NLP Project 1

Ryan Peruski, Nolan Coffey, Triton Eden, William Duff

The Assignment

Team 5:

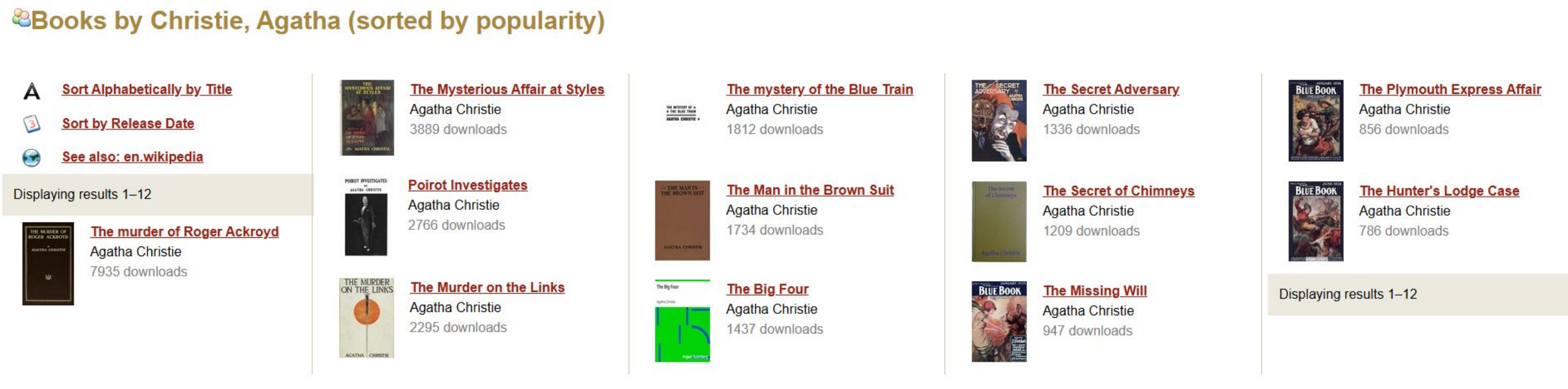
Analyze the plot of Agatha Christie novels computationally using word embeddings.

Our focus:

Attempt to predict plots by analyzing patterns in keyword proximity in Christie's works.

Data Extraction

- Used Project Gutenberg to get the Agatha Christie books
- We analyzed 3 books that seemed similar (same protagonist, same kind of plot, etc)
- After that, we manually grabbed key points (protagonist, victim, antagonist, murder weapon)



Books by Christie, Agatha (sorted by popularity)

Sort Alphabetically by Title
Sort by Release Date
See also: en.wikipedia

Displaying results 1–12

The murder of Roger Ackroyd
Agatha Christie
7935 downloads

The Mysterious Affair at Styles
Agatha Christie
3889 downloads

Poirot Investigates
Agatha Christie
2766 downloads

The Murder on the Links
Agatha Christie
2295 downloads

The mystery of the Blue Train
Agatha Christie
1812 downloads

The Man in the Brown Suit
Agatha Christie
1734 downloads

The Big Four
Agatha Christie
1437 downloads

The Secret Adversary
Agatha Christie
1336 downloads

The Secret of Chimneys
Agatha Christie
1209 downloads

The Missing Will
Agatha Christie
947 downloads

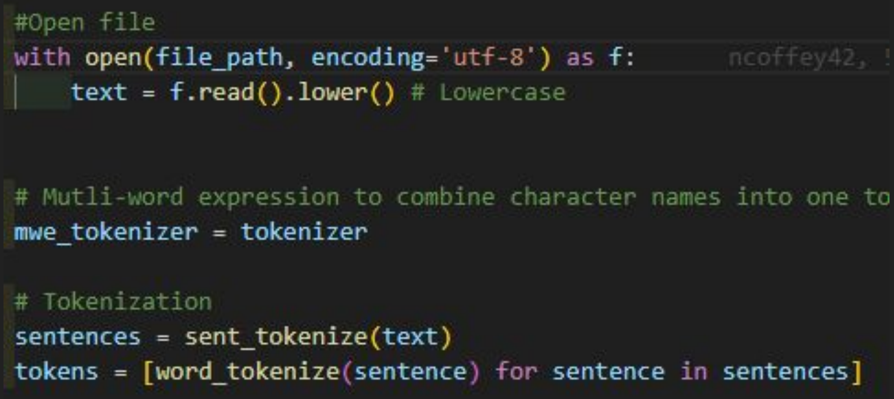
The Plymouth Express Affair
Agatha Christie
856 downloads

