AALBORG UNIVERSITY
COPENHAGEN

Semester: ICTE 4
Title: **Exploring methods to find and label latent topic-groups in a blogging environment**

Aalborg University Copenhagen
A.C. Meyers Vænge 15
2450 København SV

Semester Coordinator: Henning Olesen

Secretary: Maiken Keller

Project Period: SPRING 2017

Semester Theme: Master thesis

Supervisor(s): Jannick Kirk Sørensen

Project group no.:

Members Seynabou Ndiaye
(do not write CPR.nr.):

Abstract:
Twitter is mainly known for its tweets that have been subject for many researches. However, it has a normal size blog that is actually growing very fast that allows more room for content. With approximately 1600 blogs, twitter is a great communication channel between users and businesses that promotes themselves.
As one of the biggest social media this small "failure" raised my curiosity of trying to know how is the grouping done or could be done?
What could be the possibilities of improvement in this case?
In order to find possible groups of topics, we have crawled the twitter blog collect 100 blogs on which we have applied a cluster analysis and a topic modelling analysis.
In the context of Information retrieval (IR) modelling contextual information in documents search have been subject of several researches [22].

In this report is described the experimental design, process and results to extract hidden structure in our corpus using LSA and KMEANS that will be compared in this report. Blogs categorise blogs and compare with what twitter is actually proposing on its portal. In this research paper, we are trying to achieve a categorization that can have a positive impact on blog's search. In this report, we discuss the challenges with document clustering, through the following questions: What are the challenges cluster labelling? How to find topic labels for clusters.?

Pages: 70
Finished:

When uploading this document to Digital Exam each group member confirms that all have participated equally in the project work and that they collectively are responsible for the content of the project report. Furthermore each group member is liable for that there is no plagiarism in the report.

Exploring methods to find and label latent topic-groups in a blogging environment

Master thesis

Author
Seynabou Ndiaye

Supervisor:
Jannick Kirk Sørensen

September 2017

Acknowledgements

I would like to thank few people who have been helping me during these last 2 years. I sincerely would like to thank the School of Informatics communication technologies for all its facilities. I thank my family for their constant support.

Table of Content

Table of Contents

List of Tables

Table of Figures

List of Abbreviations

NLP	Natural language processing
API	Application programming interface
ATC	Automatic categorisation
CA	Cluster analysis
LSA	Latent semantic analysis
NLTK	Natural language tool kit
TF	Term frequency
IDF	Inverse document frequency
IF	Information retrieval
SVD	Singular value decomposition
PCA	Principal component analysis

Chapter 1: Introduction

Over the recent past years there is a great evolution in the use of social media platforms. People's use of these services can be explained by their perception of the platforms and the way they interact with them. In a hypothetical rational sorting[1], people tend to group social media platforms by looking into the proposed contents. They can subscribe to groups with special interest where specific content is available and shared within the closed group. For services that provide a quick content propagation such as twitter feeds, users tend to be more active to visualize or contribute to posts. Among other hypothesis behind user's motivation of use of social media services cited by Wilkes et al are the affordances including searches[1].

In this context of IR in social media, affordances of effective search features or recommendation are competing motivation that explain user's satisfaction feeling from a website.
Twitter is a successful online social information network launched in July 2006, it comes with an innovative way of publishing information. Its number of actives users are estimated to 328 million [1]monthly active users in 2016. With a growing number of blogs and user's histories. Twitter is today one of the most utilized social media platform for real time information that relates what is happening in the world [2].
Its blogs are gaining in popularity as they allow people to express themselves in a more extensive way unlikely tweets that are short only offering 140 characters.
Twitter blogging is rich of more than1600 of blogs, generating a huge amount of online content of diverse topic. This rich amount of data has created the need for accurate content search, filtering and recommendation.[3].

Because of the rapid growth in use of blogs, the "search" features need to be dynamic to facilitate searches through the blog site.

Here are the main observations after interaction with the twitter blog in may 2017.
1: Twitter Blog presents its collection of documents grouped into six (6) topics concerning the company itself. They are also recommending contents in 2 ways "Recent " and "Popular". After trying to search for specifics contents from the blog, the returned results were not quite accurate. The results were not really related to the search term entered.
2: The other observation is that the site is not proposing on its interface a grouping of blogs posted by the individual users.
My assumption here is that there is a need for organizing and labelling blogs posted by users into groups to help them in the search in a way that the confusion is as less as possible. Performing a clustering analysis of these blogs is a way to define the categories of topics discussed.
Still, in the process of bringing order in the web, if the blogs are clustered both users and the system will benefit from it. Twitter blog will be structured in the way to presents its contents and it will be more convenient for blogger to access resources.
With such an organization, a search query can perform faster as it does not search through the whole content but only inside the corresponding cluster.
Blogs main characteristics are that they are long text size text. In this context of creating blogs, finding similar authors, blogs written by a specific author or which tags are related to which blogs, all those

