Language Identification with Singular Value Decomposition

Sabrina Pereira and Jane Sieving

Abstract

We set out to investigate if we could use singular value decomposition to analyze reference texts for a language, and from this, extrapolate the letter patterns characteristic to a language. We found this could indeed be done - frequency of letter pairings vary greatly between languages, and a reference plot can be created that depicts an accurate representation of the true frequency of letter pairings. Based on these patterns, we were able to do a similar analysis on 'unknown' texts in several languages, and find the best match as to what language the text is in.

Introduction

Singular value decomposition is a technique employed to factorize a matrix into three matrices such that for a matrix *M*:

$$M = U\Sigma V$$

The *U* matrix is composed of the normalized eigenvectors of MM^T (called left singular vectors), *V* is composed of the normalized eigenvectors of M^TM (called right singular vectors), and Σ is a diagonal matrix containing the square roots of the eigenvalues common to of MM^T and M^TM (called the singular values).

This factorization is useful in order to be able to approximate the original matrix and analyze patterns between the matrix's values. This then leads to a variety of applications in areas such as data compression (ex. Image compression) and solving matrix equations (ex. least squares fit to lines). For our project, we will be using the matrix approximation and data compression aspect of SVD to be able to approximate letter pairing frequencies of different languages and visualize the relationships between these letters in a 2D plot.

Hypothesis

We hypothesize that there are patterns seen by using SVD as a tool to analyze the reference texts will hold for a large enough text in the same language. We are experimenting to see if we are able to identify the language a text is written in by comparing letter pattern frequencies between the reference texts and an input text. We believe that analyzing the same text translated into different languages will show us different patterns characteristic to each chosen language, giving us insight into which letters are most often used together, which letters are used most and least often, and which letters tend to start or end words.

Design Choices and Data Selection

We are choosing to test five languages: English, Spanish, Portuguese, French, German. All languages use a variant of the Latin alphabet. For each of these languages, we chose one popular text originally written in its native language and gathered translations for the remaining four. We attempted to choose texts of different genres written in different time periods, narrowing in on these five for our texts: *The Metamorphosis* by Franz Kafka, *The Little Prince* by Antoine de Saint-Exupéry, *Don Quixote* by Miguel de Cervantes, *The Alchemist* by Paulo Coelho, and The Book of Mormon.[1]

All texts in the same language were then joined together into a combined text in order to create the letter pairing matrix to be used as a reference each of our languages.

To create the reference matrices, we looked at the words in the combined texts individually (that is, letter pairs were not considered across spaces) and counted the number of instances in which a letter pair appeared. These counts were then put into 26 by 26 matrices, such that for a matrix M, the (i, j)th entry would count the number of times the letter j followed the letter i (equivalent to the number of times i precedes j), and vice versa for the (j, i)th entry. (Encountering the word "data" in the text, for example, would increase the (D,A) entry by one count, the (A,T) entry by one count, and the (T,A) entry by one count.)

To overcome the issue that most every language has a slight variation in the characters of its alphabet, we decided to impose the English alphabet on the analysis of all languages chosen. We decided to categorize any language special characters as their closest english equivalent (for example, Ñ is to be replaced by N, ß is to be replaced by S, all accented characters to be replaced by an unaccented one). We have researched letter frequencies for our reference languages and found that these special characters appear in low enough frequency such that we are comfortable with the level of error. We also believe that categorizing the special characters in this way will still allow us to confidently compare reference results to an input text as the same type of categorization and analysis will be applied to it.

Applying SVD to Matrices

We then took the reference matrices and decomposed them using the method outlined for SVD. We will be using the rank two approximation of each of our matrices to gain information on letter distribution.

A matrix decomposed with SVD can be rewritten as a summation of matrices as shown below:

$$M = u_1\sigma_1 {v_1}^T + u_2\sigma_2 {v_2}^T + ... + u_n\sigma_n {v_n}^T$$

The first term in the summation is the matrix in the summation that contains the most information about the original matrix, and is the rank one approximation to our matrix.

The the addition of the first and second terms in the summation form the rank two approximation. This second added matrix adds corrections to the first, and sheds light on how the rank one approximation most deviates from the original.

