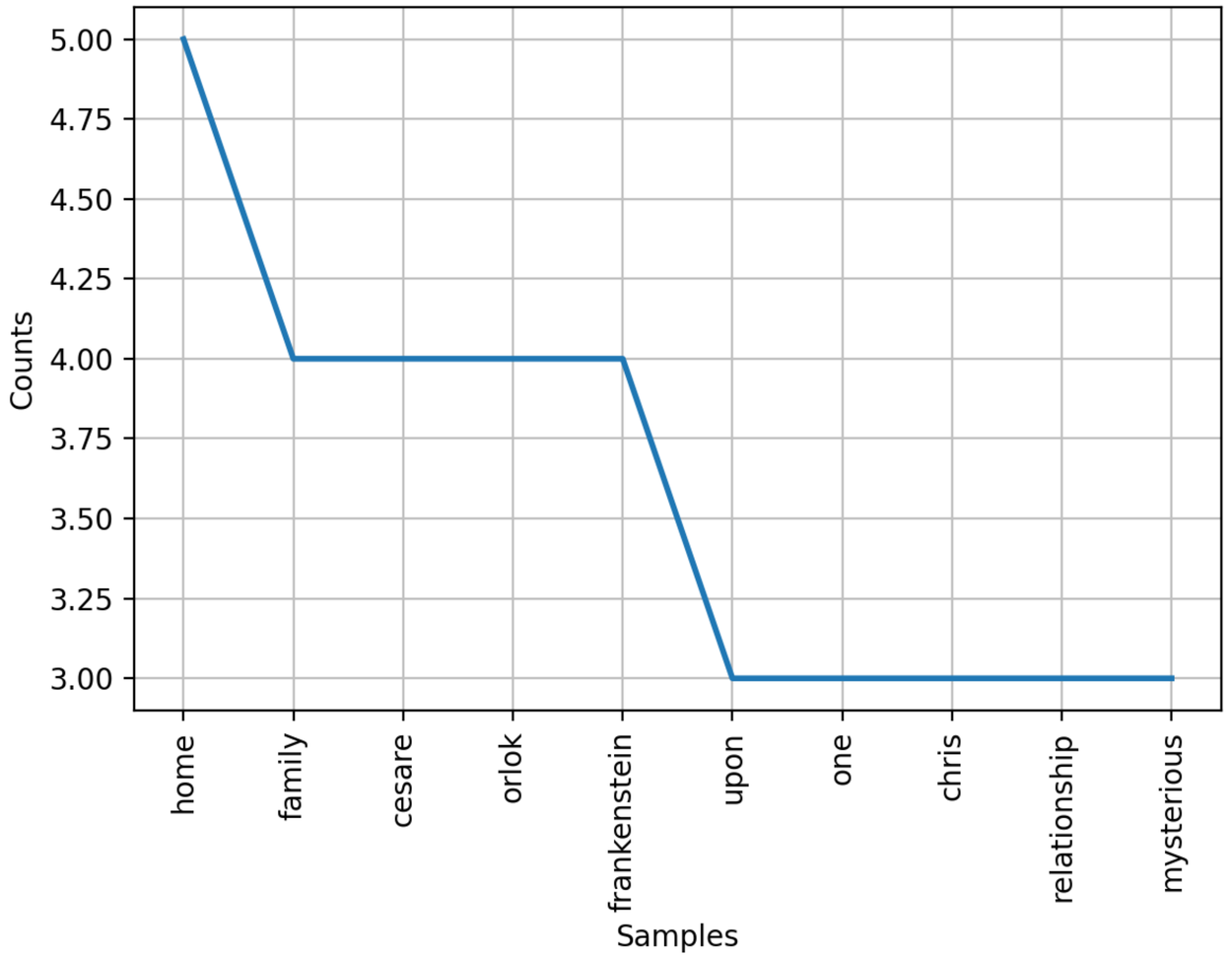
COMEDY FILE


6.0
5.5
5.0
4.5
4.0
3.5
3.0
Counts
find
millionaire
must
king
murder
peter
gets
tramp
love
life
Samples