[1] https://www.statista.com/topics/737/twitter/

linkages are highly related to a well-structured content. Several number of clustering methods exits and they offer different possibilities depending on how much we want to know from our data.
Executing such a query is time and process consuming. It could be more effective if other criteria were considered. Without a doubt, the more explicit are the search terms, the more precise will be the output.
The problem in the context of IR, if we search for the keyword "mobile ", this is ambiguous and can map in different context "Mobile Broadband", "Mobile phone" and so one". The content of short text may vary from users' daily activities to news. If the search results both text the system is still answering correct however because of polysemy mobile can in another context for example refer to "purpose".
"Documents clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms"[4]. The purpose of clustering is to group unstructured documents into meaningful classes where related topic will be in the same cluster. With a careful selection of the stop words, the system groups the reminding words that are considered to be relevant while retrieving information in categories with similar documents belonging to the same group.
In order to tackle the problem of which document should be in which group, it is where to define topics related in each document, clusters them and label them.
Currents problems in IR on blogs are term with nuance. Because the content retrieval attaches a bag of words or tags to the search term and retrieve content based on the tags, words with multiple meaning "polysemy" and words with same meaning "synonyms" cause problems.
One approach is here to reveal hidden tags from the social media [5]
Another approach takes into consideration the context in which tags is used, a semantic tagging is performed and each tag will be grouped considering its context. Even though these are relevant piece of data of the twitter blog, the tags will not be involved as they can bias the way to categorize the dat.

In order to address the problem information retrieval in blog search, a latent semantic analysis will be conducted on the set of documents collected from the twitter blogs, and this could be done with any other blogs.
Our candidate's algorithms are LSA and KMEANS to find cluster. The Elbow method can quickly determine the number of subspaces [6]. LSA have been introduced to complete the lack from VSM on polysemy and synonymy[4], therefore VSM will not be applied. However, LSA main limitations is that it is known its incapacity to handle polysemy effectively. " Due to polysemy, wrong documents could be deemed relevant." [7]. Another limitation for LSA is that it's matrix dimensions' reduction is not random, human judgement in the choice of the number of factors and polysemy can have influence on the results. In fact, because its simplicity LSA method can reveals the first shape of the clusters and lot of information about the data.
In fact, documents are clustered based on the data that we have available and the terms they contain
One step further is taken using K-means which handle polysemy. It allows a better visualization of the clusters will be applied and there we can tell the system how many clusters we would like to ned up with. Xue et al. [8] in a cluster based CF a way of recommending contents based on formed clusters. In their research, they apply k-means algorithm firstly as a pre-processing step to visualize the cluster. There, the clusters are rather limited by distance from one from point to the centroid of another cluster.
Afterword it makes sense to define topics related in those blogs. Topic modelling is a statistical model for generating the observed variables in the document based on the latent variable of the document.[9]. Words in a topic modelling have to make sense and related, for example "navy, ship, captain" and "tobacco, farm, crops. We will be applying the LDA method that is actually the most popular one model in topic modelling.

1.1 Motivation

Our motivation on the choice of the topic is animated by the evolution in the web. The WWW is today the place where everything is taking place shopping job search business marketing, tutorials, blogs and so one. For the web, this unstructured big data could benefit both users and business if they are exploited. As the use of blogs in twitter is increasing more data are posted. Companies advertise through blogs, news, educational reviews are among other contents posted. Blogs popularity is mainly attributed to the fact that they have minimal entry barriers compared to web pages. Bloggers and their followers can though that communication channel interact and foster information sharing[10] . The blogosphere is rich of data, online businesses are creating value in their organization by studying this huge amount of data. Analysing those data typically are used in improving user's satisfaction on online services.
Several searches have already been considering the need of applying semantic search based on the tags. But our approach differs in the sense that in the pre-processing the header which is the title is cut from the text. So, we only study the main text input.
However, blogs contents are not so sparse, they are thus we can study the topics they contain and group them into clusters.
Our motivation on the choice of twitter blogs for the dataset is motivated the richness of its content. Several natural limitations from the short text (tweets)that are restricted in size, have discarded them even though they are they are interesting as subject of research. The dataset candidate for our text mining have been collected from twitter blog but it could be from any other blogging site.
Previous researches have been done using short text , however they are not as interesting as blogs when it comes to apply text mining [11]. Most of bloggers take the opportunities given by twitter environment to drive traffic to their blogs. One huge problem is how to efficiently search while having big volumes of data.
As a subfield of artificial intelligence, machine learning methods such LSA, K-means and LDA can help us today to find concept in unstructured messages.

