Linköping University | Department of Computer science
Bachelor thesis, 16 ECTS | Datateknik
2016 | LIU-IDA/LITH-EX-G--16/037--SE

Cluster Analysis of Discussions on Internet Forums

Klusteranalys av Diskussioner på Internetforum

Rasmus Holm

Supervisor : Berkant Savas
Examiner : Cyrille Berger

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår. Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art. Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannenslitterära eller konstnärliga anseende eller egenart. För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances. The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Rasmus Holm

Abstract

The growth of textual content on internet forums over the last decade have been immense which have resulted in users struggling to find relevant information in a convenient and quick way.

The activity of finding information from large data collections is known as information retrieval and many tools and techniques have been developed to tackle common problems. Cluster analysis is a technique for grouping similar objects into smaller groups (clusters) such that the objects within a cluster are more similar than objects between clusters.

We have investigated the clustering algorithms, Graclus and Non-Exhaustive Overlapping *k*-means (NEO-*k*-means), on textual data taken from Reddit, a social network service. One of the difficulties with the aforementioned algorithms is that both have an input parameter controlling how many clusters to find. We have used a greedy modularity maximization algorithm in order to estimate the number of clusters that exist in discussion threads.

We have shown that it is possible to find subtopics within discussions and that in terms of execution time, Graclus has a clear advantage over NEO-*k*-means.

Acknowledgments

First and foremost, I would like to say thanks to Berkant Savas for giving me the opportunity to do my bachelor thesis at iMatrics and for being my supervisor. I have learned a lot during the few months of work.

I would also like to thank Cyrille Berger for being my examiner, giving me directions on how to solve problems that I have encountered and all the great feedback.

Finally, I would like to thank Martin Estgren and Daniel Nilsson for giving me feedback on the report.

Contents

List of Figures

List of Tables

1 Introduction

The social media and internet forums on the Internet has expanded massively in the last decade with companies such as *Facebook*, *Twitter* and *Reddit*. It contains huge amounts of textual information with various degree of relevance and for a regular user it can be incredibly hard to find what he or she is looking for. It is also difficult as a user to accommodate to a new social media without being overwhelmed by the amount of information in search for something interesting and relevant.

Information retrieval is the activity of finding information from large data collections and much research has been done in the area with development of tools and techniques to tackle common problems. *Clustering* is one technique that can be used to find groups of similar data objects in a data collection which can provide insight and understanding of the data. This insight can then be incorporated into assistance services making it easier and friendlier for users to navigate and search through data [24].

1.1 Motivation

An internet forum is a place where people are able to hold conversations in the form of posting messages and because of the anonymity the Internet brings, the conversations often bring forth internet trolls that deliberately provoke other users through posts containing abnormal or perverse content for their own amusement. Conversations can go on for a very long period of time and be composed of hundreds or thousands of posts. For a user that have not actively been participating since the beginning may find it very difficult to follow the current discussion or may be intimidated to the point where it is no longer of interest even though the user has taken an interest in the topic.

The amount of information that are put up on the Internet on a monthly basis is huge which can be observed in the figures 1.1 and 1.2 for just a small part of Reddit, more in 1.2. Computer algorithms can potentially be used to gain insight into all this data.

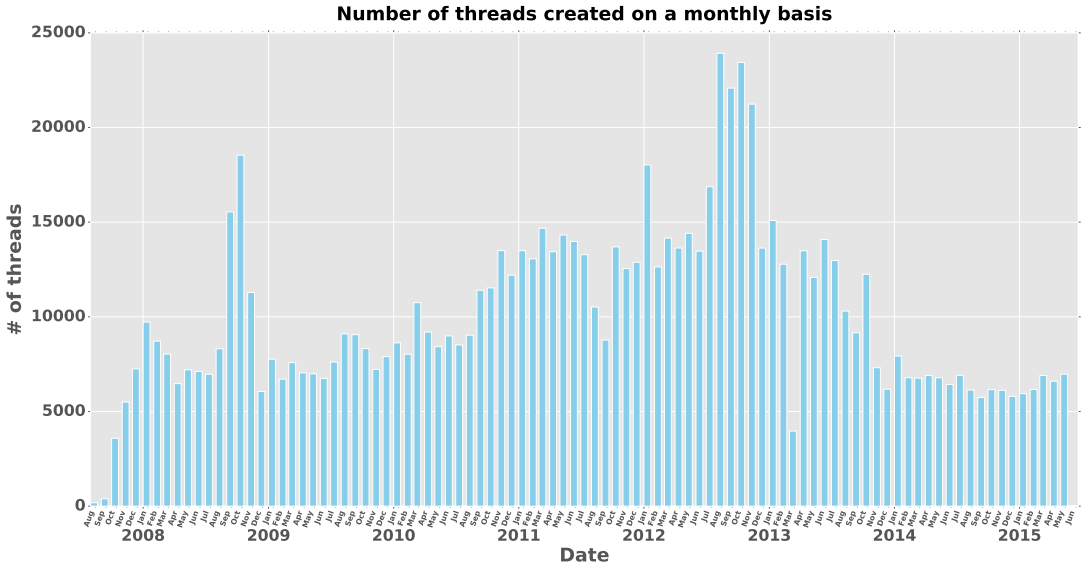


Figure 1.1: The number of threads created on a monthly basis in the politics subreddit over the period of October, 2007 and May, 2015. The two distinct spikes in 2008 and 2012 are most likely explained by the presidential election in the United States of America at the time.

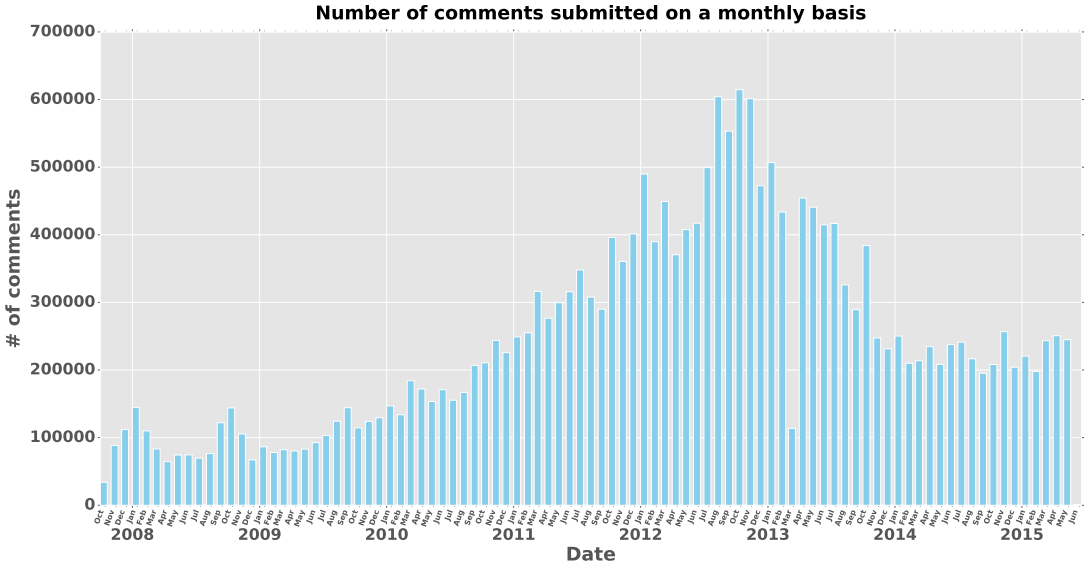


Figure 1.2: The number of comments submitted on a monthly basis in the politics subreddit over the period of October, 2007 and May, 2015. The subreddit saw a rapid increase of submitted comments until 2013 and then started to decline. This is probably due to content being pushed to another subreddit. The news subreddit started to gain popularity at the time[2].

Clustering techniques can potentially find posts by internet trolls and by using this information, automated tools could be developed that hide/delete those posts resulting in less off-topic content and reducing the amount of content shown to the user. Clustering may also be of help in finding meaningful posts and recognize users that are well involved in the conversation and are knowledgeable about the topic.

[2]http://redditmetrics.com/r/politics#compareñews

1.2 Reddit

Reddit is a social network service since 2005 and one of the most visited[3] websites on the Internet. Reddit consists of subreddits that can be described as communities discussing a certain topic of interest such as news, gaming, or politics. Today, June 30, 2016, Reddit has around 880,000[4] subreddits in total and had over 725 million comments[5] submitted in 2015. Every subreddit is composed of discussion threads, which will be referred to as threads, about a specific subject and users are able to submit posts, which will be referred to as comments, regarding the subject. Because of the sheer size of Reddit, it is very difficult and time consuming for users to navigate and find the desired information. Therefore is Reddit the ideal target to test clustering algorithms that may possibly address the problem of too much information.

The study will be using user comments from the Reddit discussion forum. The data collection[6] contains around 1.3 billion user comments between October, 2007 and May, 2015.

1.3 iMatrics

The thesis will be carried out at *iMatrics AB*, a company conducting text analysis and is developing tools to improve the user experience in online discussion forums. For instance to make it easier to navigate through text, extract relevant information, detect abusive content, and recommend content.

1.4 Aim

The purpose of this thesis is to investigate different clustering algorithms from the literature on textual data taken from Reddit and find out what kind of information that can be extracted in order to improve the user experience on internet forums.

1.5 Research questions

- Can the chosen clustering algorithms be used to find structure in textual content?
- How do the algorithms compare in terms of execution time?

1.6 Delimitations

Cluster analysis is a vast field with many methods and it is not possible to cover every single one. We have limited the choice of clustering algorithms from two families. The *k-means* algorithm and its extensions and *graph partitioning* techniques. These methods have shown great performance in practice on large scale data in terms of execution time and high quality of the clustering results [22, 11].

Using the entire available dataset is not possible because the size is too large to process within the time frame. The data used for analysis have been reduced to only include the *politics* subreddit.

[3] https://www.similarweb.com/website/reddit.com
[4] http://redditmetrics.com/history
[5] http://expandedramblings.com/index.php/reddit-stats/2/
[6] https://www.reddit.com/r/datasets/comments/3bxlg7/ i_have_every_publicly_available_reddit_comment/

2 Theory

In this chapter the theory around clustering will be presented. It starts with a brief introduction to the field of machine learning followed by an introduction to the mathematical notation. Then information about text representation, similarity metrics, and graphs will be presented. The final two sections will be about clustering algorithms and cluster validation methods.

2.1 Machine Learning

In machine learning, there are three major learning paradigms namely *supervised*, *unsupervised*, and *reinforcement* learning [30].

Supervised learning is learning by examples through inputs of "correct" answers known as the *ground truth* given the set of features to an algorithm. This process is called the *training phase*. An example could be a set of patient records with a diagnosis of some type of tumour and it is either benign (not cancerous) or malignant (cancerous). By using this data with supervised learning, it is possible to create a model based on the features in the records, e.g., the size of the tumour. This model can then be used to predict whether a new patient has cancer given its features. The rate at which a model predicts correctly depends on which algorithm is used, what features are used, and many other parameters.

In unsupervised learning there is no "correct" answer, but it may still be desirable to derive structure from the data. An example could be to find groups of customers who share similar purchase behaviour and use that information for targeted advertising.

Reinforcement learning is learn by trial-and-error and is commonly used in dynamic environments where feedback comes as rewards. For instance a robot trying to walk and gets a reward for every step it takes and no reward for falling over.

Cluster analysis is included in the unsupervised learning paradigm and is a technique for grouping or segmenting a collection of objects into smaller groups (clusters) such that the objects within a cluster are more related to each other than objects from different clusters. The clusters can be used to describe different properties in a collection of data [18]. Due to being an unsupervised technique it can be difficult to evaluate the clustering solution. Usually no one knows what kind of information the clusters will contain and domain knowledge has to be used to determine if the clusters yield useful results. There are however other evaluation methods to consider that will be presented at the end of this chapter. Clustering has for

instance been used in image segmentation to find objects and striking features [31], finding patterns in gene expression in order to understand biological processes [5], and in many other fields. Figure 2.1 demonstrate a simple example of clustering.

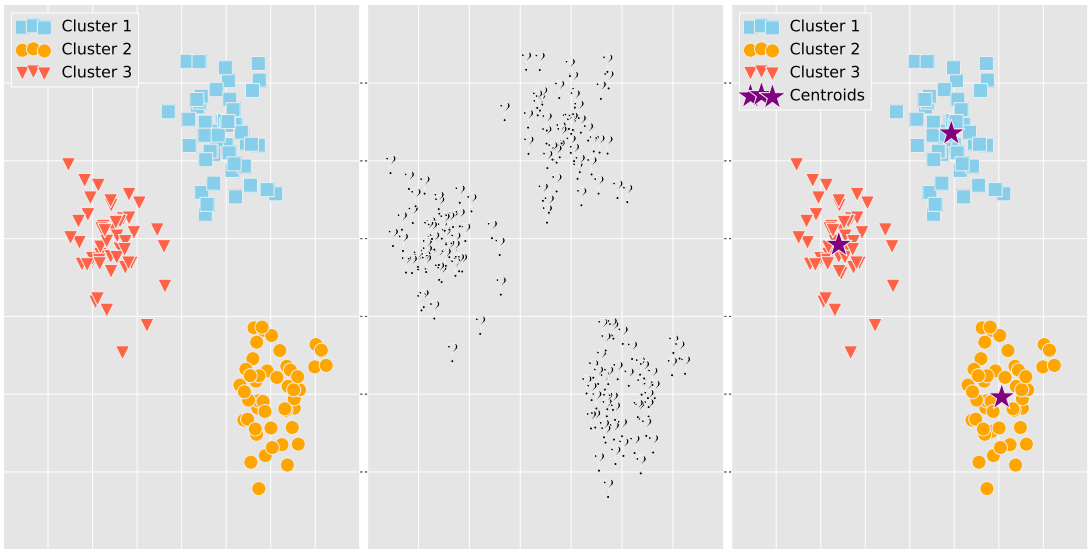


Figure 2.1: Left: The ground truth. Center: What the input looks like from the perspective of the clustering algorithm. Right: The output clusters from a run made by the k-means algorithm where the purple stars represent the centroids of the clusters.

In figure 2.1 the algorithm can perfectly distinguish the groups, but this is a very simplified example with only two dimensions and the groups are well separated into ellipsoid looking point clouds. The data is usually not that perfectly separable and can have different looking patterns such as in figure 2.2.

Figure 2.2: Left: The ground truth. Center: What the input looks like from the perspective of the clustering algorithm. Right: The output clusters from a run made by the k-means algorithm where the purple stars represent the centroids of the clusters.

In figure 2.2, the algorithm cannot distinguish between the two groups because of how the shapes almost overlap and the two groups are not linearly separable. The *k-means* algorithm

will be presented in section 2.6 together with other alternative algorithms that may be better at finding groups such as those in figure 2.2.

2.2 Mathematical Notation

Capital calligraphic letters will denote sets, e.g., $\mathcal{D} = \{d_1, \ldots, d_n\}$ and $|\mathcal{D}|$ is the cardinality of the set, i.e., the number of elements n. The same notation will be used to denote the length of a vector. Lower case letters, e.g., v or v_i are always assumed to be vectors, unless otherwise stated. The transpose of a vector is denoted v^T and the dot product between two vectors is denoted $u^T v$. Matrices will be denoted with capital letters, e.g., U or U_i. The character "#" will be used as a short hand for the word "number", e.g., "# of cars" is translated to "number of cars".

2.3 Text Representation and Transformation

Bag of words is a common representation of a text document which describes the set of words the text document contains. In order to obtain all the words in a document, a *tokenization* preprocessing step is required to split the text document into a stream of *terms*. This is done by removing punctuations and replacing non-text characters with white space. The set of all terms in the document collection is called the *dictionary* of the document collection [19]. Given the two sentences "Hello world!" and "Hello, how are you?", the dictionary is consisting of the terms "Hello", "world", "how", "are", and "you".

The *term frequency* (tf) of term t in document d with the terms t_d is defined as

$$F_{\text{tf}}(d, t) = \sum_{w \in t_d} \mathbb{1}(t = w), \tag{2.1}$$

where

$$\mathbb{1}(\text{expr}) = \begin{cases} 1 & \text{if expr is true,} \\ 0 & \text{otherwise,} \end{cases}$$

is the indicator function. Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ be a set of documents and $\mathcal{T} = \{t_1, \ldots, t_m\}$ be the set of terms that occurs in $\mathcal{D}$. The vector representation of a document d_i is then defined as

$$v_i = \left(F_{\text{tf}}\left(d_i, t_1\right), \ldots, F_{\text{tf}}\left(d_i, t_m\right)\right). \tag{2.2}$$

Term frequency-inverted document frequency (tfidf) is another term frequency metric that can be used to give less weight to frequently occurring terms in distance and similarity computations and is defined as

$$F_{\text{tfidf}}\left(d, t\right) = F_{\text{tf}}\left(d, t\right) \log\left(\frac{|\mathcal{D}|}{F_{\text{df}}\left(t\right)}\right), \tag{2.3}$$

where $F_{\text{df}}(t)$ is the number of documents the term t appears in. Then the vector representation of a document d_i is defined as

$$v_i = \left(F_{\text{tfidf}}\left(d_i, t_1\right), \ldots, F_{\text{tfidf}}\left(d_i, t_m\right)\right). \tag{2.4}$$

The tfidf can be interpreted as follows [24]:

- High when t occurs frequently within a small group of documents.
- Low when the term t occurs infrequently or occurs in a big portion of the documents.

With a set of n documents $\mathcal{D}$ consisting of the set of m terms $\mathcal{T}$, the *document-term frequency matrix* contains rows corresponding to the documents and the columns corresponding to the terms as

$$F_{\text{dtf}} = \begin{pmatrix} F(d_1,t_1) & F(d_1,t_2) & \cdots & F(d_1,t_m) \\ F(d_2,t_1) & F(d_2,t_2) & \cdots & F(d_2,t_m) \\ \vdots & \vdots & \ddots & \vdots \\ F(d_n,t_1) & F(d_n,t_2) & \cdots & F(d_n,t_m) \end{pmatrix},$$

where $F(d_i,t_j)$ is either $F_{\text{tf}}(d_i,t_j)$ or $F_{\text{tfidf}}(d_i,t_j)$.

These representations are called *vector space models* (VSMs) which have the key assumption that the ordering of the words does not matter. There are however two problems with the VSM representation, *high dimensionality* of feature space and *sparse* data. There are feature selection methods that can reduce these problems by reducing the size of the dictionary [23].

Filtering is the process of removing words from the dictionary and a standard method is by removing *stop words* which are words such as "a" and "the" that does not contribute much information about the content. Words that occur very often or very seldom can also be considered uninformative words that can be removed [23, 1].

Stemming is a method for trying to build the basic forms (stems) of words by removing the ending of the words, e.g., *producer*, *produce*, *product* and *production* becomes *produc*. This is usually done by *Porter's suffix-stripping algorithm* for the English language [23].

2.4 Similarity and Distance Metrics

Anna Huang [20] and Strehl et al. [15] have conducted studies regarding the impact of different *similarity* and *distance metrics* on text data. In this section, one metric that were found in aforementioned studies to give good results compared to human expert classification will be presented.

2.4.1 Cosine Similarity

The *cosine similarity* [20] is defined as the cosine of the angle between two vectors and can then be used when documents are represented by vectors as presented above. Given two documents v and w, their cosine similarity is expressed as

$$S_C(v,w) = \frac{v^T w}{\|v\|\|w\|}, \tag{2.5}$$

where $v, w \in \Re^m$ and $\|v\| = \sqrt{\sum_{i=1}^{|v|} v_i}$ given v_i is the value at position i in vector v. The result will be $S_C(v,w) \in [0,1]$ given $v, w \geq 0$. The output is 1 if the vectors are identical and 0 if they are perpendicular to each other. The distance metric is defined as

$$D_C(v,w) = 1 - S_C(v,w). \tag{2.6}$$

2.5 Graph

Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph with a set of vertices $\mathcal{V} = \{v_1, \ldots, v_n\}$ and a set of edges $\mathcal{E} = \{e_1, \ldots, e_m\}$. The *weighted adjacency matrix* of a graph is the matrix $W \in \Re^{n \times n}$ with $w_{ij} \geq 0$ for $i, j = 1, \ldots, n$. If $w_{ij} = 0$ then the vertices v_i and v_j are not connected by an edge. The weighted adjacency matrix is symmetric, i.e., $w_{ij} = w_{ji}$ for $i, j = 1, \ldots, n$. For example if a vertex corresponds to a geographic location the edge weights w_{ij} could correspond to the distance between the locations i and j.