The Hunter's Lodge Case
Agatha Christie
786 downloads

Displaying results 1–12

Preprocessing

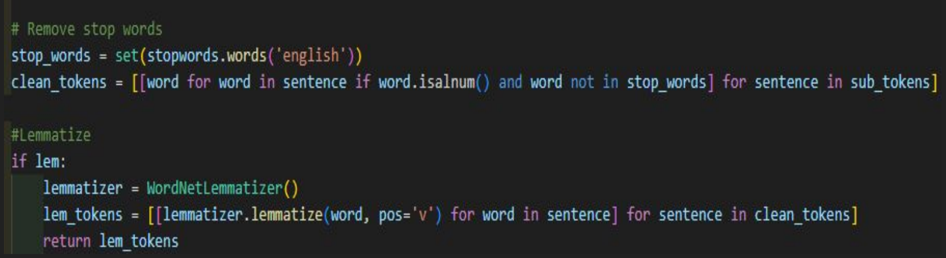
- Turns raw txt file with book into tokens we can use for the model
- We normalize, tokenize, and clean – as per usual in NLP
- Some manual MWE tokens
- We use lemmatization as an independent variable, so we test models with and without it



```
#Open file
with open(file_path, encoding='utf-8') as f:        ncoffey42,
    text = f.read().lower() # Lowercase


# Mutli-word expression to combine character names into one to
mwe_tokenizer = tokenizer

# Tokenization
sentences = sent_tokenize(text)
tokens = [word_tokenize(sentence) for sentence in sentences]
```

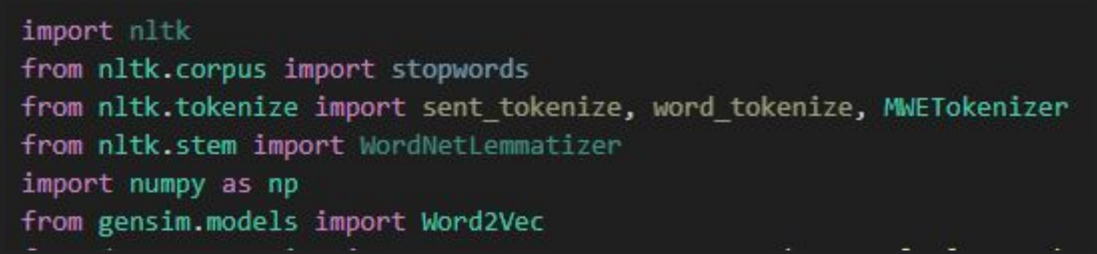


```
# Remove stop words
stop_words = set(stopwords.words('english'))
clean_tokens = [[word for word in sentence if word.isalnum() and word not in stop_words] for sentence in sub_tokens]

#Lemmatize
if lem:
    lemmatizer = WordNetLemmatizer()
    lem_tokens = [[lemmatizer.lemmatize(word, pos='v') for word in sentence] for sentence in clean_tokens]
    return lem_tokens
```

Methods, Code, and Libraries

- Our Model that we chose was Word2Vec
- From there, we created separate W2V models with CBOW and Skip-gram
- NLTK for preprocessing (lemmatization, tokenization, and cleaning)!
- Requests library for data extraction from Project Gutenberg!