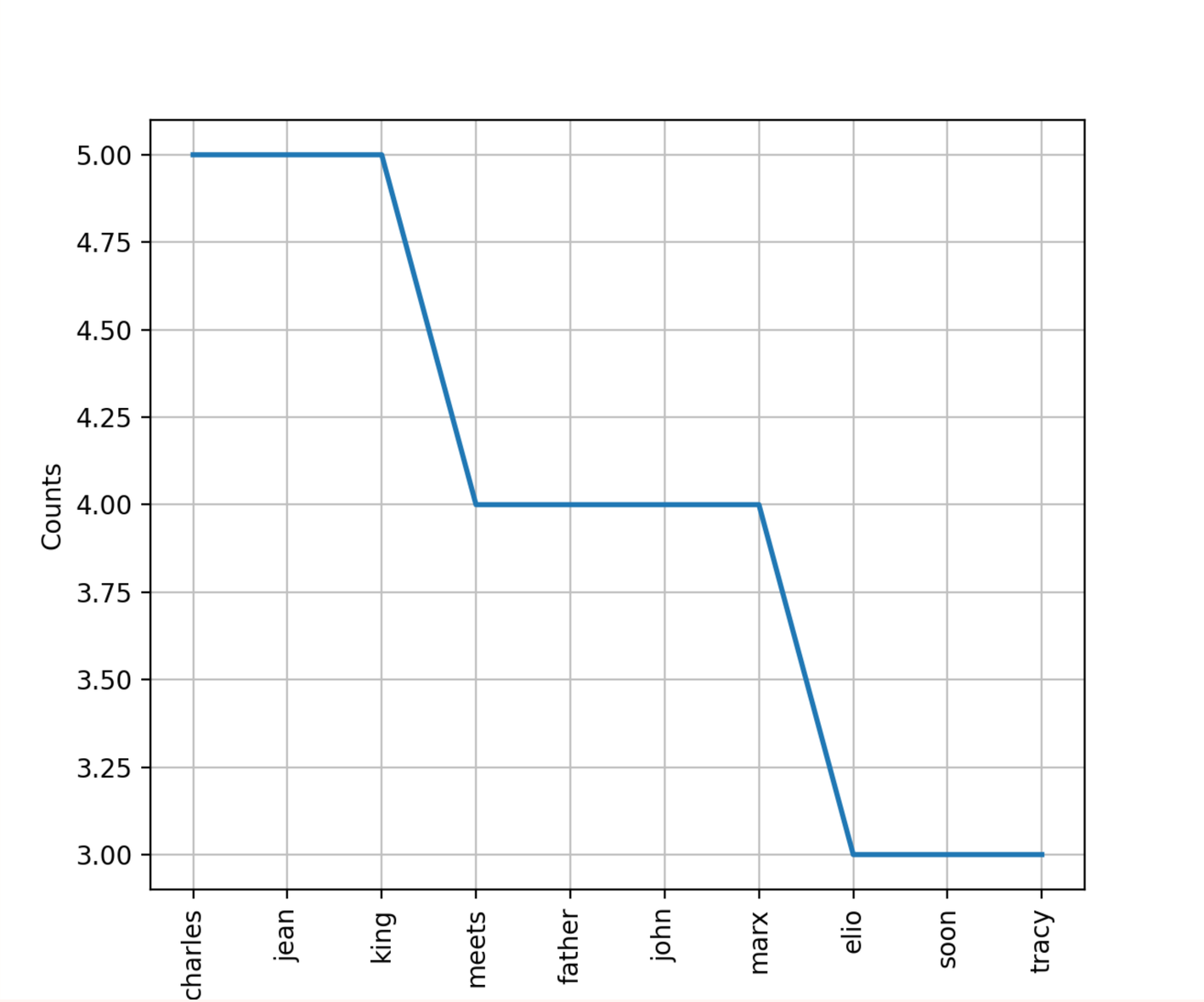
DRAMA FILE


4.0
3.8
3.6
3.4
3.2
3.0
Counts
man
soon
jackson
one
adonis
relationship
kane
friend
mistress
life
Samples