The choice of the blogs documents is mainly due to the size and the unstructured aspect big data of its content, the information stream is huge and users may be overwhelmed, therefore the need of having methods of grouping twitter huge range of blogs into groups[2].

Our main motivation in this research is that applying logs content analysis can somehow reveals topic discussed topics and organized the content in a way to check the relation between documents and words they contain. Several text analytics algorithms have been implemented to be able to collect store analyse and search for data.

In general, we as users would like to be positively surprised by the service we request for. This raises the following problem are the document clustered efficiently? can we rank the blogs in other innovative ways than most popular and most recent? While considering the clustering methods and topic modelling several literatures have named LSA, and KMEANS as good candidate.
The motivation is that we have here many blogs 1600 approximately and we cannot read them all to know what they are about. We would like to know which topics are discussed in the blogs. How can we can gain knowledge in topics discussed in the blogs? Our contribution would be as a proof of concept is to give quite equal chance to all documents in the blogs search, latent Dirichlet LDA is applied to the dataset

1.2 Problem statement

Looking into document clustering in a blogging environment, we raise few of the questions related to it. In an information retrieval, when as a user, we search blogs related to a given term, it can happen that documents that are retrieved do not contains the search term. In fact, the system takes into account words that are semantically connected. In particular, in the twitter blog, the ways contents are organised should make it easier to access to the content and give equal visibility chances to all the blogs.
Furthermore, the linkage between group segmentation for the blogs and the topic discussed in each of the clusters leads us to review literature for document clustering and topic modelling approaches.
Before we go through the clustering challenges investigations, few questions have been raised. The following research question is addressed by this report:

How to find latent topic-group and label them in a blogging environment?

- How do we organize a document collection clustering into semantically connected keywords?
- How can we evaluate similarity between blogs?
- If we apply semantic analysis strategy, do we get similar grouping as on twitter blogs?
- What are the implications of clustering documents in different ways?
- What is the relation between topics and cluster labelling?
- While applying K-means and LSA which one perform better?

1.3 Project delimitation

Due to the broadness of IR and machine learning it is imperative to fence this work. The project is not dealing with any automated or online machine learning process. It is more about collecting and processing data collected from twitter blogs. It covers two methods of clusters while finding patterns in the data and ensure usability.

- ❖ The project is not addressing search as such, it not about how search engine optimization or implementation design works, but more about the algorithms behind mainly for clustering and finding topics discussed in blogs, in order to give equal chance to blogs in the search process.
- ❖ The project is not only focusing on social media, the experiments is applicable to any other blogs. The choice of twitter is more driven by the need of a concrete case and due its actual importance big data generation.
- ❖ In the context of semantic analysis, we will not go in detail about synonyms, and polysemy however, it will be briefly discussed in relation with the model we have adopted.

1.4 Research objectives

This research expectation is a proof of concept identifying latent structurers in blog posts and define topic developed. Similar blogs should belong to the same group. It also has a focus on the implication of the choice of stop words while finding concept in words.
Along this report, the following contributions could be expected:

- ✓ Application of the Latent semantic method to visualize latent structure in our dataset
- ✓ Apply K-means and evaluate its clustering output
- ✓ Learning the topics through Experiments by applying the LDA
- ✓ Discuss the challenges in documents clustering using the two methods LSA / K-means

From the research questions stated above, one can see that Choosing or finding the right dataset could be tricky as if it is too small then it might be difficult to find patterns in it. For study purposes, we fix a limit of 100 documents so that the computation can be performed on a single machine.

- How can we measure the relevance of the document clustering?
- Does the clusters labels reflects its content?
- Because of time constraints the solution may not be evaluated on a large scale. However, a discussion and a thorough analysis of the clusters of the meaningfulness of the discovered clusters. On the other hand, we will be using algorithm to assess the topic

1.5 Thesis outline

three main aims in this project, first is a state of the art of blogs grouping into topics using unsupervised and unsupervised learning. Second is through experiments to evaluate document clustering, topic modelling algorithms and last bring a discussion on the findings.

The thesis is organized as following: In chapter 1 we present the actual evolution of twitter blogging. In this section, we present our motivations for blogs content analysis in IR. It also underlines basis of our motivation to apply topic modelling and cluster labelling.
In chapter 2 is presented the methodology adopted in this research to collect, process and study the data. Chapter 3 we outline related work to this topic. It revisits information retrieval with a focus on blog search in general. Other scientific papers in the field of topic modelling and cluster labelling will be investigated. Challenges and limitations related to their applications.
Chapter 4 brings an analysis of the problem and here we make decisions about which tools to utilize. Chapter 5 is presented experiments as main contribution, from the dataset collection, processing storage, clustering and the labelling. It eventually in a way of a recommender system rank the search results. Chapter 6 is a summary of results and findings.
The sixth part of this research Chapter 7 is dedicated to discussion about our findings with a comparison with the state of the art algorithm. Chapter 8 conclusion and eventual future perspective are presented in Chapter 9.