Using this information, we can create plots to help us visualize the relative frequency of letter pairings in a language. The first singular vector plot is created by plotting ($\boldsymbol{u}_{1k}, \boldsymbol{v}_{1k}$) for all letters k in the alphabet, and will give us the most insight into the frequency of letter pairs for each language. The second singular plot, plotting ($\boldsymbol{u}_{2k}, \boldsymbol{v}_{2k}$), will give us the most information on how letter pairing frequency deviates from what would be expected from the first singular vector plot.

[Then we will plot the SVD of another text of unknown language and compare it to the known plots to identify the best match language. We will probably do this by calculating the distance of each point from the same point in the other languages, and the language with the least total distance will be chosen as the closest match.]

Experiment Results

Below are the plots generated from the reference texts for each language. The first two graphs are from one of our ‘unknown’ texts which we identified the language of. The text is actually from the emancipation proclamation, which is in English. [2]

Analysis

From the plots, the first singular vector plots may not accurately give us a the letter frequency for these characters in every language - this mistake should not be made. Although it does give a rough idea of the true letter frequency, the data collected the counted letter pairs per word, and in doing so, the letters in the middle of the word are counted as both a prefix and a suffix, while the letters on the end are only counted once. This skews the apparent letter frequency towards letters most appearing in the middle of words.

In the second singular vector plot, we see a rough classification of letters into vowels and consonants. This is due to the fact that, in general, vowels are followed consonants and vice versa, while vowels following vowels and consonants following consonants are used less frequently. In languages where this holds true more often, the classification is more distinct. Out of all of the chosen languages, German presents the least amount of classification of this type - it may be attributed to to the fact that it is a language in which consonants are paired with consonants quite often.

By multiplying the $\boldsymbol{u}_{1i}$ and $\boldsymbol{v}_{1j}$ values for any letters *i* and *j*, we find the approximate frequency of this letter pairing *ij*. By multiplying the $\boldsymbol{u}_{2i}$ and $\boldsymbol{v}_{2j}$ values for the same letters chosen above, we get an idea of how much of a deviation there is in the actual letter pairing as estimated by the first singular vector plot.

For each ‘unknown’ text, we calculated the distances on the u, v plane between each letter in both the first and second singular vector plots. Individually for the first and second plot distances, we squared and summed the per-letter distances to get a summary of the overall deviation between the plots. The language which showed the lowest deviation values was concluded to be the language of the unknown text. For the unknown text plotted above, these were the deviation values:

From this, we can see that the text closely resembled the letter patterns of English, and it is in fact an English text.

Conclusion

We successfully applied the methods for SVD to analyze the letter pairings for each of our combined reference texts for all of our chosen languages. With the plots, we can begin to see what letters pairings are most characteristic in each language, and the main differences in languages that use such similar alphabets.

We found that our plots were a good estimate for the true frequency of letter pairings, if not true letter frequency. When compared to a reference text by calculating the deviations in distance of the singular vector plots, we were able to find the closest match language for ‘unknown’ input texts from several languages.

Appendix

[1] *Don Quixote* by Miguel de Cervantes was originally written in Spanish in 1615, and is a satiric novel. *The Alchemist* by Paulo Coelho was originally written in Portuguese in 1988 and is a quest/adventure novel. *The Metamorphosis* by Franz Kafka was originally written in German in 1915, a short story belonging in the absurdist fiction genre. *The Little Prince* by Antoine de Saint-Exupéry was originally written in French in 1943 and is a fable/novella. The Book of Mormon is a religious text originally published in 1830 by Joseph Smith.

[2] Reference text files, adjacency matrices, more results, and Python and MATLAB code used in this project can be found at https://github.com/jsieving/svd-languages.

References

Papers:

Singular Vectors' Subtle Secrets
Authors: David James, Michael Lachance and Joan Remski
https://drive.google.com/file/d/19_EKtSxaEW8g_1iZX92TPwCxmI8uVJiZ/view
Aided us in better understanding singular value decomposition and in beginning to understand how to analyze our texts. Showed us how to create the first and second singular vector plots and how to interpret them.

