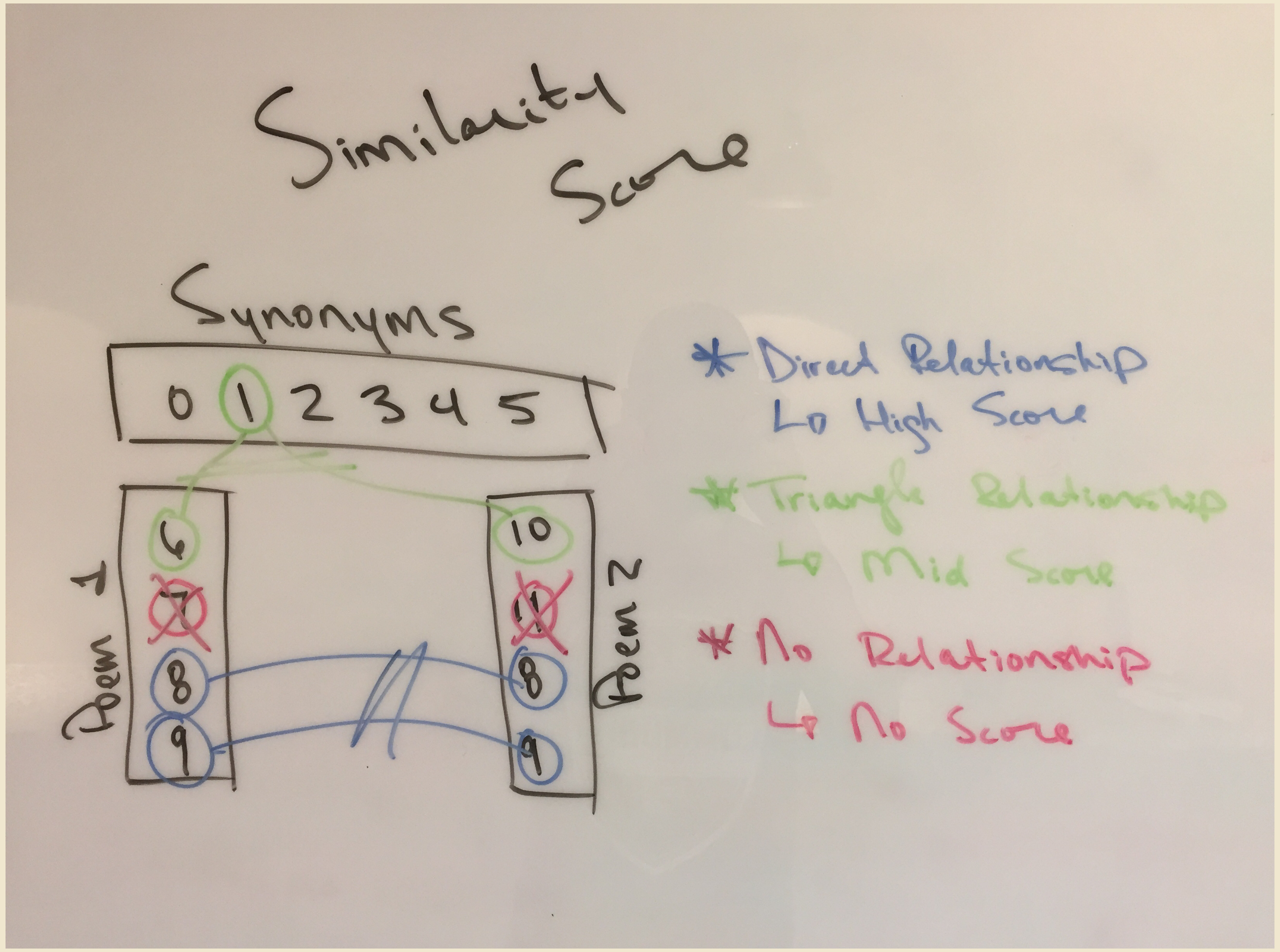
Getting a bit in depth..

Steps To Completion

1. Get data
2. Clean data
3. Perform TF-IDF over poems
4. Establish initial network graphs
5. Add synonyms to base graphs, attaching edges
6. Create graphs for poem combinations
7. Establish similarity matrix, perform similarity logic
8. Get top 3 scores per poem
9. Create basic flask app

