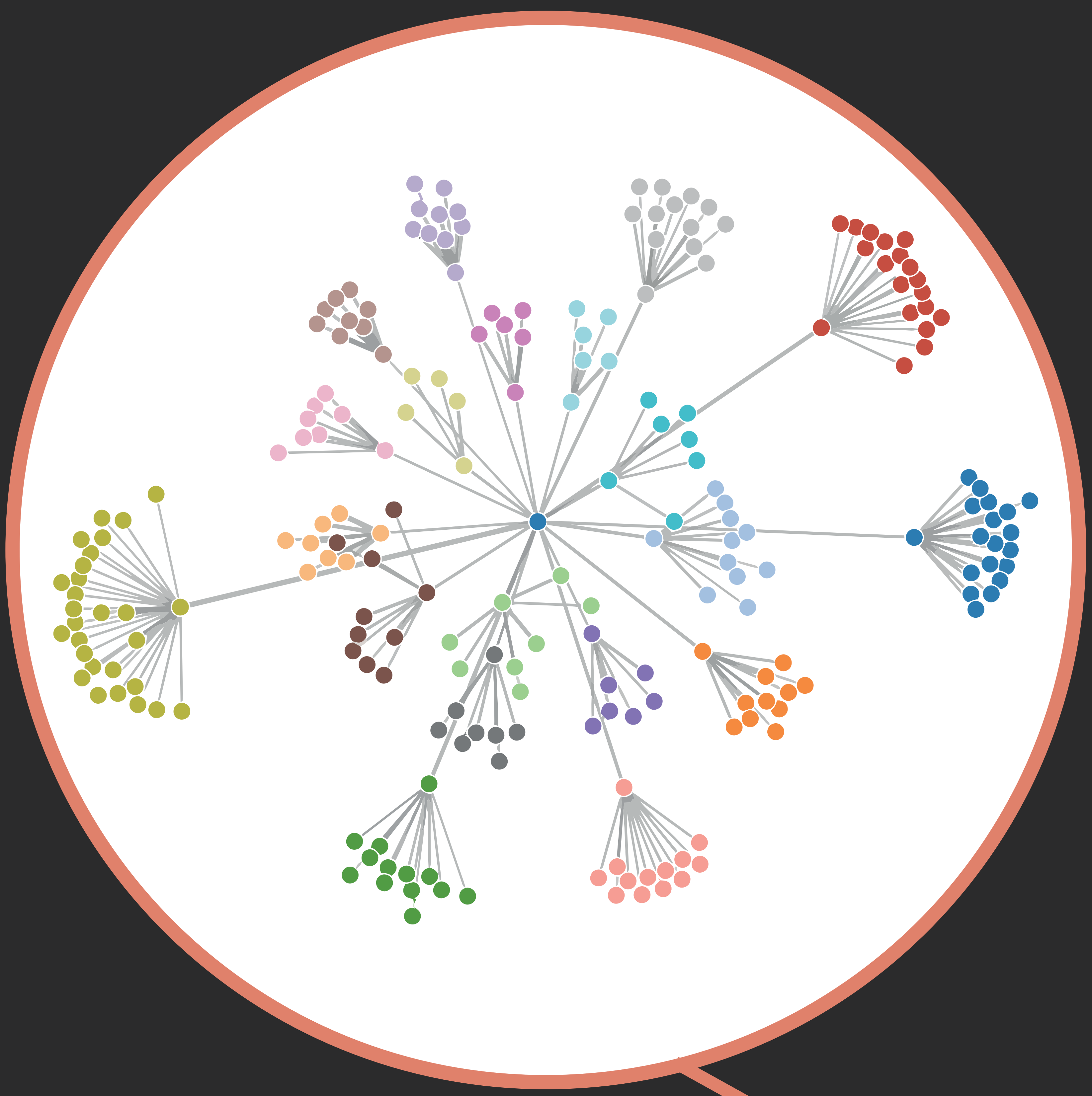
clusterr

Grouping newspaper articles



Clusterr is a web application that clusters newspaper articles and visualises them intuitively. Clusterr allows the user to retrieve a specified number of newspaper articles within a date range. It then groups the articles with a common topic into the same clusters, and displays the clusters using interactive and intuitive visualisation.