The adjacency matrix will be denoted A and has the same properties as W with the exception that $a_{ij} \in \{0, 1\}$. Assume vertices correspond to users in a social network then the value of a_{ij} could be 1 if users i and j are friends, 0 otherwise.

The degree of a vertex v_i is defined as $d_i = \sum_{j=1}^{n} w_{ij}$ and the *degree matrix* D is defined as the diagonal matrix with degrees $d_1, \ldots, d_n$ on the diagonal.

2.5.1 Graph Partitioning

The graph partitioning problem aims to find k disjoint vertex partitions $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k$ such that $\mathcal{V}_1 \cup \mathcal{V}_2 \cup \ldots \cup \mathcal{V}_k = \mathcal{V}$ and some measurement is minimal/maximal. To be able to accomplish this task various objective functions have been defined to evaluate a set of partitions. In this section a few such objectives will be formally defined.

2.5.1.1 Cut

Given the weighted adjacency matrix W and $W(U, V) = \sum_{i \in U, j \in V} w_{ij}$, the *mincut* is defined as

$$\text{cut}(\mathcal{V}_1, \ldots, \mathcal{V}_k) = \min_{\mathcal{V}_i, \ldots, \mathcal{V}_k} \frac{1}{2} \sum_{i=1}^{k} W(\mathcal{V}_i, \mathcal{V} \setminus \mathcal{V}_i), \tag{2.7}$$

where $\mathcal{V}_i \subset \mathcal{V}$ and $\mathcal{V} \setminus \mathcal{V}_i$ is the set difference, i.e, all the elements in $\mathcal{V}$ that are not in $\mathcal{V}_i$. The mincut does not yield satisfactory partitions in practice because the solution often results in separating individual vertices from the graph. Some extensions to it have therefore been developed known as *normalized cut* and *ratio cut* that constrains the size of the partitions to be more reasonable [33]. They are defined as

$$\text{Ncut}(\mathcal{V}_1, \ldots, \mathcal{V}_k) = \min_{\mathcal{V}_i, \ldots, \mathcal{V}_k} \sum_{i=1}^{k} \frac{W(\mathcal{V}_i, \mathcal{V} \setminus \mathcal{V}_i)}{\text{vol}(\mathcal{V}_i)}, \tag{2.8}$$

$$\text{RatioCut}(\mathcal{V}_1, \ldots, \mathcal{V}_k) = \min_{\mathcal{V}_i, \ldots, \mathcal{V}_k} \sum_{i=1}^{k} \frac{W(\mathcal{V}_i, \mathcal{V} \setminus \mathcal{V}_i)}{|\mathcal{V}_i|}, \tag{2.9}$$

where $\text{vol}(\mathcal{V}) = \sum_{i \in \mathcal{V}} d_i$.

2.5.1.2 Ratio Association

The *ratio association* objective does the opposite of the ratio cut and tries to maximize the within-cluster association relative to its size. It is defined as

$$\text{RAssoc}(\mathcal{V}_1, \ldots, \mathcal{V}_k) = \max_{\mathcal{V}_i, \ldots, \mathcal{V}_k} \sum_{i=1}^{k} \frac{W(\mathcal{V}_i, \mathcal{V}_i)}{|\mathcal{V}_i|}. \tag{2.10}$$

2.5.1.3 Modularity

Another type of measure is the *modularity* by Newman and Girvan [26] which looks at the edge distribution in the graph and compares it to the expected edge distribution of a random graph known as the *null model*. A null model is a graph which matches some of the structural features from a specific graph, but is otherwise taken as an instance of a random graph. A random graph is described by a probability distribution from which the graph was generated. The null model is expected to not possess any particular structure, hence it can be used to check if the studied graph displays structure or not. A common null model, proposed by Newman and Girvan [26], adds edges at random under the constraint that the expected degree of each vertex matches the ones in the original graph.

Let E be defined as a $k \times k$ symmetric matrix whose element

$$e_{ij} = \frac{\sum_{n \in \mathcal{V}_i, m \in \mathcal{V}_j} a_{nm}}{|\mathcal{E}|},$$

where $\mathcal{V}_i$ and $\mathcal{V}_j$ are partitions. The trace $Tr(E) = \sum_i e_{ii}$ is the fraction of edges that connect vertices in the same partition, and a good partitioning of the graph should obviously have a high value of the trace. This is however not enough because the optimal value would be to have all vertices in a single connected component. To address this issue, the modularity is defined as

$$Q = \sum_i (e_{ii} - a_i^2), \tag{2.11}$$

where $a_i = \sum_j e_{ij}$, the fraction of edges that connect to vertices in c_i.

2.6 Clustering Algorithms

Clustering algorithms can have different properties, some are the following [24]:

- *Hard clustering* where every data point is assigned to excatly one cluster
- *Overlapping clustering* where every data point can be assigned to more than one cluster.
- *Flat clustering* creates clusters without relationship between clusters.
- *Hierarchical clustering* creates a hierarchy of clusters.

2.6.1 k-Means

The *k-means* algorithms is a hard flat clustering algorithm and can be summarized in 3 steps given a dataset $\mathcal{X}$ [21].

1. Select k initial cluster centroids. Repeat step 2 and 3 until convergence.
2. Assign each data point $x \in \mathcal{X}$ to its closest cluster centroid.
3. Compute new cluster centroids by averaging over all assigned data points for each cluster.

The objective of k-means can be seen as minimizing the sum of the squared error over all k clusters and is expressed as

$$J(\mathcal{C}) = \min_{\mathcal{C}} \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|^2, \tag{2.12}$$

where $\mathcal{C} = \{c_1, \ldots, c_k\}$ is the set of k clusters and $\mu_i = \sum_{x \in c_i} \frac{x}{|c_i|}$ is the centroid of c_i.

k-means is a simple algorithm, however it requires difficult tuning of usage-specific parameters. Those are the number of clusters k, selection of the initial k cluster centroids, and the distance metric. The distance metric is usually the *Euclidean distance* which results in finding ellipsoid looking clusters like those in figure 2.1. The number of clusters can be domain specific, e.g., trying to find three different shirt sizes (S, M, L) based on customer heights and weights. There is no universal way of knowing how many clusters to choose and lastly the initial positions of the cluster centroids are very important since the algorithm converges to

a local minimum. A naive approach is to run the algorithm with different initial cluster centroids and pick the cluster centroids with the least squared error, but there are more advanced methods like the *k-means++* algorithm [4] that can improve in terms of speed and the objective value.

2.6.2 Non-Exhaustive Overlapping *k*-Means

The *Non-Exhaustive Overlapping k-means* (NEO-*k*-means) algorithms is *non-exhaustive* meaning it addresses the issue of *outliers* by not assigning every data point to atleast one cluster. The NEO-*k*-means is an extension to the *k*-means algorithm described above with a modified objective function [34].

The NEO-*k*-means algorithm consists of a set of clusters $\mathcal{C} = \{c_1, \ldots, c_k\}$ and given a set of data points $\mathcal{X} = \{x_1, \ldots, x_n\}$, an assignment matrix $U \in \Re^{n \times k}$ is constructed such that $u_{ij} = 1$ if x_i belongs to cluster c_j, 0 otherwise. The objective function is defined as

$$J(\mathcal{C}) = \min_U \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \|x_i - m_j\|,$$

$$\text{where } m_j = \frac{\sum_{i=1}^{n} u_{ij} x_i}{\sum_{i=1}^{n} u_{ij}},$$

$$\text{s.t. } \mathrm{Tr}(U^T U) = (1 + \alpha)n, \quad (1)$$

$$\sum_{i=1}^{n} \mathbb{1}((U\mathbf{1})_i = 0) \leq \beta n. \quad (2)$$

$\mathbf{1}$ is a vector of length k having all elements set to 1, therefore $(U\mathbf{1})_i$ equals the number of clusters x_i belongs to. Constraint (1) limits the number of total cluster assignments and constraint (2) specifies the maximum number of outliers. α and β are user defined parameters to control the size of the overlapping region and the maximum percentage of outliers respectively. It is required to have $0 \leq \alpha \leq (k-1)$ and $\beta n \geq 0$ and setting $\alpha = 0$ and $\beta = 0$ equals the regular *k*-means algorithm.

2.6.3 Kernel *k*-Means

As shown in figure 2.2, *k*-means cannot always separate groups of data points. To allow nonlinear separators, a *kernel* is used denoted Φ which is a function that maps data points to a higher dimensional feature space. Then the regular *k*-means algorithm can be applied in this new feature space which corresponds to nonlinear separators in the input space.

The kernel *k*-means objective function is

$$J(\mathcal{C}) = \min_{\mathcal{C}} \sum_{m=1}^{k} \sum_{x_i \in c_m} \|\Phi(x_i) - \mu_m\|^2, \quad (2.13)$$

where $\mathcal{C} = \{c_1, \ldots, c_k\}$ is the set of k clusters and $\mu_m = \sum_{x_i \in c_m} \frac{\Phi(x_i)}{|c_m|}$ is the centroid of c_m. $\|\Phi(x_i) - \mu_m\|^2$ can be rewritten as

$$\|\Phi(x_i) - \mu_m\|^2 = \Phi(x_i)^T \Phi(x_j) - \frac{2\sum_{x_j \in c_m} \Phi(x_i)^T \Phi(x_j)}{|c_m|} + \frac{2\sum_{x_j, x_l \in c_m} \Phi(x_j)^T \Phi(x_l)}{|c_m^2|}. \quad (2.14)$$

Only inner products are calculated with the kernel function implying a *kernel matrix K* can be created where $k_{ij} = \Phi(x_i)^T \Phi(x_j)$.

By using kernels it is possible to optimize the graph theoretic objectives defined in 2.5.1 with the kernel *k*-means algorithm and more generally using the *weighted kernel k-means* algorithm, for a detailed explanation and examples of common kernels see [11].

2.6.4 Non-Exhaustive Overlapping k-Means on Graphs

Kernel k-means can optimize graph theoretic objectives and so there is a natural transition of the NEO-k-means algorithm to work on graphs as well. Let Y be the assignment matrix such that $y_{ij} = 1$ if vertex v_i belongs to partition c_j, $y_{ij} = 0$ otherwise. Also let y_j denote the jth column of Y, then the non-exhaustive overlapping graph clustering objective is defined as

$$\begin{aligned} J(G) &= \max_{Y} \sum_{j=1}^{k} \frac{y_j^T A y_j}{y_j^T D y_j}, \\ &\text{s.t. } \mathrm{Tr}(Y^T Y) = (1+\alpha)n, \\ &\sum_{i=1}^{n} \mathbb{1}\{(Y\mathbf{1})_i = 0\} \leq \beta n. \end{aligned} \tag{2.15}$$

α and β control the degree of overlap and exhaustiveness respectively. By setting $\alpha = 0$ and $\beta = 0$, the objective is equivalent to the normalized cut. It is possible to adjust to other objectives as well. The implementation of the algorithm by Whang et al. [34] uses the multilevel framework which will be explained in the context of *METIS* and *Graclus* below.

2.6.5 METIS

The METIS[1] software includes a set of serial programs for partitioning graphs and much more. The algorithm that will be described is built upon the multilevel framework and tries to optimize the k-way partitioning problem. The k-way partitioning problem is defined as finding subsets $\mathcal{V}_1, \ldots, \mathcal{V}_k$ such that $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$ for $i \neq j$, $|\mathcal{V}_i| = |\mathcal{V}|/k$, and $\mathcal{V}_1 \cup \ldots \cup \mathcal{V}_k = \mathcal{V}$ given the graph $G = (\mathcal{V}, \mathcal{E})$. The objective is to minimize the number of edges incident to vertices belonging to different subsets called the *edge-cut*.

The basic structure of multilevel framework is to take a graph G and coarsen it down to a graph consisting of relatively few vertices, partition the smaller graph, and project the result back towards the original graph. These steps correspond to three phases that make up the multilevel framework and those will be described next, for a more extensive description of METIS see [22].

2.6.5.1 Coarsening

The *coarsening phase* transforms the graph G_0 into a sequence of smaller graphs $G_1, \ldots, G_m$ such that $|\mathcal{V}_0| > |\mathcal{V}_1| > \ldots > |\mathcal{V}_m|$. A basic scheme for doing this is to combine vertices into *multinodes* and preserve all the edge information by setting the edges to the union of the edges.

One of the techniques METIS incorporates is the *heavy edge matching* (HEM) and it works as follows:

1. Set all vertices to unmarked.
2. Visit random vertex v and merge it with the adjacent unmarked vertex y that corresponds to the highest edge weight among all its adjacent vertices.
3. Set x and y to marked.
4. Repeat step 2 until all vertices have been marked.

[1] http://glaros.dtc.umn.edu/gkhome/metis/metis/overview

2.6.5.2 Partitioning

In the *partitioning phase* $G_m = (\mathcal{V}_m, \mathcal{E}_m)$ is partitioned into two parts, $\mathcal{P}_m$, each containing half the vertices of the original graph G_0. A simple approach to bisect a graph is by using a graph growing algorithm that selects a random vertex and grows a region in breath-first fashion until half the vertices are in the region.

A greedy extension to this is actually used in METIS that defines the edge-cut gained by inserting a vertex v into the growing region and the algorithm then picks the vertex with the largest gain, i.e., largest decrease in edge-cut. Multiple runs are made since it is sensitive to the starting vertex and the partitions that yield the least edge-cut are selected.

2.6.5.3 Refinement

The final phase is called the *refinement phase* where the partitions $\mathcal{P}_m$ are projected back up through intermediate partitions $\mathcal{P}_{m-1}, \mathcal{P}_{m-2}, \ldots, \mathcal{P}_1, \mathcal{P}_0$ until reaching the granularity of the original graph. Partitions P_i entails partitions in $\mathcal{P}_{i-1}$ so given a supernode in a partition of $\mathcal{P}_i$, all vertices that formed the supernode from $\mathcal{P}_{i-1}$ will be in the same partition. Since there is greater granularity in $\mathcal{P}_{i-1}$, a refinement algorithm is used to increase the edge-cut by swapping subsets of vertices between the partitions as to decrease the edge-cut. METIS uses a variation of the *Kernighan-Lin* refinement algorithm [22] which is an iterative algorithm that swaps vertices until no further edge-cut reduction is possible. One problem with the Kernighan-Lin algorithm is that is forces the partition to be almost equal sized which is not always true in practice and that is a major limitation of METIS.

2.6.6 Graclus

Graclus[1] [11] is another algorithm that uses the multilevel framework, one of the motivations behind the framework is that *spectral* clustering methods are commonly used for graph clustering. Those methods are based on the graph Laplacian matrix and its eigenvectors/eigenvalues to construct good partitions, the problem is however that the calculations are very expensive and are limited to relatively small graphs. By grouping vertices together and decompose the graph into smaller graphs, it is possible to increase both performance and memory usage. For a good introduction to spectral methods see [33].

For the coarsening step, Graclus uses a more general procedure by merging a vertex v with one of its adjacent unmarked vertex w such that it maximizes

$$\frac{e(v,w)}{w(v)} + \frac{e(v,w)}{w(w)}, \tag{2.16}$$

where $e(v,w)$ corresponds to the edge weight between v and w and $w(\cdot)$ corresponds to the vertex weight. For instance, the weight of a vertex is its degree in the normalized cut objective.

Graclus has implemented several algorithms for the initial clustering phase at the coarsest level, for instance the region growing algorithms used by METIS or a spectral method with detailed description in [10].

The refinement step of Graclus uses the kernel k-means algorithm making it more flexible in terms of choosing what objective function to optimize. It is just a matter of changing the kernel to the appropriate one. At each refinement step, the initial clusters are those induced at the previous step. The upside of using the kernel k-means algorithm is that is does not prohibit varying sizes of the partitions and is therefore more general.

[1] https://www.cs.utexas.edu/users/dml/Software/graclus.html

2.6.7 Hierarchical

Hierarchical clustering algorithms have the advantage of not having a user-defined parameter controlling the number of clusters to find as the algorithms described so far have, but at the cost of less computational efficiency. There are two types of hierarchical clustering algorithms, *agglomerative* and *divisive*. Agglomerative algorithms are bottom-up treating each data point as a single cluster and successively merge the most similar pairs of clusters until a single cluster contains all the data points. Divisive algorithms are based on a top-down approach and are less common, no such algorithm will be presented in this thesis [24].

There are various similarity metrics between clusters and some common ones are:

- *Single-link* calculates the similarity of two clusters as their most similar members.
- *Complete-link* calculates the similarity of two clusters as their most dissimilar members.
- *Average-link* calculates the similarity of two clusters as the average of all similarities between their members.

Hierarchical clustering algorithms are usually visualized as *dendrograms* and figure 2.3 shows an example using a gene expression dataset known as NCI-60 [9].

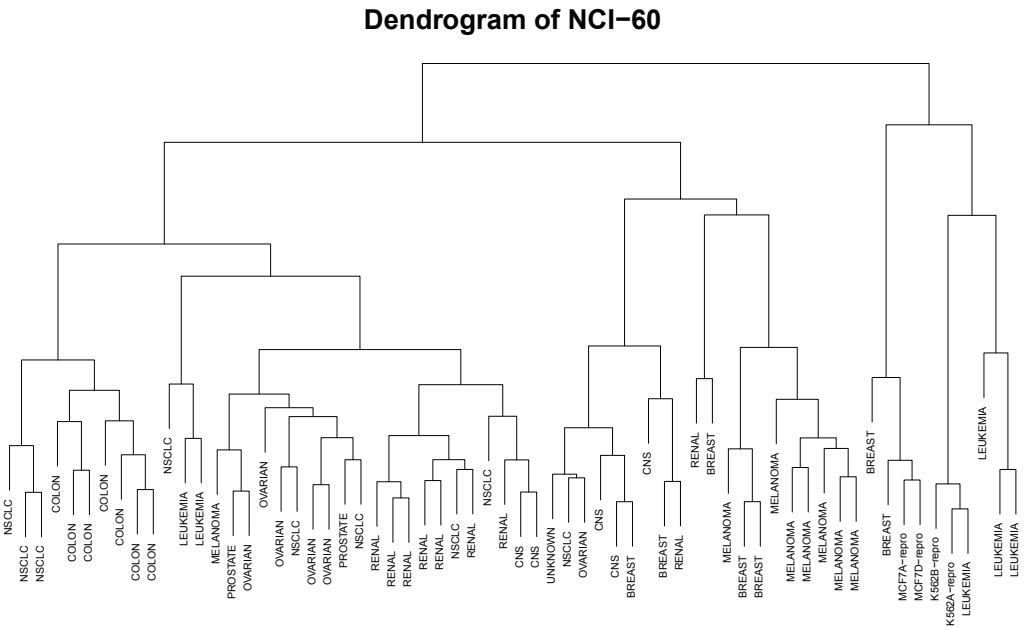


Figure 2.3: A dendrogram of the gene expression dataset NCI-60 from the National Cancer Institute (NCI) using complete-linkage.

2.6.8 Modularity Maximization

The modularity maximization algorithm proposed by Clauset et al. [8] is a hierarchical agglomerative algorithm that maximizes the modularity Q (eq. 2.11) by greedily merging clusters that produces the largest modularity score. The way the algorithm operates is to represent a cluster with a single vertex. The internal edges are represented as self-edges and edges between clusters are bundled and connect one vertex to another, i.e., connect different clusters. The algorithm is working as follows:

1. Calculate the initial values for ΔQ_{ij} and a_i.
2. Select largest ΔQ_{ij}, merge the two clusters, update ΔQ matrix, and increase Q by ΔQ_{ij} .
3. Repeat step 2 until there is only one cluster remaining.

