Measuring Similarity Between the Works of English Language Poets

John Koenig
George Washington University

Natural Language Processing - Fall 2018

Intro - Project Statement

This paper will describe an experimental deployment of Natural Language Processing techniques and algorithms on an interesting type of text data - poetry. The researcher will use a publicly available English language poetry dataset to search for data insights, hidden among the lines of the poems, and then visualize the results.

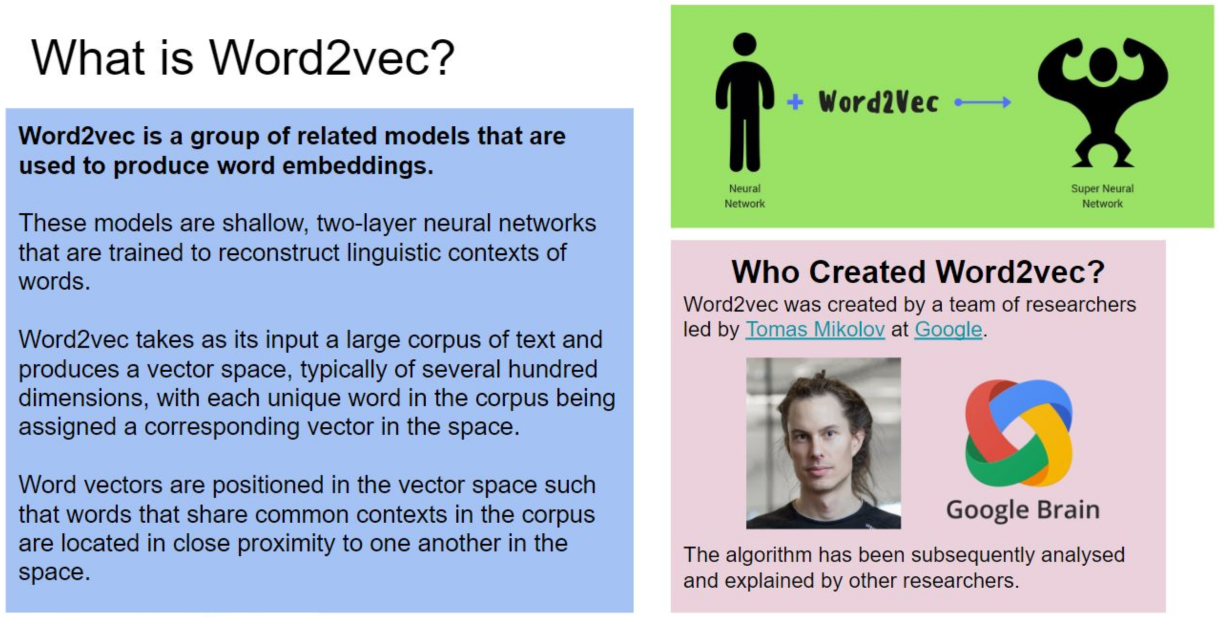
The researcher will train a Word2vec model on selected English language poetry from the Renaissance and Modern periods in order to compare similarities between poems, poets, and time periods.

About Natural Language Processing

Natural Language Processing (NLP) consists of a collection of methodologies that can be used to extract meaningful information from text data with the assistance of powerful computing resources. The NLP methodology that the researcher will deploy for this project will use a neural network to encode "word vectors" for each word in a poem so that each word can be compared and analyzed. This popular algorithm is called "Word2vec."

About Word2vec

Word2vec is an NLP algorithm that takes a "corpus" of documents and the extracts/creates unique vector for each word in vector space. The number of dimensions that represents each word is a feature of how the algorithm is designed.



What is Word2vec?

Word2vec is a group of related models that are used to produce word embeddings.

These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.

Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space.

Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

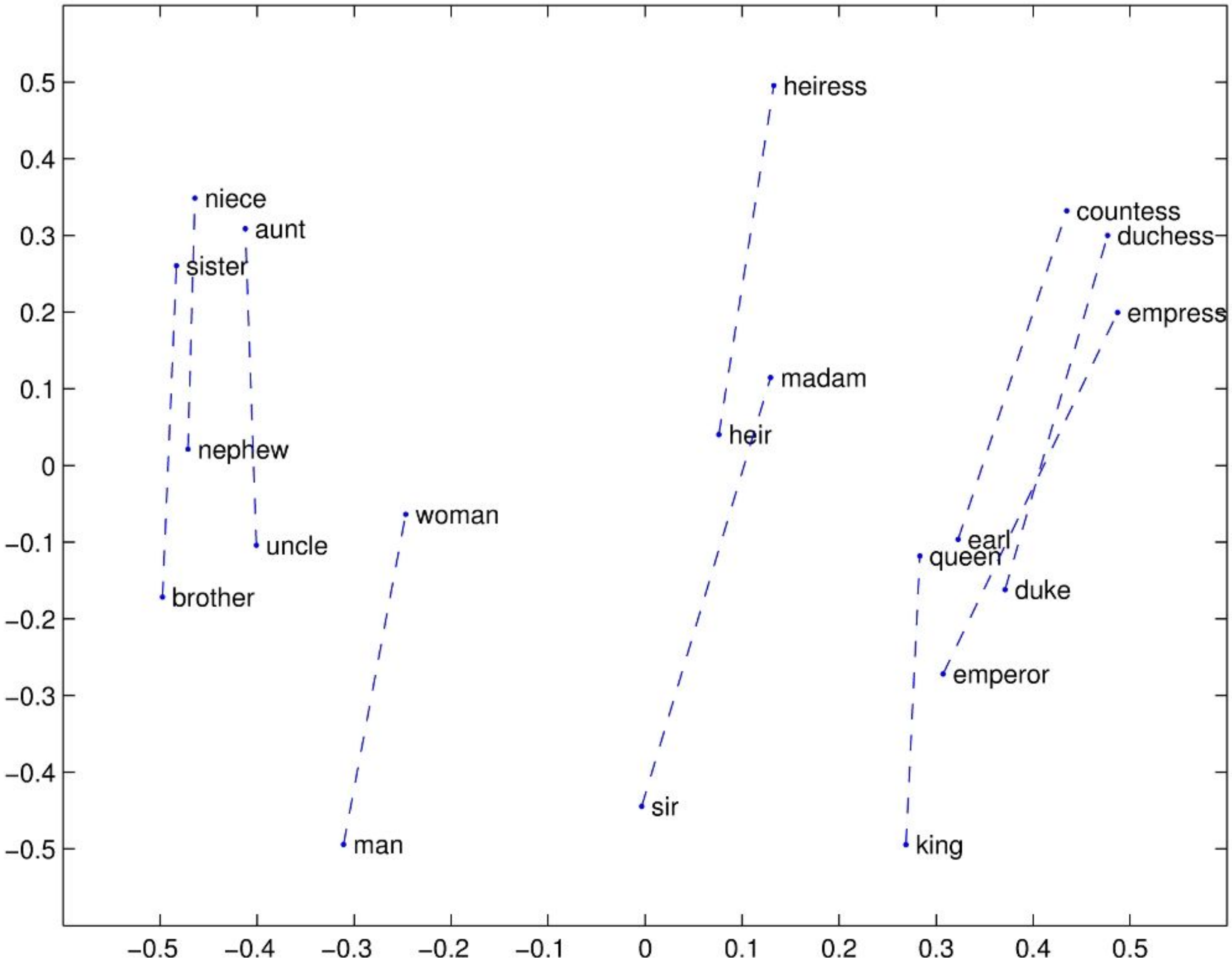
Who Created Word2vec?

Word2vec was created by a team of researchers led by Tomas Mikolov at Google.

The algorithm has been subsequently analysed and explained by other researchers.