Try it yourself: clusterr.zoecai.com

Motivation

Newspaper articles are usually grouped under categories like "Business" and "National", but never by topics over time. A news reader interested in a particular topic would have to manually sift through articles to find those of interest. We wanted to automate this.

Related Systems

Carrot^2 is an open-source framework for building search clustering engines that group small collections of documents into categories, with a focus on search engine results. It is implemented in Java and uses the Lingo clustering algorithm.

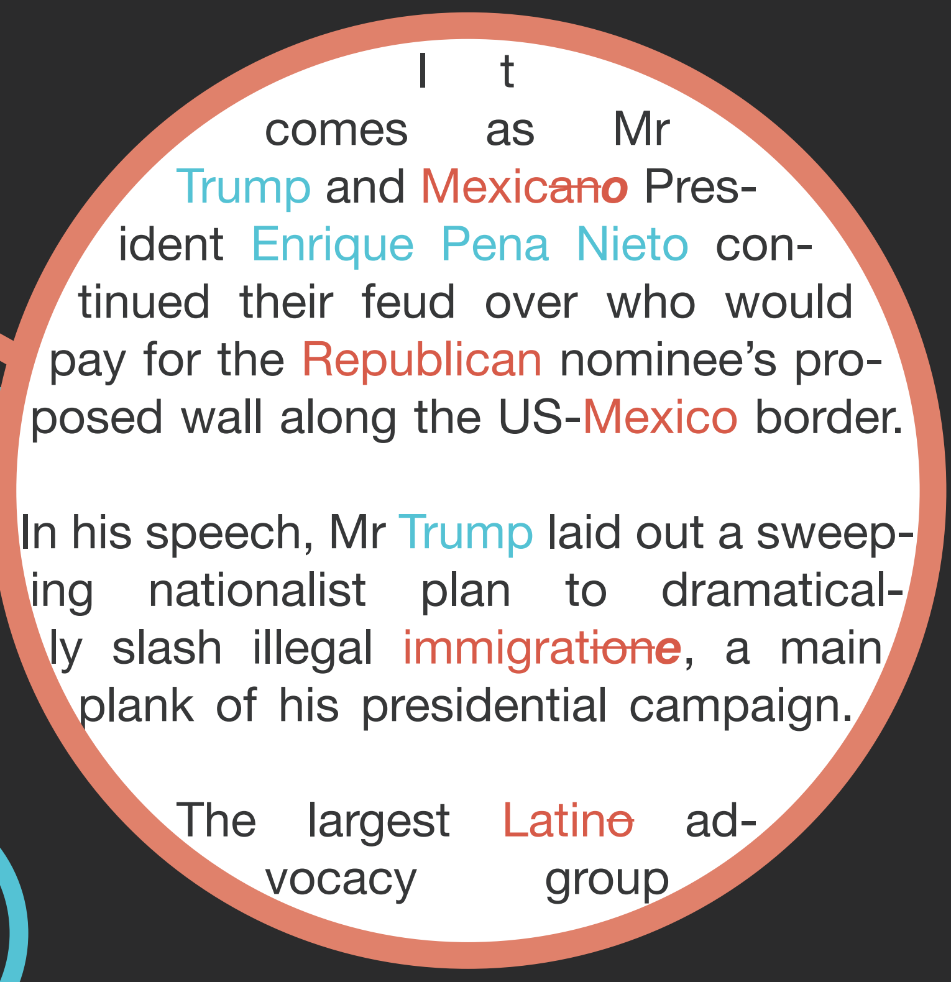
Google Trends visualises the most popular search phrases within a particular time period, allowing users to see what people care about at a particular time. Clusterr does this from the media's point of view.

Preprocessing

The documents are first preprocessed by removing stopwords and lemmatising the remaining words.