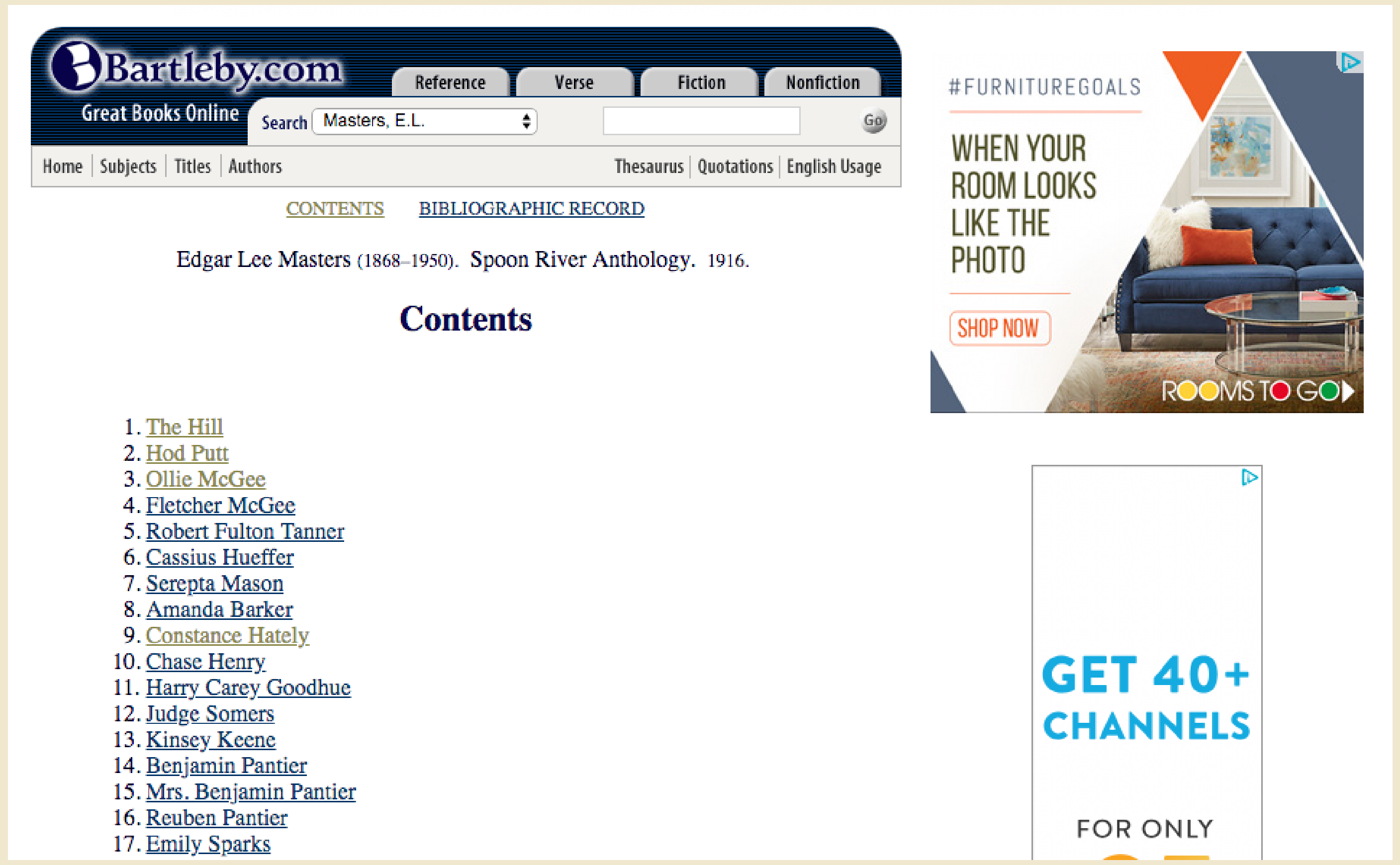


MVP
Final Outlook

Get Data

Libraries Used

- BeautifulSoup (text scraping)



Bartleby.com Great Books Online | Reference | Verse | Fiction | Nonfiction | Search | Home | Subjects | Titles | Authors | Thesaurus | Quotations | English Usage

