Home Assignment 3

Javier Galindos Vicente, 214373IV

The student would like to present the work offline in class.

I. Text Mining

In this section, some speeches of President Biden have been selected as the text dataset. The first step is to pre-process the dataset. In particular, the punctuation symbols, numbers, capitalization and the words ('ll, 're, 've, the, and) have been removed towards having a more precise analysis. The histogram of the most common words is presented in Fig. 1 and the word-cloud with the most frequent words is presented in Fig. 2. One could note observing this analysis that the dataset is composed of speeches from an American politician due to present of words like america, people, nation, country, democracy, united, etc.

The documents are represented in a Document-Term Matrix (DTM) for simplicity. Before clustering the data, it is necessary to remove sparse terms. Specifically, the maximal allowed sparsity is 0.15 in this dataset. The similarity measure used to compute the distances is the cosine similarity for every algorithm due to the better performance than other metrics (i.e. euclidean, manhattan).

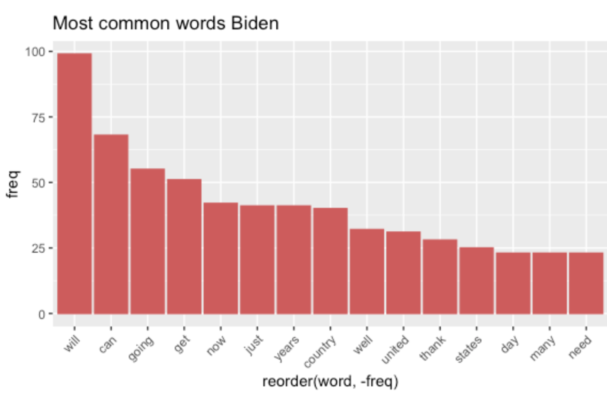

Fig. 1: Histogram of the text dataset.

To cluster the dataset, several techniques such as k-means, hierarchical clustering or k-medioids have been performed. Other probabilistic algorithms like EM or density based methods such as HDBSCAN have been explored, but the results are not optimal. The three cluster generated by the aforementioned algorithms agree in having 3 centroids between the most common words. One of them is formed from key words where Pres. Biden wants to persuade some ideas with words like state or troops. Another centroid depicts words that he use to engage his followers with examples (e.g. god, bless, thank, work, today, etc.) The last centroid represent a bigger cluster with the frequent words Biden mention to different himself and

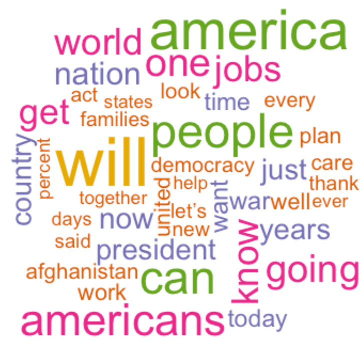

Fig. 2: Word-cloud of the text dataset.

set a political line with words (e.g. states, american, united, together, nation, people, protect, families, country, etc.).

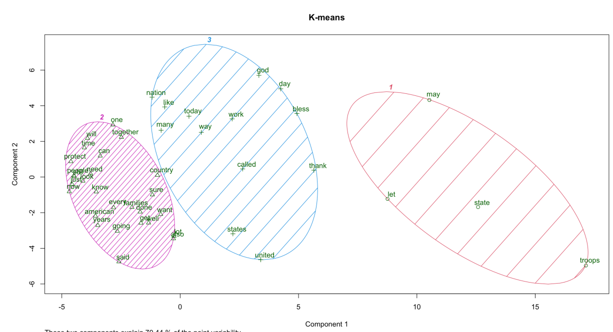

Fig. 3: K-means clustering.

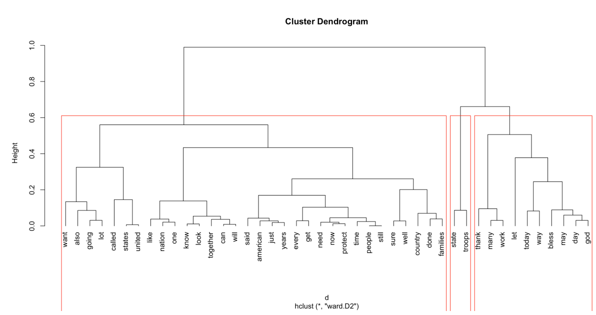

Fig. 4: Hierarchical clustering

Numerically, both the centroids and the decision boundaries could be explained in term of two components function of the frequencies (see Fig. 3 and Fig. 5). These components explain 70.33 % of the point variability. In particular, for the k-means clustering, it could be observed that the component 1 is almost capable of depict a decision boundary, whether the component

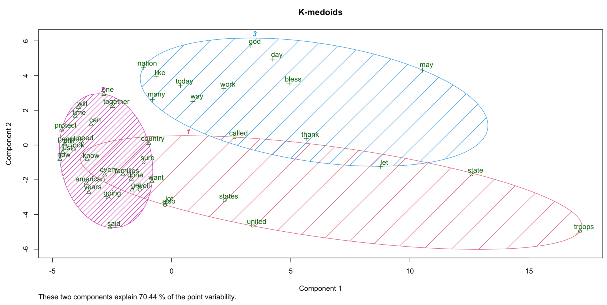

Fig. 5: K-medioids clustering.