Source: https://en.wikipedia.org/wiki/Word2vec

Each word is represented by the same number of dimensions. The reader will notice that the relationships between similar words are closer together in this 2 dimensional Word2vec example:



Source: https://nlp.stanford.edu/projects/glove/

About Poetry

Poetry is the use of language in order to communicate in a terse and emotive manner. Poetry comes in many forms and styles vary from culture to culture and from time period to time period. However, there are several characteristics of poetry that are important:[1]

- ★ Economy of Language - Communicating with carefully selected words for clarity
- ★ Intense Emotion - Provoking emotion from the reader
- ★ Continual Evolution - Poetry continues to develop and change

[1] https://www.thoughtco.com/what-is-poetry-852737

About Renaissance Poetry

From Wikipedia: Renaissance Literature

> "Renaissance literature refers to European literature which was influenced by the intellectual and cultural tendencies associated with the Renaissance. The literature of the Renaissance was written within the general movement of the Renaissance which arose in 14th-century Italy and continued until the 16th century while being diffused into the rest of the western world. It is characterized by the adoption of a humanist philosophy and the recovery of the classical Antiquity. It benefited from the spread of printing in the latter part of the 15th century. For the writers of the Renaissance, Greco-Roman inspiration was shown both in the themes of their writing and in the literary forms they used. The world was considered from an anthropocentric perspective. Platonic ideas were revived and put to the service of Christianity. The search for pleasures of the senses and a critical and rational spirit completed the ideological panorama of the period. New literary genres such as the essay (Montaigne) and new metrical forms such as the Spenserian stanza made their appearance."

Renaissance poetry was an important subset of the literature produced during this time in Western Europe.

Here is an example:

William Shakespeare -- Sonnet 147: My love is as a fever, longing still

My love is as a fever, longing still
For that which longer nurseth the disease,
Feeding on that which doth preserve the ill,
Th' uncertain sickly appetite to please.
My reason, the physician to my love,
Angry that his prescriptions are not kept,
Hath left me, and I desperate now approve
Desire is death, which physic did except.
Past cure I am, now reason is past care,
And frantic-mad with evermore unrest;
My thoughts and my discourse as madmens are,
At random from the truth vainly expressed:
  For I have sworn thee fair, and thought thee bright,
  Who art as black as hell, as dark as night.

About Modern Poetry

From Wikipedia: Modernist Poetry in English

> "Modernist poetry in English started in the early years of the 20th century with the appearance of the Imagists. In common with many other modernists, these poets wrote in reaction to the perceived excesses of Victorian poetry, with its emphasis on traditional formalism and ornate diction.
>
> Modernists saw themselves as looking back to the best practices of poets in earlier periods and other cultures. Their models included ancient Greek literature, Chinese and Japanese poetry, the troubadours, Dante and the medieval Italian philosophical poets (such as Guido Cavalcanti), and the English Metaphysical poets.
>
> Much of early modernist poetry took the form of short, compact lyrics. As it developed, however, longer poems came to the foreground. These represent the modernist movement to the 20th-century English poetic canon."

"Modern Poetry" does not mean "Modern" - as in "current." This period of poetry is over and lasted from the late 19th-century to the mid-20th century. The poetry of today would be known as "Post-Modern Poetry."

Here is an example:

Wallace Stevens -- Fabliau of Florida

Barque of phosphor
On the palmy beach,

Move outward into heaven,
Into the alabasters
And night blues.

Foam and cloud are one.
Sultry moon-monsters
Are dissolving.

Fill your black hull
With white moonlight.

There will never be an end
To this droning of the surf.

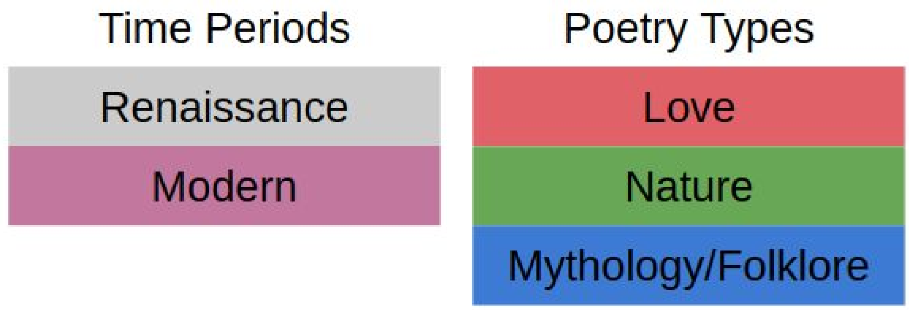
Methodology

The researcher will apply a variety of Natural Language Processing (NLP) techniques in order to process and analyze the Project Dataset:

1. Custom Segmentation, Tokenization, and Document Processing
2. Extract Key Metrics Document Metric
3. Train Word2vec Model
4. Compare Similarities Between Poems and Poets
5. Create Custom Data Visualizations of Results

Data Source

The data source for this project is the “Poems by PoemsFoundation.org”[2] dataset. The dataset (after preliminary cleaning) has 67 poets, 420 poems, and 11,330 lines of poetry. Categorical fields include “Time Periods” and “Poetry Types.”



There are 315 Renaissance poems and 160 Modern poems. There are 297 “Love” poems, “132” Nature poems, and 46 “Mythology & Folklore” poems.

Project Results

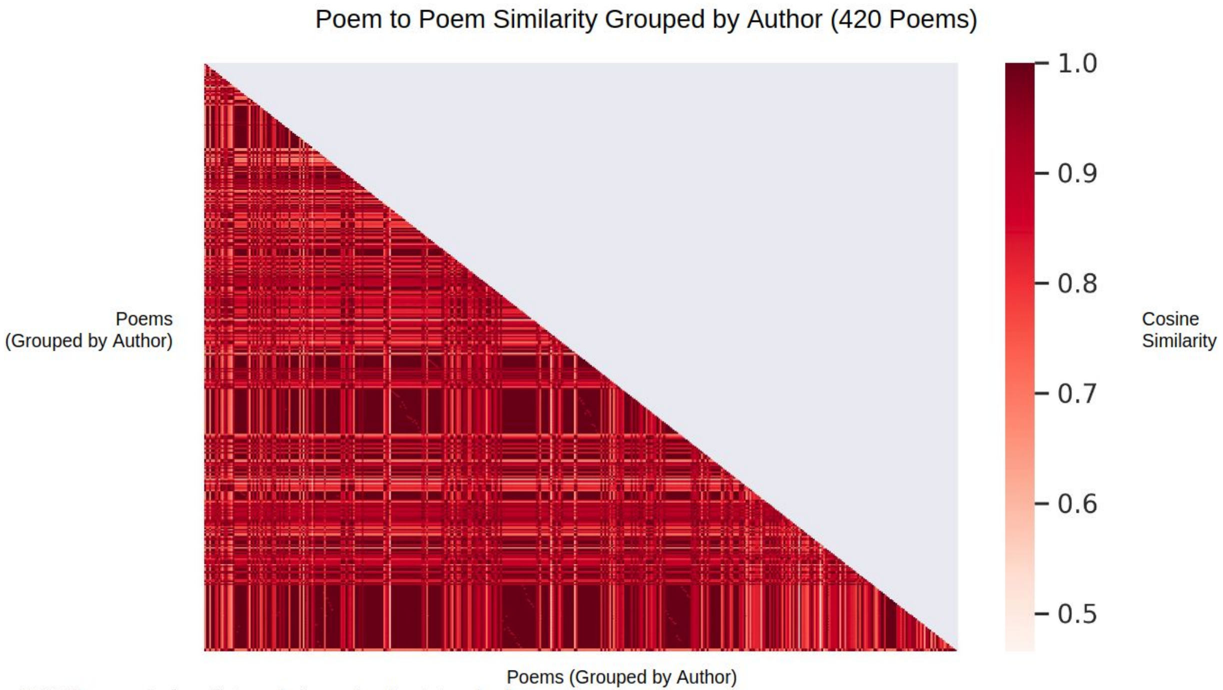
The researcher completed 3 custom scripts in order to meet the software development goals for this project:

- **NLP_poetry_v1.py**
  - Use pandas and spacy to tokenize and extract metrics by poem line
- **NLP_word2vec_v1.py**
  - Use gensim to train word2vec model on project dataset
- **NLP_poetry_similarities_v1.py**
  - Calculate similarities and visualize using seaborn