CONTENTS BIBLIOGRAPHIC RECORD

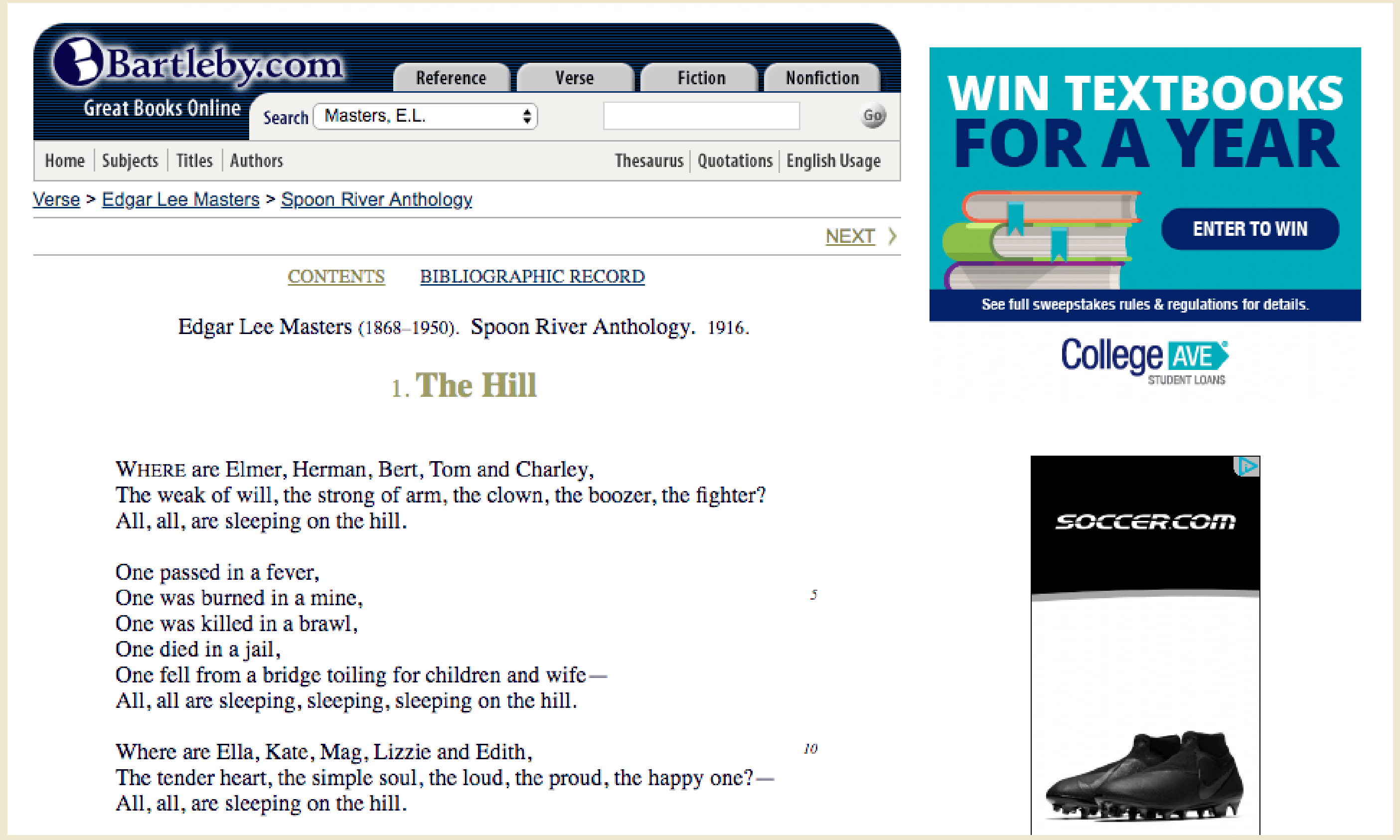
Edgar Lee Masters (1868–1950). Spoon River Anthology. 1916.

Contents

1. The Hill
2. Hod Putt
3. Ollie McGee
4. Fletcher McGee
5. Robert Fulton Tanner
6. Cassius Hueffer
7. Serepta Mason
8. Amanda Barker
9. Constance Hately
10. Chase Henry
11. Harry Carey Goodhue
12. Judge Somers
13. Kinsey Keene
14. Benjamin Pantier
15. Mrs. Benjamin Pantier
16. Reuben Pantier
17. Emily Sparks

#FURNITUREGOALS WHEN YOUR ROOM LOOKS LIKE THE PHOTO SHOP NOW ROOMS TO GO

GET 40+ CHANNELS FOR ONLY



Bartleby.com Great Books Online | Reference | Verse | Fiction | Nonfiction | Search | Home | Subjects | Titles | Authors | Thesaurus | Quotations | English Usage

Verse > Edgar Lee Masters > Spoon River Anthology

NEXT

CONTENTS BIBLIOGRAPHIC RECORD

Edgar Lee Masters (1868–1950). Spoon River Anthology. 1916.

1. The Hill

WHERE are Elmer, Herman, Bert, Tom and Charley,
The weak of will, the strong of arm, the clown, the boozer, the fighter?
All, all, are sleeping on the hill.

One passed in a fever,
One was burned in a mine,
One was killed in a brawl,
One died in a jail,
One fell from a bridge toiling for children and wife—
All, all are sleeping, sleeping, sleeping on the hill.

Where are Ella, Kate, Mag, Lizzie and Edith,
The tender heart, the simple soul, the loud, the proud, the happy one?—
All, all, are sleeping on the hill.

WIN TEXTBOOKS FOR A YEAR ENTER TO WIN See full sweepstakes rules & regulations for details. College AVE STUDENT LOANS

SOCCER.COM

Clean Data

Libraries Used

- Gensim

Process

1. Remove ascii and punctuation
2. Remove basic stop-words
3. Ignore words less than 3 characters in length
4. Convert words to vector (gensim Dictionary.doc2bow)

Perform TF-IDF Over Poems

Libraries Used

- Gensim.models.TfidfModel

$$weight_{i,j} = frequency_{i,j} * log_2 \frac{D}{document\_freq_i}$$

*For this project, tf-idf is appropriate. It determines "important" words in a document (poem) while considering the length of the corpus (Spoon River). I am not looking for a specific category, but am instead drawing out important words within a poem to help determine the overall score.