Chapter 2. Methodology

In this section is described the procedure in proposing our different topic-groups in twitter blogs. As the starting point of process, the data collection is described in (chap.3.1). The main procedure in collecting dataset. The dataset is collected from the Twitter blogs.

2.1 Web scrapping

The tweeter blogs have been collected first manually for the firsts tests. Since the web page have been updated and now it is presented in four categories but only twitter contents. Therefore, the problem is still actual the blogs we can still ask how are other ways to group the contents posted by users on the blogs so that it will be easier to find around. The first collection has been done in may 2017 since the site have been updated.
Twitter itself has an API to collect tweets but not yet an API to collection blogs data this moment and perhaps this is due to the fact that twitter blogs are quite recent.

2.2 System flow

The system is following these six procedures figure 1 system flow. It starts with a pre-processing of the data and take out elements that is not relevant in our process. The document term matrix includes the stemming and cleaning the stop words this process is can be done several times just to be sure that the reminding words are relevant.

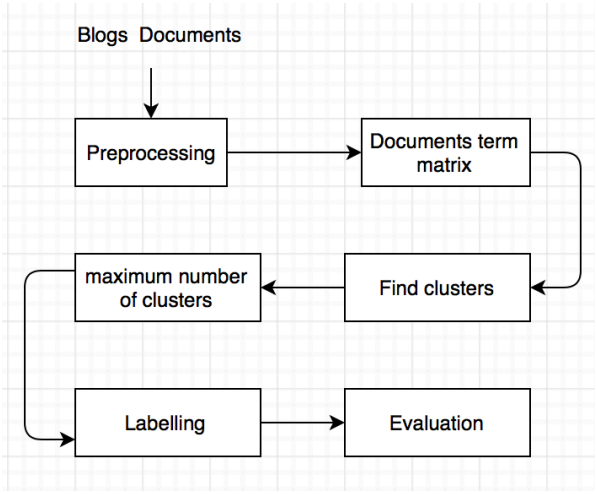


Figure 1:System Flow

After the data blogs documents have been collected then a set of algorithms are applied in order to define the categories in which those blogs could be grouped. For a proof of concept a set of 100 documents was used in our clustering analysis and our topic modelling described in the experiments in chapter 5. This is attempt to group similar document so that when a search is performed using those keywords. The document related to the cluster is part of the output. What about word combination what need to be considered? Can I conduct an unsupervised learning: case tweeter users can search for histories and pages from the tweeter blogs? in this case, it will store all the blogs with their titles and texts" It allows a structured grouping of the documents index for example: indexing document using the same name as its group name. Dataset(blogs.csv). that might be interesting for users. However, it

could be said that the method applied here is quite basic as it is mainly based on word count. Compared to our algorithm the limitation in the term frequency

2.3 Results evaluation

Assessing clustering results is said to be a difficult [2]. However, among other ways to evaluate the cluster, an external evaluation is done with using data from outside that have not been used in the clustering process. A benchmark set which is a pre-grouped class and a comparison is done to see how this one is close to the one we actually finished as an example there is the purity measure. New content can be classified by looking to its nearest neighbours KNN.

[2] https://en.wikipedia.org/wiki/Cluster_analysis

Chapter 3. Related work on cluster analysis

In this section, we review researches algorithms and methods that are available algorithms within blogs clustering. We review information retrieval with a focus on topic grouping in blogs in general, topic modelling and clustering methods. Papers that have been collected in relation to the cluster labelling and topic modelling discuss about how they deal with the choice of number of clusters? How to find the topics? How the labelling is done? And what are the topic modelling challenges. What are the major consideration that we could eventually take into consideration in our thesis.

3.1 Blogging environment

Blogs as place containing unstructured data therefore the need for organisation. They are also viewed as cyberspace where its users come to express their thoughts, pictures and urls. The collective wisdom or wisdom of the crowd is that aspect that users commenting that visitors commixing to the site and commenting and contributing to its content.[12]

Since its starts in march 2006, Twitter social media, it's have been sent 350 milliard tweets, all the tweets theme have been grouped into hashtags from 2007. It has been ranked as the most viewed website by TV2 in 2013[13].
Blogs are compared to tweets allowing more space in the message. Here can online users stay up to date with what is happening in the world wide, they can follow "Persons ". the service can be accessed from the main twitter page twitter.com or from the twitter blog. It is possible to respond, retweet, like or send a direct message to the blogger. Tags and @username example, indicate respectively the context of the message #Twitter and the username @twitter Safety.