Document Representation

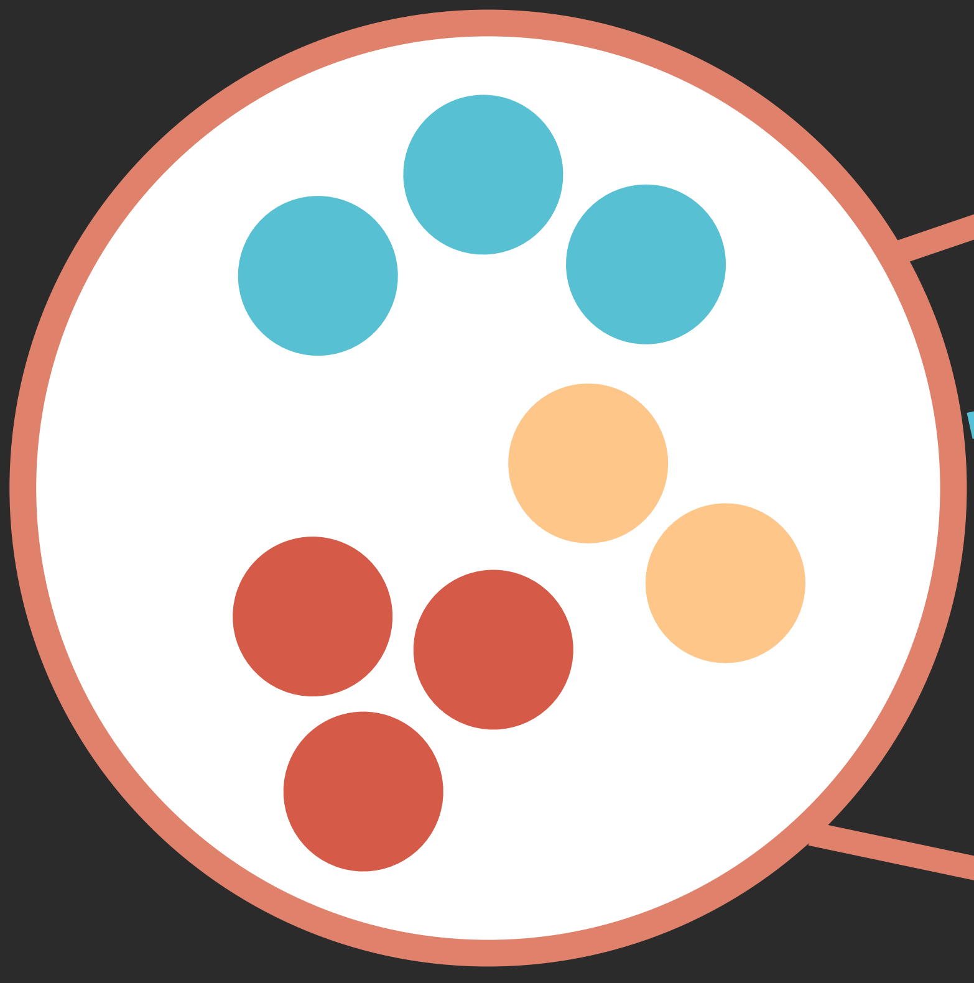
To compare the similarity of documents, Term Frequency and Inverse Document Frequency (tf-idf) weights are used to represent the importance of words as numbers. [1]



It comes as Mr Trump and Mexic~~an~~*o* President Enrique Pena Nieto continued their feud over who would pay for the Republican nominee's proposed wall along the US-Mexico border.

In his speech, Mr Trump laid out a sweeping nationalist plan to dramatically slash illegal immigrati~~on~~**e**, a main plank of his presidential campaign.

The largest Latin~~o~~ advocacy group



Clustering

Two methods of clustering are integrated for clustering — K-means and Agglomerative clustering.