Recall that the degree of a vertex v_i is defined as $d_i = \sum_{j=1}^{n} w_{ij}$ and m is the number of edges in the graph, then the increase in modularity by merging two clusters is defined as

$$\Delta Q_{ij} = \begin{cases} \frac{1}{2m} - \frac{d_i d_j}{4m^2} & \text{if } v_i \text{ and } v_j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

The update rules for ΔQ are the following:

$$\Delta Q'_{jl} = \begin{cases} \Delta Q_{il} + \Delta Q_{jl} & \text{if } v_l \text{ is connected to } v_i \text{ and } v_j, \\ \Delta Q_{il} - 2a_j a_l & \text{if } v_l \text{ is connected to } v_i \text{ but not } v_j, \\ \Delta Q_{jl} - 2a_i a_l & \text{if } v_l \text{ is connected to } v_j \text{ but not } v_i, \end{cases} \tag{2.18}$$

where v_j is the merged cluster, $a_i = \frac{d_i}{2m}$, and a_j updates to $a'_j = a_j + a_i$.

2.7 Cluster Validation

The procedure of evaluating the resulting clusters from a clustering algorithm is known as *cluster validity* and there are in general three approaches to go about doing so.

External criteria is one such approach which implies to evaluate the clusters by comparing it to already known structure in the data, e.g., having access to the ground truth. Since no such data has been available in this study, this approach will not be used and therefore not described in any further detail.

Internal criteria is another approach by measuring some quantitative measurement based on the vectors of the dataset itself. This is the main approach used in this study to evaluate the clustering solutions and below are the formal definitions of those validity indices used.

The third approach is *relative criteria* that builds upon the idea of evaluating by comparing results from different clustering algorithms or from the same clustering algorithm but with a different set of parameters.

Internal and relative criterion can be accomplished by comparing the *compactness*, that is, the members of a cluster should be as close to each other as possible and *separation* meaning the clusters should be well separated.

Be aware of that these methods are just indicators of the quality of the clusters and can be used as a tool to help evaluation. In the end, it is up to expert opinions to decide whether the clusters are appropriate based on the application [17, 25, 29].

2.7.1 Internal Validity Index

Many different internal validity indices have emerged through decades of research and there is no proven optimal measurement that always gives a good indication whether the clustering solution is good or bad. In this study, three validity indices were chosen that have shown good result according to the study conducted by Arbelaitz et al. [3].

2.7.1.1 Notation

Given the dataset $\mathcal{X}$ of n samples, the centroid of the whole dataset is defined as $\overline{x} = \frac{1}{n}\sum_{x_i \in \mathcal{X}} x_i$. The centroid of a cluster c_l is defined as $\overline{c_l} = \frac{1}{|c_l|}\sum_{x_i \in c_l} x_i$, $c_l \in \mathcal{C}$, where $\mathcal{C} = \{c_1, \ldots, c_k\}$ is the set of clusters and $|\mathcal{C}| = k$. And finally let the Euclidean distance between objects x_i and x_j be denoted $D_E(x_i, x_j) = \|x_i - x_j\|$.

2.7.1.2 Calinski-Harabasz

The Calinski-Harabasz index estimates the cluster cohesion based on the within-cluster variance and the cluster separation is based on the overall cluster variance from the centroid of the whole dataset. It is defined as

$$CH(\mathcal{C}) = \frac{n-k}{k-1} \frac{\sum_{c_l \in \mathcal{C}} |c_l| D_E(\overline{c_l}, \overline{x})}{\sum_{c_l \in \mathcal{C}} \sum_{x_i \in c_l} D_E(x_i, \overline{c_l})}. \tag{2.19}$$

Well-defined clusters should have low within-cluster variance and high between-cluster variance, the objective is therefore to achieve a high Calinski-Harabasz index value.

2.7.1.3 Davies-Bouldin

The Davies-Bouldin index estimates the cluster cohesion based on the distance from points within a cluster to its cluster centroid and the separation is based on the between-cluster distances. It is defined as

$$\begin{aligned} \mathrm{DB}(\mathcal{C}) &= \frac{1}{k} \sum_{c_l \in \mathcal{C}} \max_{c_m \in \mathcal{C} \setminus c_l} \frac{S(c_l) + S(c_m)}{D_E(\overline{c_l}, \overline{c_m})}, \\ \text{where } S(c_l) &= \frac{1}{|c_l|} \sum_{x_i \in c_l} D_E(x_i, \overline{c_l}). \end{aligned} \tag{2.20}$$

Because of the calculation of the within-cluster distances is in the nominator, the Davies-Bouldin index value should be aimed to be as low as possible. There is also an alternative variation of the Davis-Bouldin index which is defined as

$$DB^*(\mathcal{C}) = \frac{1}{k} \sum_{c_l \in \mathcal{C}} \frac{\max_{c_m \in \mathcal{C} \setminus c_l} S(c_l) + S(c_m)}{\min_{c_m \in \mathcal{C} \setminus c_l} D_E(\overline{c_l}, \overline{c_m})}. \tag{2.21}$$

This has the property of augmenting the absolute worst possible combinations where the ratio is between the maximum within-cluster distances and the least between-cluster distances.

2.7.1.4 Silhouette

The silhouette index estimates the cluster cohesion based on the distance between all points in the same cluster and the cluster separation by computing the nearest neighbour distance. It is defined as

$$\mathrm{Sil}(\mathcal{C}) = \frac{1}{n} \sum_{c_l \in \mathcal{C}} \sum_{x_i \in c_l} \frac{b(x_i, c_l) - a(x_i, c_l)}{\max(a(x_i, c_l), b(x_i, c_l))}, \tag{2.22}$$

where

$$\begin{aligned} a(x_i, c_l) &= \frac{1}{|c_l|} \sum_{x_j \in c_l} D_E(x_i, x_j), \\ b(x_i, c_l) &= \min_{c_m \in \mathcal{C} \setminus c_l} \frac{1}{|c_m|} \sum_{x_j \in c_m} D_E(x_i, x_j). \end{aligned}$$

Given the silhouette value for a single point, $\frac{b(x_i,c_l)-a(x_i,c_l)}{\max(a(x_i,c_l),b(x_i,c_l))} \in [-1,1]$, $a(x_i,c_l)$ measures the average distance from the point x_i to other points in its cluster c_l and $b(x_i,c_l)$ measures the average distance from point x_i to points in a different cluster, minimized over clusters. This can be interpreted as an increasing value indicates that the point x_i matches poorly with other clusters and is a good fit with its own cluster. A low value of the silhouette index indicates that there are too few or too many clusters.

3 Method

In the preliminary study phase it was possible to find information about previous studies that are comparable to what is being done in this study. Aysu Ezen-Can et al. [12] have used unsupervised modeling for understanding discussion forums for Massive Open Online Courses (MOOCs). The aforementioned study laid the groundwork for how the experiments were conducted in this study.

3.1 Data Collection

The data used for the analysis was taken from the politics subreddit from Reddit which is in the top 100 largest[1] subreddits with over 3 million subscribers. The data collection contained about 900,000 threads, 22.5 million user comments, and 800,00 unique users contributing either by submitting at least one comment or by creating at least one thread. The data was stored in a MySQL database.

The following desirable data about threads was not present in the data collection:

- Thread title
- Thread body
- Creator's username
- Number of comments
- Submission date
- Score
- Gold

Due to the limited number of requests per second with the Reddit *application programming interface* (API) Wrapper PRAW[2], the *Scrapy*[3] 1.0.5 framework was used to develop a web spider using Python 2.7.6 to extract the information about threads from the Reddit website.

[1] http://redditlist.com/
[2] https://praw.readthedocs.io/en/stable/
[3] http://scrapy.org/

3.2 Data Processing

Every comment contained other side information[1] and not all the data was of interest and was therefore filtered out. Below are the data contained for every comment after filtration of redundant information.

- The author's username (author)
- The text content (body)
- The submission date (created_utc)
- # of down votes (downs)
- # of up votes (ups)
- Total score (score)
- Gold count (gilded)
- The unique thread identifier where the comment is located (link_id)
- Unique identifier (name)
- Identifier of what the comment refers to, either a comment or a thread (parent_id)

The algorithms require the data to be in either a vector space model or a graph. To accomplish this, a pipeline was built with various text preprocessing operations and every user comment was processed by the pipeline. The pipeline consisted of 6 operations operating in the following order:

1. Remove all the Uniform Resource Locators (URLs).
2. Remove all punctuations given by the Python string library.
3. Remove all numbers.
4. Transform everything to lower case.
5. Remove stop words given by the *Natural Language Toolkit*[2] (NLTK) for the English language
6. Normalize all words to their stem using the Snowball (Porter2) stemmer from NLTK.

Figure 3.2 shows the pipeline.

[1] https://github.com/reddit/reddit/wiki/JSON
[2] http://www.nltk.org/

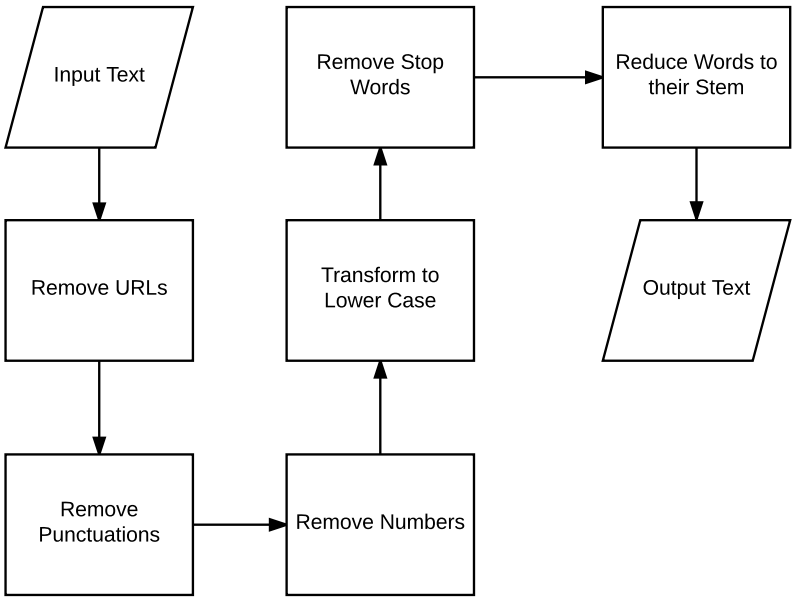


Figure 3.1: The text preprocessing pipeline used to process all the text content.

3.3 Text Transformation

In order to cluster comments from a thread, the document-term frequency matrix has to be constructed where every comment is considered a document. The *scikit-learn* 0.17.1 framework [28] provides functionality to transform text into the two representations presented in section 2.3 using their *CountVectorizer* and *TfidfVectorizer*.

The non-exhaustive overlapping k-means algorithm can use the document-term frequency matrix directly. Apart from it, we used the Graclus software, the METIS software, and the NEO-*k*-means on graphs. The NEO-*k*-means on graphs was acquired by requesting it from Joyce Jiyoung Whang [34]. These algorithms are expected to work with a graph representation and the document-term frequency matrix is a vector space model. To transform the matrix into a graph using igraph[1] 0.7.1, every row is considered a vertex. The graph is generated by computing the pairwise cosine distance (eq. 2.6) between all rows and then specify a threshold at which the distance has to be below in order to add an edge between two vertices. Only the largest connected component of the graph acted as input to the clustering algorithms. For the algorithms to get reasonable execution time it is important that the graph is *sparse*, i.e., $|\mathcal{E}| = O(|\mathcal{V}|)$ [13].

3.4 Experimentation

Performance. In order to answer, *How do the algorithms compare in terms of execution time?*, this experiment tests the performance on a large scale with threads of various sizes using all the algorithms.

Cluster Sizes. This experiment aims to gain insight in how the cluster sizes change with the number of samples in the data and with different clustering algorithms.

Edge Density. By varying the average number of edges incident to a vertex, the graph becomes more or less connected. This experiment provides insight in how this may affect both

[1] http://igraph.org/python/

the modularity maximization estimate and the clustering solution. To perform this experiment, we defined a *low degree* graph as $\frac{|\mathcal{E}|}{|\mathcal{V}|} \in [4, 8]$ and a *high degree* graph as $\frac{|\mathcal{E}|}{|\mathcal{V}|} \in [12, 16]$.

Edge Weight. By using edge weights, some measurement between two samples is encoded into the graph and this experiment inspects how the modularity maximization and the graph clustering algorithms get affected by it. We used the edge weight to correspond to the cosine similarity (eq. 2.5) between two samples.

Text Transformer. The intent of this experiment is to see the impact of using term frequency and term frequency-inverse document frequency.

Overlap. The NEO-k-means algorithm can be tuned to generate clusters with overlap and this experiment aims to find how this changes the objective values and if the kind of content that overlaps is reasonable.

Modularity Maximization Estimate. All the algorithms are parametrized by the number of clusters to find and this experiment aims to provide insight in how good the results are when using the estimated optimal cluster count found by the modularity maximization algorithm. This is done by using more and less number of clusters than estimated and determine if some sort of sweet spot is found.

Structure. To answer, *Can the chosen clustering algorithms be used to find structure in textual content?*, the content of the clusters have to be analysed and this experiment clusters a few manually chosen threads to be studied more extensively with and without overlap.

3.5 Evaluation

The objective of clustering is to discover present patterns in a data collection and this means searching for clusters whose members are similar to each other and different clusters are well separated.

There are in general three different evaluation criterion and those are the following [17]:

- External criteria base the quality on already known information about the dataset.
- Internal criteria measure the quality by quantify the compactness within clusters and the separation of different clusters.
- Relative criteria compares results from different clustering algorithms or results from the same clustering algorithm with distinct set of parameters.

All the experiments aside from the one analysing the structures used internal and relative criterion since no ground truth data was accessible. The objective functions used are those described in section 2.7.1. To determine the structures found, visualization and the text content was the key tools to see if it make sense to a human being. Analysing the content and using visualization is however not practical on a large scale so the assumption that the parameters generalize well was made and that the results found on just a few examples give atleast some insight in what the algorithms can find.

4 Results

In this chapter the results generated by the experiments will be presented. It begins by presenting results showing how the behaviour of the algorithms changes with different parameters and how certain parameters affects the clustering results. After that a few clustering results from hand picked threads are presented to see what structures can be found.

4.1 Algorithmic Behaviour

All the experiments were performed in VirtualBox with Linux Mint 17.1 on a laptop with an Intel Core i7-6700HQ CPU and 4GB RAM.

4.1.1 Performance

The time includes only the time it took to run the clustering part and not constructing the vector space model or graph. In the case for Graclus, the time includes the time it took to read the clustering solution from file which was generated by the Graclus software.

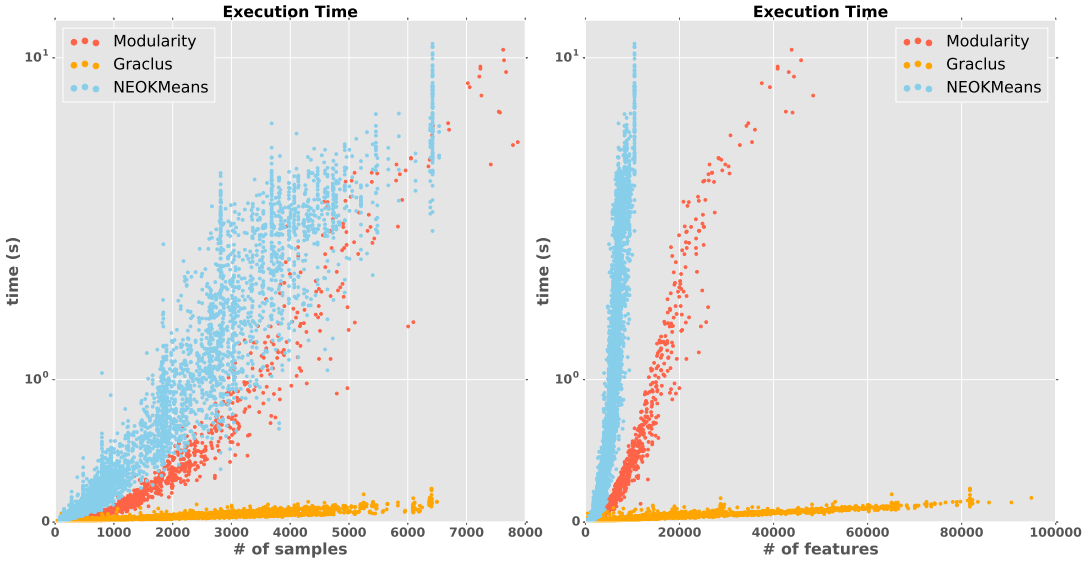


Figure 4.1: Left: Comparison of the performance in terms of execution time in relation to the number of samples. The sample size corresponds to the number of vertices in the graph for modularity maximization and Graclus. Right: The execution time in relation to the number of features which corresponds to the number of edges for modularity maximization and Graclus.

4.1.2 Modularity Maximization

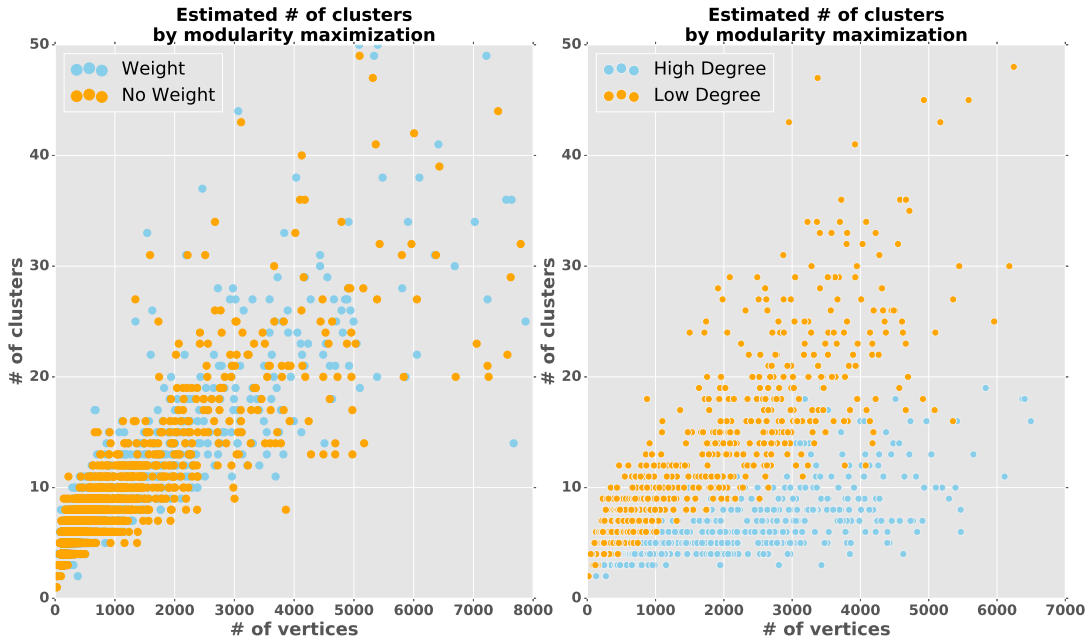


Figure 4.2: Shows the number of clusters estimated by modularity maximization in relation to the number of vertices in the graph. Left: The parameter deciding whether to use edge weights was varied. The graphs were all of low degree. Right: The parameter whether to use high or low degree was varied. The graphs contained edge weights.

4.1.3 Cluster Sizes

In the following two diagrams, the number of clusters were estimated by the modularity maximization algorithm.

