Measuring Similarity Between the Works of English Language Poets

John Koenig
George Washington University
Natural Language Processing - Fall 2018

Contents

Project Statement

The researcher will determine what patterns and similarities exist within a custom dataset comprised of English language poems.

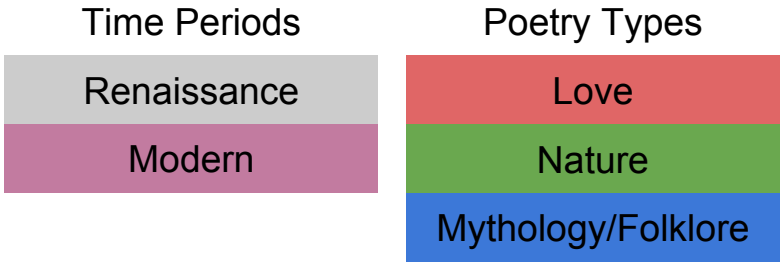
The Data

The “Project Dataset” includes 500 English language poems written by 50 unique authors over 2 time periods. Poems are categorized as 1 of 3 poetry types.

13,244 Lines

475 Poems

67 Poets



Methodology

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

- Custom Segmentation, Tokenization, and Document Processing
- Extract Key Metrics Document Metric (using Pandas)
- Train Skip-Gram Model (using Gensim)
- Compare Similarities Between Authors
- Create Custom Data Visualizations of Results (using 3Data)

What is Gensim?

gensim

topic modelling for humans

Gensim is a robust open-source vector space modeling and topic modeling toolkit implemented in Python.

It uses NumPy, SciPy and Cython (for performance).

Gensim is specifically designed to handle large text collections, using data streaming and efficient incremental algorithms, which differentiates it from most other scientific software packages that only target batch and in-memory processing.

Selected content from Wikipedia

Main Features

- TF_IDF
- Random Projections
- Word2vec
- Document2vec
- Hierarchical Dirichlet Processes (HDP)
- Latent Semantic Analysis (LSA, LSI, SVD)
- Latent Dirichlet Allocation (LDA)

Who Created Gensim?

Gensim was created by Radim Řehůřek and a group of other contributors:

Founder at RARE Technologies, creator of Gensim. SW engineer since 2004, PhD in AI in 2011. Lover of geology, history and beginnings in general. Occasional travel blogger.

Selected content from https://radimrehurek.com/gensim

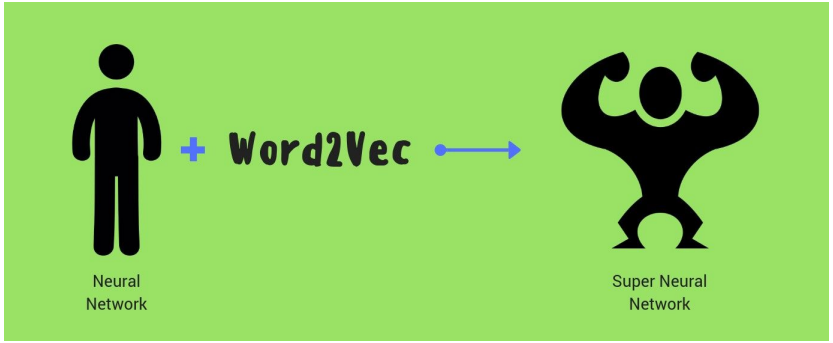
What is Word2vec?

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.

