INTRO TO STATISTICAL LEARNING COURSEWORK

ANDREW MELVILLE 01200147

1. Introduction

It is much easier for an Italian to learn French than for an Englishman to learn Vietnamese, as French is much 'closer' to Italian. While the similarity between two languages is able to illustrate this advantage, a deeper understanding of the spread of language and people can be explored using scaling analysis by creating a spatial configuration of the estimated 'distances' between languages.
From this, we do not expect to recreate a world map as this is equivalent to the assumption that the spread of language is purely geographical. Instead a configuration of this type may be influenced much more by cultural similarities and historical migration patterns.

2. Dataset

Our data set consists of the top $1,000$ words in the English Language and their translations in 45 other languages across Europe and Asia. To ensure that similarity is not obscured by different alphabets, we must first transliterate each of these languages into the 26 letters of the English alphabet. This will allow us to compare words without regional differences such as accents or punctuation.
The implicit assumption in the way we encode this data is that words from two languages are similar if they share many of the same letters. Then, we assume that two languages are similar if many of their words are similar. So, the cleaned words are now encoded into a 50 by $26,000$ dataframe. The columns are made up of the English alphabet repeated $1,000$ times, once for each word. A cell encodes the presence of a letter in a word with a 1, or its absence with a 0. Finally, we compute pairwise distances between languages using the Jacard distance. This is calculates the ratio of dissimilar letters across languages to the number of letters present. This is a natural expression of the assumption we made above. The distance matrix is then used for scaling analysis in R.

3. Classical Multi-Dimensional Scaling

Using the inbuilt function in R, we applied Classical Multi-dimensional Scaling to our distance matrix, with 10 dimensions to ensure sufficient capture of the variance in the data.
Analysis of the eigenvalues of our distance matrix (see Figure 1) shows that a large part of the variance lies along the first three principal components. This is promising as it shows that it may be able to capture some of the geography of the data.
The first two principal components of the scaling solution appear to have distinguished the languages into distinct clusters. In the top left are the Romance languages (French, Italian, Spanish), while the bottom left are mainly Germanic and Northern

Date: 16 February 2020.

European languages (English, German, Swedish, and the far-right isolates Slavic languages (Russian, Ukrainian, Czech). In the centre cluster are languages that do not appear to fall into these three distinct partitions of Europe (Welsh, Turkish, Greek). This third axis appears to measure some notion of 'isolation', as each language in this region is either a minority language (Welsh and Finnish) or a language which has influenced each of the other languages in equal measure (Greek).
Therefore the first principal component appears to largely distinguish between Eastern and Western Europe, while the second principal component divides the Western Europe languages into those with Romance roots and those with Germanic roots. This hypothesis is supported by the position of English almost in between the two groups, as it is the product of several hundred years of migration of people from these areas.
The close syntactic link between Galician and Portuguese and their similarities between Spanish and Catalan are also reported as expected. A similar but initially surprising pairing is that of Welsh and Basque. Despite the large geographical distance between the two areas and apparent lack of migration or cultural link, both languages appear to share many aspects according to our metric. This link is supported by Genomic research which has found links between the ancestral DNA of males in both populations. This is further evidence to suggest a common migration pattern.
In the centre cluster are languages that do not appear to fall into these three distinct partitions of Europe (Welsh, Turkish, Greek). As these are not explained by the first two major axes, we would expect them to have large values in the third dimension. This is confirmed by Figure 3.

4. K-Means Clustering

We would now like to cluster this configuration using K-means clustering to attempt to recover the families of European languages described above. Accordingly, we compute the Calinski-Harabasz score, which is a heuristic measure for finding the 'optimal' number of clusters for a given configuration. The result (see Figure 4) shows a high-value for $k = 4$, while the score rises for larger values. This is most likely due to overfitting on subsets of the languages. We therefore choose $k = 4$ for the optimal clustering, the results of which are shown in Figure 5. The clustering has picked out the four language families as we might expect.

5. Conclusion

The decision to identify distances between languages using their spelling differences across the top $1,000$ most common words and associated Jacard metric is vindicated by the recovery of the 3 large language families in Europe. Instead of similarity in spelling, one may choose to instead standardize each language by its phonetic spelling. For example a soft 'c' in English would be linked to the same pronunciation of an 's' in French. In this way one would hope that common roots would be recovered more reliably.