Mini Batch K-Means

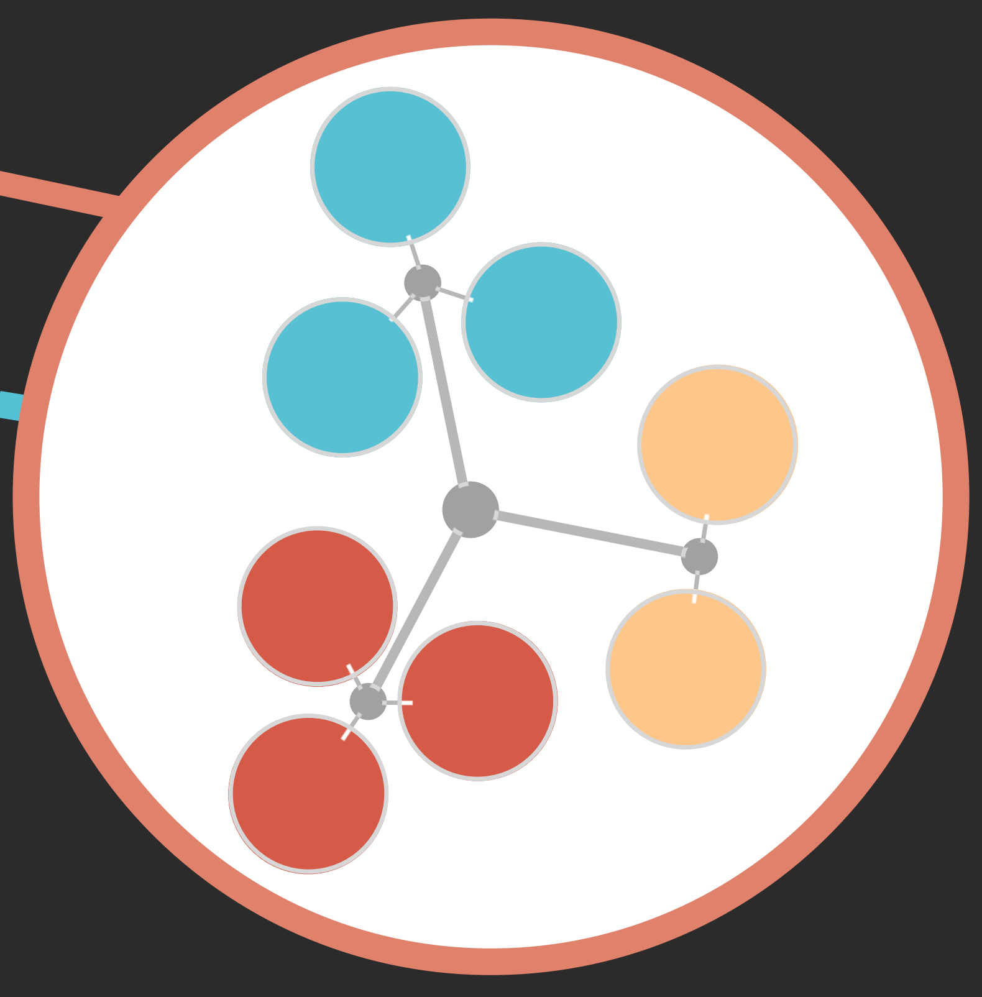
Mini Batch K-means is an iterative algorithm with two steps:

1. Assign all sample articles to its closest cluster
2. Move each cluster centroid to the center of each cluster

The process is repeated until convergence. [2]

Agglomerative Clustering

Agglomerative clustering initialises each article as its own cluster, then merges the clusters together successively by combining the two clusters with the smallest inter-cluster distance, where the inter-cluster distance is the maximum distance between any two points in the clusters. [1]



Web Application & Visualisation

The Guardian API is used to retrieve the newspaper articles, which are processed, clustered, and displayed on the web page. The distance of each node from its centroid represents how similar the article represented is to the rest of the articles in the same cluster.

Future Work

- To make current clustering algorithms faster and more accurate
- To use semantic meaning of the article text in addition to tf-idf
- To try supervised learning and other clustering algorithms

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Project 72

Ruoyi (Zoe) Cai, Chanjun Park
Supervisor: Gill Dobbie

References

[1] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.
[2] D. Sculley, "Web-Scale K-Means Clustering," in WWW 2010, Raleigh, 2010.