Google Brain

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

Selected content from Wikipedia

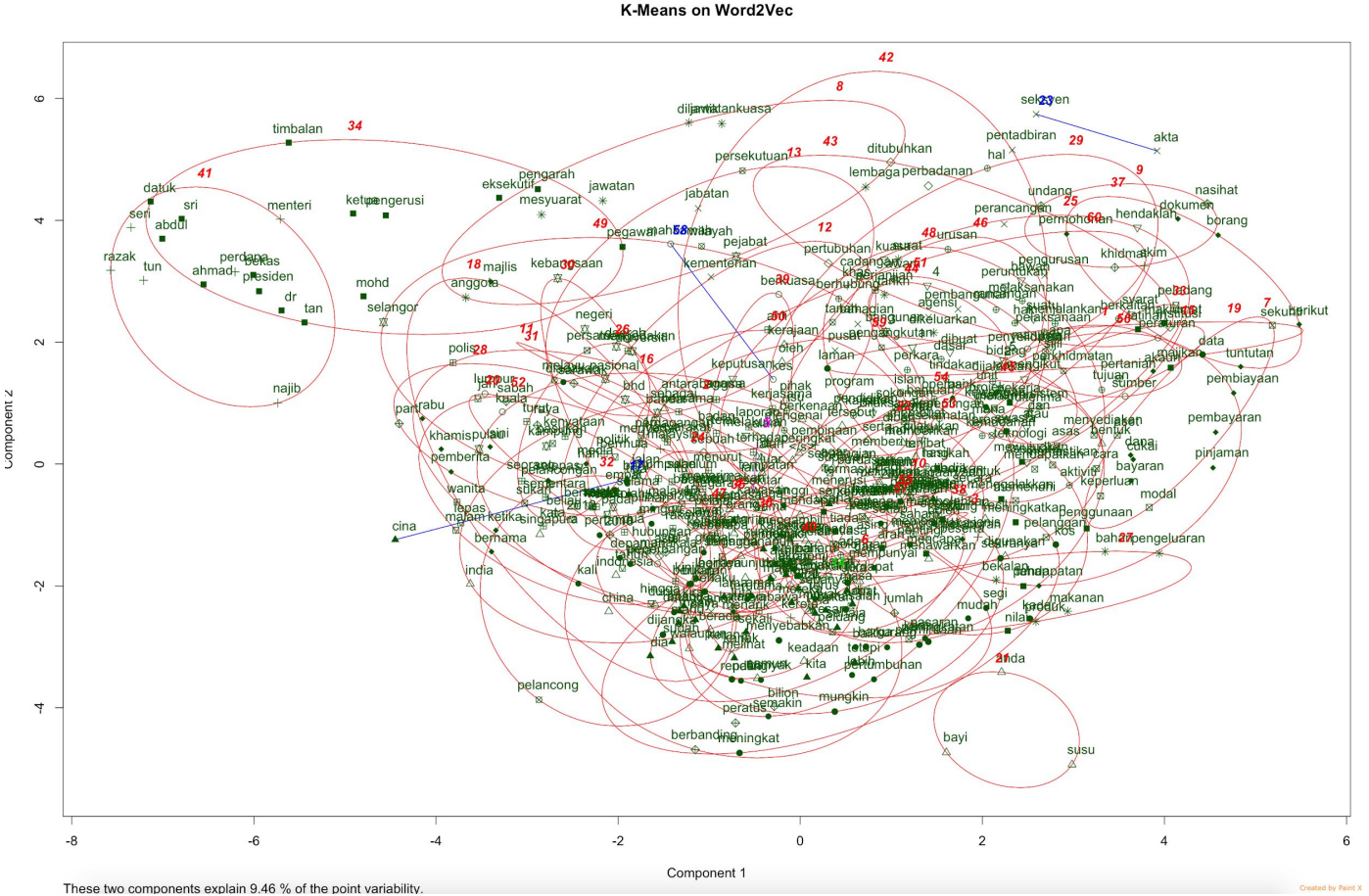
Continuous Bag-of-Words (CBOW)	Context Words predict Center Word	Order Doesn't Matter
Skip-Gram	Center Word predicts Context Words	Order Matters

Center Word	Context Words
[1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0] [0, 0, 1, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0] [0, 0, 1, 0, 0, 0, 0] [0, 0, 0, 1, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0] [0, 1, 0, 0, 0, 0, 0] [0, 0, 0, 1, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 1, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0] [0, 0, 1, 0, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 1, 0, 0]	[0, 0, 1, 0, 0, 0, 0] [0, 0, 0, 1, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 1, 0, 0, 0] [0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 1, 0, 0] [0, 0, 0, 0, 0, 1, 0]