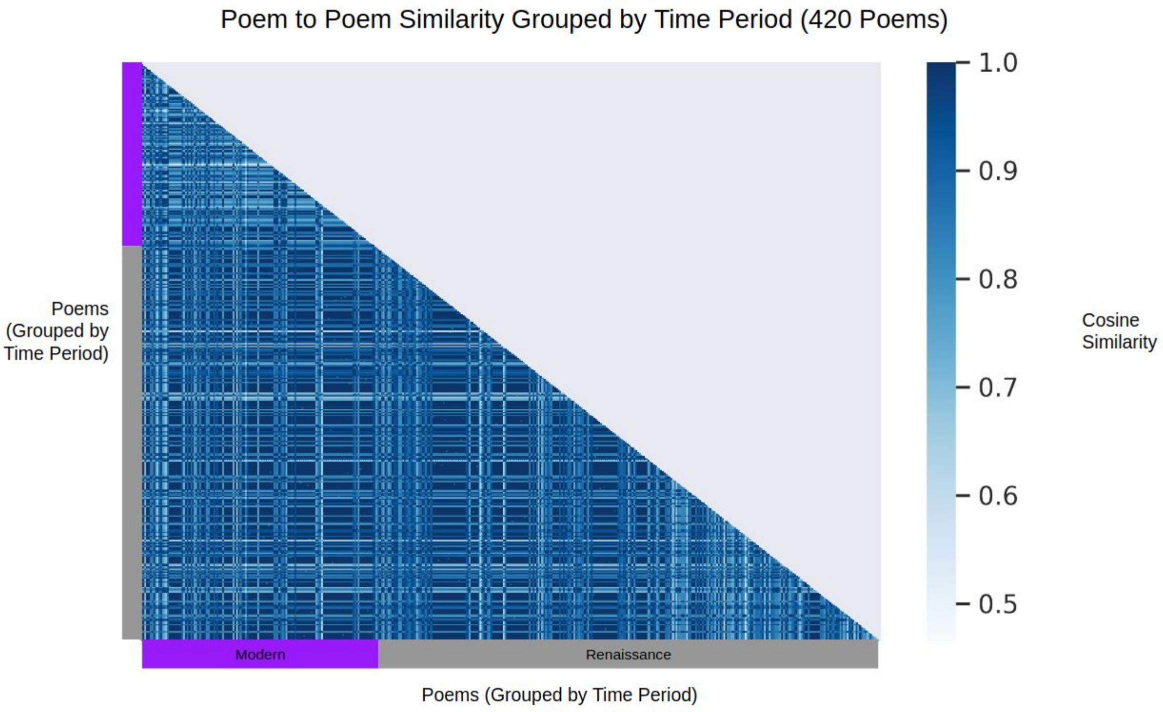
Custom visualizations are included on the next page...