6. Source

https://1000mostcommonwords.com/

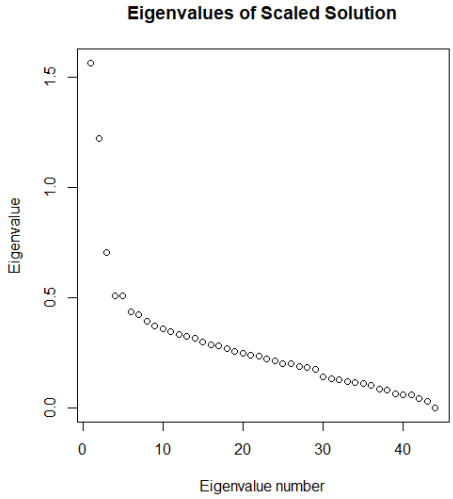

FIGURE 1. Eigenvalues of the distance matrix

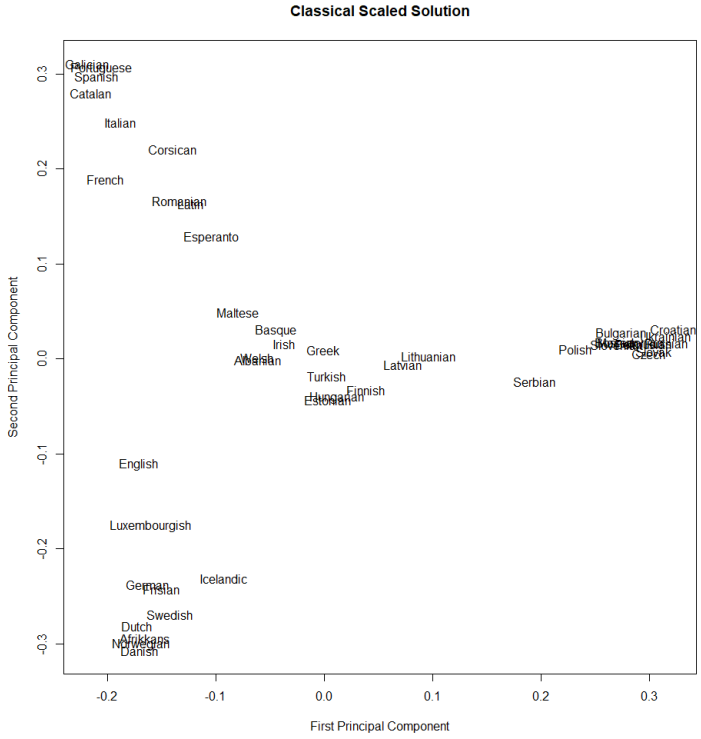

FIGURE 2. Solution of Classical Multi-dimensional Scaling

7. FIGURES

8. SOURCE

https://1000mostcommonwords.com/

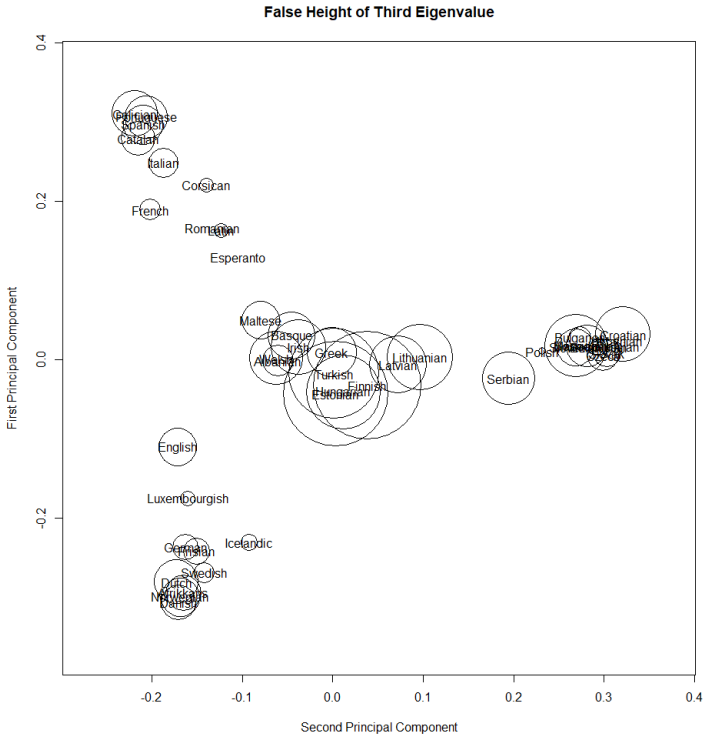

FIGURE 3. False height map displaying the value of each language in the third principal component