Skip-Gram is slower but works better with infrequent words

Word2vec

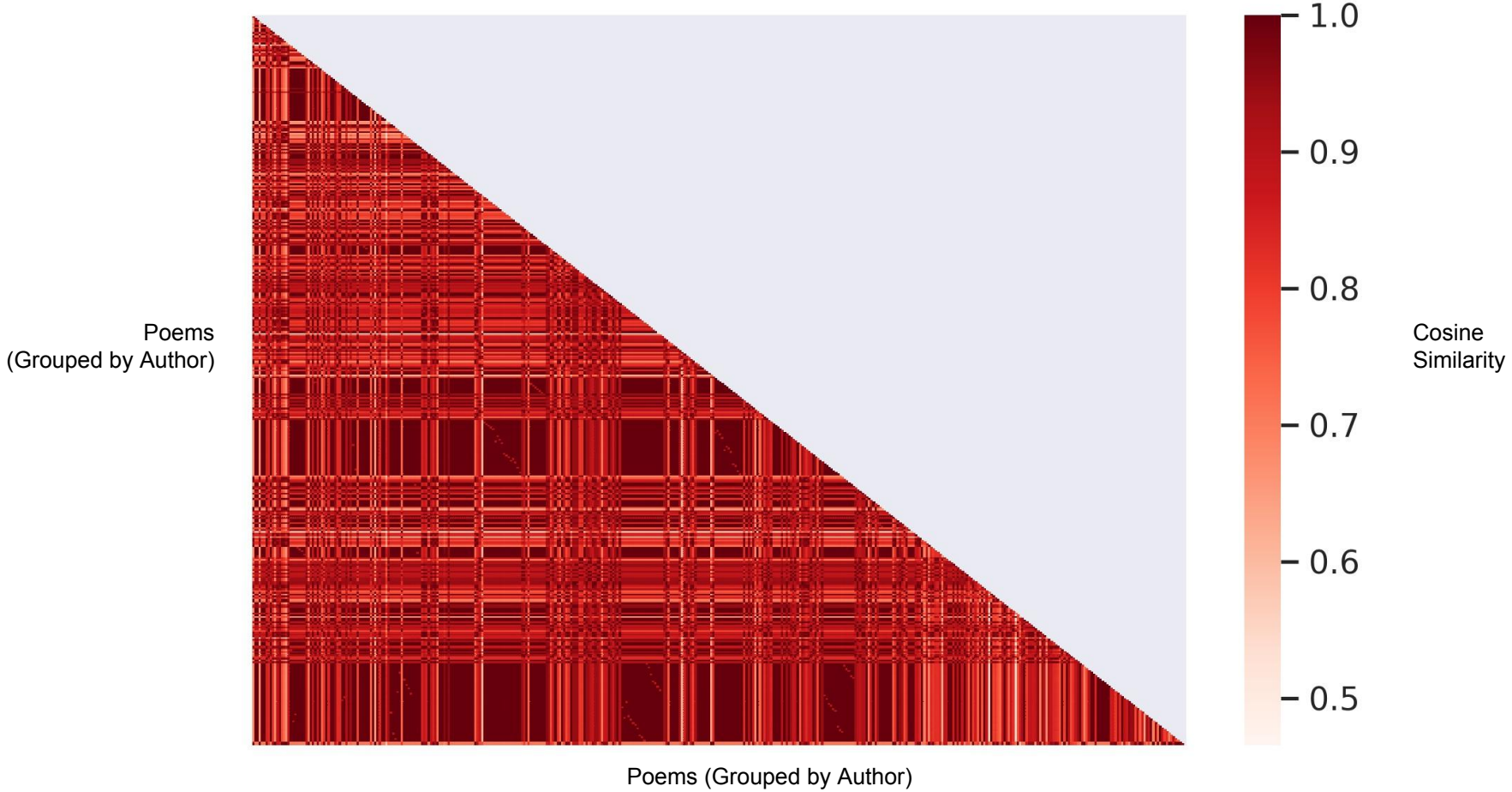
Word embeddings can be used to analyze documents and cluster them based on various similarity measures



