**Cluster File Descriptions**

This document contains a description of any files that contain cluster information.

When we’ve determined the “best” parameters for a specific clustering technique, we should highlight that cluster’s description in this document.

File: Clusters\_1.txt

Original data vectorization: 20,000 2-grams from each document (Vectors\_1.csv, Vectors\_1\_Names.txt)

Dimensionality reduction: PCA to 2 dimensions (Reduced\_1.csv)

Clustering method: Lloyd’s, 8 centers, initialized with k-means++

File: Clusters\_2.txt

Original data vectorization: 20,000 2-grams from each document (Vectors\_1.csv, Vectors\_1\_Names.txt)

Dimensionality reduction: PCA to 2 dimensions (Reduced\_1.csv)

Clustering method: Lloyd’s, 10 centers, initialized with k-means++

Notes:

Genre-wise, I would classify cluster 0 as containing pulp and children’s books (with the exception of *The Great Gatsby*), cluster 1 as romantic and epic literature 1, cluster 2 as romantic and gothic literature, cluster 3 as short novels and poetry, cluster 4 as adventure stories and Shakespeare, cluster 5 as ancient poetry 1, cluster 6 as romantic and epic literature 2, cluster 7 as ancient poetry 2, cluster 8 as pulp and children’s books 2, and cluster 9 as romantic literature (with the exception of the Odyssey).

Author-wise, H.G Wells’s *The Invisible Man* and *The Time Machine* are both in cluster 0, Leo Tolstoy’s *War and Peace* and *Anna Karenina* are both in cluster 1, Charles Dickens’s *Great Expectation* and *A Tale of Two Cities* are both in cluster 2, Jane Austin’s *Sense and Sensibility, Emma*, and *Pride and Prejudice* were all in cluster 2, Jules Verne’s *A Journey to the Center of the Earth* and *20,000 Leagues under the Sea* were in cluster 4, Shakespeare’s *Macbeth, Hamlet*, and *Romeo and Juliet* were in cluster 4, and Mark Twain’s *The Adventures of Tom Sawyer* and *The Adventures of Huckleberry Finn* were both in cluster 8. Unfortunately, Charles Dickens’s *A Christmas Carol* was not with his other works; instead, it was with the other short novellas. Also, Jules Vernes’s *Around the World* was with pulp and children’s books, not with Vernes’s other books. Finally, Alexandre Dumas’s *The Three Musketeers* and *The Count of Monte Cristo* were in separate clusters (although they were both with other romantic books).

File: Clusters\_3.txt

Original data vectorization: 20,000 2-grams from each document (Vectors\_2.csv, Vectors\_2\_Names.txt)

Dimensionality reduction: PCA to 5 dimensions (Reduced\_2.csv)

Clustering method: Lloyd’s, 7 centers, initialized with k-means++

Notes:

I’ve marked this as the best Lloyd clusters. Generally, the clusters fall into good genre divisions, and authors are grouped together.

Genre-wise, I’d classify cluster 0 as poems and short novels, cluster 1 as romantic literature (with the exception of *The Odyssey*), cluster 2 as adventure and romance (with the exception of Plutarch’s *Lives*), cluster 3 as children’s and pulp books (with the exception of *The Great Gatsby*), cluster 4 as ancient poetry, cluster 5 as epics, and cluster 6 as Jane Austen.

Author-wise, Charles Dickens’s *Great Expectation* and *A Tale of Two Cities* were in cluster 1, Jules Verne’s *Around the World in 80 Days*, *A Jounrey to the Ceter of the Earth*, and *20,000 Leagues under the Sea* were in cluster 2, Mark Twain’s *The Adventures of Tom Sawyer* and *The Adventures of Huckleberry Finn* were in cluster 3, H.G. Welles’s *The Invisible Man* and *The Time Machine* were in cluster 3, Shakespeare’s *Macbeth, Hamlet*, and *Romeo and Juliet* were in cluster 4, Leo Tolstoy’s *War and Peace* and *Anna Karenina* were in cluster 5, and Jane Austen’s *Sense and Sensibility, Emma,* and *Pride and Prejudice* were together in cluster 6. Unfortunately, Alexandre Dumas’s *The Three Musketeers* and *The Count of Monte Cristo* were in different clusters (of course, both of these were translations).

File: Clusters\_4.txt

Original data vectorization: 20,000 3-grams from each document (Vectors\_3.csv, Vectors\_3\_Names.txt)

Dimensionality reduction: PCA to 2 dimensions (Reduced\_3.csv)

Clustering method: Lloyd’s, 6 centers, initialized with Gonzalez’s algorithm

Notes:

Genre-wise, I’d classify cluster 0 as children’s and pulp (with the exception of *The Great Gatsby*), cluster 1 as Emily Dickenson, cluster 2 as epics and ancient texts, cluster 3 as adventure, cluster 4 as romantic (except *The Odyssey*), and cluster 5 as short novels and poems.

Author-wise, *Around the World in 80 Days* is separated from the other Verne works, *The Three Musketeers* and *The Count of Monte Cristo* were separate, *Anna Karenina* and *War and Peace* were separated, and, of course, *A Christmas Carol* was separate from the other Dickens novels.

File: Gauss\_Clusters\_1.txt

Original data vectorization: 20,000 2-grams from each document (Vectors\_1.csv, Vectors\_1\_names.txt)

Dimensionality reduction: PCA to 2 dimensions (Reduced\_1.csv)

Clustering method: mixture of Gaussians, 6 centers, initialzed with Lloyd’s algorithm

Notes:

I’ve marked this as the best mixture of Gaussians clustering. Truth be told, I haven’t noticed too much difference between different runs of mixture of Gaussians.

In this clustering, I might draw attention to the fact that books like *A Tale of Two Cities* and *The Three Musketeers* can be grouped with 19th century romantic literature like *Pride and Prejudice*, Edgar Allen Poe, *Dracula*, and *Gulliver’s Travels*, but they can also be grouped with more pulpy, adventure-driven novels like *Around the World in 80 Days*, H.G. Wells, and *Tarzan*.