Text & Graph Data Mining (**in-class defense**)

Omar El Nahhas, 214316IV

I. Introduction

The goal of this assignment is to analyse text and graph data and draw conclusions based on the implemented functions. For the text data, 8 random speeches of Barack Obama were selected from Best Speeches of Barack Obama through his 2009 Inauguration. The graph data was selected to be the first media example proposed by the professor. All functions have been developed from scratch in R, except for functions allowed by the teaching staff. Relevant sources for code (besides course slides/code) have been mentioned in the scripts. All self-written functions have been compared and validated with R packages to confirm their correctness.

II. Pre-processing

In order to perform analysis on the data, it should be processed in the correct format and potentially filtered to perform calculations on them.

A. Text data

For the text data, the 8 separate text documents filled with Obama's speeches are loaded to form a corpus (i.e. set of texts). Afterwards, the corpus is filtered by removing punctuation, numbers, capitalised letters, English stop-words and other commonly used words such as 'the' and 'and', and white-spaces. Also, 'â€“' (one of the most commonly occurring words) has been filtered out, as this is a decoding mistake related to UTF-8. Afterwards, sparse (low frequency) items are filtered from the dataset with a threshold of 10%. All operations are performed sequentially as described above, and is then used for further analysis.

B. Graph data

For the graph data, the nodes and edges files are loaded in as CSV, after which edges data columns are appended and named in order to make it in the right dataframe format to be turned into a graph. Afterwards, the links and edges are used to convert the nodes and edges dataframes into a network. Both an indirected and directed network are created from the same original dataset. Afterwards, duplicate edges and loops (i.e. connections to the same node) are removed.

III. Results

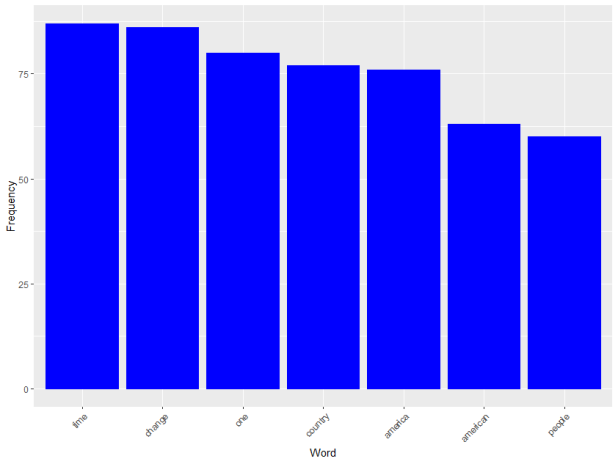
For text analysis, clustering is performed on the most frequent words (post-filtering) based on term similarity to determine the meaning of centroids and its decision boundaries. For graph analysis, 8 different measures are implemented in order to determine the role of nodes and identify the implications of changes in the network. This work limits itself to analysis on an abstract level.

A. Text analysis

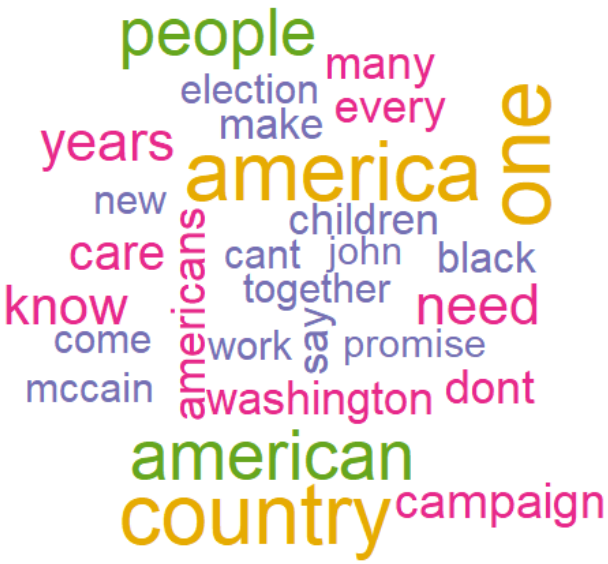
In order to understand underlying relations between the documents, clustering is performed on the most frequent words. A histogram and word cloud is seen in Fig. 1. In case of the 8 documents from Obama's speeches pre-2009, the frequent terms are identified, clustered and plotted using PCA-2. K-means, PAM, Hierachical clustering and Hierarchical DBSCAN have been implemented to identify which method(s) yield which clusters according to the chosen metric function. The best clustering method for this problem found through experimenting, K-means, is the only method visualised in this report due to size limitations. The goal is to depict the meaning of the centroids and decision boundaries in order to get a better understanding what each cluster could classify as, and until which limit (decision boundary) it would belong to this particular cluster. The 2-means clustering method with Manhattan distance in Fig. 2a yields the best separation between clusters, with the two PCA components explaining 84% of the point variability. The two cluster centroids depict the different phases of the presidential speech. The right-most red cluster shows the most common words of the hook and final statement, while the left-most blue cluster show the middle part. The edges of the clusters depict the decision boundaries, red contains catchy phrases which are similar across candidates (addressing the American people, promising change, people-oriented) with a boundary of the hook and final statement, while the blue cluster contains the actual explanation of how the catchy promises will be made a reality. Therefore, the blue cluster is what specifically identifies Barack Obama and his presidential campaign. An interesting observation is the success of Manhattan distance in this use-case, as the theory suggests that the Cosine distance is better suited for similar text data mining problems, but is giving very poor results when clustering the two PCA components using Cosine distance as seen in Fig. 2c.

