Supplementary Information - Identification of Literary Movements Using Complex Networks to Represent Texts

Diego Raphael Amancio, Osvaldo N. Oliveira Jr., Luciano da Fontoura Costa

Contents

1 List of Stopwords 2
2 Database 2
3 Dimensionality Reduction 6
4 Quantifying Clustering Quality 6
5 Clusters 8
1. List of Stopwords

The list below shows the stopwords that were removed from the text in the pre-processing step. Thus, all of these words are disregarded in the analysis of writing style.

I A ABOARD ABOUT ABOVE ACROSS AFTER AFTERWARDS AGAINST AGIN AGO AGREED-UPON AH ALAS ALBEIT ALL ALL-OVER ALMOST ALONG ALONGSIDE ALTHO ALTHOUGH AMID AMIDST AMONG AMONGST AN AND ANOTHER ANY ANYONE ANYTHING AROUND AS ASIDE ASTRIDE AT ATOP AVEC AWAY BACK BE BECAUSE BEFORE BEFOREHAND BEHIND BEHYNDE BELOW BENEATH BESIDE BESIDES BETWEEN BEWTEEN BEYOND BI BOTH BUT BY CA. DE DES DESPITE DO DOWN DUE DURIN DURING EACH EH EITHER EN EVERY EVERYONE EVERYTHING EXCEPT FAR FER FOR FROM GO GODDAMN GOODY GOSH HALF HAVE HE HELL HER HERSELF HEY HIM HIMSELF HIS HO HOW HOWEVER IF IN INSIDE INSOFAR INSTEAD INTO IT ITS ITSELF LA LE LES LEST LIEU LIKE ME MINUS MOREOVER MY MYSELF NEAR NEAR-BY NEARER NEAREST NEITHER NEVERTHELESS NEXT NO NOR NOT NOTWITHSTANDING O O’ER OF OFF ON ONCE ONE ONESelf ONLY ONTO OR OTHER OTHERS OTHERWISE OUR OURS OURSELVES OUT OUTSIDE OUTTA OVER PER RATHER REGARDLESS ROUND SE SHE SHOULD SINCE SO SOME SOMEONE SOMETHING THAN THAT THE THEIR THEM THEMSELVES THEN THERE THEREFORE THESE THEY THINE THIS THOSE THOU THOUGH THROUGH THROUGHOUT THRU TILL TO TOGETHER TOWARD TOWARES TOWARDS UH UNDER UNDERNEATH UNLESS UNLIKE UNTIL UNTO UP UPON UPPON US VIA VIS-A-VIS VIS-à-VIS WE WELL WHAT WHATEVER WHATSOEVER WHEN WHenever WHERE WHEREAS WHEREFORE WHEREUPON WHETHER WHICH WHICHEVER WHILE WHO WHOEVER WHOM WHOSE WHY WITH WITHAL WITHIN WITHOUT YE YEA YEAH YES YET YONDER YOU YOUR YOURS YOURSELF YOURSELVES

2. Database

The database comprises 77 books available online at the Gutenberg project repository [1], whose publication date ranged from 1590 to 1922. The details of the books are given in Tables S1, S2 and S3.
**Figure S1.** Books used in the experiments based on the variation of style according to the publication date. Literary periods are distinguished by a double horizontal line in the table.

| Publication | Book                                           | Author                  |
|-------------|------------------------------------------------|-------------------------|
| 1590        | The Faerie Queene                              | Edmund Spenser          |
| 1605        | Don Quixote                                    | Miguel de Cervantes     |
| 1605        | The Advancement of Learning                    | Francis Bacon           |
| 1612        | True Christianity                              | Johann Arndt            |
| 1624        | Devotions upon Emergent Occasions              | John Donne              |
| 1651        | Leviathan                                      | Thomas Hobbes           |
| 1653        | The Love Letters of Dorothy Osborne            | Dorothy Osborne         |
| 1664        | Experiments and Considerations Touching Colours| Robert Boyle            |
| 1678        | The Pilgrim’s Progress                         | John Bunyan             |
| 1684        | Love-Letters Between a Nobleman and His Sister | Aphra Behn              |
| 1711        | The Spectator                                   | Joseph Addison          |
| 1719        | Robinson Crusoe                                | Daniel Defoe            |
| 1726        | Gulliver’s Travels                             | Jonathan Swift          |
| 1731-50     | Selected Sermons of Jonathan Edwards           | Jonathan Edwards        |
| 1735-48     | The Letters of Horace Walpole                  | Horace Walpole          |
| 1740        | Pamela, or Virtue Rewarded                     | Samuel Richardson        |
| 1744        | The Fortunate Foundlings                       | Eliza Haywood           |
| 1748        | Clarissa, or, the History of a Young Lady      | Samuel Richardson        |
| 1748        | Life’s Progress through the Passions           | Eliza Haywood           |
| 1749        | The History of Tom Jones, a Foundling          | Henry Fielding          |
| 1761        | The Vicar of Wakefield                         | Oliver Goldsmith        |
| 1767        | The Life and Opinions of Tristram Shandy, Gentleman | Laurence Sterne        |
| 1770        | Thoughts on the Present Discontents            | Edmund Burke            |
| 1771        | The Expedition of Humphry Clinker              | Tobias Smollett         |
| 1778        | Evelina                                        | Frances Burney          |
| 1782        | Cecilia, Memoirs of an Heiress                 | Frances Burney          |
| 1791        | A Vindication of the Rights of Woman           | Mary Wollstonecraft     |
| 1793        | Autobiography of Benjamin Franklin             | Benjamin Franklin        |
| 1794        | The Mysteries of Udolpho                       | Ann Radcliffe           |
**Figure S2.** Books used in the experiments based on the variation of style according to the publication date (continuation). Literary periods are distinguished by a double horizontal line in the table.

