Text as Data: Homework 4

Due: 11/7 at 5pm

In this homework assignment we will analyze Machiavelli's *The Prince*. Download `Mach.tar` from the course website and expand the compressed folder. (This is relevant `http://xkcd.com/1168/`).

Each file represents a subset of the manuscript. We will analyze its contents using principal components, multidimensional scaling, and clustering methods.

Create a Document-Term Matrix

Using the sections from the Machiavelli text, create a document term matrix.

- Discard punctuation, capitalization
- Apply the porter stemmer to the documents
- Remove stop words from the following list (remembering to stem the stop words): '`http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`'
- Identify the 500 most common unigrams
- Create a $N \times 500$ document term matrix $\boldsymbol{X}$, where the columns count the unigrams and the rows are the documents

We will work with a normalized version of the term document matrix. That is we will divide each row by the total number of words in the top 500 unigrams used:

$$\begin{aligned}\boldsymbol{x}_i^* &= \frac{\boldsymbol{x}_i}{\sum_{j=1}^{500} x_{ij}} \\ \boldsymbol{X}^* &= \begin{pmatrix} \boldsymbol{x}_1^* \\ \boldsymbol{x}_2^* \\ \vdots \\ \boldsymbol{x}_N^* \end{pmatrix}\end{aligned}$$

Low Dimensional Embeddings with Principal Components

1) Wise Will (WW), your friend with a weird name, notices you looking at the slides about principal component analysis (PCA). WW casually remarks that the variance of the eigenvalues of the variance-covariance matrix is a useful heuristic for knowing if PCA can be fruitfully applied to some document-term matrix. WW, completely unsolicited, explains that as the variance of the eigenvalues goes up, the more useful PCA will be. He then laughs and leaves your office. WW is kind of a jerk.

   Let's formalize WW's suggestion. Suppose document-term matrix $\boldsymbol{X}$ has variance-covariance matrix $\boldsymbol{\Sigma} = \frac{\boldsymbol{X}^{'}\boldsymbol{X}}{N}$. And suppose that $\boldsymbol{\Sigma}$ has eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d > 0$. Then we calculate the variance of the eigenvalues as

   $$\sigma^2 \;=\; \frac{1}{d}\sum_{j=1}^{d}(\lambda_j - \bar{\lambda})^2$$

   where $\bar{\lambda}$ is $\frac{1}{d}\sum_{i=1}^{d}\lambda_i$. WW is saying that as σ^2 gets bigger, a low-dimensional embedding via PCA will provide a better summary of our data.

   Does WW have a good point? Why or why not? (Hint: what do the eigenvalues represent?)

2) Apply the function `prcomp` to $\boldsymbol{X}^{*}$. Be sure to set use a scaled version of the data, by setting `scale = T`, which will ensure that each column has unit variance.

   a) Create a plot of variance explained by each additional principal component. What does this plot suggest about the number of components to include?

   b) Plot the two-dimensional embedding of the text documents. Label the texts with their number. (Each file is `Mach_XX.txt`, where `XX` is the chunk number)

   c) Label the two largest principal components. What does this embedding suggest about the primary variation this representation of the Prince? (Hint: if your `embed` is your object with principal components, examine `embed$rotation`)

3) An alternative method—discussed at the end of the seventh lecture—is multidimensional scaling (MDS). Classic MDS attempts to preserve distances between objects in a low dimensional scaling.

   a) Calculate the Euclidean distance between each document using $\boldsymbol{X}^{*}$. Call this matrix $\boldsymbol{D}(\boldsymbol{X}^{*})$ (Hint: use `R`'s built in function `dist`)

   b) Apply the classic MDS to $\boldsymbol{D}(\boldsymbol{X}^{*})$ using the `R` function `cmdscale`. That is, execute the code

   ```
   mds_scale<- cmdscale(DISTANCE_MATRIX, k = 2)
   ```

c) Apply PCA to $\boldsymbol{X}^*$, but this time do not use `prcomp`'s scaling option. That is, use `prcomp` with `scale = F`.

d) Compare the first dimension of the output from classic MDS to the first dimension of the embedding from principal components. What is the correlation between the embeddings?

d) Now use `dist` to create a distance matrix using the `manhattan` metric, apply Classic multidimensional scaling to the distance matrix based on manhattan distance, and compare the first dimension of this embedding to the first dimension from PCA. What is the correlation?

e) What do you conclude about the relationship between PCA and MDS?

Clustering Methods

1) Using the `kmeans` function, create a plot of the `kmeans` objective function as the number of clusters varies from 2 to $N-1$.

2) Apply K-Means with 6 clusters, being sure to use `set.seed` to ensure you can replicate your analysis

3) Label each cluster using computer and hand methods:

   i) Suppose $\boldsymbol{\theta}_k$ is the cluster center for cluster k and define $\bar{\boldsymbol{\theta}}_{-k} = \frac{\sum_{j \neq k} \boldsymbol{\theta}_j}{K-1}$ or the average of the centers not k. Define

   $$\text{Diff}_k \quad = \quad \boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}}_{-k}$$

   Use the top ten words from Diff_k to label the clusters

   ii) Sample and read texts assigned to each cluster and produce a hand label

4) Measure the cohesiveness and exclusivity of each cluster. What is the most exclusive and most cohesive topics? What are the least? Does this align with your reading of the texts?