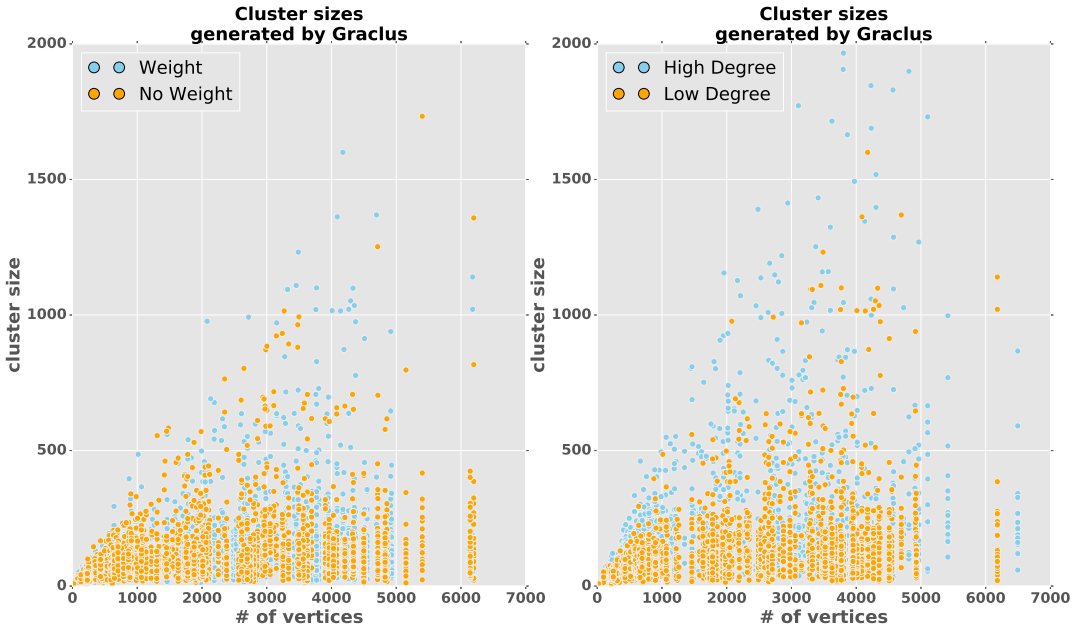


Figure 4.3: Shows how the cluster sizes changes with increasing number of vertices in the graph using Graclus. Left: Varying the weight parameter. Right: Varying the degree parameter.

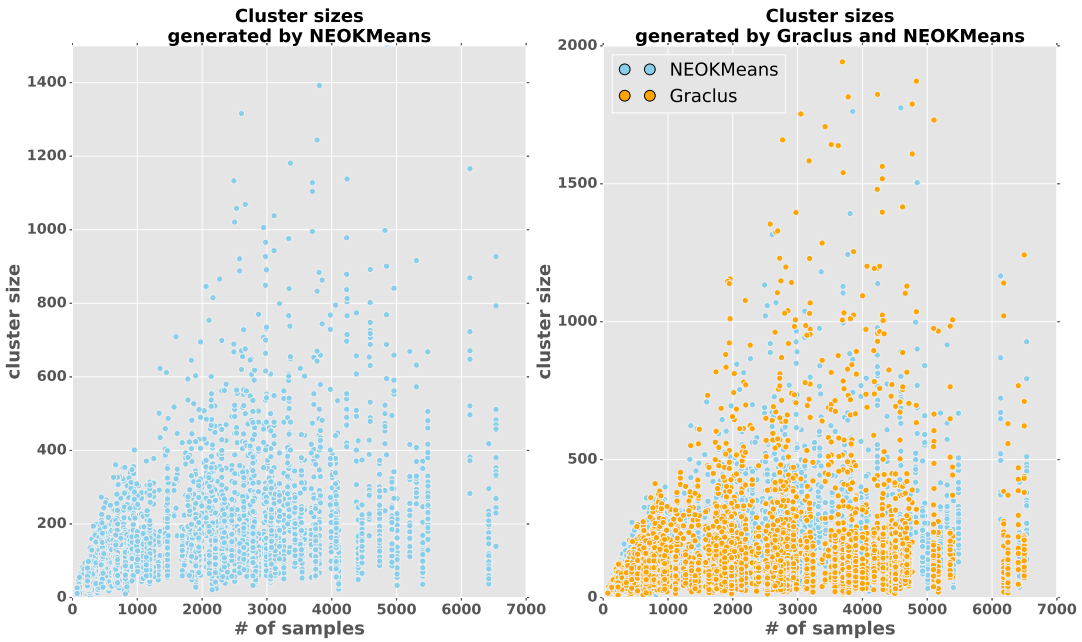


Figure 4.4: NEO-k-means having $\alpha = 0$ and $\beta = 0$. Left: Shows how the cluster sizes changes with increasing number of samples using NEO-k-means. Right: Compares the cluster sizes generated by NEO-k-means and Graclus. The graphs have varying values of the degree and weight parameters.

In the following diagrams, (High) means the objective should be aimed to be as high as possible and (Low) the opposite. Equal coloured lines means the result was generated from the same data but with a varying parameter. The number of clusters have been increased and decreased from the modularity estimate.

4.1.4 Edge Density

This experiment used the term frequency-inverse document frequency transformer and edge weights. The number of samples for each result are the following:

155	6095	281	935	2720	472

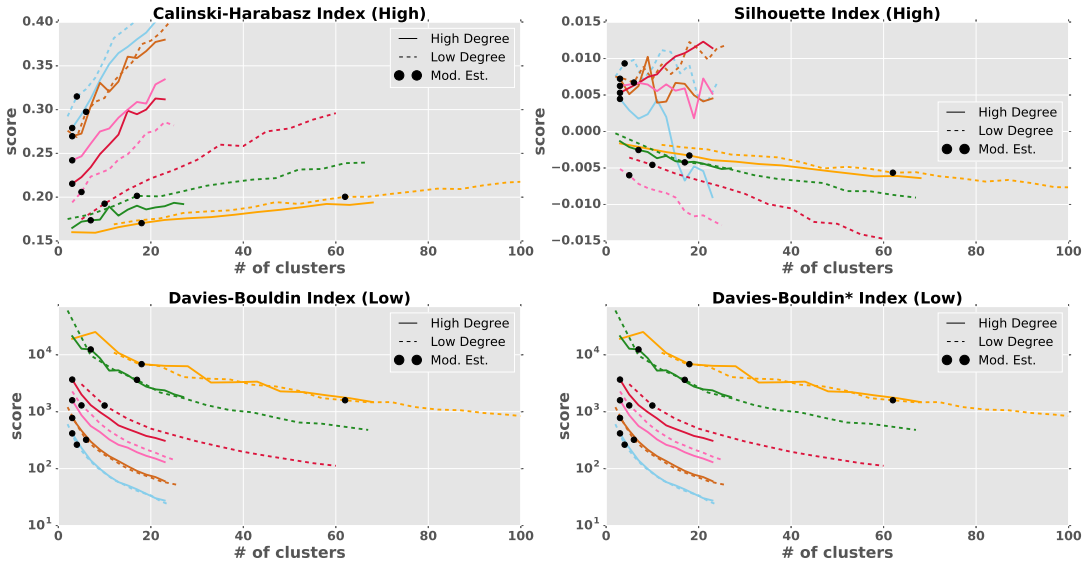


Figure 4.5: A comparison of the objective functions varying the number of edges in the graph using Graclus on threads of various sizes.

4.1.5 Edge Weight

This experiment used the term frequency-inverse document frequency transformer and high degree graphs. The number of samples for each result are the following:

392	498	708	1402	2664	4458

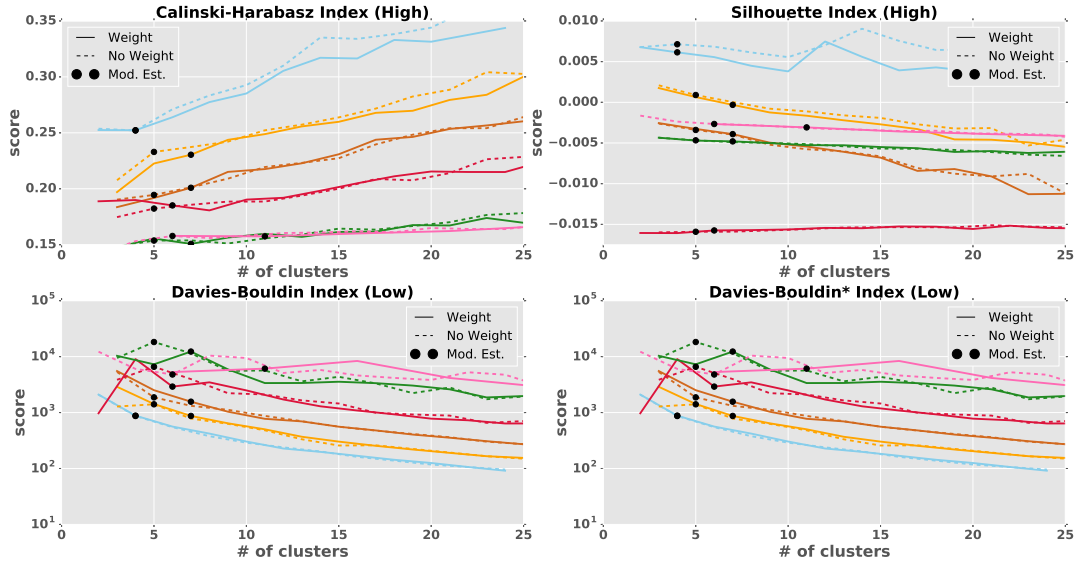


Figure 4.6: A comparison of the objective functions varying the weight parameter using Graclus on threads of various sizes.

4.1.6 Text Transformer

Term frequency and term frequency-inverse document frequency are denoted tf and tfidf respectively. This experiment used low degree graphs and edge weights. The number of samples for each result are the following:

155	3317	278	906	1143	389

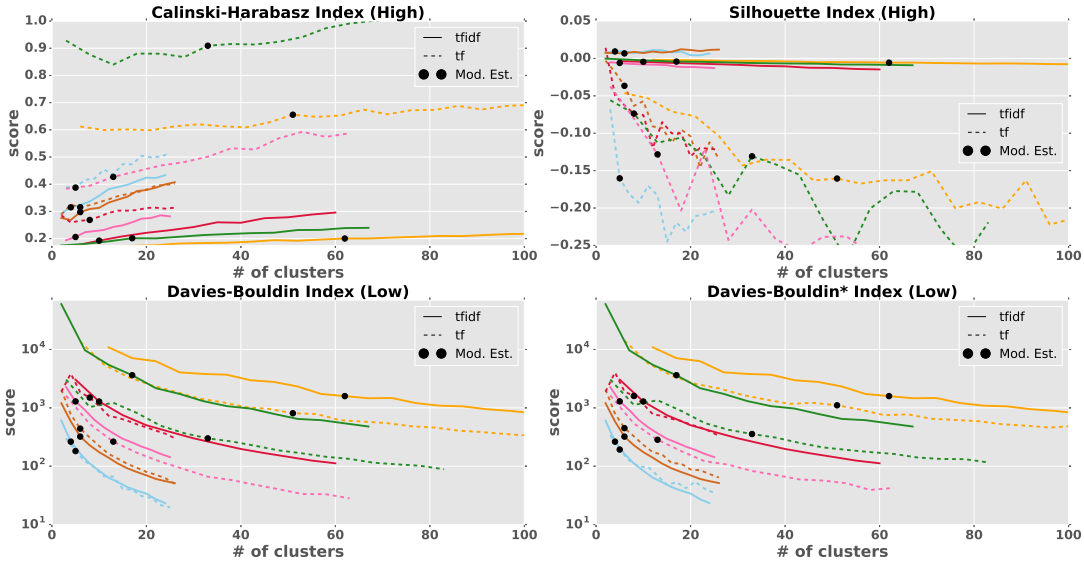


Figure 4.7: A comparison of the objective functions varying the text transformer using Graclus on threads of various sizes.

In this experiment, the number of samples for each result are the following:

155	6422	281	945	2816	478

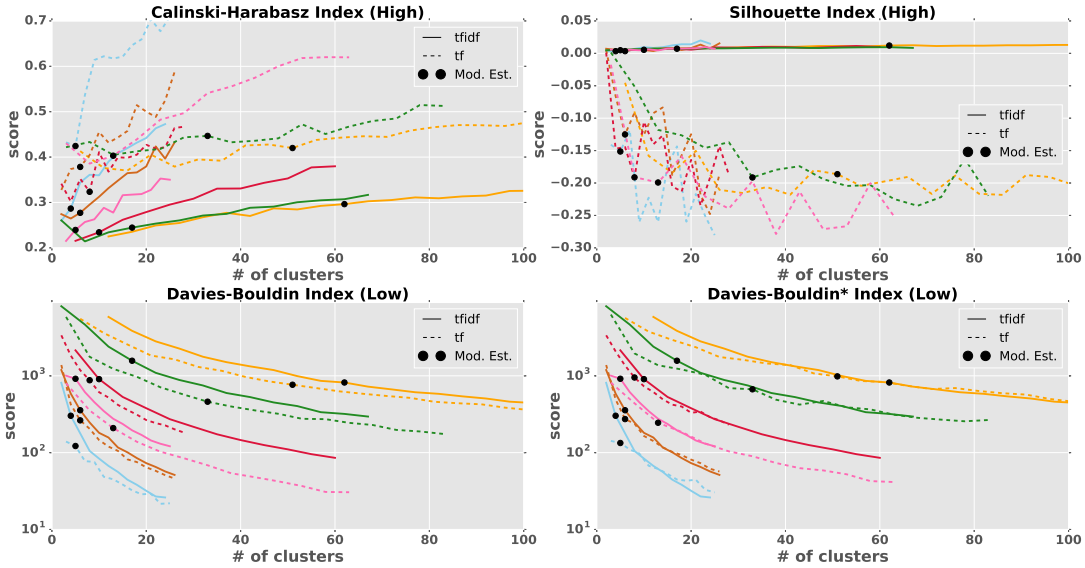


Figure 4.8: A comparison of the objective functions varying the text transformer using NEO-k-means with $\alpha = 0$ and $\beta = 0$ on threads of various sizes.

In this experiment, the number of samples for each result are the following:

151	419	1175	261	607	874

.

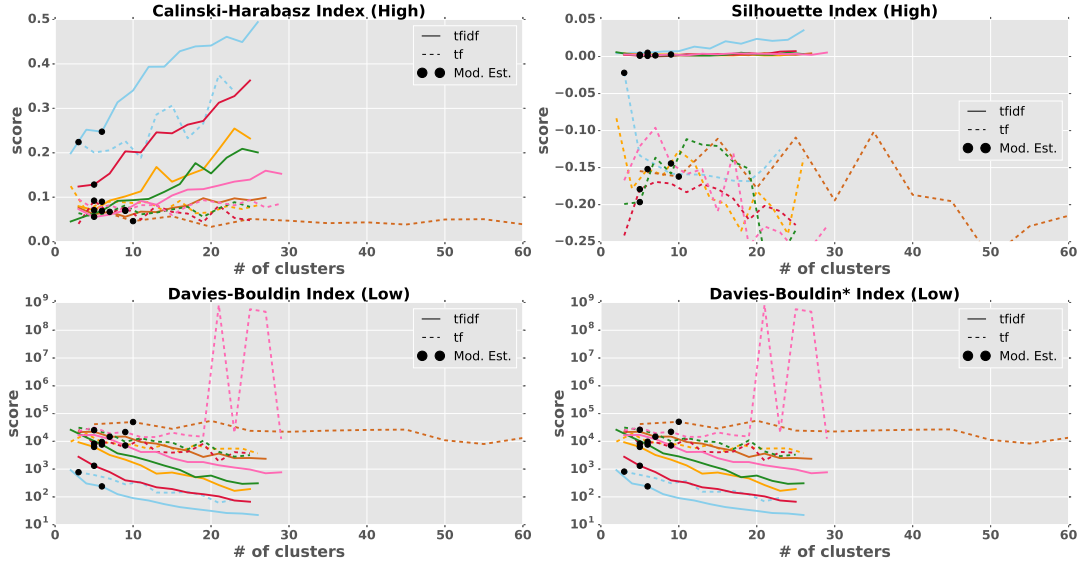


Figure 4.9: A comparison of the objective functions varying the text transformer using NEO-k-means with $\alpha > 0$ and $\beta = 0$ on threads of various sizes. The alpha values were chosen according to the first strategy by [34] with $\delta = 1.25$.

4.1.7 Overlap

This experiment used the term frequency-inverse document frequency transformer. The number of samples for each result are the following:

232	529	658	972	1670	3023	3319	5484

.

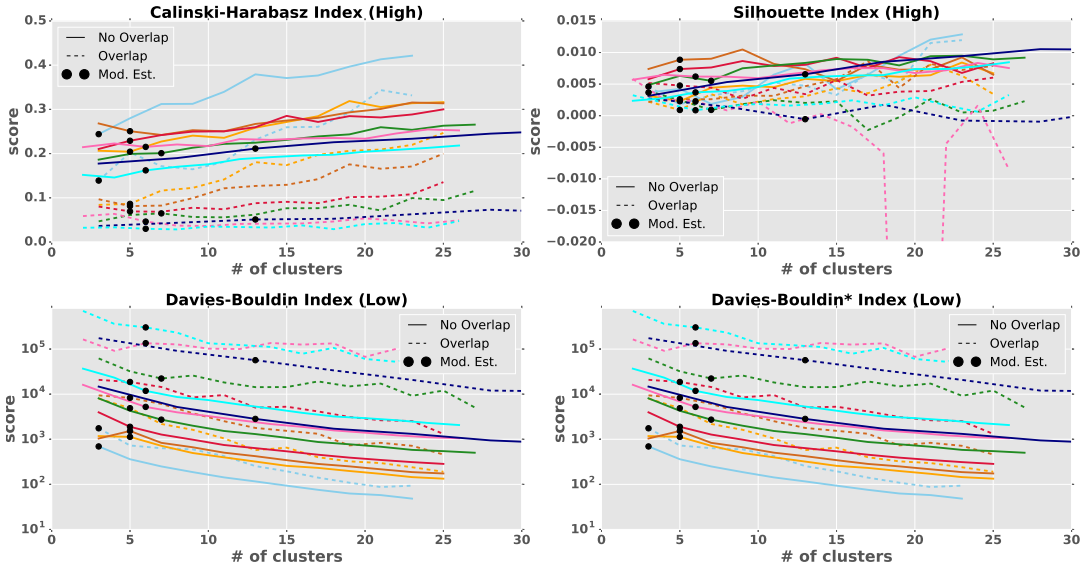


Figure 4.10: A comparison of the objective functions using NEO-k-means with overlap, i.e., $\alpha > 0$ and without, i.e., $\alpha = 0$ and $\beta = 0$ on threads of various sizes. The alpha values were chosen according to the first strategy by [34] with $\delta = 1.25$.

4.1.8 Modularity Maximization Estimate

The following result used the term frequency transformer, edge weights, and low degree graphs. The number of samples for each experiment are the following:

176 | 7230 | 305 | 1101 | 3056 | 517.

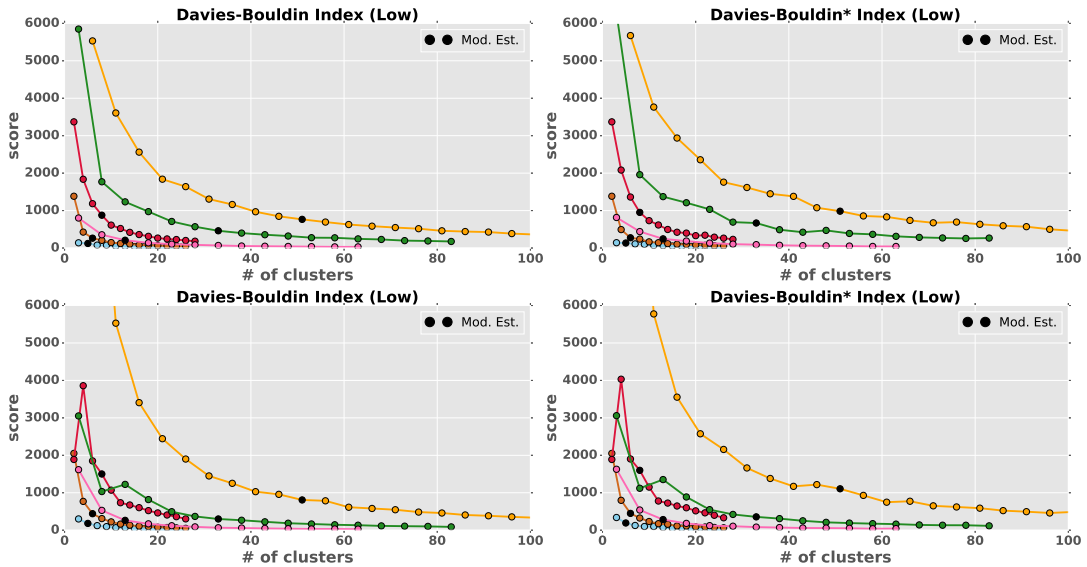


Figure 4.11: A look at how good the modularity maximization estimate is compared to other cluster counts. Top: Generated by NEO-k-means. Bottom: Generated by Graclus.