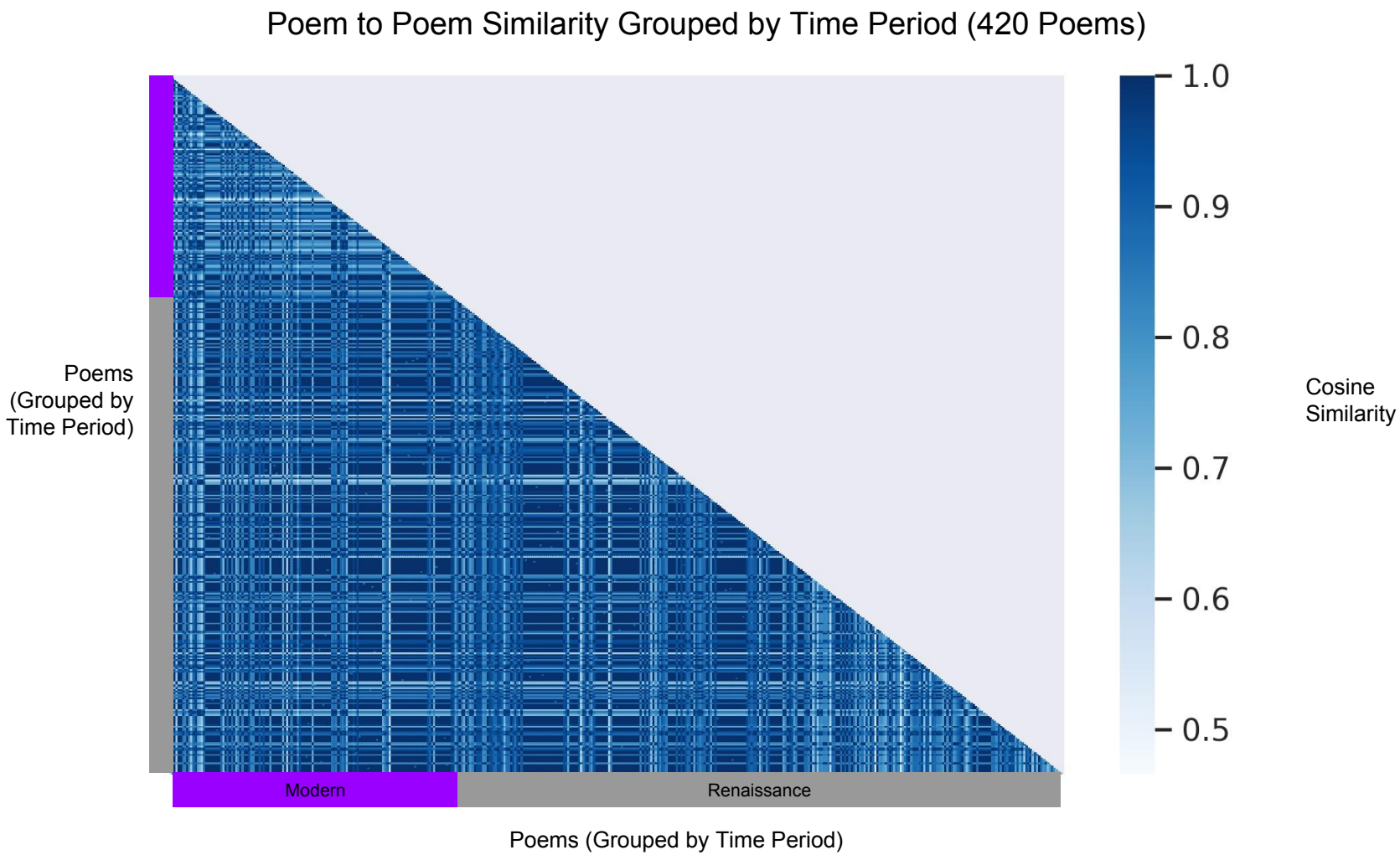
Example from https://kimusu2008.github.io/Clustering_word2vec/