Brook and Montanez on their research paper titled "collective wisdom based blog clustering "based on singular value decomposition SVD where they present its advantages and disadvantages. on blogs clustering emphasis on the similarity between blogs[12].One existing blog search employ traditional the clustering based on the text data have been accepted as an accurate to cluster data. The CWBBC (collective wisdom based blog theory is accurate in clustering blogs). The focus was here on the graph theory exploiting communication links based on the words labels [12].

3.2 Supervised and unsupervised machine learning

"Unsupervised machine [3]learning methods are a way to describe hidden patterns from unlabelled data". It is part of machine learning, figure 2 shows the two types of learning. If an algorithm is given a set of input {y1, y2, ….yn} and some output parameters are not known in advance then an unsupervised learning task is performed[14].
The accuracy of the output could be discussed as the data is unlabelled. A popular algorithm for topic modelling in unsupervised ML is LDA. However a problem of modelling short Text with this method have been underlined. [11] Ramage et al from their research on LDA in a blogging context argue LDA can be applied to solve the challenge of topic modelling, content filtering and recommendation[15]. They apply a labelled LDA with a focus on hashtags, contrary to our system that only focus on latent parameters and manually label the groups the clusters that have been depicted.

[3] https://en.wikipedia.org/wiki/Unsupervised_learning

Still, we will be investigating both supervised and unsupervised techniques. Supervised learning (SL) tasks develop predictive model and are based on both the input and the output data. When the ML algorithm have set of inputs {x1, x2, …. xn} and a set of output {y1, y2, …..y_n} then the and our system is trying to predict the output of a new input then is this is a SL classification.
The performance of the classification can is evaluated by measuring the accuracy or misclassification rate [14]

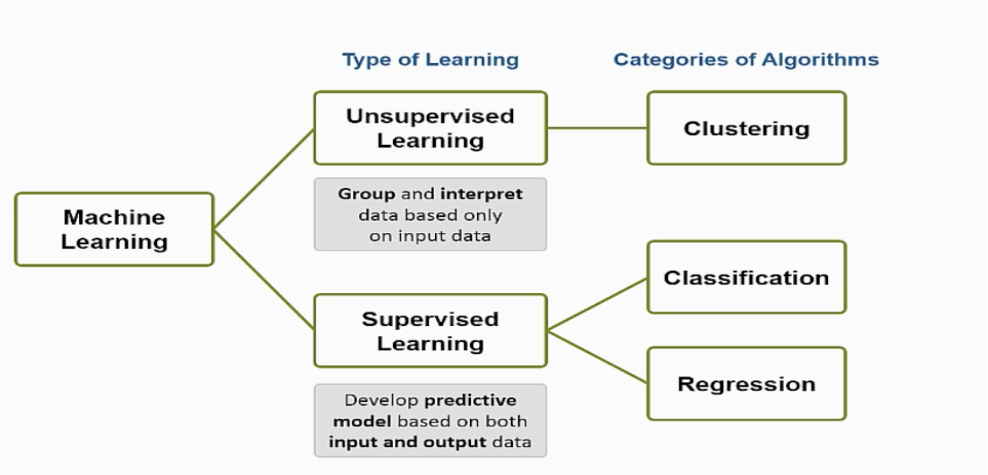


Figure 2 Clustering algorithm a subset of machine learning Clustering[16]

- Classification

It is a task in the machine learning environment that helps in the prediction of data with the aim to group unlabelled document.
As a part of unsupervised machine learning it can be used in several contexts to predict features for example in "user 'segmentation", "image classification ".

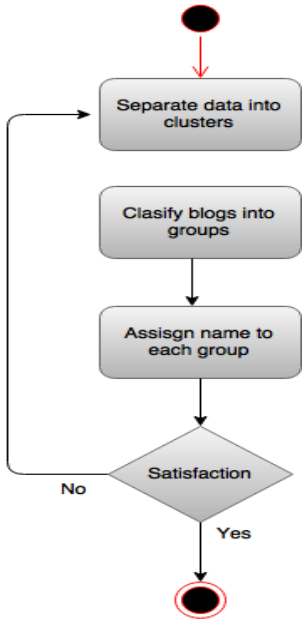


Figure 3 Classification steps

Figure 3 shows the steps for an unsupervised classification where we suppose that our data is unlabelled and we do not to know a lot about the initial data. A specific Classification problem is to guess which group a given observation belongs to.