4.2 Clustering Solutions

In this section, the clustering solutions of two manually picked threads will be more thoroughly examined. The titles of the threads are "Elementary school mass shooting took place in Kindergarten classroom. At least 27 dead, 14 children."[1] with over 14,000 comments and "Marijuana Has Won The War On Drugs"[2] with around 350 comments.

In the following tables, the key terms refer to the 5 most frequently occurring terms and LDA terms are terms extracted by Latent Dirichlet Allocation (LDA) [6], a method for topic extraction. The sample comments shown are all picked out by NEO-k-means with $\alpha = 0$ and $\beta = 0$ and the comments chosen have been limited to around 15-20 words. For every cluster centroid, the sample with the least cosine distance was picked. The number of clusters have been estimated by the modularity maximization algorithm for all the examples.

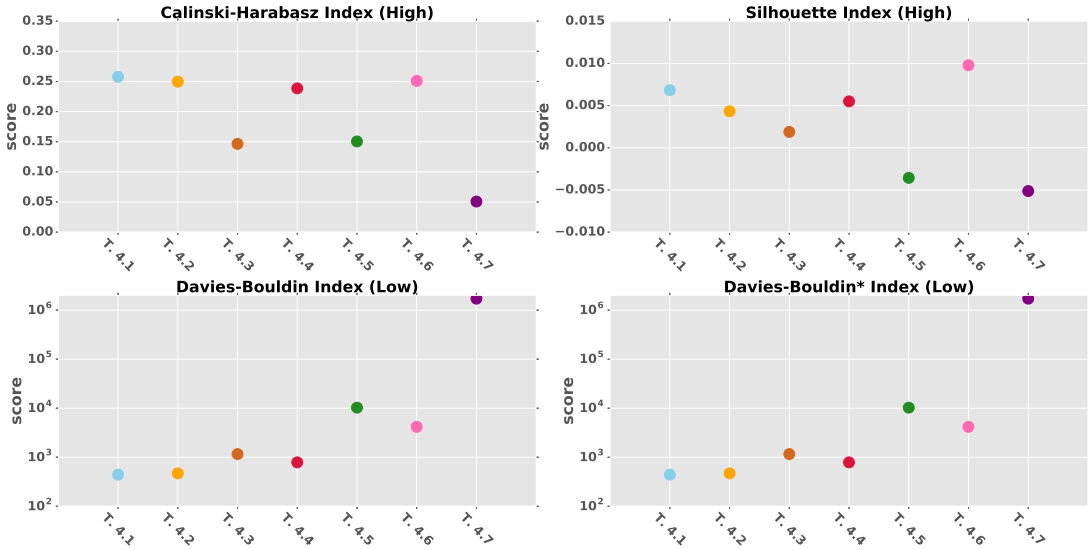


Figure 4.12: A comparison of the objective functions of the clustering solutions. Table is denoted T. and tables 4.1 - 4.4 are referring to clustering solutions from the thread about drugs on war. Tables 4.5-4.7 are referring to clustering solutions from the thread about the school shooting.

[1] https://www.reddit.com/r/politics/comments/14uoel
[2] https://www.reddit.com/r/politics/comments/1boemk

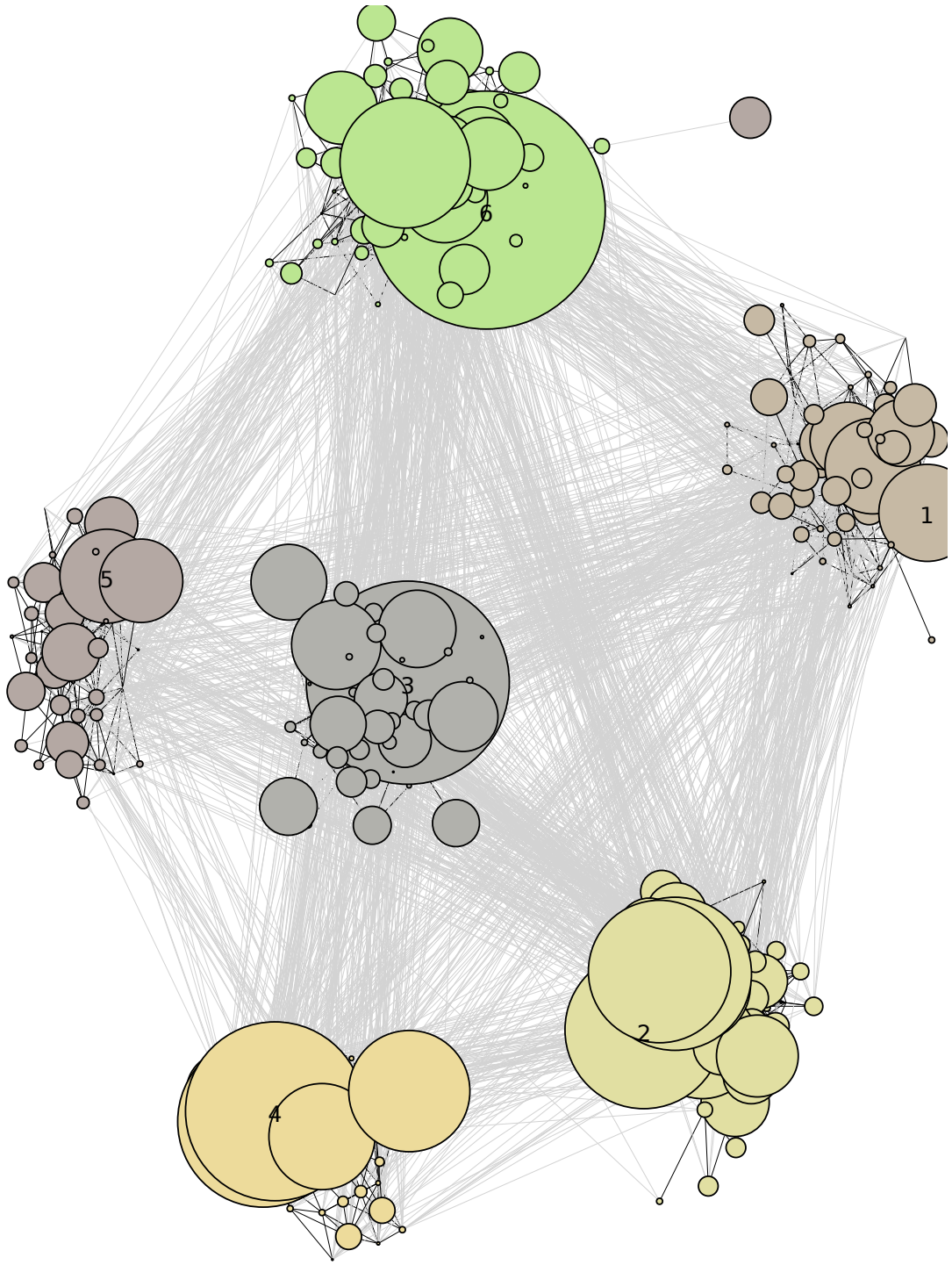


Figure 4.13: A graph representation of the discussion about marijuana and the war on drugs where the clusters have been found by Graclus. The graph have low edge density, edge weights, and the size of a vertex corresponds to the number of words in the comment. Black edges are edges within clusters and gray edges are edges between clusters.

Table 4.1: Marijuana has won the war on drugs

Cluster (size)	Key Terms	LDA Terms	Samples
1 (56)	would, legal, pot, make, cartel	that, yeah, would, way, time	1. "Time to break out the "MISSION ACCOMPLISHED" banners, I guess." 2. "That's not much of a distinction." 3. "The same way I would feel about someone running a speakeasy during prohibition." 4. "Yeah what's the deal with that? "
2 (65)	drug, addict, war, use, legal	drug, war, win, substanc, bad	1. ""The continued banning of addictive, non-social substances (i.e not cannabis) is not a bad thing."" 2. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/" 3. "While that is certainly a persuasive argument for never arrested users, how do you feel about arresting dealers?" 4. "Cannabis and hemp will be legal, it's not a matter of if but when. "
3 (45)	marijuana, state, legal, think, like	marijuana, state, still, problem, illeg	1. "Marijuana is not a drug. It's a plant!" 2. "The states isn't very good at winning wars." 3. "It would be a meme to spread, Truman surrendered to cannabis why can't we. lol. " 4. "The problem is a lack thereof. Seriously, *worse*?"
4 (38)	prison, peopl, go, drug, im	im, still, number, though, sure	1. "Neither, look it up. And I've done them all too." 2. "I did a few months ago. They change their numbers all the time though. " 3. "Tell that to the millions of people still incarcerated for Pot charges" 4. "Aussie here. I'm still doubting if it'll happen here in my lifetime :("
5 (41)	cop, say, law, get, dont	cop, friend, right, name, id	1. "I'd just like to say - greatest title of any article/post ever." 2. "They shouldn't. If they are not educated in the topic, the should have no right to speak, same goes for men." 3. "Cops are never your friend, but sometimes your friends are cops, which is totally different." 4. "What an awful name for an article. Just not true"
6 (81)	peopl, weed, fuck, think, drug	fuck, mean, tomato, compromis, weed	1. "Sincere enough to be a politician. " 2. "I think felons can't vote in most places. Correct me if I am wrong." 3. "Does someone have a restrictive monopoly on tomatoes?" 4. "You say democracy means compromise. Fuck compromise and fuck democracy."

Sample comments from the clusters shown in figure 4.13.

Table 4.2: Marijuana has won the war on drugs

Cluster (size)	Key Terms	LDA Terms	Samples
1 (27)	keep, way, bong, champion, ive	way, weve, without, phrase, run	1. "I did a few months ago. They change their numbers all the time though. " 2. "But some do. Either way, that's not a reason for banning tomatoes." 3. "I don't know what I'd do without ketchup." 4. "Some of our most prominent families have the roots of their fortunes rooted in rum running. " 5. "Yo tell Rayray I said supppppp"
2 (54)	state, marijuana, legal, one, drug	marijuana, state, win, sure, one	1. "I for one would like to welcome our new overlord, Marijuana. All hail marijuana. " 2. "The states isn't very good at winning wars." 3. "i'm sure someone at monsanto is working on that." 4. "Who's the one to stop them from talking? " 5. "Why does this article say that California is the biggest state in the nation?"
3 (47)	drug, addict, war, use, harm	drug, war, win, substanc, continu	1. "Drugs are bad, m'kay?" 2. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/" 3. "Yeah what's the deal with that? " 4. ""The continued banning of addictive, non-social substances (i.e not cannabis) is not a bad thing."" 5. "While Marinol, the more legal "substitute" is far more dangerous and can result in overdose."
4 (60)	would, legal, make, cartel, say	would, still, lifetim, point, legal	1. "It would probably just be illogical." 2. "No it hasn't. Still illegal.. " 3. "The point wasn't the cost of my pot habit, it was the cost of prohibition. " 4. "Cannabis and hemp will be legal, it's not a matter of if but when. " 5. "Kind of like how bootleggers are still a huge problem. Oh wait."
5 (92)	peopl, prison, drug, think, go	friend, vote, right, never, cop	1. "A better analogy might be Philip-Morris, who people hate but who have not been arrested for their actions." 2. "That's why women shouldn't have right to vote" 3. "It's very upsetting if you're a decent human being. But yes... even more so for dog lovers. :(" 4. "Cops are never your friend. just remember that." 5. "Sad thing is the money will be spent the same day it's cut."
6 (46)	fuck, make, long, weed, give	fuck, tomato, websit, true, link	1. "Wow your really smart." 2. "According to a link in the article, Obama invented the smokers' game Chicago." 3. "Unarmed plant - 1, largest military/ paramilitary industrial complex in the world - 0" 4. "viva marijuana! long live pot!" 5. "What in the fuck is wrong with this website? "

Sample comments from clusters generated by NEO-k-means with $\alpha = 0$ and $\beta = 0$.

Table 4.3: Marijuana has won the war on drugs

Cluster (size)	Key Terms	LDA Terms	Samples
1 (85)	drug, marijuana, war, addict, legal	drug, war, marijuana, win, name	1. "They misspelled his name. It's on his name plate as Kerlikowske. " 2. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/" 3. "Marijuana is not a drug. It's a plant!" 4. "Well, obviously if it seems implausible to you it cannot be right, what was I thinking."
2 (82)	drug, addict, prison, use, legal	win, war, drug, friend, substanc	1. "The point wasn't the cost of my pot habit, it was the cost of prohibition. " 2. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/" 3. "Cops are never your friend, but sometimes your friends are cops, which is totally different." 4. "The punishments for having this substance are worse than what the substance can do to you even in the extreme."
3 (106)	peopl, drug, dont, legal, would	that, dont, peopl, shouldnt, still	1. "They shouldn't. If they are not educated in the topic, the should have no right to speak, same goes for men." 2. "But some do. Either way, that's not a reason for banning tomatoes." 3. "Don't forget all the other drugs. " 4. "Tell that to the millions of people still incarcerated for Pot charges"
4 (112)	drug, peopl, prison, legal, war	drug, war, win, make, longer	1. "Technically, it's not longer a drug and so the war continues." 2. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/" 3. "It's not rambling if you have a point to make." 4. "Cops are never your friend. just remember that."
5 (54)	drug, addict, war, use, problem	war, drug, win, sound, plant	1. "While that is certainly a persuasive argument for never arrested users, how do you feel about arresting dealers?" 2. ""The continued banning of addictive, non-social substances (i.e not cannabis) is not a bad thing."" 3. "Kind of like how bootleggers are still a huge problem. Oh wait." 4. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/"
6 (74)	would, legal, drug, peopl, marijuana	fuck, would, tomato, yeah, im	1. "Why does this article say that California is the biggest state in the nation?" 2. "Didn't think it would be possible in my lifetime. Yay" 3. "What in the fuck is wrong with this website? " 4. "haven't you heard of the killer tomatoes? "

Sample comments from clusters generated by NEO-k-means with $\alpha = 0.57362$ and $\beta = 0$. The alpha value was chosen according to the first strategy by [34] with $\delta = 1.25$.

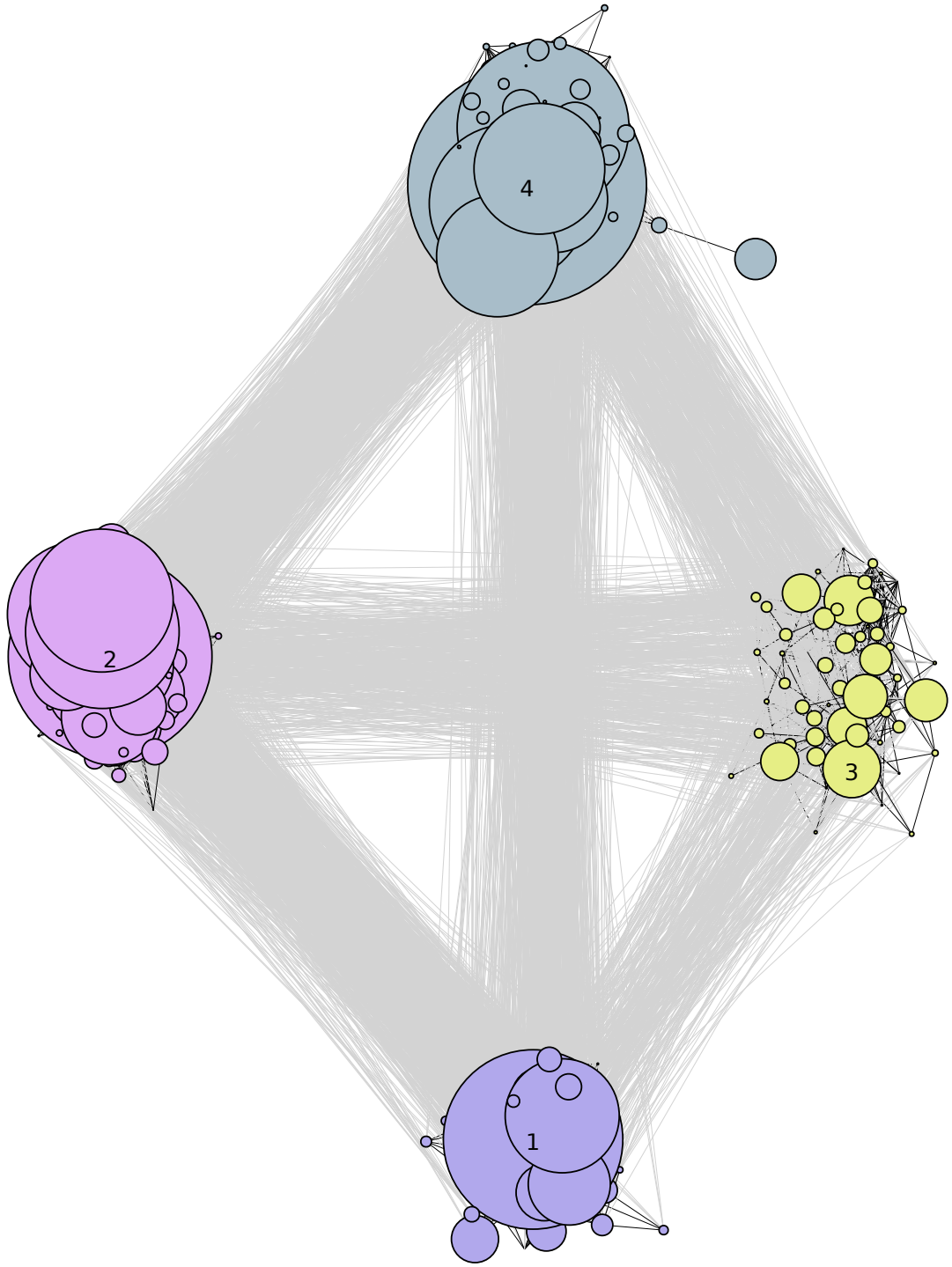


Figure 4.14: A graph representation of the discussion about marijuana and the war on drugs where the clusters have been found by Graclus. The graph have high edge density, edge weights, and the size of a vertex corresponds to the number of words in the comment. Black edges are edges within clusters and gray edges are edges between clusters.

Table 4.4: Marijuana has won the war on drugs

Cluster (size)	Key Terms	LDA Terms	Samples
1 (68)	drug, war, marijuana, win, like	drug, war, win, marijuana, plant	1. "i'm old enough to remember when we lost the war on poverty. " 2. "'Drugs Win Drug War' http://imageshack.us/f/242/drugwarmv5.jpg/" 3. "are you aware of the fact that marijuana overdose is impossible?" 4. "The point wasn't the cost of my pot habit, it was the cost of prohibition. " 5. "Drugs are bad, m'kay?" 6. "I hope he dies of cancer without access to medical marijuana"
2 (105)	legal, would, drug, use, alcohol	would, still, legal, probabl, cannabi	1. "No it hasn't. Still illegal.. " 2. "I would love to try my hand at doing an indoor grow." 3. "You need to assess your approach to the world." 4. "Smoke weed, probably. " 5. ""Support for legalization is at an all time high"" 6. "This sounds like less of a cop thing and more of a sexism thing."
3 (59)	fuck, im, articl, say, even	fuck, titl, articl, name, websit	1. "Is there a link after the jump? I fucking hate businessinsider...." 2. "It's very upsetting if you're a decent human being. But yes... even more so for dog lovers. :(" 3. "I'd just like to say - greatest title of any article/post ever." 4. "I did a few months ago. They change their numbers all the time though. " 5. "Does someone have a restrictive monopoly on tomatoes?" 6. ">Democracy means compromise No it doesn't. Democracy is a tyranny of the majority."
4 (94)	peopl, go, dont, prison, get	vote, right, that, yeah, tomato	1. "They shouldn't. If they are not educated in the topic, the should have no right to speak, same goes for men." 2. "Cops are never your friend, but sometimes your friends are cops, which is totally different." 3. "51% of people cannot agree on anything without compromising with each other, in some way." 4. "I vote at least twice a year and go to quarterly city council meetings, because I can!" 5. "Pretty sure people prefer weed to tomatoes.." 6. "So because we've created a monster, we should keep doing the same stupid shit? "

Sample comments from the clusters shown in figure 4.14.