B. Graph analysis

For graph analysis, the Local Clustering Coefficient (LCC), Degree Centrality (DC), Degree Prestige (DP), Gregariousness (GRG), Closeness Centrality (CC), Proximity Prestige (PP), Betweenness Centrality (BC), Common Neighbour Based Measure (CN) and Jaccard Measure (JM) have been implemented. All analysis results except for the CN and JM (due to their analysis results a large 17x17 matrix) are shown in Table I with a + indicating a directed net and - an undirected net. Based on the analysis, it can be concluded that node s03 has great importance for the information flow of the network as seen in the high betweenness centrality, among other measures. This conclusion is aligned with the visual observation of node s03 in the undirected graph plot in Fig. 3.

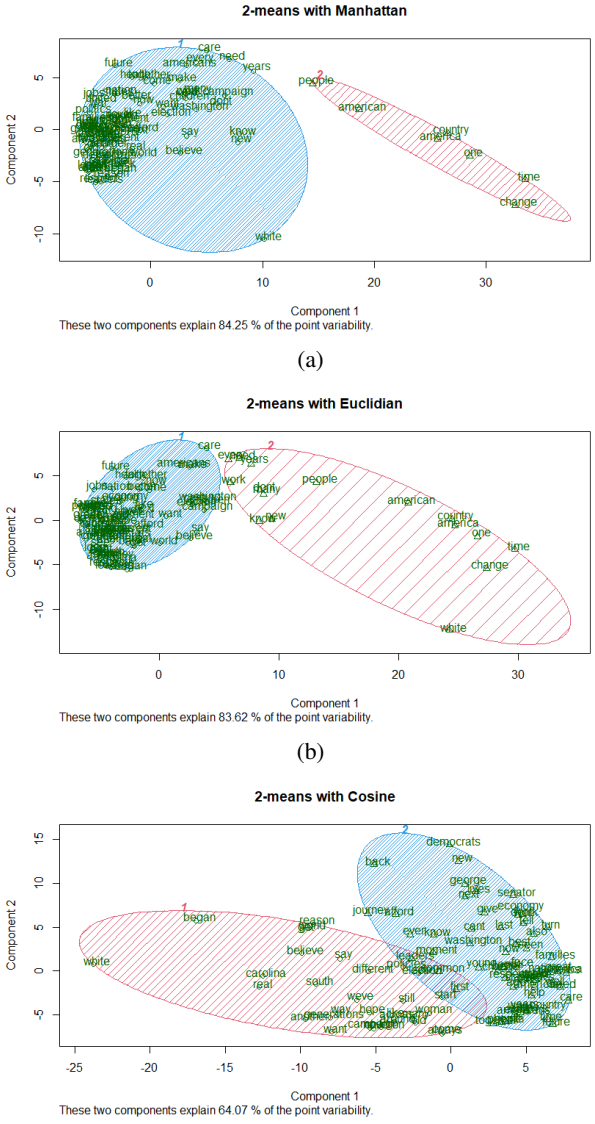


(a) Word frequency histogram.



(b) Word cloud (frequency > 35).

Fig. 1: Most frequently occurring words in Obama's speeches.



(a)

(b)

(c)

Fig. 2: K-Means with L1 (a), L2 (b) and Cosine (c).

TABLE I: Directed (+) and undirected (-) graph analysis.

Node	-LCC	-DC	+DP	+GRG	-CC	+PP	+BC
s01	0.60	0.31	0.25	0.25	8.9e-3	0.01	0.06
s02	0.60	0.31	0.13	0.25	7.4e-3	0.02	0.00
s03	0.25	**0.56**	**0.38**	**0.44**	**1.1e-2**	0.01	**0.53**
s04	0.33	0.44	0.25	0.31	9.2e-3	0.02	0.36
s05	0.50	0.31	0.06	0.25	1.0e-2	**1.00**	0.18
s06	0.40	0.31	0.25	0.13	9.8e-3	0.03	0.08
s07	0.33	0.25	0.06	0.25	9.6e-3	0.05	0.00
s08	0.33	0.19	0.13	0.19	5.6e-3	0.04	0.05
s09	0.33	0.25	0.19	0.06	8.1e-3	0.04	0.00
s10	0.50	0.25	0.25	0.06	6.9e-3	0.02	0.11
s11	0.33	0.19	0.19	0.00	9.8e-3	0.04	0.00
s12	0.30	0.31	0.19	0.19	1.1e-2	0.04	0.15
s13	0.33	0.19	0.13	0.13	5.2e-3	0.02	0.08
s14	0.16	0.25	0.13	0.13	8.6e-3	0.04	0.00
s15	0.50	0.25	0.13	0.19	8.1e-3	0.02	0.02
s16	**1.00**	0.13	0.06	0.12	2.0e-3	0.05	0.00
s17	0.33	0.25	0.25	0.06	5.7e-3	0.02	0.24

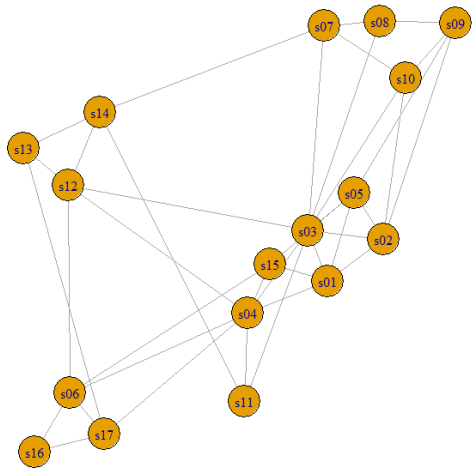


Fig. 3: Undirected graph plot of the data.