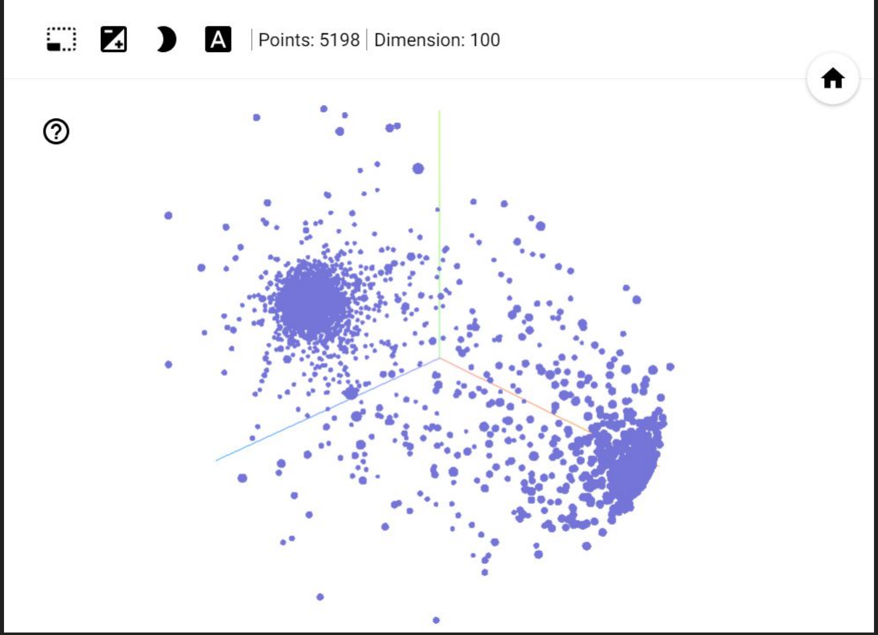


```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize, word_tokenize, MWETokenizer
from nltk.stem import WordNetLemmatizer
import numpy as np
from gensim.models import Word2Vec
```