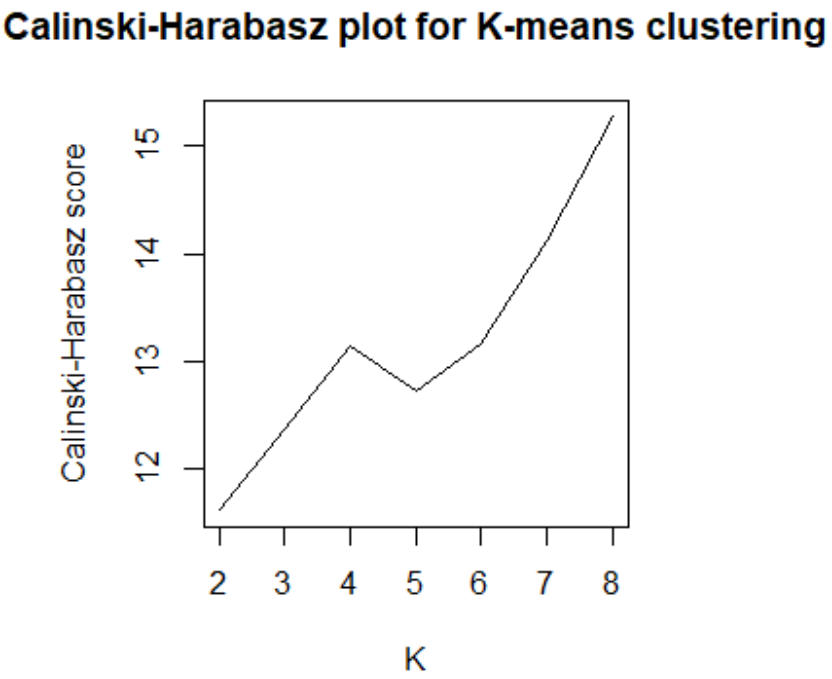

FIGURE 4. Calinski-Harabasz score for several different values of K in K-means clustering

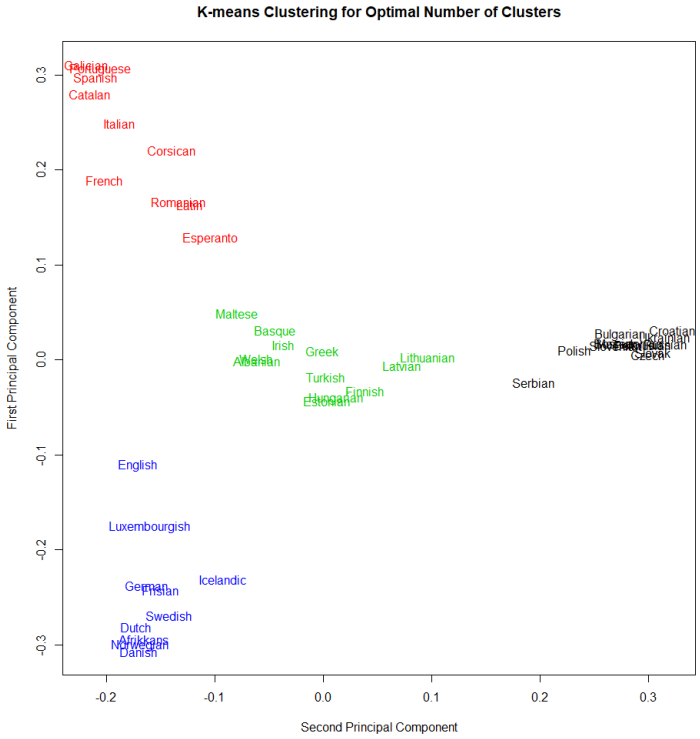

FIGURE 5. K-means clustering for the 'optimal' value of K