Establish Base Network

Libraries Used

- Gensim
- Networkx

Process

Each unique term in a poem represents a node in a graph network. The tf-idf has been calculated, so the freq. doesn't matter. No edges, just nodes.

Add Synonyms

Libraries Used

- Networkx
- nltk (wordnet)

Process

For each term in a poem, get the synonyms and add them the the graph. Only keep unique synonyms, and keep track of the history of previous poems. Attach edges between synonyms and current terms in the graphs.

Merge Poem Combinations

Libraries Used

- Networkx

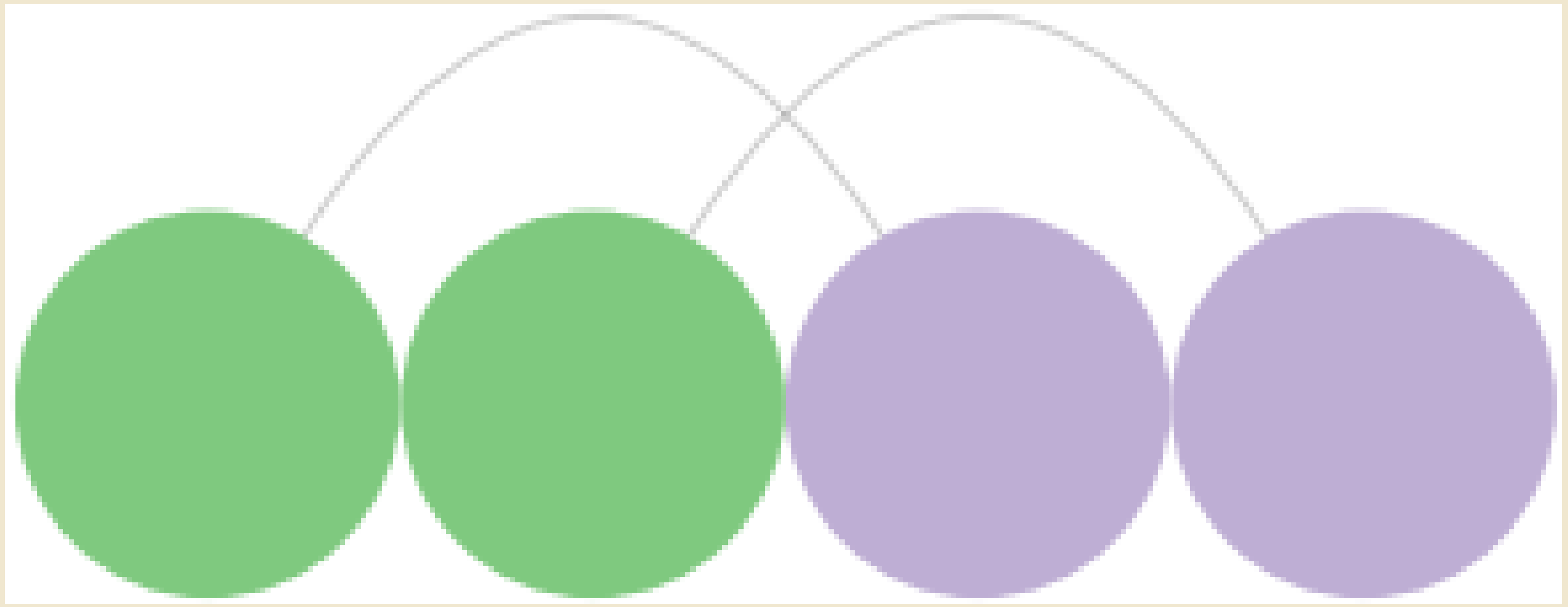
Process

For every two poems, draw an edge between the same terms, then combine the synonym terms. Remove all terms that are not contributing, and synonyms that point only one way.

William Goode (P1) - Minerva Jones (P2)

P1T1 == P2T1 --> Direct Relationships

"with" - "village"



"TO all in the village I seemed, no doubt.."