1 is negative, the words belong to cluster 2, if the component 1 is between 0 and 7, the words belong to cluster 3 and finally, if the component 1 is higher than 7, the words belong to cluster 1 (observe Fig. 3). The decision boundaries for the k-medioids algorithm are fuzzier. The output of hierarchical clustering algorithm does not provide a clear decision boundary.

II. Graph Mining

The dataset used for validation of the different function is obtained from Kateto and presented in Fig. 6 (undirected graph) and in Fig. 7 (directed graph). To pre-process the graphs, both loops and duplicates has been simplified. Furthermore, both an *igraph* graph object and the adjacency matrix of the graphs have been computed for implementing the functions of the different measures.

The Local Clustering Coefficient (LCC), Degree Centrality (DC) and Closeness Centrality (CC) are computed for the undirected graph. On the flip side, the Degree Prestige (DP), Gregariousness (Greg) , Proximity Prestige (Prox) and Betweenness Centrality (BC) are computed for directed graphs. The common neighbor and Jaccard measure could be indistinctly computed for directed or undirected graphs.

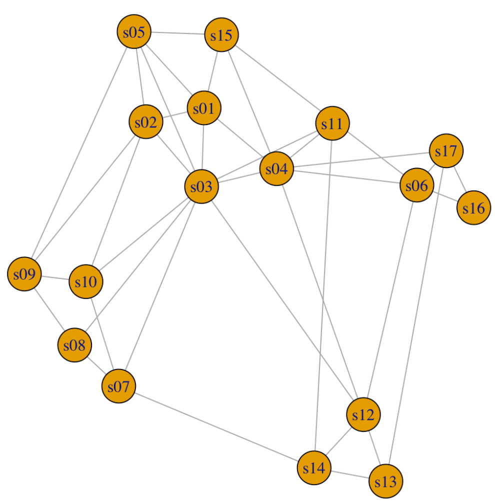

Fig. 6: Undirected graph.

The output of each measure has been cross-checked with the native functions of the library *igraph* to validate that the result is correct. The results are presented in Table I. It is

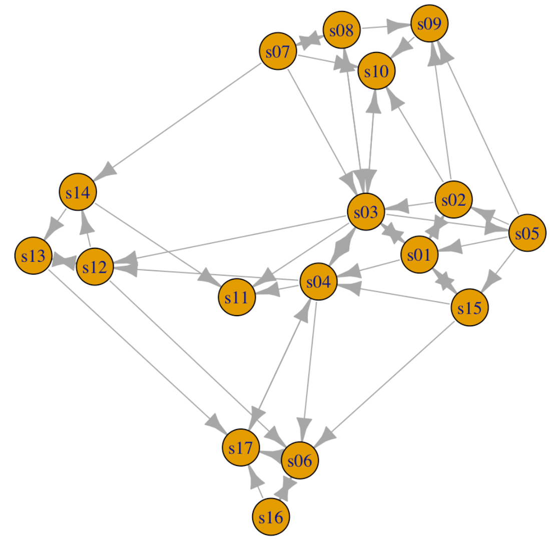

Fig. 7: Directed graph.

noteworthy that the node with highest clustering coefficient is s16 and the node with most influence is s03 according to the other measures, both when the graph is directed and when it is undirected. This phenomena could also be cross-validated visually in Fig. 6 and Fig. 7 as node s03 seems to be a "hub" and is crossed by many connections.The nodes with the most common neighbors are s03 and s09. S08 and S10 are the pair of nodes with the highest Jaccard measure.

TABLE I: Measure output

Node	LCC	DC	DP	Greg	CC	Prox	BC
s01	0.60	0.31	0.25	0.25	0.53	0.60	0.06
s02	0.60	0.31	0.12	0.25	0.48	0.54	0.00
s03	0.25	**0.56**	**0.37**	**0.43**	**0.66**	**0.75**	**0.53**
s04	0.33	0.43	0.25	0.31	0.61	0.69	0.35
s05	0.50	0.31	0.06	0.25	0.51	0.58	0.18
s06	0.40	0.31	0.25	0.12	0.50	0.56	0.07
s07	0.33	0.25	0.06	0.25	0.48	0.54	0.00
s08	0.33	0.18	0.12	0.18	0.45	0.51	0.05
s09	0.33	0.25	0.18	0.06	0.41	0.46	0.00
s10	0.50	0.25	0.25	0.06	0.47	0.53	0.10
s11	0.33	0.18	0.18	0.00	0.51	0.58	0.00
s12	0.30	0.31	0.18	0.18	0.57	0.64	0.15
s13	0.33	0.18	0.12	0.12	0.43	0.48	0.07
s14	0.16	0.25	0.12	0.12	0.47	0.53	0.00
s15	0.50	0.25	0.12	0.18	0.48	0.54	0.01
s16	**1.00**	0.12	0.06	0.12	0.35	0.40	0.00
s17	0.33	0.25	0.25	0.06	0.45	0.51	0.24