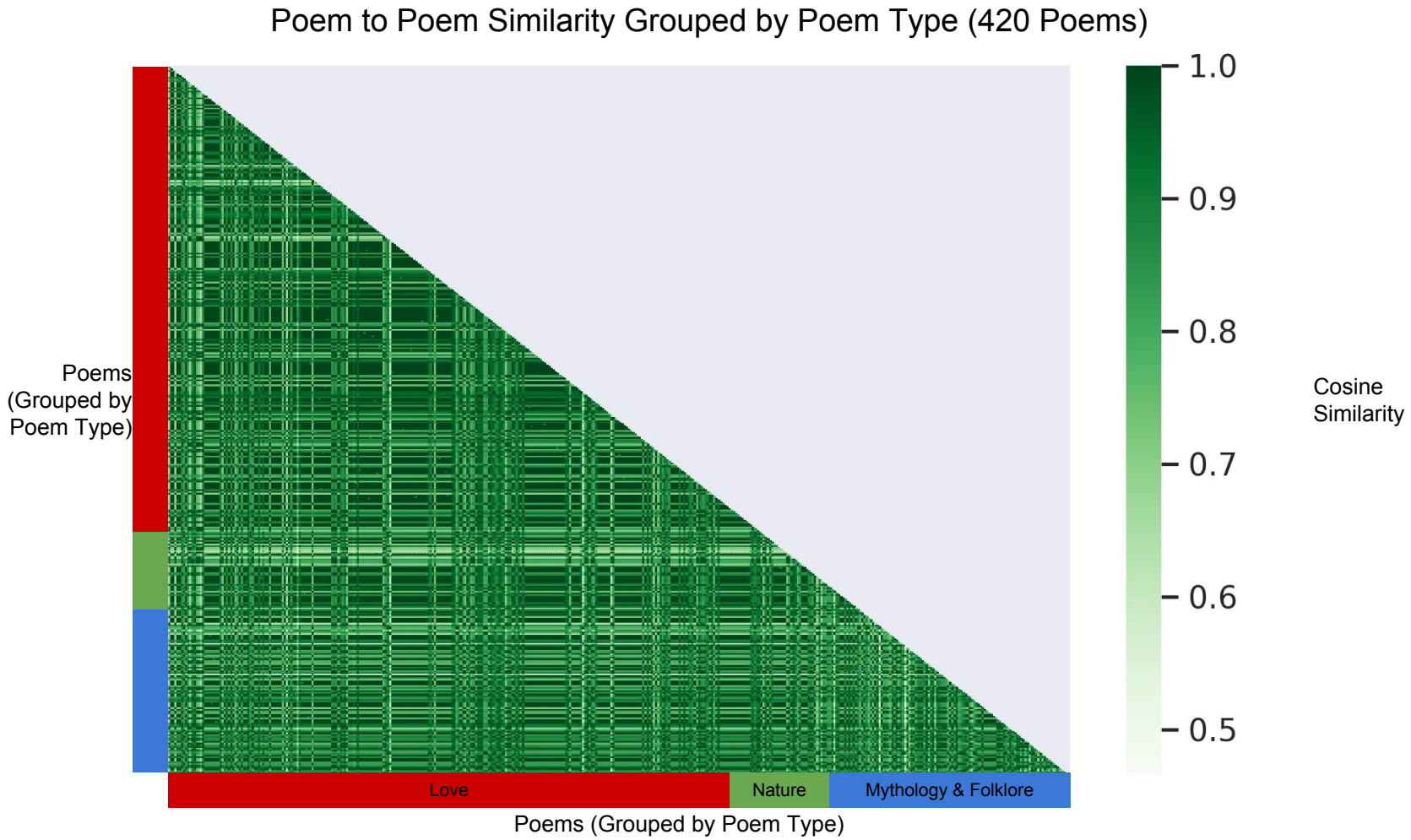
Project Results

Poem to Poem Similarity Grouped by Author (420 Poems)