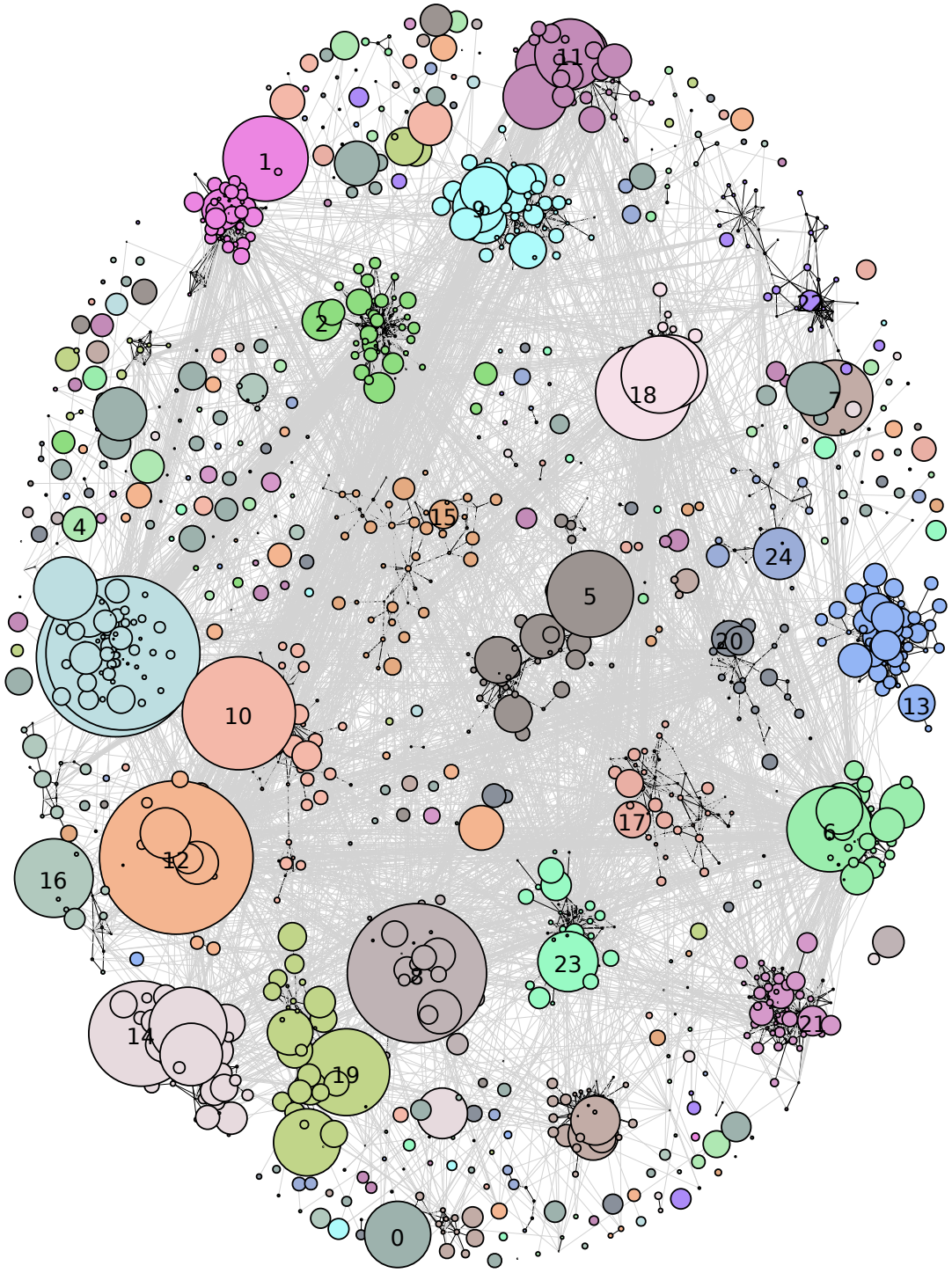


Figure 4.15: A graph representation of the discussion about a school shooting where the clusters have been found by Graclus. The graph have high edge density, no edge weights, and the size of a vertex corresponds to the number of words in the comment. Black edges are edges within clusters and gray edges are edges between clusters. The graph does not show every single vertex but rather a subset from each cluster.

Table 4.5: School shooting 2012 in America.

Cluster (size)	Key Terms	LDA Terms	Samples
0 (55)	like, make, gun, dont, go	boxcutt, headcount, risen, stimmt, guunss	1. "Woppidy doo Basil, what does it mean?" 2. "Guns, stahpit. Guunss. STAPH." 3. "the headcount has risen to 20 children." 4. ">Metal Health CUM ON FEEL THE NOIZE"
1 (74)	game, video, violent, blame, peopl	game, video, palin, sarah, blame	1. "i guess CoD could be crossed off the list of games to play" 2. "How long before the media blames Sarah Palin (again)?" 3. "I think it was violent video games."
3 (329)	mental, health, gun, peopl, issu	health, mental, care, issu, gun	1. "It would help if mental health services were as easily accessible as guns." 2. "We have state mental hospitals." 3. "How about gun control *and* mental health?"
6 (553)	peopl, kill, gun, dont, use	kill, peopl, gun, knife, dont	1. " They killed themselves, guns kill other people. " 2. "Because it's just as easy to kill someone with a knife?" 3. "A guns only purpose is to kill or maim. A knife has more purposes than to harm. "
11 (527)	gun, illeg, crimin, peopl, get	illeg, gun, crimin, buy, state	1. "Do you know where to buy a gun illegally? " 2. "Only people with guns over there are the criminals. So what does that solve?" 3. "You can go buy a gun from a different member of the gang that provides the weed. "
17 (341)	dead, mother, kill, shooter, school	mother, brother, dead, shooter, stole	1. "Update - 18 children dead." 2. "He killed his mother. She was a teacher at the school." 3. "Shooters brother [source](http://www.foxnews.com/us/2012/12/14/police-respond-to-shooting-at-connecticut-elementary-school/)"
18 (122)	lanza, ryan, adam, brother, shooter	lanza, ryan, adam, brother, name	1. "How old was this Adam Lanza kid? Edit: Jesus fuck, he was 20.... why" 2. "So it was a Ryan Lanza just not the one they linked?" 3. "Adam Lanza is the shooter not his brother Ryan Lanza"

Sample comments from a few clusters generated by Graclus from the thread about a school shooting 2012 in America. 24 clusters were found in total corresponding to those in fig. 4.15 and cluster number 0 contains samples outside any cluster. Those can be considered outliers but were lost in the transformation from vector space model to graph, i.e., vertices not in the largest connected component.

Table 4.6: School shooting 2012 in America.

Cluster (size)	Key Terms	LDA Terms	Samples
7 (432)	gun, control, peopl, one, like	control, gun, talk, time, america	1. "This is why we need gun control " 2. "So is now the time to talk about gun control?" 3. "Murder is pretty tightly controlled in this country. We have pretty stiff penalties too." 4. "If it helps to get some real gun control in the US then I fully support it. " 5. "America is collapsing on itself it appears.."
10 (394)	news, media, like, stori, peopl	news, reddit, media, agenda, fox	1. "The media will never change..." 2. "These are the types of stories we should be focusing on after such a tragedy. " 3. "How in the world can this be down voted? Explain!" 4. "this isn't politics... this shouldn't be politics. why is it in /r/politics?" 5. "And Fox News blames this on Obama in 3 ... 2 ... 1 ..."
12 (210)	drug, gun, illeg, war, peopl	drug, war, noth, illeg, work	1. "America also has a dirty history with prohibition - it has never worked. For anything." 2. "A little is better than nothing." 3. "Drug users still get their illegal drugs don't they?" 4. "See: civil war" 5. "why isn't murder illegal?"
13 (645)	peopl, gun, dont, like, kill	peopl, fortun, less, rise, like	1. "It's not like we have people who are beyond poor making bombs in the middle east." 2. "also people like guns." 3. "WEAPONS DON'T KILL PEOPLE, PEOPLE DO!!!!!!!!!!!!!!!!!!!!" 4. "people quickly forget history" 5. "Pretty sure lots of people care."
18 (570)	dont, know, im, think, gun	dont, im, know, think, realli	1. "I'm a Christian. I'm pretty sure you just got your wish. " 2. "But kids. Kids don't deserve this" 3. "I don't even know what to say anymore." 4. "If you have one you don't need the other." 5. "I really don't think so."
24 (631)	mental, health, ill, peopl, gun	perk, slight, care, hand, take	1. "That does not make them mentally ill. " 2. "instead, it should be a story about mental health and reaching out those you are worried about" 3. "It would help if mental health services were as easily accessible as guns." 4. "no its not, its time for him to do something about more effective mental health care." 5. "How about gun control *and* mental health?"

Sample comments from a few clusters generated by NEO-k-means with $\alpha = 0$ and $\beta = 0$ from the thread about a school shooting 2012 in America. 29 clusters were found in total.

Table 4.7: School shooting 2012 in America.

Cluster (size)	Key Terms	LDA Terms	Samples
5 (2022)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. " They killed themselves, guns kill other people. " 2. "This is why we need gun control " 3. "this would happen if people wanted it, people dont" 4. "There is already so many guns out there. Like 85% of the people I know own a gun."
9 (2032)	gun, peopl, would, control, get	gun, control, kill, peopl, dont	1. "Those children didn't die. They would have if he had a gun." 2. " They killed themselves, guns kill other people. " 3. "Guns don't kill people; people with guns kill people." 4. "This is why we need gun control "
13 (2026)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "This is why we need gun control " 2. "If those kids had guns, this wouldn't have happened." 3. "if there are over 300 million guns in america, doesn't that tell you something? Americans like guns." 4. "Yeah! Guns don't kill people, bullet do."
14 (2736)	gun, peopl, would, get, dont	gun, assault, rifl, use, ban	1. "So we should ban assault rifles?" 2. "This is why we need gun control " 3. "No assault weapons were used in this crime." 4. "Do you know where to buy a gun illegally? "
15 (7701)	gun, peopl, like, would, dont	thank, im, kid, like, dont	1. "I feel like the news should not be interviewing the little children about the shooting. It just seems wrong to me." 2. "Adam Lanza is the shooter not his brother Ryan Lanza" 3. "As an atheist, it's because of people like him that I hope hell exists." 4. "I can't even tell if you are serious right now."
20 (2017)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Sort out your gun control laws" 2. "You don't have to take guns away from people, but they should be MUCH harder to get. " 3. "Do you know where to buy a gun illegally? " 4. "Guns don't kill people; people with guns kill people."
23 (2501)	gun, peopl, mental, would, get	mental, gun, ill, peopl, kill	1. "So you're saying gun crime wouldn't be reduced by making guns illegal?" 2. "How about gun control *and* mental health?" 3. "I completely agree. It's a very complex social issue." 4. "Guns don't kill people; people with guns kill people."

Sample comments from a few clusters generated by NEO-k-means with $\alpha = 4.18399$ *and* $\beta = 0$ *from the thread about a school shooting 2012 in America. The alpha value was chosen according to the first strategy by [34] with* $\delta = 1.25$ *and 29 clusters were found in total.*

5 Discussion

The experiments were focused on finding structures within discussion threads. It can be shown in figure 5.1 that most threads in the dataset contain less than 100 comments and we claim that using these are not much of interest when trying to find structures within a single discussion because of the lack of data volume. These might be more appealing for finding similar threads or consider them as a very large thread. Instead the experiments were conducted on mostly random selected threads of various sizes given the size ≥ 100.

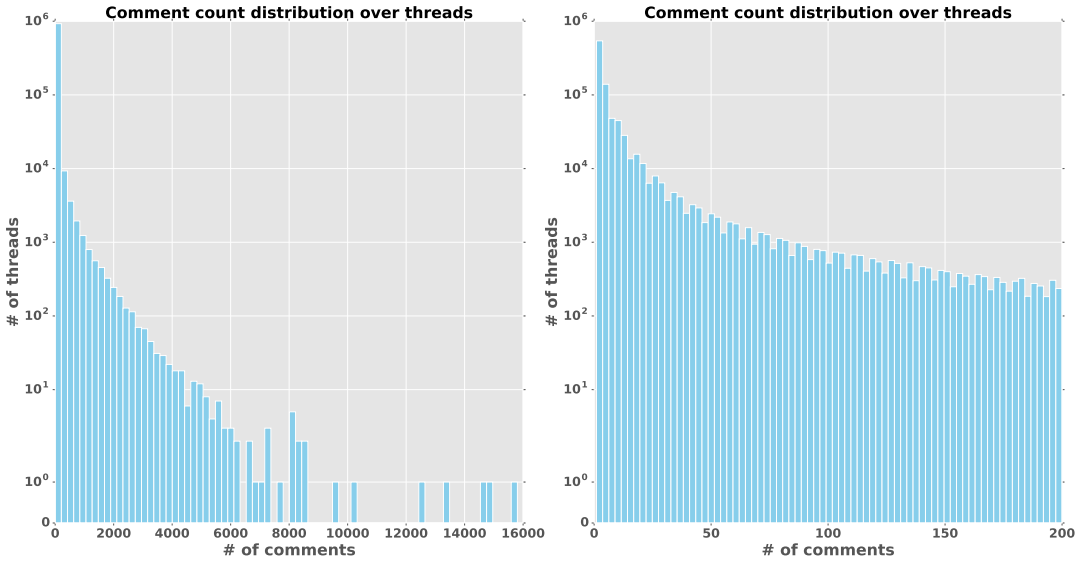


Figure 5.1: How the comments are distributed over threads. Left: Shows the distribution over all threads. Right: Zoomed in at the distribution over threads with 200 comments or less. It is apparent that most threads contain less than 100 comments, 910,731 threads, compared to 35,232 threads with ≥ 100 comments.

We limited the use of algorithms to Graclus and NEO-*k*-means for the experiments mainly because from internal experimentation with METIS the assumption of having equal sized clusters did not seem appropriate in the context of a human discussion. The NEO-*k*-means

with graph as input was not completely stable in its implementation and since we already had NEO-*k*-means working on vector space models we decided not to use it in our experiments.

5.1 Results

5.1.1 Performance

The comparison in terms of execution time between the algorithms (fig. 4.1) shows a significant gain in using Graclus. It looks to scale well and both modularity maximization and NEO-*k*-means look substantially less desirable. Be mindful of that both modularity maximization and NEO-*k*-means used Python implementations while Graclus is a C++ program which may give Graclus an advantage. Another point is that NEO-*k*-means was implemented by ourselves and may not yet be fully optimized.

Modularity maximization and Graclus have another preprocessing step in creating the graph representation which NEO-*k*-means can omit and this operation is quite slow because it computes the pairwise distances running in $O(n^2)$, where n is the number of samples. The comparison is not completely fair because of this, but constructing the graph is something that can be performed once and then saved if it needs to be used more than once and that is the reason it was excluded from the execution time. Likewise, the construction of the document-term frequency matrix needs to be performed for all the algorithms and is therefore not of interest to add to the execution time.

5.1.2 Modularity Maximization

Figure 4.2 shows the estimated number of clusters by the modularity maximization algorithm and these results follow directly from the theory. Since the modularity equation (eq. 2.11) does not consider edge weights, it should not influence the estimation whether edge weights are used or not which is the case and can be seen in figure 4.2. Having a more connected graph, i.e., higher edge density give rise to a lower cluster count estimate which is a logical consequence of the equation as well. Since the modularity is to be maximized, having higher edge density means that the clusters have to be larger for e_{ii} to reduce the effect of a_i^2 and for lower edge density they should be more compact which indicate more clusters. Figures 4.13 and 4.14 shows this quite clear when comparing the edge densities between different clusters.

There is a positive correlation between the size of the graph in terms of vertices and the number of clusters estimated, i.e., the more vertices in the graph the more clusters are estimated to exist.

5.1.3 Cluster Size

Looking at the cluster sizes generated by Graclus (fig. 4.3), the results follow what the modularity estimated. Using edge weights or not do not affect the cluster sizes but the number of edges do. This is evident by looking at the modularity estimation (fig. 4.2) and observe that a less connected graph increase the estimated cluster count. This means the vertices are distributed over more clusters which entail a decrease in the average cluster size. Edge weights do not affect the modularity estimate, hence do not affect the cluster sizes. Cluster sizes between Graclus and NEO-*k*-means (fig. 4.4) are not significantly different which may indicate that the algorithms find similar clusters.

5.1.4 Quantitative Comparisons

Here is a short summary of how to interpret the score of the objective functions.

Davies-Bouldin: A lower score indicates that the distances from points within a cluster to its cluster centroid are lower, i.e., more compact and/or the distances between cluster centroids are higher, i.e., more separated.

Calinski-Harabasz: A higher score indicates lower within-cluster variance, i.e., more compact and higher between-cluster variance, i.e., more separated.

Silhoutte: A higher score indicates that more points are well-matched to its corresponding cluster.

5.1.4.1 Edge Density

From figure 4.5 we can observe the following:

Davies-Bouldin: The results generate similar curves, but at the modularity estimates a less connected graph give better score in all cases.

Calinski-Harabasz: In most cases (4/6), the score is better using a lower degree at the modularity estimates. Most results except the red and pink ones give similar curves.

Silhoutte: The score is tied and the curves are more diverse between the results. The yellow, green and brown ones give similar results but the rest are different. The score is however not changing much and stays close to 0. This indicates that the clustering solutions are overall quite poor.

From this experiment, we can determine that using a less connected graph yield better results in most cases, especially when using the modularity estimate, according to the objective functions and is therefore the recommended choice. Note though that our choice of edge densities was arbitrary using $\frac{|\mathcal{E}|}{|\mathcal{V}|} \in [4, 8]$ and $\frac{|\mathcal{E}|}{|\mathcal{V}|} \in [12, 16]$, but since a lower edge density is recommended $\frac{|\mathcal{E}|}{|\mathcal{V}|} \in (0, 4]$ may be an even better choice.

Using a lower edge density do affect the size of the largest connected component, the part of the graph acting as input to the graph clustering algorithms, impacting the number of outliers, i.e., vertices that are not connected to the largest connected component. This may be one of the reasons that the objective scores are better overall for less connected graphs because outliers are only considered when computing the centroid of the whole dataset for the Calinski-Harabasz index. One way of dealing with this is to cluster several connected components that are larger than some threshold separately and consider all the clusters to be the clustering result.

5.1.4.2 Edge Weight

Using edge weights or not results in similar curves for all the objectives in all the cases (fig. 4.6) which indicate that the choice does not matter. However, it should intuitively become better clusters with weights since the similarity between two samples is explicitly encoded in the data and looking at eq. 2.16 it should have an effect. This may be the case where the content within the clusters are quite different but the scores have limitations and cannot implicate it. A qualitative study would have to be conducted to determine if that is true.

From an intuitive point of view and following equation 2.16, using edge weights should be preferred when applying Graclus.

5.1.4.3 Text Transformer

From figures 4.7, 4.8 and 4.9 one can observe the following:

Davies-Bouldin: Both transformers give quite similar curves, but at the modularity estimates the term frequency transformer give better score in most cases. Using term frequency with overlap did make the objective explode, see the pink curve in fig. 4.9, to the point it became unstable and gave infinite score when the samples size was over 1200.

Calinski-Harabasz: The term frequency transformer is better in all cases except when using NEO-k-means with overlap, i.e., $\alpha > 0$.

Silhoutte: Term frequency transformer is worse in every single case.