| Publication | Book                                                | Author                      |
|-------------|-----------------------------------------------------|----------------------------|
| 1811        | Sense and Sensibility                              | Jane Austen                |
| 1813        | Pride and Prejudice                                | Jane Austen                |
| 1815        | Emma                                                | Jane Austen                |
| 1818        | Frankenstein; or, The Modern Prometheus             | Mary Shelley               |
| 1818        | Persuasion                                         | Jane Austen                |
| 1826        | The Last of the Mohicans: A Narrative of 1757       | James Fenimore Cooper      |
| 1839        | The Voyage of the Beagle                           | Charles Darwin             |
| 1844        | The Black Tulip                                     | Alexandre Dumas            |
| 1847        | Wuthering Heights                                   | Emily Bronte               |
| 1850        | David Copperfield                                   | Charles Dickens            |
| 1861        | Great Expectations                                  | Charles Dickens            |
| 1865        | The Headless Horseman                               | Mayne Reid                 |
| 1869        | The Innocents Abroad, or The New Pilgrims’ Progress | Mark Twain                 |
| 1876        | Daniel Deronda                                      | George Eliot               |
| 1882        | Two on a Tower                                      | Thomas Hardy               |
| 1894        | Life’s Little Ironies                               | Thomas Hardy               |
| 1897        | Dracula                                             | Bram Stoker                |
| 1897        | Jane Eyre                                           | Charlotte Bronte           |
| 1897        | The Invisible Man                                   | Herbert George Wells       |
| 1903        | The Jewel of Seven Stars                            | Bram Stoker                |
| 1905        | The House of Mirth                                  | Edith Wharton              |
| 1906        | The Jungle                                          | Upton Sinclair             |
| 1909        | The Lady of the Shroud                              | Bram Stoker                |
| 1911        | The Lair of the White Worm                          | Bram Stoker                |
| 1913        | Pollyanna                                           | Eleanor H. Porter          |
| 1913        | The New Freedom                                     | Woodrow Wilson             |
| 1915        | Dear Enemy                                          | Jean Webster               |
| 1920        | The Age of Innocence                                | Edith Wharton              |
| 1922        | Ulysses                                             | James Joyce                |
Figure S3. Books used in the experiments on the variation of the style of Charles Darwin and Edith Wharton.

| Publication | Book                                                      | Author            |
|-------------|-----------------------------------------------------------|-------------------|
| 1839        | Voyage of the Beagle                                      | Charles Darwin    |
| 1844        | Volcanic islands                                          | Charles Darwin    |
| 1846        | Geological observations                                   | Charles Darwin    |
| 1854        | A Monograph of the Sub-class Cirripedia                   | Charles Darwin    |
| 1859        | On the Origin of Species                                  | Charles Darwin    |
| 1868        | Variation of Plants and Animals Under Domestication       | Charles Darwin    |
| 1871        | The Descent of Man and Selection in Relation to Sex       | Charles Darwin    |
| 1872        | The Expression of Emotions in Man and Animals             | Charles Darwin    |
| 1875        | Insectivorous Plants                                      | Charles Darwin    |
| 1881        | The Formation of Vegetable Mould Through the Action of Worms | Charles Darwin    |
| 1901        | Crucial Instances                                         | Edith Wharton     |
| 1902        | The Valley of Decision                                    | Edith Wharton     |
| 1905        | The House of Mirth                                        | Edith Wharton     |
| 1907        | The Fruit of the Tree                                     | Edith Wharton     |
| 1910        | Tales of Men and Ghosts                                   | Edith Wharton     |
| 1912        | The Reef                                                  | Edith Wharton     |
| 1913        | The Custom of the Country                                 | Edith Wharton     |
| 1917        | Summer                                                    | Edith Wharton     |
| 1920        | The Age of Innocence                                      | Edith Wharton     |
| 1922        | The Glimpses of the Moon                                  | Edith Wharton     |
3. Dimensionality Reduction

In order to visually examine the dynamics of style and remove correlations between attributes we applied two strategies to reduce the data dimensionality: the Principal Component Analysis (PCA) [2] and the Canonical Variable Analysis (CVA) [3]. Both methods are widely used for high dimensional databases, and therefore we provide only a brief description of their strategies.