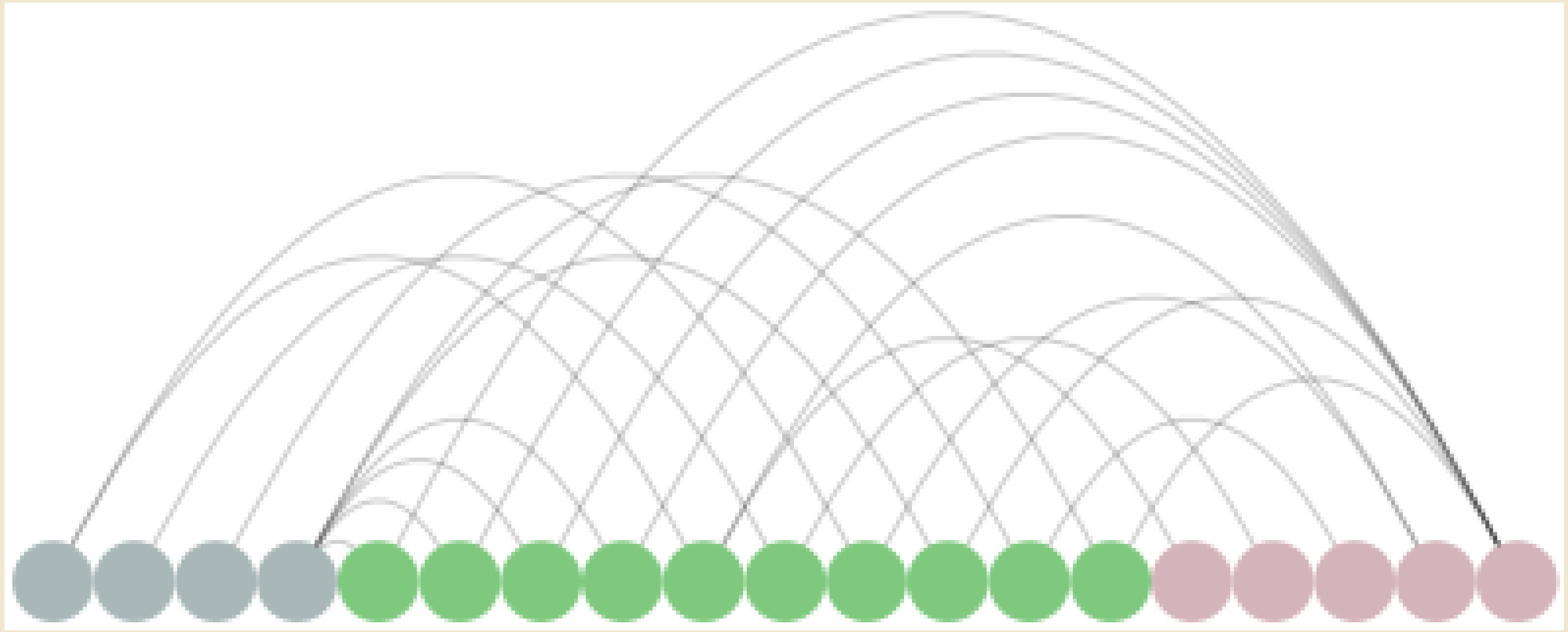
"I AM Minerva, the village poetess.."

"Will some one go to the village newspaper.."

William Goode (P1) - Minerva Jones (P2)

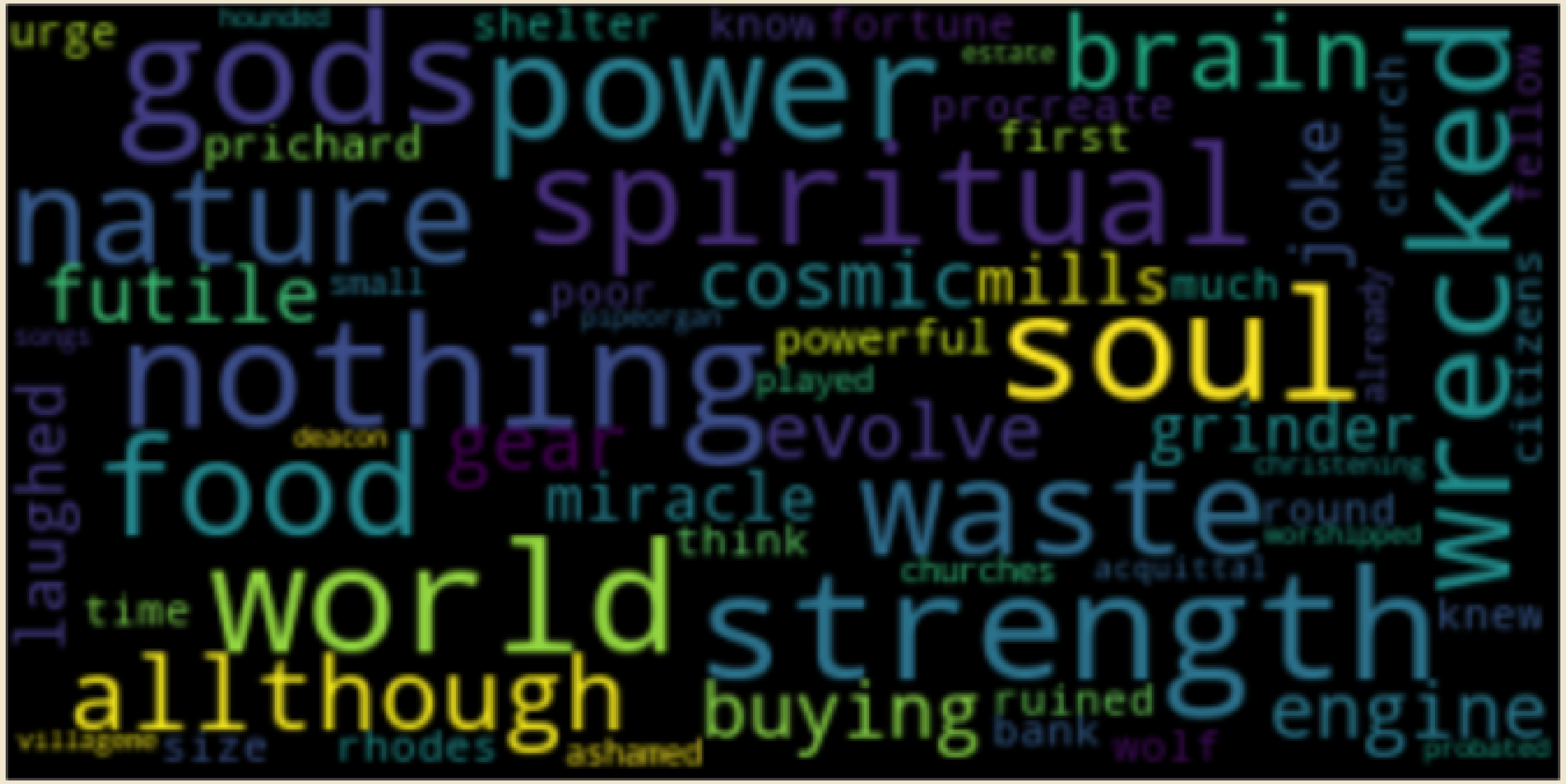
P1T1 * S1 * P2T1 --> Synonym Relationship