3.3 Cluster analysis

- Information retrieval IR

The main general flow from user query from document retrieval is shown of figure 4. As motioned in the introduction efforts have been existing in IR in the web in general , "Google's mission is to organize the world's information and make it universally accessible and useful" [17]. It claims in its ordering efforts "The democracy on the web" ref. Votes on a hyperlink is the basis of the PageRank algorithm, as they also underline, those votes do not install tyranny of the majority on the search engine. One can think that older website will accumulate more votes and be more promoted.
Twitter in its way to perform the" global conversation" is contrary to google, identifying most discussed topic and value them more. Perhaps, this gives higher rank to pages that receives most post from friends and surround. However, "twitter algorithm values the diversity of actors".
They emphasis more on the diversity of short text contents and people participating on a topic, there friendship will be counted negative in the algorithm.

We usually want to search for the specific content when we are one blogs site. In such social media, we are interested in to topic associated with a blog, or similar author or blogs that deals with the same topics but are not closely similar. Among consulted papers, researches admit the existence of "aggregated strategies", where researches have applied topic modelling on aggregated and individual topics[15] Ramage et Al argue that it is important to have a better representation of blogs data to help user in finding who to follow and have a better filtering of the topics. LDA is presented here as a promising method that is said that this method models efficiently similarity document.
The main difference about the two approaches is not really discussed [11], however, Deerwester et Al propose a new approach of automatic indexing and document retrieval using Latent semantic analysis LSA. Their techniques is a solution to the conceptual retrieval where users were trying to match words of queries with words of documents [18]. while retrieving data a word can always be used in literally different ways based on the concept and also based on users. A fundamental challenge of current information retrieval methods is that the words stored in the system might not be the same from the search query. Synonyms can be used to express the same object. In fact, describes the fact that there are many different way to define the same given object. Depending on the context, the need, the education level, the same word can be referred to in different words. Studies have also shown that controlling the vocabularies using human capabilities is not fully innocent and is driven by the induvial capabilities.

Polysemy refers here to the fact that generally a word has more than one meaning and its meaning sometime depend on it context.

The implication can be that a research will return contents with documents that might not all interest the user. Moreover, Salton in his book titled "The SMART Retrieval System-Experiments in Automatic Document Processing" state that document clustering can also be used as a means of improving the efficiency of an IR system by pre-clustering the entire corpus and retrieving clusters rather than document[4]

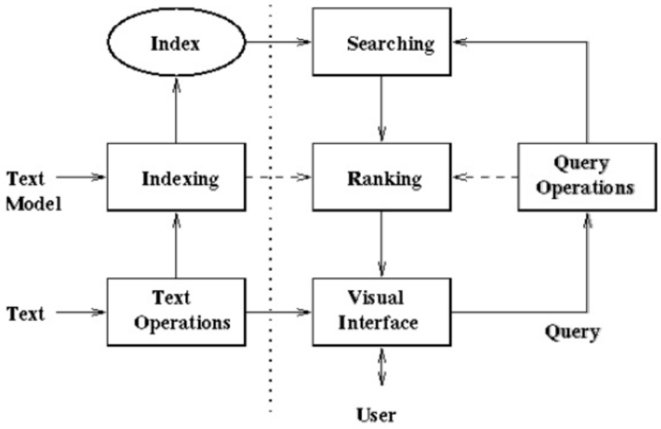


Figure 4[4] Information retrieval flow

- Tag Crowd

Created by Daniel Steinbock a PhD student from Stanford University in 2006 by, Tag Crowd is actually a very popular word count cloud generator. With this tool, online users are able to upload or paste a text and then visualize its word frequency [19]. The words that are listed in the visualization depend on their frequency weight in the document therefore the higher the term frequency the bigger is written the word.
Usually the purpose of a tag cloud is to is to present in a visual format a given content.
Nowadays among other well-known tag cloud providers, we can cite Tagline Generator, Wordle that also like Tag Crowd. They work as a table of content and give the reader an overview of major set of key words and tells about the content.
More, other researches in this area have enhanced this work by putting tags into smaller cloud using spatial algorithm. Those works have been involving clustering algorithms to group tags with similar meaning together.
A research ways to cluster content conducted by Torniai et al uses the size of the tags and colour to described the information. [20].
Another research that tries to address the usability in the way to read clusters propose a delimitation and layout setting by deleting whitespaces between clusters. The way it works is that the layout is created in a way to put together polygons in order to bound the document terms[21].