PCA is based on the diagonalization of the covariance matrix describing the database. Mathematically, this method is implemented as a rotation of the original axes so that the new reduced axes coincide with the directions of maximum data variation. Further details regarding PCA can be found in Refs. [4, 5]. The CVA method is similar to PCA because its implementation involves a procedure based on computation of eigenvectors and eigenvalues of matrices. It also generates new axes that are linear combinations of the original axes, which makes it possible to quantify the contribution of each original feature to the identification of the new axes. This is done by calculating the weights attribute to the original variables (the weights are given by the components of the eigenvectors). CVA differs from PCA in that it takes into account that each instance in the database has a label associated to the class the instance belongs to. With this information, the method minimizes the distance among objects belonging to the same cluster and maximizes the distance among objects belonging to different clusters. The derivation of the CVA projection algorithm is described in Refs. [5, 6].

The compression rate $\eta$ for these projection techniques is obtained by computing the eigenvalues $\lambda$ of the corresponding matrices to derive the new axes. If $j$ is the number of components after applying the method, then the compression rate is:

$$\eta = \frac{\sum_{i=1}^{j} \lambda_i}{\sum_{k=1}^{\kappa} \lambda_k}.$$  

(1)

Since we projected the 11-dimensional data into a 2-dimensional subspace, $j = 2$ and $\kappa = 11$.

4. Quantifying Clustering Quality

To analyze how the writing style varied over time, we came across the problem of finding the partition of the data that generate the best clustering. Because we assessed all possible partitioning schemes we needed an automatic measure to determine the best clustering. We employed the two strategies described below.

The first criterion of quality is referred to as simplified silhouette (SWC) [7], defined as follows. Let $\Delta_i^s$:

$$\Delta_i^s = \frac{1}{N_k} \sum_{i \in \text{cluster } k} \| \vec{v}_i - \vec{\mu}_k \|$$  

(2)
be the average dissimilarity between the object \(i\) belonging to the cluster \(k\), where \(N_k\) is the number of objects in the cluster and \(\mu_k\)

\[
\bar{\mu}_k = \frac{1}{N_k} \sum_{i \in \text{cluster } k} v_i
\]  

(3)

is the centroid of the cluster. Let \(\delta_i^s\):

\[
\delta_i^s = \frac{1}{N_p - 1} \sum_{j \neq P} \|v_i - v_j\|
\]

(4)

be the average dissimilarity between the object \(i\) belonging to the cluster \(k\) and the \(N_p\) objects of the nearest cluster \(P\). The cluster quality according to the SWC criterion is computed as:

\[
SWC = \frac{1}{m} \sum_{i} \frac{\delta_i^s - \Delta_i^s}{\max\{\delta_i^s, \Delta_i^s\}},
\]

(5)

where \(m = \sum N_k\). SWC is restricted to the interval \([-1,1]\) so that high values mean well-defined clusters.

Also used was the Dunn index (\(DN\)) [7], defined as the ratio between the highest inter-group distance \(\delta\) and the shortest intra-group distance \(\Delta\):

\[
DN = \min_{p,q \in \{1 \ldots n\} \text{ and } p \neq q} \left\{ \frac{\delta_{p,q}}{\max_l \{1 \ldots n\} \Delta l} \right\}.
\]

(6)

The distance \(\delta_{p,q}\) between the clusters \(p\) and \(q\) and intra-group distance \(\Delta_l\) of the cluster \(l\) are given by:

\[
\delta_{p,q} = \|\mu_p - \mu_q\|
\]

(7)

\[
\Delta_l = \frac{2}{N_l} \sum_{i \in l} \|v_i - \mu_l\|
\]

(8)

Similarly to the definition of \(SWC\), the \(DN\) measurement is directly proportional to the cluster quality.
5. Clusters

Figure S1 illustrates the clustering obtained using the proposed methodology for 3, 4, 5 and 7 clusters.

Figure S1. Scatter plot representing the best clustering for the writing style. The number of clusters considered was: (a) 3 clusters, (b) 4 clusters; (c) 5 clusters; and (d) 7 clusters. The clustering with 6 groups is shown in the manuscript.

References

[1] http://www.gutenberg.org/
[2] Jolliffe I T 2002 *Principal component analysis* Springer New York
[3] McLachlan G J 2004 *Discriminant Analysis and Statistical Pattern Recognition* Wiley
[4] Bishop C M 2006 *Pattern Recognition and Machine Learning* Springer-Verlag New York
[5] Costa L F, Rodrigues F A, Travieso G and Villas Boas P R 2005 *Advances in Physics* 56 167
[6] Duda R O, Hart P E and Stork D G 2001 *Pattern Classification* John Willey and Sons.
[7] Vendramin L, Campello R J G B and Hruschka E R 2010 *Statistical Analysis and Data Mining* 3 209