"capture" - "get" - "catch"
"hunt" - "search" - "wanderings"



William Goode (P1) - Minerva Jones (P2)

P1T1 !! P2T1 --> Nodes that did not contribute



While these words do provide overall value to the text.. None of these words are either shared across poems, or connected through synonyms. So, they do not provide any benefit to this analysis.

Create Similarity Matrix

Libraries Used

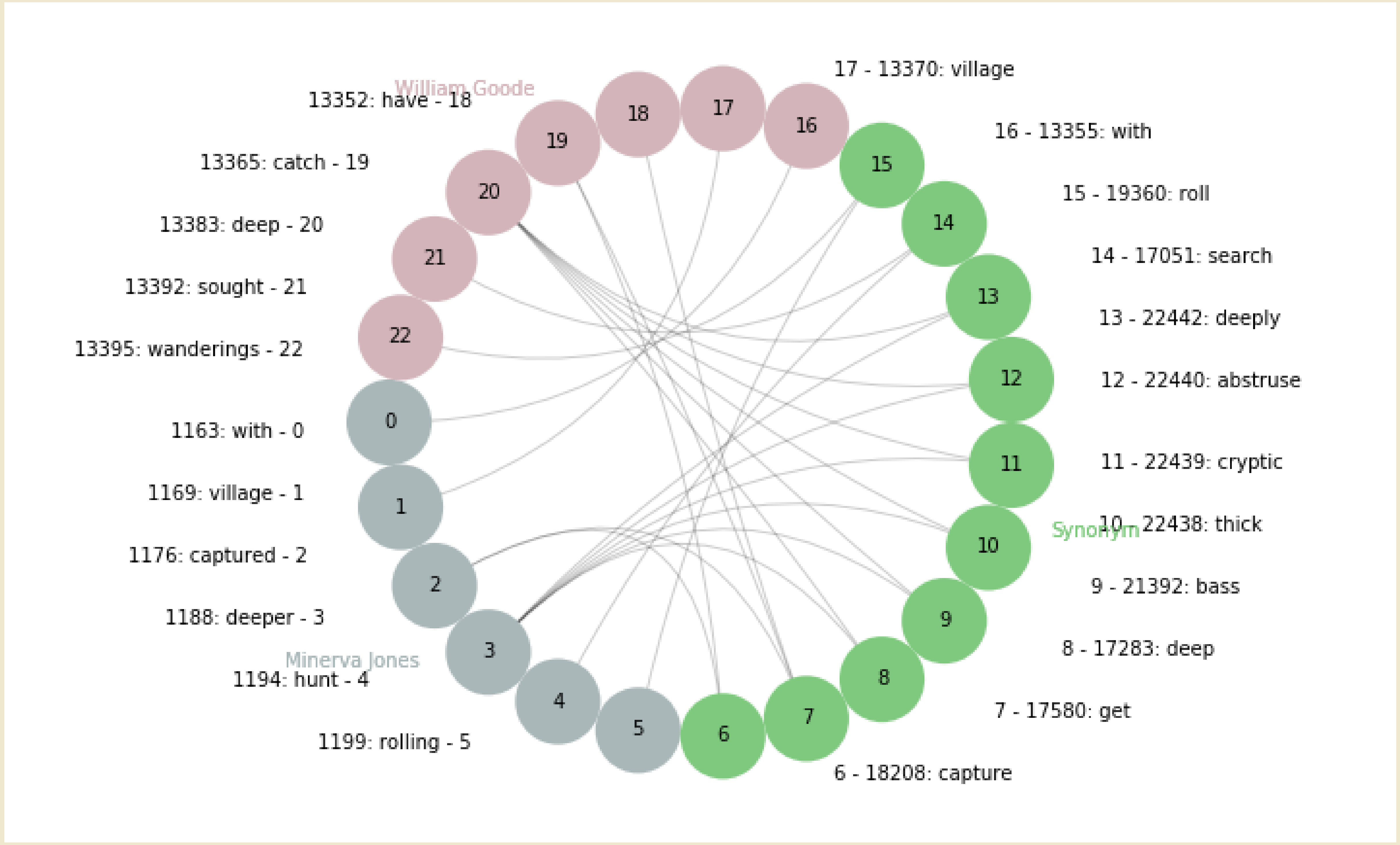
- Networkx
- Pandas

Process

For each combination, determine similarity score.

```
Overall Score +=
(
  sum(Same Terms tf-idf) +=
  (sum(Syn Relationships tf-idf) / len(contributing syns))
) / max_score
```

William Goode (P1) - Minerva Jones (P2)