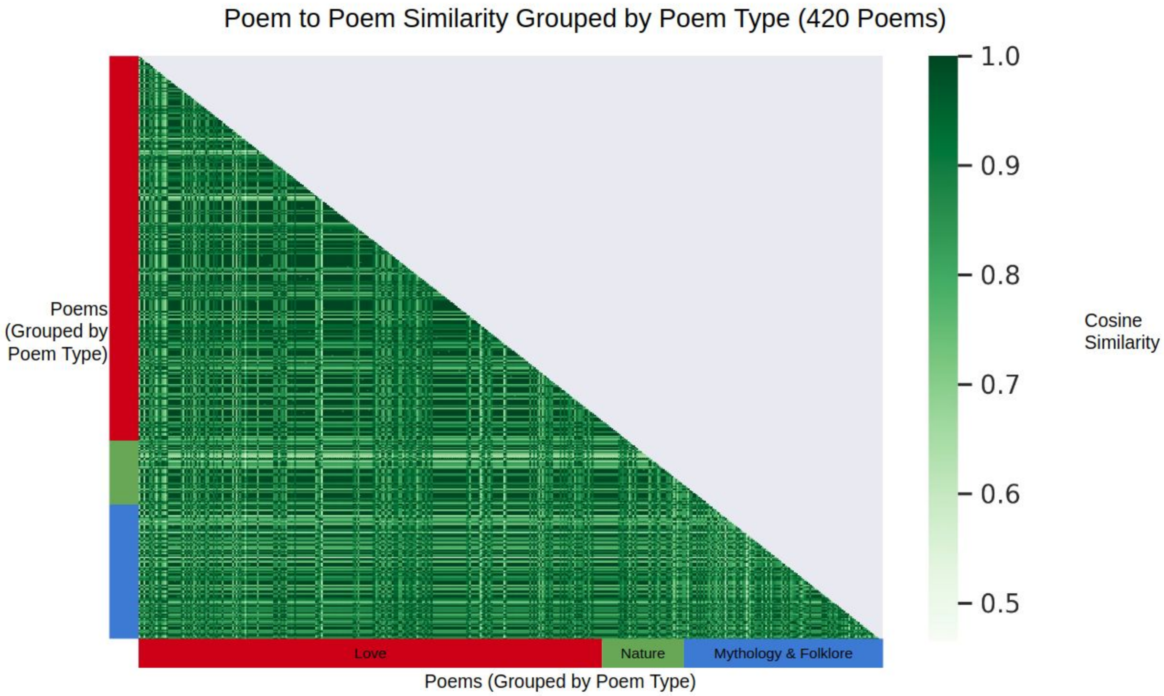
[2] https://www.kaggle.com/ultrajack/modern-renaissance-poetry