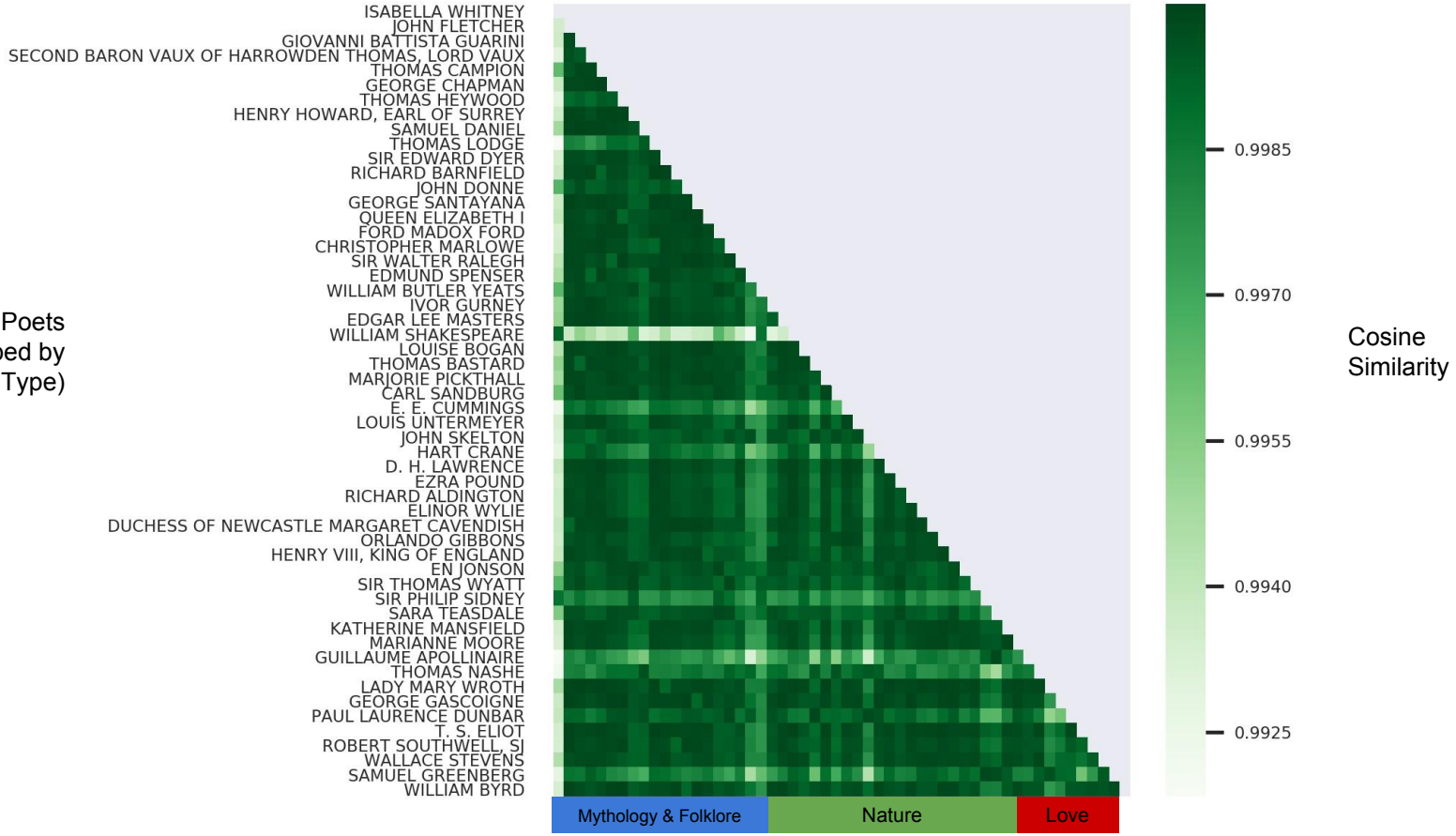
NOTE: The researcher is unable to mark where each author starts and ends


Poem to Poem Similarity Grouped by Time Period (420 Poems)
Poems
(Grouped by
Time Period)
Modern
Renaissance
Poems (Grouped by Time Period)
1.0
0.9
0.8
0.7
0.6
0.5
Cosine
Similarity


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 Grouped by Poet Time Period (54 Poets)

Poet to Poet Similarity Grouped by Poet Type (54 Poets)



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”

Contact

For inquiries about this research project, please contact:

John Koenig
jpk11830@gwmail.gwu.edu