TensorBoard

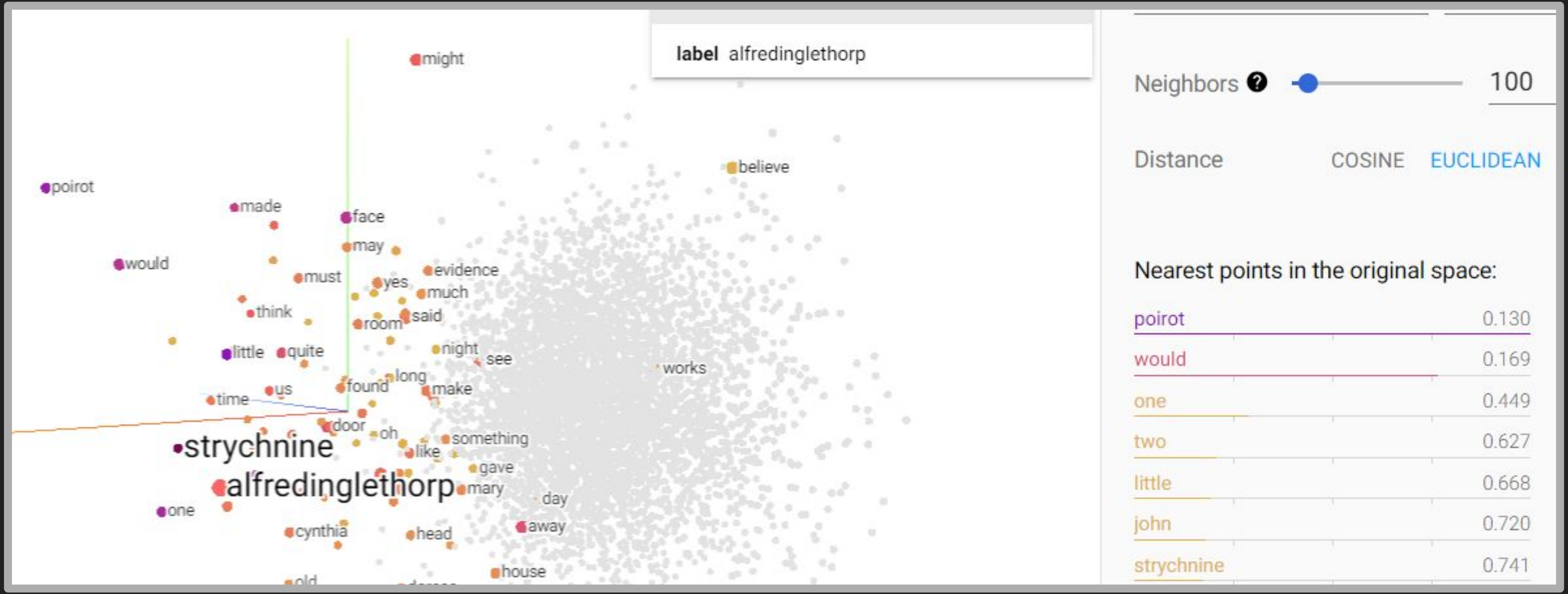
Methodology

- Our Process:
  Get the book -> find important key points in book (manually) -> preprocess book -> run the model (CBOW or Skip-gram) -> write to tensor and display specific Euclidean distances



Results

- Mixed Results:
  - Success with *The Mysterious Affair at Styles* - CBOW without lemmatization
  - Not as clear with *The Murder on the Links* and *The Murder of Roger Ackroyd*
  - Potential applications to some of Christie's other novels



label alfredinglethorp

Neighbors ❓ 100

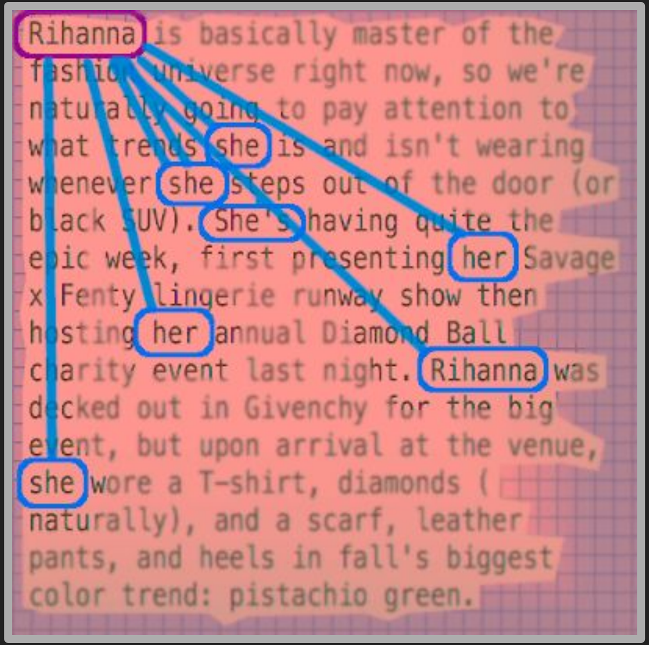
Distance COSINE EUCLIDEAN

Nearest points in the original space:

poirot	0.130
would	0.169
one	0.449
two	0.627
little	0.668
john	0.720
strychnine	0.741

Future Prospects

- Implement Coreference Resolution using spaCy’s NeuralCoref library
  - Ex: Poirot went home. Then he went to bed. == Poirot went home. Then Poirot went to bed.
- Implement Named Entity Recognition (NER)
  - Ex:
    - Albert - Character,
    - Styles - Place
  - Use of NER will allow us to make conclusion by comparing the euclidean distance of all of the characters
- Perform a chapter by chapter analysis to incorporate a temporal aspect



Rihanna is basically master of the fashion universe right now, so we're naturally going to pay attention to what trends she is and isn't wearing whenever she steps out of the door (or black SUV). She's having quite the epic week, first presenting her Savage x Fenty lingerie runway show then hosting her annual Diamond Ball charity event last night. Rihanna was decked out in Givenchy for the big event, but upon arrival at the venue, she wore a T-shirt, diamonds (naturally), and a scarf, leather pants, and heels in fall's biggest color trend: pistachio green.

https://lvngd.com/blog/coreference-resolution-python-spacy-neuralcoref/

Questions? (And Demonstration)