Poem to Poem Similarity

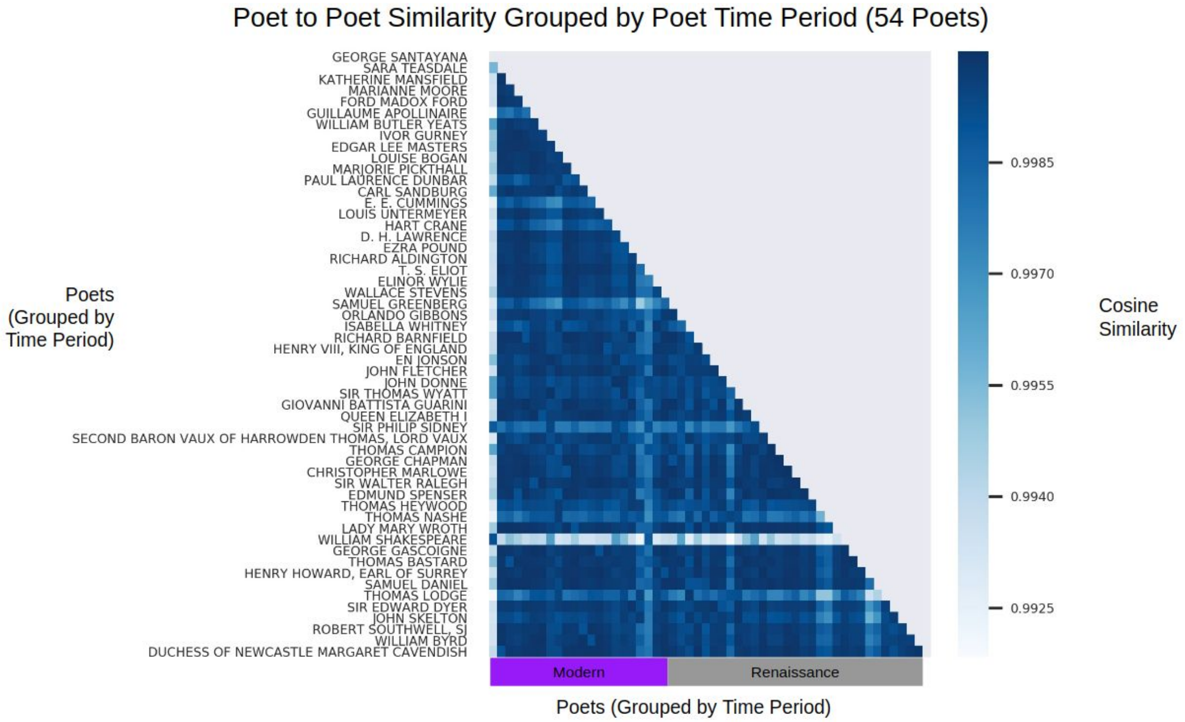


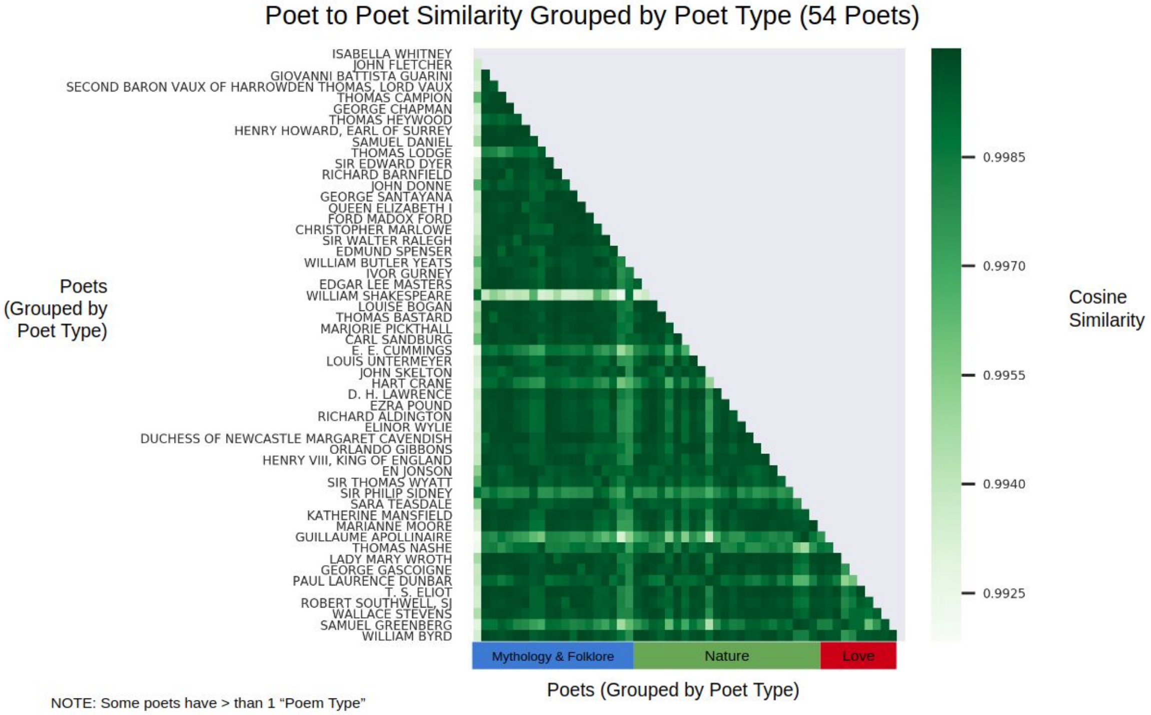
NOTE: The researcher is unable to mark where each author starts and ends




Poem to Poem Similarity Grouped by Poem Type (420 Poems)
Poems (Grouped by Poem Type)
Love
Nature
Mythology & Folklore
Poems (Grouped by Poem Type)
1.0
0.9
0.8
0.7
0.6
0.5
Cosine Similarity

Poet to Poet Similarity





NOTE: Some poets have > than 1 "Poem Type"

Conclusions

Clear patterns and similarities do exist between the different poems, poets, poet time periods, and poem types involved in this project. The strongest similarities exist between poems authored by the same author (this makes sense). Similarity is high among most poems and poets, although the project did show that some poems and and poets are “unique” (low similarities to other authors). Though some patterns exist, there does not seem to be strong similarity between poems of the same “poet time period” or “poem type”

Opportunities exist for further research and work on creating more advanced visualizations for this project.