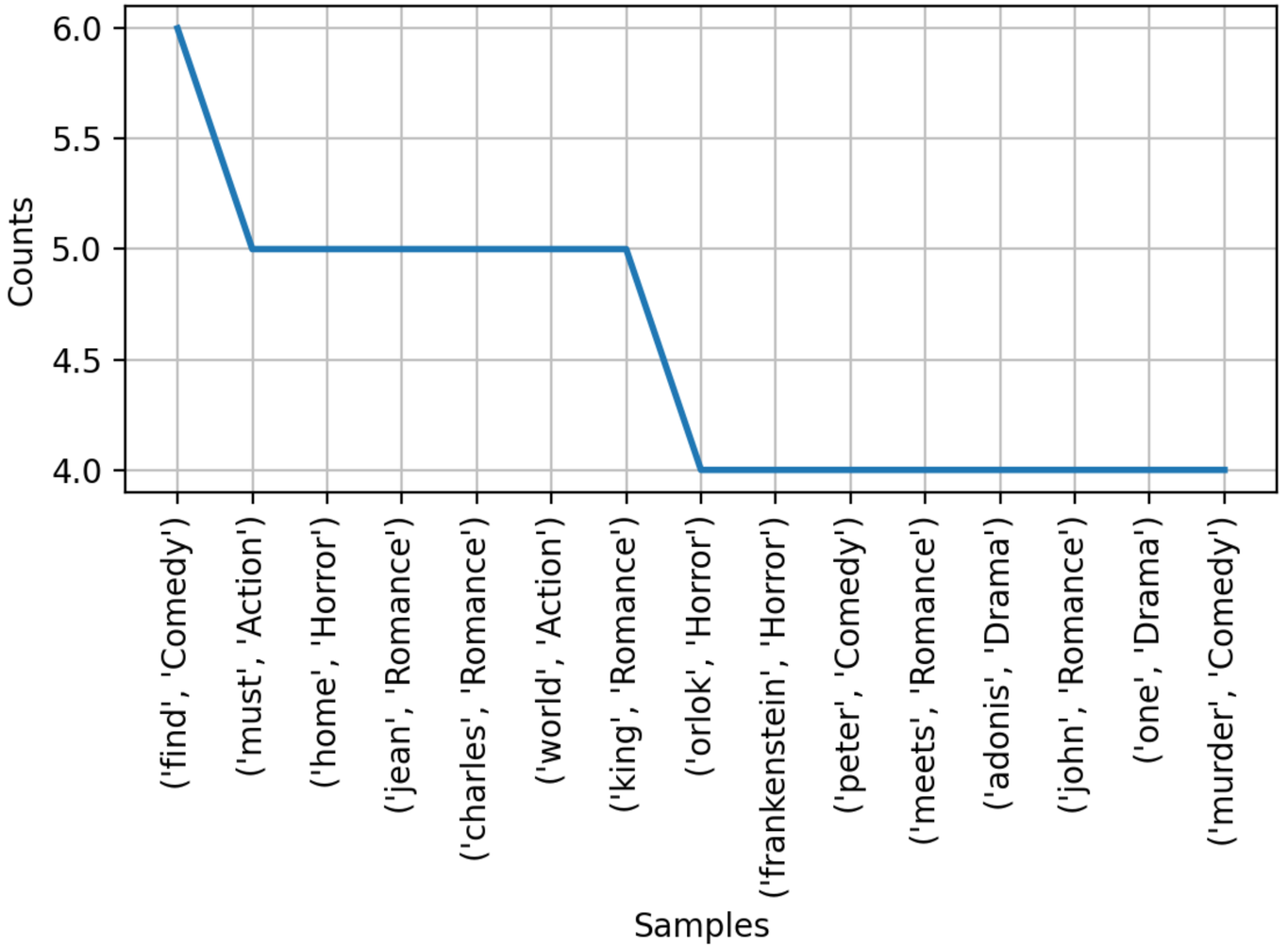
HORROR FILE


5.00
4.75
4.50
4.25
4.00
3.75
3.50
3.25
3.00
Counts
home
family
cesare
orlok
frankenstein
upon
one
chris
relationship
mysterious
Samples

ROMANCE FILE



TOP 15 COMBINED



ACCURACY

- The training data was not 100% accurate. It was around 85% accuracy through multiple runs.
- The test data was even less accurate. Because the data was randomized each time and the full set wasn’t used, the test data tended to average between 20-25% accuracy in predicting the words that are more likely to be associated with that genre.

REFERENCES

- Tutorials were followed from this link: https://www.nltk.org/book/ch06.html