Similarity Score - 0.25

# P1 Contributing Nodes - 7
# P2 Contributing Nodes - 6
# Synonyms - 10

# Direct Relationships - 2
# Unique Triangles - 11

Get Top 3 Scores

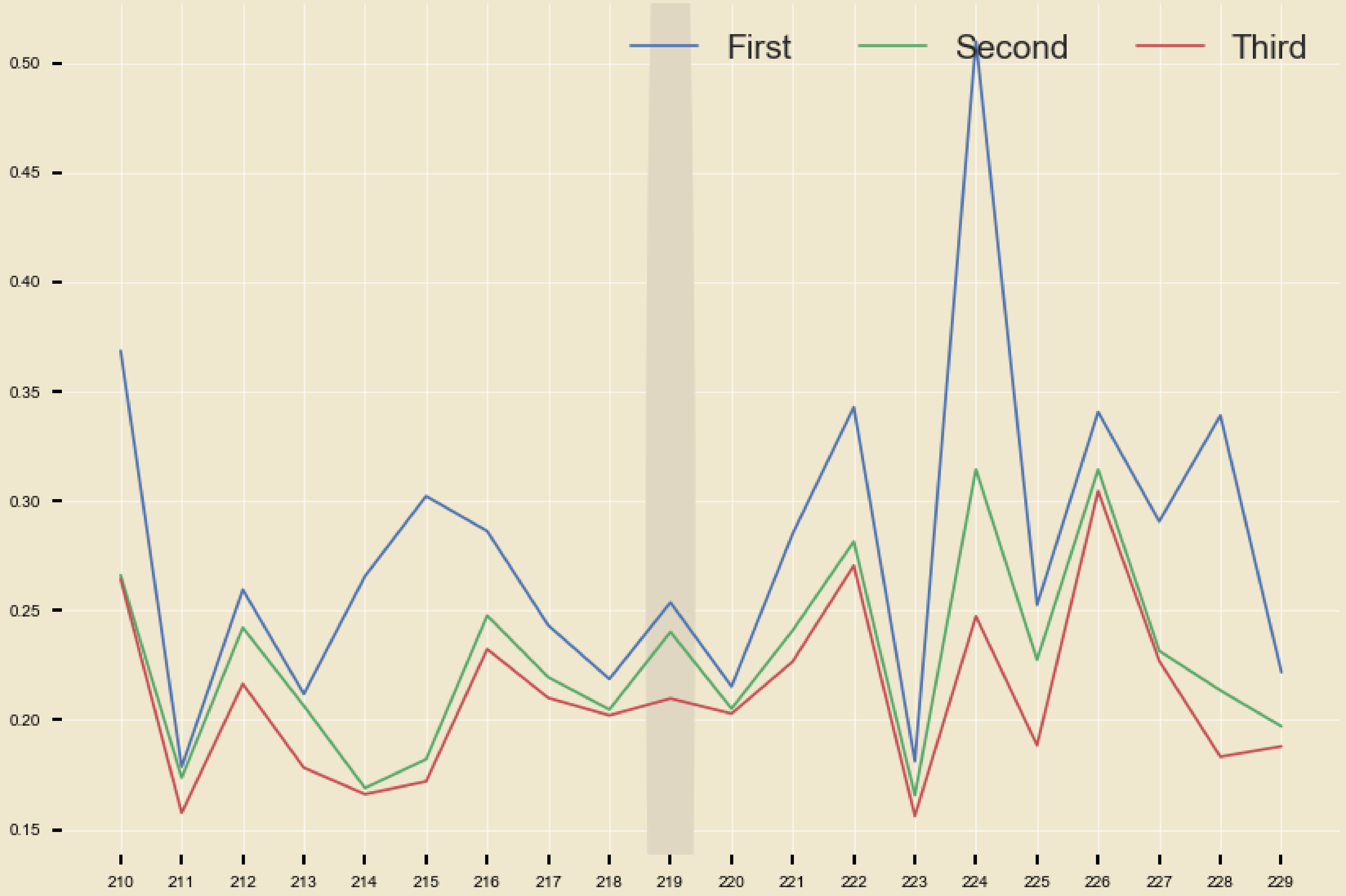
Libraries Used

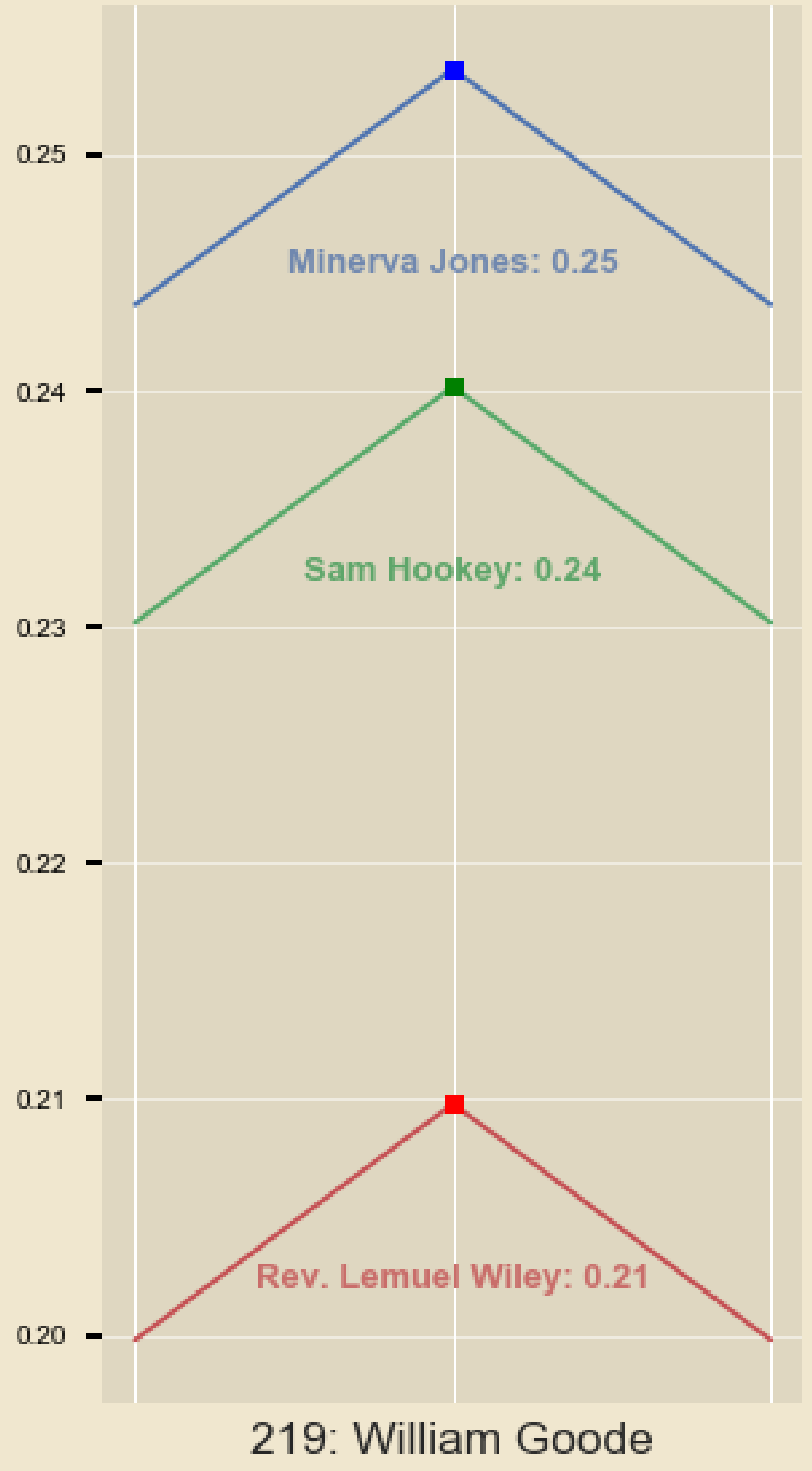
- Pandas

Process

For every poem, determine the top 3 scoring poems for easy comparison.

Top 3 Scores Per Poem





Basic App for Comparison

Libraries Used

- Flask

Process

Create an application that shows two poems side by side, their similarity score, and final network for easy reference.

Now Let's Compare!