- Document clustering

"Modelling contextual information in search queries and documents is a long-standing research topic in IR" [22]. Document clustering has a high importance in search in general. It has several application and existing methods in the domain of information retrieval and data mining. As an unsupervised machine learning, document clustering is carried out to group document with the same concept.
Due to the fact that different words can be used to define the same thing, for example "stole" and "chair", semantic analysis is what is presented on most search engine to step over and solve that problem in information retrieval.
Chen & al approach the problem of finding context in document clustering by applying a convolutional pooling structure mechanism CLSM[22] . They propose latent semantic analysis approach (LSA) in a

[4] https://www.slideshare.net/kanimozhiu/tdm-information-retrieval

way that takes into consideration words that are organized on a convectional structure considering that words in the document are organised. The purpose of their research is to define low dimension and semantics vectors in a context of information retrieval.
They have been capturing the contextual information in a sequence of word and at a certain level capture the essence of the sentence.
They apply a nonlinear transformation on the word to collect high level semantic and represent all the word in a full text string on a single vector.
In document clustering, terms usually are represented by independent vectors, in a search the context of the word has a great relevance as in their context [22].

Document clustering is what is in many research paper is applied to find out topics in documents and their context. Clusters will just work as categories which at the search time can be accessed from the label of the category (in) or from the words in each category (out). The cluster label can guide the user and let him know that this is the Concept behind that bag of words.
Assuming that when performing a search a number of criteria are involved then we can consider that it is threefold: the clustering, the topic modelling and the labelling. Each of them as its own can be a full topic, however they will be investigated but not fully developed.
Nowadays, because of the popularity of the search engines, documents clustering is gaining more consideration. Brooks & Montanez research done in the domain of blog clustering is the Blog Tree Clustering have been forming blogs clusters based on the similarity between blogs[23]. They study the similarity of articles within that have the same tag. They focused on tag and their relations with articles. The side that interest us in research is that where "show that clustering algorithms can be used to reconstruct a topical hierarchy ". They look into articles that share most relevant keywords using TFIDF.

Hieratical clustering is among other methods utilized in two ways and agglomerative and a divisive. We are grouping in an agglomerative way to find the subclasses in an already existing group, a divisive segmentation is applied on subclasses to find the highest root. This method could be useful if at the level where already have found the topic and maybe interested in finding related sub topics.

- Latent semantic analysis (LSA)

Document classification is subject of research in different fields. However most of studies were oriented toward tweets or short text.
PLSI and LDA have been used to find latent features in data[24]. The background of applying LSA? When an online search is performed, the user might not even enter the same word at every search query. Example today you can search for automobile and tomorrow you search for car or vehicle or Mercedes, differ word can be used to express the same thing.

Data pre-processing: In this initial step of LSA, the data is cleaned from irrelevant and common English words, that list of words is stored in the stop words text file. The stemming process is taking the root of word to avoid redundancy. This step can be more selective if the search is interested is interested in specific words.
Vectorisation: after a pre-processing the corpus, they are left out with word that have a relevance for their context. Once each document has been cleaned, the remaining words for each document are what define the vector. The document term matrix is matrix with the document in columns and the rows the significant words and their occurrence.
TFIDF: the computation of the word occurrence is an important step in LSA. How many times a word appear in document is define by the term frequency TF. However, knowing that the importance

of a word in document is not only defined by the number of time a word appear in a text. IDF, the inverse document frequency weight the words that appear a lot in the document inversely.

Term Frequency (TF)

$$TF_{i,j} = \frac{f_{i,j}}{\max_j f_{i,j}}$$

where $f_{i,j}$ = frequency of term i in collection j. Collection consists of 11 documents (1 MA and 10 SAs).

Inverse Document Frequency (IDF)

$$IDF_t = \log(1 + \frac{N}{n_t})$$

where $N = 11$ documents consisting of 1 MA and 10 SAs

n_t = the number of documents in 11 answers that contain term, t.

SVD: With Singular value decomposition (SVD), documents are modelled into a document term matrix .[18]. Let's take C= NxD, in this configuration of the documents, N is the number of collections, D is the number of word and C They are terms and documents vectors. is a matrix factorisation applied in LSA for dimension reduction? The initial matrix is reduced by breaking it into three matrix Z, W and Z transpose. The diagonal elements are in descending order.

For a lighter computation, a low rank matrix is used. It is approximately generated by making a selection of singular values that are considered enough to do the selection. The reduced matrix is as following:

Mk = Uk * Sk * Vk'

Where **Uk** matrix only uses the largest values.

It analyses the semantic in the matrix M that have been organized in a term document matrix. The main purpose of this part of the algorithm is to unveil latent structure in the data. "SVD is defined in linear algebra as the factorisation of real and complex matrix". It can be applied to any matrix M as on the example below. The decomposition below is showing the decomposition of that rectangular matrix into three matrices each with specifics characteristics. This analysis begins with document term matrix with the documents on the row and the terms in the columns. When the matrix is broken several elements of the output are very small and do not contains enough information to be used. Those small data could be ignored and set focus on limited but essentials elements. Those values can even be visualized in a spatial or geometrical configuration.