The silhouette index always gave better results to the term frequency-inverse document frequency transformer, but it did not give indication of any good clustering results and should

not be considered. The term frequency transformer should therefore be used when working with non-overlapping clusters and term frequency-inverse document frequency transformer when using NEO-k-means with overlap, i.e., $\alpha > 0$.

5.1.4.4 Overlap

Figure 4.10 shows that in almost every case for all objective functions using overlap yield worse score. This is not surprising though since the objectives was not made for overlapping clusters in mind. There are more cluster assignments with overlap which results in more calculations and increasing values.

5.1.4.5 Modularity Maximization Estimate

A common method for determining the optimal cluster count is to use the elbow or knee criterion [32] by plotting a monotonically decreasing or increasing objective on the y-axis and the number of clusters on the x-axis. At the cluster count where the objective stops increasing/decreasing significantly and starts converging can be considered the optimal cluster count.

The Davies-Bouldin indices we have used follow this kind of objective pretty well and looking at figure 4.11, one can observe that the modularity estimates give a reasonable good estimate in most cases. This implicate that the modularity maximization algorithm can be used to find a decent estimate of how many clusters exist in a discussion thread or atleast be used to find an initial estimate that can be increased/decreased until the solution is acceptable. It might even be possible to use supervised learning to find the optimal cluster count by observing how the objective function behave around the modularity estimate.

5.1.4.6 Summary

To summaries the findings, the graph should be constructed such that it is less connected in order to get better clustering results according to the objective functions. Whether to use edge weights or not need further investigation, but it is probably wise to use it. The term frequency transformer should be used to find non-overlapping clusters and the term frequency-inverse document frequency transformer for finding overlapping clusters.

The modularity maximization algorithm seem to find reasonable cluster counts and can be used to find a starting point rather than having to guess the number of clusters in a discussion.

5.1.5 Qualitative Comparisons

In this section, we look at the samples from some of the clusters generated from the threads about the school shooting in America 2012 and war on drugs.

5.1.5.1 War on drugs

In this section, we refer to tables 4.1, 4.2,4.3 and 4.4 as t1, t2, t3 and t4 respectively.

The tables t1 and t4 shows clusters varying the edge density. Reading the samples does not convey much useful information about the discussion. Neither results have any really distinct clusters, but the content shown from clusters 1 and 6 in t1 and cluster 3 in t4 can be considered uninformative. From the terms in t1, one can get more insight in what topics the discussion contains compared to t4 meaning t1 can be regarded as a "better" solution. Figure 4.12 shows that t1 gives better objective scores compared to a t4 which follow the reasoning that t1 is a "better" solution.

Cluster 2 in t1 and cluster 3 in t2 are quite similar, but other than that it is not apparent that they find the same topics. Maybe this is because the discussion is quite small, around 350 comments, and not enough content is present to form well defined clusters.

By looking at the terms in t3, we can at least observe the terms “drug”, “war”, and “addict” are common in many of the clusters when overlap is used. This indicate that those words are important to the discussion. The overlapping comments that is shown in t3, like “’Drugs Win Drug War’ ...”, contain terms common in many of the clusters which indicate that the overlapping samples are good.

In figure 4.13 the vertex sizes look to be evenly distributed with a few extreme cases and in figure 4.14 the sizes are more related to each other. The larger ones do hide some of the smaller ones and we have not seen the content of these larger comments making it difficult to draw any conclusions from the aforementioned figures. What one can tell is that the size of the comments do vary.

The results are not very helpful in understanding what the discussion is all about. Whether this depends on how the information is presented, if the content in the discussion is lacking, or if the algorithms are not good enough for the task is unknown.

5.1.5.2 School shooting

In this section we refer to tables 4.5, 4.6 and 4.7 as t5, t6 and t7 respectively.

In t5 one can see that topics about gun control, mental health, the perpetrator, and video games were found. Similarly, in t6 one can observe topics about the gun control, news, mental illness, and drugs.

Cluster 0 in t5, containing outliers, can in fact be considered “bad” content which is to be expected when they are not similar enough to be part of the largest connected component.

For instance cluster 3 in t5, clusters 24 and 7 in t6 and clusters 5, 9, 13, 14, 20, and 23 in t7 all contain information about health care and gun control. This means the algorithms find similar topics within the discussion and note that the tables only show a few of the clusters. There are more overlapping topics when comparing all the clusters, see appendix.

Looking at the sizes of the clusters generated by NEO-k-means with overlap, t7, those become very large compared to no overlap because the number of cluster assignments is over 4 times more since $\alpha > 4$. The words “gun” and “people” are common in all clusters shown which indicate that those words are part of a consistent theme in the discussion. However, the clusters are not as well defined as those in t5 and t6 making it a worse clustering solution. This is mainly because the overlapping region is too large which means the way it is chosen have to be addressed and looked into further.

We can observe in fig. 4.12 that the overlapping cluster solution perform worse in all objective functions which is not surprising. What is somewhat surprising is the fact that NEO-k-means without overlap have better score in all objectives. This may be explained by the 29 clusters found compared to 24 clusters found by Graclus.

The results are decent and one is able to gain insight in what some of the discussion topics are. It also shows that the algorithms can find similar clusters, but the choice of the size of the overlapping region have to be more thoughtful in order to have potential to be useful.

5.1.5.3 Summary

The clustering solutions discussed above have given mixed results. The smaller thread about war on drugs gave rather poor results while the thread about the school shooting had much more promising results.

Further investigation is needed through tests of more discussions with varying properties like length, number of unique users, and topic to understand when the algorithms are appropriate to use and their limitations. It is also important to look into how to best present the clustering content to convey the information in a more beneficial way rather than just showing a few short comments and common terms.

5.2 Data Storage

The data was in its original state stored in JSON files and it turned out to be a problem due to bad performance when running statistics on the data and having to deal with the file organization which meant tailor tools for a specific file organization. Moving files around resulted in having to change the file processing tools accordingly which is time consuming.

The reason we constructed a MySQL database was to solve these problems. The MySQL Database Management System (DBMS) provides an abstraction layer of where the data is stored, how the data is stored, and how the data is accessed. All of these properties are beneficial and for instance getting the number of unique users in the data took over 230 seconds using JSON compared to around 11 seconds using MySQL. This was a very easy task, but the Structured Query Language (SQL) which MySQL provides makes it possible to run heavy analysis on large data with ease and let the developer focus on the data rather than the tools.

5.3 Method

We have applied various objective functions to determine what parameters seem to generate better clustering results but we have not analysed what those objective functions actual tell us about the cluster content itself. There may or may not be a relationship between the objective scores and what a human would consider a more or less beneficial clustering result. This is one area that is lacking within this study and should be considered important to analyse in future studies.

The features used in the clustering process determine what structures can be found in the data. We have limited to only use the textual content and not any side information such as the up-vote/down-vote score which causes the clusters to contain similar comments, but nothing about their importance. Depending on what the goal is in using clustering, the features must be conformed to the goal, e.g., to see what terms co-occur in user comments it would be more appropriate to cluster terms instead of comments, i.e., the rows of the document-term frequency matrix correspond to terms and the columns correspond to documents.

The quantitative tests comparing overlapping clusters to non-overlapping clusters are unfair, similar to comparing oranges and apples, due to the objective functions being meant to compare non-overlapping cluster solutions. In [34], they used the ground truth to evaluate the result which we unfortunately did not have. We used one of the strategies in [34] to determine how large the overlapping region should be and that is an area that can be experimented with to find either a more appropriate strategy in this context or experiment more with the parameters.

A much more extensive qualitative study have to be conducted in order to understand when the algorithms are working and which parameters are important to get good results. The same with how to create a more interpretable approach for presenting the information to users.

5.4 The work in a wider context

Algorithms are a huge part of the modern society in assisting decision making for people [16]. It is for instance very likely a person is seeking information about a subject using *Google Search* and infer the top results as being reliable sources without much consideration. There are other search engines such as *Bing* and *DuckDuckGo* that may give different sources when searching for the same subject.

The point is that algorithms show different information and control what information is shown depending on the parameters used by the algorithms and their internal behaviour.

The term *filter bubble* was coined by Eli Pariser [27] describing the potential of online personalization to isolate people from diverse viewpoints and content.

By using clustering algorithms, we have shown that it is possible to find subtopics within a discussion but how can this information be used? The way the obtained information is presented to the users is an important part of addressing issues that can arise when algorithms do decisions for us. Issues like filter bubbles as aforementioned, but still be valuable to the users in terms of improving the experience.

We have only used the textual content without incorporate any semantic meaning of the text, but imagine using more features that are collected from the text content such as opinions and/or personalized features from *cookies*. This could potentially become a tool for filtering out information that does not match the users personal viewpoint or interest and that may polarize the discussions. This phenomenon is called *echo chambers* [14] where users are selectively exposed to similar beliefs.

By knowing how algorithms reason, users could potentially exploit that for their own gain [16]. It is therefore important for companies employing tools that utilize algorithms to do decisions and users of these tools to be conscious about their impact and limitations.

5.5 Source Criticism

All the sources have been evaluated with respect to their credibility which was established by checking the authors educational backgrounds and research areas. The number of citations was also taken into consideration but for more recent papers this had less of an effect.

6 Conclusion

The purpose of this thesis was to determine if the chosen clustering algorithms can find structure within discussions on internet forums. We have discussed the results found in the experiments and can now conclude the findings. To do so we need to go back to the research questions.

Can the chosen clustering algorithms be used to find structure in textual content? We have seen that the algorithms can find subtopics within a discussion given the textual content. The length of the discussion may have an impact on how distinct the topics within the clusters are or it may be the discussion itself not having well defined topics. It could also have to do with how the information is presented and so further steps for modeling the topic to get more interpretable results should be performed.

How do the algorithms compare in terms of execution time? The comparison of execution times was rather one-sided and Graclus is the clear winner. The modularity maximization algorithm and non-exhaustive overlapping k-means showed fairly similar performance.

6.1 Future Work

This section presents a few areas that may be interesting to analyse for future work.

6.1.1 User Feedback

One of the most important thing that was not considered in this study is getting a deeper understanding of the clustering solution from a user perspective. To address this issue, it would be beneficial if users were able to interactively use the algorithms or analyse pre-generated solutions by the algorithms and rate the solutions themselves since it is up to the users of the tools to determine the usefulness. This could be incorporated inside a website and by using the user feedback strengthen the view of what works and does not work. It would also be possible to find out if the objective functions correlate with the users opinions.

6.1.2 Features

There are other feature selection methods for textual data [1] than those that were used in this study and more advanced feature transformation methods to consider such as *Latent Semantic*

Indexing (LSI), *Probabilistic Latent Semantic Analysis* (PLSA), and *Non-negative Matrix Factorization* (NMF) which are known as dimension reduction techniques. It would be interesting to know how methods like these affect the clustering result and the scalability of the algorithms.

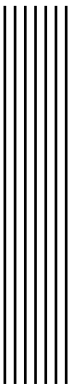
Cluster analysis does not necessarily need to be a complete unsupervised method but can be used as a *semi-supervised* method using *side-information* [2].

Reddit provides a voting system allowing users to up-vote and down-vote comments and threads and by using that information it may be possible to guide a clustering algorithm to find "good" content assuming votes correspond to content quality. It also lets users be part of the algorithms decision making.

The text content can be used to construct other features such as describing the readability, e.g., *Automated Readability Index* (ARI) that have been used in finding antisocial behaviour [7].

More sophisticated methods may be able to find features to guide a clustering algorithm. For instance using *uneddit*[1] that shows the content of deleted comments on Reddit and train a supervised learning algorithm to predict how likely a comment is to be deleted and use that to determine how appropriate the content of a comment is.

[1] https://uneddit.com/



Bibliography

[1] Charu C. Aggarwal and ChengXiang Zhai. "An introduction to text mining". In: *Mining Text Data* (2013), pp. 1–10. ISSN: 20403372. DOI: 10.1007/978-1-4614-3223-4_1.

[2] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu. "On the use of side information for mining text data". In: *IEEE Transactions on Knowledge and Data Engineering* 26.6 (2014), pp. 1415–1429. ISSN: 10414347. DOI: 10.1109/TKDE.2012.148.

[3] Olatz Arbelaitz et al. "An extensive comparative study of cluster validity indices". In: *Pattern Recognition* 46.1 (2013), pp. 243–256. ISSN: 00313203. DOI: 10.1016/j.patcog.2012.07.021.

[4] D. Arthur and S. Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* 8 (2007), pp. 1027–1035. ISSN: 0898716241. DOI: 10.1145/1283383.1283494. URL: http://portal.acm.org/citation.cfm?id=1283494.

[5] A. Ben-Dor, R. Shamir, and Z. Yakhini. "Clustering gene expression patterns." In: *Journal of computational biology : a journal of computational molecular cell biology* 6.3-4 (1999), pp. 281–297. ISSN: 1066-5277. DOI: 10.1089/106652799318274.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022. ISSN: 15324435. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1.

[7] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "Antisocial Behavior in Online Discussion Communities". In: *Proceedings of the Ninth International Conference on Web and Social Media, 2015, University of Oxford, Oxford, UK, May 26-29, 2015* (2015), pp. 61–70. arXiv: 1504.00680. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10469.

[8] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. *Finding community structure in very large networks*. 2004. DOI: 10.1103/PhysRevE.70.066111. arXiv: 0408187 [cond-mat]. URL: http://arxiv.org/abs/cond-mat/0408187.

[9] Aedín C. Culhane, Guy Perrière, and Desmond G. Higgins. "Cross-platform comparison and visualisation of gene expression data using co-inertia analysis." In: *BMC bioinformatics* 4 (2003), p. 59. ISSN: 1471-2105. DOI: 10.1186/1471-2105-4-59.

[10] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. "A Unified View of Kernel k-means , Spectral Clustering and Graph Cuts". In: *Computational Complexity* 25.5 (2005), pp. 1–20. DOI: citeulike-article-id:486970. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.1701{\&}rep=rep1{\&}type=pdf.

[11] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. "Weighted graph cuts without eigenvectors a multilevel approach". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.11 (2007), pp. 1944–1957. ISSN: 01628828. DOI: 10.1109/TPAMI.2007.1115.

[12] Aysu Ezen-Can et al. "Unsupervised modeling for understanding MOOC discussion forums". In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (2015), pp. 146–150. DOI: 10.1145/2723576.2723589. URL: http://dl.acm.org/citation.cfm?id=2723576.2723589.

[13] Santo Fortunato. *Community detection in graphs*. 2010. DOI: 10.1016/j.physrep.2009.11.002. arXiv: 0906.0612.

[14] R. Kelly Garrett. "Echo chambers online?: Politically motivated selective exposure among Internet news users". In: *Journal of Computer-Mediated Communication* 14.2 (2009), pp. 265–285. ISSN: 10836101. DOI: 10.1111/j.1083-6101.2009.01440.x.

[15] Joydeep Ghosh, Raymond Mooney, and Alexander Strehl. "Impact of Similarity Measures on Web-page Clustering". In: *In Workshop on Artificial Intelligence for Web Search (AAAI 2000)* (2000), pp. 58–64. DOI: 10.1.1.29.2377. URL: https://www.aaai.org/Papers/Workshops/2000/WS-00-01/WS00-01-011.pdf.

[16] Jutta Haider and Olof Sundin. "Algoritmer i samhället". In: *Kansliet för strategi-och samtidsfrågor, Regeringskansliet* (2016). URL: http://lup.lub.lu.se/record/8851321/file/8851333.pdf.

[17] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. "On clustering validation techniques". In: *Journal of Intelligent Information Systems* 17.2-3 (2001), pp. 107–145. ISSN: 09259902. DOI: 10.1023/A:1012801612483.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "The Elements of Statistical Learning". In: *Springer 2001* 18.4 (2009), p. 746. ISSN: 00111287. DOI: 10.1007/b94608. URL: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20{\&}path=ASIN/0387952845.

[19] Andreas Hotho, S. Staab, and G. Stumme. "Wordnet improves Text Document Clustering". In: *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on* 03 (2003), pp. 541–544. DOI: 10.1.1.8.8026. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.8026{\&}rep=rep1{\&}type=pdf.

[20] Anna Huang. "Similarity measures for text document clustering". In: *Proceedings of the Sixth New Zealand* April (2008), pp. 49–56. URL: http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual{_}Papers/pg049{_}Similarity{_}Measures{_}for{_}Text{_}Document{_}Clustering.pdf.

[21] Anil K. Jain. "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8 (2010), pp. 651–666. ISSN: 01678655. DOI: 10.1016/j.patrec.2009.09.011. arXiv: 0402594v3 [arXiv:cond-mat]. URL: http://dx.doi.org/10.1016/j.patrec.2009.09.011.

[22] George Karypis and Vipin Kumar. "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 359–392. ISSN: 1064-8275. DOI: 10.1137/S1064827595287997. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.106.4101$\backslash$nhttp://epubs.siam.org/doi/abs/10.1137/S1064827595287997.

[23] Luying Liu et al. "A comparative study on unsupervised feature selection methods for text clustering". In: *... . IEEE NLP-KE'05. Proceedings of ...* 00 (2005), pp. 597–601. DOI: 10.1109/NLPKE.2005.1598807.

[24] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Vol. 1. c. 2008, p. 496. ISBN: 0521865719. DOI: 10.1109/LPT.2009.2020494. arXiv: 052186571997805218657 15. URL: http://dspace.cusat.ac.in/dspace/handle/123456789/2538.

[25] Louis Massey. "Evaluating and comparing text clustering results". In: *Proceedings Computational Intelligence (CI 2005)* (2005), pp. 85–90. URL: https://www.actapress.com/PDFViewer.aspx?paperId=21103.

[26] M. E. J. Newman and M. Girvan. "Finding and evaluating community structure in networks". In: *Physics* (2003), p. 16. DOI: 10.1103/PhysRevE.69.026113. arXiv: 0308217 [cond-mat]. URL: http://arxiv.org/abs/cond-mat/0308217.

[27] Eli Pariser. "The Filter Bubble: What the Internet Is Hiding from You". In: *ZNet* (2011), p. 304. ISSN: 1863-2300. DOI: 10.1353/pla.2011.0036. arXiv: arXiv:1011.1669v3. URL: http://www.amazon.com/dp/1594203008.

[28] Fabian Pedregosa and G. Varoquaux. "Scikit-learn: Machine learning in Python". In: *... of Machine Learning ...* 12 (2011), pp. 2825–2830. ISSN: 15324435. DOI: 10.1007/s13398-014-0173-7.2. arXiv: arXiv:1201.0490v2. URL: http://dl.acm.org/citation.cfm?id=2078195.

[29] Eréndira Rendón et al. "Internal versus External cluster validation indexes". In: *International Journal of Computers and Communications* 5.1 (2011), pp. 27–34. URL: http://w.naun.org/multimedia/UPress/cc/20-463.pdf.

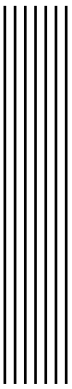
[30] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 2010, p. 1132. ISBN: 0137903952. DOI: 10.1017/S0269888900007724. arXiv: arXiv:1011.1669v3. URL: http://amazon.de/o/ASIN/0130803022/.

[31] J. Shi and J. Malik. "Normalized Cuts and Image Segmentation". In: *Ieee Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905. ISSN: 0162-8828. DOI: 10.1109/34.868688. arXiv: 0703101v1 [cs]. URL: http://www.computer.org/portal/web/csdl/doi?doc=abs/proceedings/cvpr/1997/7822/00/78220731abs.htm$\backslash$$npapers3://publication/uuid/268FC197-AF47-4C7C-887F-BEDB94A81320.

[32] Lucas Vendramin, Ricardo J. G. B. Campello, and Eduardo R. Hruschka. "Relative clustering validity criteria: A comparative overview". In: *Statistical Analysis and Data Mining* 3.4 (2010), pp. 209–235. ISSN: 19321872. DOI: 10.1002/sam.10080. arXiv: 1206.3552.

[33] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4 (2007), pp. 395–416. ISSN: 09603174. DOI: 10.1007/s11222-007-9033-z. arXiv: arXiv:0711.0189v1.

[34] Joyce Jiyoung Whang, Inderjit S. Dhillon, and David F. Gleich. "Non-exhaustive, Overlapping k-means". In: *SIAM International Conference on Data Mining (SDM)*. 2015, pp. 936–944.