Singular Value Analysis of Cryptograms
Authors: Cleve Moler and Donald Morrison
https://pdfs.semanticscholar.org/c78e/9a7d30cb416c8eef9ceac5aeb4d95f3ab829.pdf
Referenced in *Singular Vectors' Subtle Secrets*, introduced this type of textual analysis in the context of cryptography. Elaborated on the groupings of letters in the second singular plot, explaining them to be classifications of letters as vowels, consonants, or neuter.

Introduction to Singular Value Decomposition with Image Compression Application
Authors: Kevin Cheng, Sabrina Thompson, John Watson
https://drive.google.com/file/d/1NrOp6mEDLkjUWzyhT0mV8qSxXK8S9Mlw/view
Further helped our understanding of SVD, in particular the component matrices, and served as our main reference for associated vocabulary and equations. Provided an example of imaging data compression that served to understand a function of SVD.

Web:

Cool Linear Algebra: Singular Value Decomposition
Author: Andrew Gibiansky
http://andrew.gibiansky.com/blog/mathematics/cool-linear-algebra-singular-value-decomposition
Gave an explanation of SVD by going over diagonalization and drawing the parallels. Walked through the matrix multiplication of the component matrices.

Reference Texts:
Below are the following links containing the reference texts used.

The Book of Mormon
English: http://media.ldscdn.org/pdf/lds-scriptures/book-of-mormon/book-of-mormon-34406-eng.pdf
French: http://media.ldscdn.org/pdf/lds-scriptures/book-of-mormon/book-of-mormon-59012-fra.pdf

German: http://media.ldscdn.org/pdf/lds-scriptures/book-of-mormon/book-of-mormon-34406-deu.pdf
Portuguese: http://media.ldscdn.org/pdf/lds-scriptures/book-of-mormon/book-of-mormon-59012-por.pdf
Spanish: http://media.ldscdn.org/pdf/lds-scriptures/book-of-mormon/book-of-mormon-59012-spa.pdf

Don Quixote by Miguel de Cervantes
English: http://www.gutenberg.org/cache/epub/996/pg996.txt
French: https://www.ebooksgratuits.com/html/cervantes_don_quichotte_1.html
German: http://www.gasl.org/refbib/Cervantes__Don_Quixote.pdf
Portuguese: http://www.ebooksbrasil.org/adobeebook/quixote1.pdf
Spanish: http://parnaseo.uv.es/Lemir/Revista/Revista19/Textos/Quijote_1.pdf

The Alchemist by Paulo Coelho
English: https://archive.org/stream/TheAlchemist_410/TheAlchemist_djvu.txt
French: http://www.oasisfle.com/ebook_oasisfle/Coelho,%20Paulo%20-%20L'alchimiste.pdf
German: http://liebevoll-wei.se/Der_Alchimist.pdf
Portuguese: https://archive.org/stream/o-alquimista-paulo-coelho/o-alquimista-paulo-coelho_djvu.txt
Spanish:http://www.itvalledelguadiana.edu.mx/librosdigitales/Paulo%20Coelho%20-%20El%20alquimista.pdf

The Metamorphosis by Franz Kafka
English: https://www.gutenberg.org/files/5200/5200-h/5200-h.htm
French: https://beq.ebooksgratuits.com/classiques/Kafka_La_metamorphose.pdf
German: http://www.gutenberg.org/cache/epub/22367/pg22367.txt
Portuguese: http://www.culturabrasil.org/zip/metamorfose.pdf
Spanish: https://es.wikisource.org/wiki/La_Metamorfosis:_Cap%C3%ADtulo_Uno

The Little Prince by Antoine de Saint-Exupéry
English, French, German, and Spanish: http://www.malyksiaze.net/us/ksiazka
Portuguese:
http://www.buscadaexcelencia.com.br/wp-content/uploads/2010/08/O_Pequeno_Pr%C3%ADncipe_Ilustrado.pdf

Reference Texts:

Below are the following links used for test input texts.
http://spanish.yourdictionary.com/spanish-language/learning-spanish/poems-in-spanish.html
http://www.learn-portuguese-now.com/alternative-education.html
https://lingua.com/french/reading/
https://lingua.com/german/reading/
https://www.poetryfoundation.org/