Appendix

Clustering Solutions

Here are all the clustering solutions from the thread about the school shooting in America, 2012.

Graclus

Cluster (size)	Key Terms	LDA Terms	Samples
0 (55)	like, make, dont, gun, go	boxcutt, headcount, risen, stimmt, staph	1. "Woppidy doo Basil, what does it mean?" 2. "Dewey v Truman? Obamacare anyone?" 3. "Guns, stahpit. Guunss. STAPH." 4. "the headcount has risen to 20 children."
1 (74)	game, video, violent, blame, violenc	game, video, palin, sarah, blame	1. "we need to ban video games. " 2. "How long before the media blames Sarah Palin (again)?" 3. "I think it was violent video games."
2 (249)	mental, ill, peopl, dont, gun	ill, mental, shooter, problem, untreat	1. "Again you have a misconception of what mental illness is. It is not simply being capable of terrible things." 2. "You are scapegoating, and lack understanding as to what mental illness is." 3. "That does not make them mentally ill. "
3 (329)	mental, health, gun, peopl, issu	health, mental, care, issu, gun	1. "How about gun control *and* mental health?" 2. "no its not, its time for him to do something about more effective mental health care." 3. "We have state mental hospitals."

4 (4926)	gun, peopl, like, would, dont	gun, dont, your, think, would	1. "Maybe you're wrong?" 2. "Exactly. Well said. " 3. "You see, if those kids all had guns this never would have happened."
5 (190)	bomb, car, peopl, gun, drive	bomb, car, driver, licens, homemad	1. "A car isn't built to hurt other people, guns are. " 2. "people die in car crashes... LEGISLATE AGAINST CARS!!!" 3. "A homemade bomb? A fire? "
6 (553)	peopl, kill, gun, dont, use	kill, peopl, gun, knife, dont	1. " They killed themselves, guns kill other people. " 2. ""Guns don't kill people, but they sure as fuck make it a lot easier!"" 3. "A guns only purpose is to kill or maim. A knife has more purposes than to harm. "
7 (152)	carri, gun, conceal, school, shoot	carri, free, conceal, zone, gun	1. "You should allow children to carry weapons. That would resolve all these school shootings." 2. "That sucks gun free zones work so well to..." 3. "...who are also carrying concealed guns."
8 (282)	rifl, assault, weapon, gun, use	assault, rifl, weapon, automat, ban	1. "Ban assault weapons now. " 2. "What makes a rifle an assault rifle? " 3. "You can own most of the same weapons, but they aren't fully automatic. It's still very simple to make them automatic though."
9 (209)	gun, rate, homicid, per, us	rate, homicid, per, murder, us	1. "Look at Japan's suicide rate. " 2. "UK has 0.03 Gun related murders per 100,000 compared to 2.93 for the USA" 3. "UK murder rate: 1.2 US murder rate: 4.2 percent difference: 350"
10 (166)	drug, gun, war, would, peopl	drug, war, walmart, civil, sell	1. "Drug users still get their illegal drugs don't they?" 2. "See: civil war" 3. "trust me, there is cocaine at walmart."
11 (527)	gun, illeg, crimin, peopl, get	illeg, gun, crimin, buy, state	1. "My state considers any magazine over 10 rounds to be high capacity. " 2. "Do you know where to buy a gun illegally? " 3. "You can go buy a gun from a different member of the gang that provides the weed. "
12 (1207)	gun, peopl, would, law, get	gun, ban, law, check, owner	1. "violent crime != crime homicide == violent crime" 2. "Why are guns not banned already? " 3. "As in, having background checks of some sort."

13 (80)	door, school, lock, classroom, drill	door, lock, classroom, nashvill, riddl	1. "He started in the main office, so he likely walked right in the front door." 2. "CBS is reporting that the teachers who locked down their classrooms had locks on their doors" 3. "this is what happens when liberals take over education. Most places you need to be let in the doors of a school."
14 (227)	teacher, gun, arm, school, would	teacher, arm, children, shoot, happen	1. "That's just what we need. Hundreds of thousands of underpaid, overstressed teachers packing heat. " 2. "No, see they think the teachers should all be armed..." 3. "If the children were armed they could have defended themselves..."
15 (361)	parent, children, famili, cant, kid	parent, christma, goe, heart, famili	1. "I wonder what the families will do with all of the Christmas presents they bought for their kids?" 2. "The news shouldn't do it, but its the parents who are giving them permission to do it. Take it up with the parents too." 3. "Such an unfortunate event. My thoughts and prayers goes out to all of the families affected by this. "
16 (585)	post, news, peopl, name, like	post, wow, news, name, facebook	1. "What does this have to do with politics? This was already posted under news 2 hours before you posted this. " 2. "They showed HIS face and HIS facebook profile. Even with the same name, they basically fucked him over. " 3. "Wow I cant believe this happened just a town over..."
17 (341)	dead, mother, kill, shooter, school	mother, brother, dead, shooter, stole	1. "He killed his mother. She was a teacher at the school." 2. "The killer is one of the 27 dead, according to BBC news..." 3. "Because if those children had guns, only the shooter would be dead.../s"
18 (122)	lanza, ryan, adam, brother, shooter	lanza, ryan, adam, brother, name	1. "So it was a Ryan Lanza just not the one they linked?" 2. "Yeah, they are now saying it's his brother, Adam, not Ryan" 3. "Adam Lanza is the shooter not his brother Ryan Lanza"

19 (307)	right, gun, peopl, freedom, arm	pleas, freedom, right, tell, lol	1. "You don't have freedom, if you give up your guns?" 2. "Right of the people â right of the militia" 3. "Can you (America) please, please **please** have a national conversation about gun control?"
20 (588)	gun, control, law, talk, peopl	control, gun, talk, polit, law	1. "I feel like if the kids want to talk, let them talk." 2. "This is why we need gun control " 3. "this isn't politics... this shouldn't be politics. why is it in /r/politics?"
21 (223)	china, attack, gun, knife, children	china, stab, attack, today, knife	1. "but...but... Wal Mart..." 2. "Like China that had 22 kids stabbed today? " 3. "Nobody was killed in the China attack."
22 (329)	thank, comment, upvot, downvot, im	thank, comment, upvot, downvot, thread	1. "There are 2.2 million subscribers on this subreddit alone, and it has less than 40k votes. That is barely 2% of this subreddit." 2. "Thank you, thank you, thank you." 3. "If I could upvote this comment ten more times, I would. This. Exactly this. "
23 (365)	fuck, shit, gun, peopl, go	fuck, shut, shit, your, serious	1. "Go fuck yourself you piece of shit." 2. "Fuck you, fuck him, fuck humanity man, fuck." 3. "Holy shit that's fucked"
24 (641)	like, god, im, peopl, go	god, oh, read, troll, cri	1. "Oh god, I'm so sorry." 2. "I'm crying having just read that now." 3. "Who the FUCK would kill Kindergarteners??? It's times like this I hope there IS a heaven and hell..."

NEO-K-Means

Non-Overlapping

Cluster (size)	Key Terms	LDA Terms	Samples
1 (216)	thank, im, gun, god, like	thank, sibl, younger, mine, elementari	1. "God damn it. God *damn* it." 2. "Thank you, thank you, thank you." 3. "Can't upvote this enough!"
2 (168)	game, violent, video, blame, violenc	video, blame, game, violent, crime	1. "I think it was violent video games." 2. "violent crime != crime homicide == violent crime" 3. "Don't blame the reporters, blame the parents and school officials who allow it to happen."

3 (682)	live, parent, cant, feel, kid	famili, parent, cant, live, rest	1. "Parents, hug your kids today." 2. "I can't even imagine the scene." 3. "The difference between 22 injuries and 22 lives... is 22 lives."
4 (391)	gun, rate, us, homicid, countri	us, rate, countri, murder, homicid	1. "Can you source the knifings for us?" 2. "Than how come it happens in this country more than almost any other country? " 3. "UK murder rate: 1.2 US murder rate: 4.2 percent difference: 350"
5 (1067)	like, make, one, peopl, thing	like, make, point, thing, time	1. "no, more like nothing to live for, but they had a reason to die, those men gave them a reason." 2. "That doesn't even make sense. " 3. "It's at times like these we see the best and worst of humanity coalescing at one point."
6 (1101)	gun, peopl, would, get, like	gun, illeg, legal, ban, use	1. "stfu gun nut. recycle all the guns" 2. "GIVE ALL TEACHERS GUNS NOW! THE ANSWER TO GUN VIOLENCE IS MORE GUNS!" 3. "Do you know where to buy a gun illegally? "
7 (432)	gun, control, peopl, one, like	control, gun, talk, time, america	1. "and yet there will be people on here advocating against gun control. fucking nutjobs." 2. "This is why we need gun control " 3. "So is now the time to talk about gun control?"
8 (536)	fuck, gun, peopl, go, shit	fuck, shit, holi, sick, shut	1. "Holy shit that's fucked" 2. "fuck you, you don't know shit" 3. "Fuck you, fuck him, fuck humanity man, fuck."
9 (205)	sad, word, im, go, sorri	sad, word, sorri, thought, littl	1. "It make me feel more sad, but you has the point ... i feel really sad" 2. "There simply aren't words for this." 3. "I'm sorry, I thought this was /r/politics? "
10 (394)	news, media, like, stori, peopl	news, reddit, media, agenda, fox	1. "this isn't politics... this shouldn't be politics. why is it in /r/politics?" 2. "Respectfully, the media focuses on the shooter only because it's what sells. The media has no benevolent intentions here." 3. "Mass media news kills."

11 (316)	someth, like, peopl, happen, make	someth, hell, cri, like, obama	1. "I cried. I didn't even try to hold it back. I just cried. " 2. "Something something Right of the PEOPLE something something. Shall not be infringed." 3. "He is already in hell.... those actions are what his mind is doing in hell..."
12 (210)	drug, gun, illeg, war, peopl	drug, war, noth, illeg, work	1. "See: civil war" 2. "A little is better than nothing." 3. "Drug users still get their illegal drugs don't they?"
13 (645)	peopl, gun, dont, like, kill	peopl, fortun, less, rise, like	1. "Some people are just too crazy for this world." 2. "WEAPONS DON'T KILL PEOPLE, PEOPLE DO!!!!!!!!!!!!!!!!!!!!" 3. "I think the worst is the one where the most people died."
14 (466)	right, gun, peopl, arm, amend	right, troll, know, smaller, hth	1. "Right of the people â right of the militia" 2. "B..but second amendment." 3. "So 2nd amendment should protect our right to bear arms and bulk fertilizer."
15 (419)	shooter, lanza, ryan, brother, post	post, lanza, brother, ryan, shooter	1. "I posted this on facebook (before seeing your post)... yours was slightly better received." 2. "I know the guy, it was his brother Adam." 3. "Adam Lanza is the shooter not his brother Ryan Lanza"
16 (233)	gun, check, background, wait, state	check, long, wait, haha, palin	1. "haha as long as those guns are killing white ppl who cares. http://i.minus.com/ilkynRf19EnyN.gif NIGGAS RULE." 2. "As in, having background checks of some sort." 3. "It's been working for decades now! Oh wait..."
17 (307)	that, well, gun, peopl, think	that, well, said, yeah, one	1. "Oh I guess that's okay then :) /s" 2. "That's not a fact. That's a shitty guess." 3. "Well... That's a problem..."
18 (570)	dont, know, im, think, gun	dont, im, know, think, realli	1. "I really hope you are just a troll. " 2. "I'm a Christian. I'm pretty sure you just got your wish. " 3. "it is. Too bad the people that need it don't know it. "

19 (535)	would, gun, peopl, think, like	would, think, never, happen, gun	1. "Who the hell would downvote this?" 2. "So then what would you do? How would any change happen? " 3. "No one died in that attack in China. If he had a gun you could bet they would have."
20 (443)	gun, law, peopl, control, would	law, gun, control, chang, stricter	1. "what exactly are sane gun control laws?" 2. "Changing the law will change the lifestyle. " 3. "It's time to do something about gun laws. "
21 (414)	children, dead, school, die, gun	children, dead, yesterdaynobodi, ireland, fallen	1. "Update - 18 children dead." 2. "I am not for it. But its better than having children die." 3. "And 22 children were stabbed in China today http://www.cbc.ca/news/world/story/2012/12/14/china-knife-attack-school.html Fuck this world "
22 (569)	kill, peopl, gun, knife, dont	kill, peopl, gun, knife, kid	1. "A guns only purpose is to kill or maim. A knife has more purposes than to harm. " 2. "He killed his mother, father, and brother." 3. " They killed themselves, guns kill other people. "
23 (325)	your, gun, like, say, right	your, agre, complet, right, fuck	1. "I am from the USA and I COMPLETELY agree." 2. "You're right. Fuck." 3. "They taste the same whether you pull the wings off or not. You're all about wasted effort."
24 (631)	mental, health, ill, peopl, gun	perk, slight, care, hand, take	1. "That'll solve all our problems!" 2. "That does not make them mentally ill. " 3. "How about gun control *and* mental health?"
25 (153)	yes, gun, peopl, dont, like	yes, gun, that, talk, realli	1. "Where is your god now, theists? Is he still all loving? Yes, this is the time to bring this up. It always is." 2. "Yes, but their primary purpose is still to do harm to something. " 3. "Yes. I agree, and that's what we should have."
26 (421)	get, gun, peopl, need, one	get, need, help, gun, coverag	1. "Fired? I wish, they'll probably all get raises for it. (edits:grammatical)" 2. "I don't think you get it." 3. "To get a legal gun you need to be 21? "
27 (401)	rifl, assault, weapon, gun, use	rifl, assault, weapon, use, ban	1. "So we should ban assault rifles?" 2. "What makes a rifle an assault rifle? " 3. "The shooter used two pistols, not an assault rifle."

28 (257)	comment, read, gun, im, peopl	comment, read, thread, pleas, articl	1. "Oh wow. The comments... the YouTube comments..." 2. "I'm crying having just read that now." 3. "I just read it in this article. I don't know if it's the original source http://gma.yahoo.com/breaking-conn-school-district-locked-down-shooting-report-151955384–abc-news-topstories.html?.tsrc=yahoo"
29 (581)	school, teacher, kid, shoot, gun	school, teacher, kid, arm, kinder-garten	1. "He killed his mother. She was a teacher at the school." 2. "In my elementary school you could." 3. "Like China that had 22 kids stabbed today? "

Overlapping

Cluster (size)	Key Terms	LDA Terms	Samples
1 (2028)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Guns don't kill people; people with guns kill people." 2. "So is now the time to talk about gun control?" 3. "There is already so many guns out there. Like 85% of the people I know own a gun."
2 (2026)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Sort out your gun control laws" 2. "This is why we need gun control " 3. "Guns don't kill people; people with guns kill people."
3 (2019)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. " They killed themselves, guns kill other people. " 2. "This is why we need gun control " 3. "Guns don't kill people; people with guns kill people."
4 (2023)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "So you're saying gun crime wouldn't be reduced by making guns illegal?" 2. "Guns don't kill people; people with guns kill people." 3. "So is now the time to talk about gun control?"
5 (2022)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "You see, if those kids all had guns this never would have happened." 2. "Guns don't kill people; people with guns kill people." 3. "This is why we need gun control "

6 (2022)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Do you know where to buy a gun illegally? " 2. "Guns don’t kill people; people with guns kill people." 3. "So is now the time to talk about gun control?"
7 (2024)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Guns don’t kill people; people with guns kill people." 2. "So is now the time to talk about gun control?" 3. "Guns don’t kill people; people with guns kill people."
8 (2251)	gun, peopl, would, control, dont	gun, control, peopl, right, kill	1. "Right of the people â right of the militia" 2. "Guns don’t kill people; people with guns kill people." 3. "So is now the time to talk about gun control?"
9 (2032)	gun, peopl, would, control, get	gun, control, kill, peopl, dont	1. "This is why we need gun control " 2. "How about gun control *and* mental health?" 3. "Guns don’t kill people; people with guns kill people."
10 (2260)	gun, peopl, control, would, law	gun, control, law, kill, peopl	1. "Guns don’t kill people; people with guns kill people." 2. "So is now the time to talk about gun control?" 3. "This is why we need gun control "
11 (2019)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "So much for the "if somebody there had a gun, they could have killed the shooter" argument. " 2. "So is now the time to talk about gun control?" 3. "Guns don’t kill people; people with guns kill people."
12 (2026)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Guns don’t kill people; people with guns kill people." 2. "So is now the time to talk about gun control?" 3. "There is already so many guns out there. Like 85% of the people I know own a gun."
13 (2026)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "things like this happen in countries were guns are illegal." 2. "This is why we need gun control " 3. "Guns don’t kill people; people with guns kill people."
14 (2736)	gun, peopl, would, get, dont	gun, assault, rifl, use, ban	1. "So we should ban assault rifles?" 2. "This is why we need gun control " 3. "Guns don’t kill people; people with guns kill people."

15 (7701)	gun, peopl, like, would, dont	thank, im, kid, like, dont	1. "Well yeah, that's why it's not allowed." 2. "I never said I'm not part of it." 3. "28 dead, 20 children. I don't even know what to say."
16 (2021)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "So is now the time to talk about gun control?" 2. "Guns don't kill people; people with guns kill people." 3. "There are already 250+ million guns in America, there is no practical way to outlaw access to guns here."
17 (2023)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "How about gun control *and* mental health?" 2. "Guns don't kill people; people with guns kill people." 3. "This is why we need gun control "
18 (2037)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "This is why we need gun control " 2. "So is now the time to talk about gun control?" 3. "Guns don't kill people; people with guns kill people."
19 (2360)	gun, peopl, would, dont, get	peopl, gun, control, kill, dont	1. "Guns don't kill people; people with guns kill people." 2. "This is why we need gun control " 3. "Guns don't kill people; people with guns kill people."
20 (2017)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "There are already 250+ million guns in America, there is no practical way to outlaw access to guns here." 2. "Guns don't kill people; people with guns kill people." 3. ""Guns don't kill people, but they sure as fuck make it a lot easier!""
21 (2606)	gun, peopl, would, control, dont	gun, control, kill, peopl, us	1. "How is their [firearm homicide rate](http://en.wikipedia.org/wiki/List_of_countries_by_firearm-related_death_rate) more than 3 times the US if guns are illegal?" 2. "Guns don't kill people; people with guns kill people." 3. "Gun control works. Banning guns does not."
22 (2031)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Guns don't kill people; people with guns kill people." 2. "So is now the time to talk about gun control?" 3. "There is already so many guns out there. Like 85% of the people I know own a gun."

23 (2501)	gun, peopl, mental, would, get	mental, gun, ill, peopl, kill	1. "That does not make them mentally ill. " 2. "How about gun control *and* mental health?" 3. "Guns don't kill people; people with guns kill people."
24 (2412)	gun, peopl, kill, would, dont	kill, gun, peopl, control, knife	1. "Yeah, I mean just look at the UK and Australia and all the horrible mass killings they have over there..." 2. "This is why we need gun control " 3. " They killed themselves, guns kill other people. "
25 (2023)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Guns don't kill people; people with guns kill people." 2. "So is now the time to talk about gun control?" 3. "This is why we need gun control "
26 (2419)	gun, peopl, would, dont, get	fuck, gun, holi, need, your	1. "Guns don't kill people; people with guns kill people." 2. "Fuck you, fuck him, fuck humanity man, fuck." 3. "This is why we need gun control "
27 (2127)	gun, peopl, would, control, get	gun, control, peopl, kill, dont	1. "You know what they should do to stop gun crimes? Make killing illegal." 2. "Guns don't kill people; people with guns kill people." 3. "So is now the time to talk about gun control?"
28 (2029)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "Do you know where to buy a gun illegally? " 2. "So is now the time to talk about gun control?" 3. "Guns don't kill people; people with guns kill people."
29 (2027)	gun, peopl, would, control, dont	gun, control, kill, peopl, dont	1. "You see, if those kids all had guns this never would have happened." 2. "Guns don't kill people; people with guns kill people." 3. "This is why we need gun control "