Even though SVD has wide aspect in information retrieval, we will be interested in eigenvector value decomposition and the minimum values that ca represent the data.

The lower dimension matrix contains small number of factors. In this representation, terms are represented as Eigen's values and documents represented as eigenvectors. In IR, the dimension reduction is done thanks to SVD.

Dimension reduction: The main purpose of the SVD mathematical computation is to reduce the dimension of the initial matrix. With this mathematical computation, it is possible to reduce the size of the matrix to its principal components, in other depending on the context it is called PCA or eigen values decomposition.

Semantic categorisation: By semantic characterisation is that terms or documents belong to the same concept. Tan Ping et Al propose a semantic characterisation (SemC) where they make use of the LSI algorithm to pull out information that are semantically connected. The research is based on an ***explicit labelling of the text categories***, which is based on expert's vison of semantic structure. The LSI model has a great emphasis on the knowledge contained in the training set[25] category label are derived from expert label categorization. However, could this be applied generally on any data set.? A supervised text categorization is applied on the with their study they improved their categorisation results.

Main challenge with this cluster labelling method is that this procedure count highly on human knowledge of implicit category labelling.

Stop word choice: Reorganizing text documents into clusters require clean the data from words with little meaning. Words that are used in the modelling processing are heavily dependent on what the choice of the stop word. St
A research conducted by Inderjit et Al[26] state that in parallel the document processing, words that occur < 0:2% and > 15% of the documents are removed.

- K-means: Finding the k clusters

This algorithm is a very common clustering method in text clustering. It Is initialized by guessing random points and assign them as class, the process can be repeated [27]. Several algorithm to cluster exists , but Fry et al in their researches show that "K-means clustering performs better than Peak-searching clustering in terms of grouping similar reviews based on topics" [28].
"K-means is based on the idea that a centroid can represent a cluster. After selecting k initial centroids, each document is assigned to a cluster based on a distance measure, and then k centroids are recalculated. This step is repeated until an optimal set of k clusters are obtained"[29].
As a supervised machine learning algorithm. We start by defining in the input the number of clusters we want to end up.
For each elements of the dataset all the computation from the data to the centroid is done to see in which cluster the data belongs to. If the Euclidean distance is applied, then the closer a data is to a given centroid then the it belongs to that cluster.
Finding the right **k** using K-means is an iterative partition algorithms that is suitable in handling large datasets due to their relatively low computational requirements [4]. One main challenge underlined when applying K-means is the choice of the number of cluster[30] .

- K-NN

With this algorithm the unknown sample is classified by the closest neighbour [27]. Some previous work that have been done on this simple linkage method.
In pattern recognition, an unknown blog can be classified by looking to its closest neighbours going through the whole computation is not needed. The nearest neighbour can be computed based on similarity measures as the Euclidean distance[31].

- Elbow method: Finding Optimal number of clusters

The elbow [5]method is an algorithm that is relevant in finding the optimal number of clusters **k.** It is a 2 dimensions' plot where the Percentage of Variance Explained PVE is a function of a number of clusters. the PVE is a ratio between the group variance and total variance. Zhang et al, in their research paper have been combining elbow with LBF an hybrid linear model to determine number of clusters. The "elbow" of the curve , defines a K number of cluster which adding more clusters has no effect in decreasing the error[6].

- Cluster labelling

A recent research on topic category analysis shows a clear difference between the collected data and the meaningful data is underlined [24].
Indeed, the application of documents clustering improve the effectiveness in IR, Contents is organized in a way to minimize query processing time. Defined as "a method for finding and tracing clusters of words (called "topics" in shorthand) in large bodies of text[32]. the main purpose is to group documents this could be in the context of business analysis where the enterprise needs to learn about customer insight or to be relevant in the search results displayed after a search query. The relevance of a document is not solely based on the terms in that given document. In fact, the semantic analysis helps here to also retrieve documents that even do not contain the search query but is found to be semantically connected to search term entered.
In traditional document retrieval environment, a search query will fail in retrieving which do not include the search term.
As stated, earlier the main motivation is to give equal chance and visibility to all blogs documents. Document dealing with a specific topic do not necessarily have those specific words in the text.
In this context, a number of algorithms have been proposed to cluster and label the groups.
Other great challenges in cluster labelling are finding the title for the cluster group and how to address the problem of overlapping topics.
There are two possibilities, one is directly involve the user in a requirement document the topic groups that can be of importance. In this case, some system features unknown from the user perspective might be unconsidered which can lead to a partial solution of the problem.
By involving, the user, a card sorting activity could be introduced as [1], where 59 participants invited to the workshop were asked through a drag and drop interface to group and label
The other one is to mathematically address the problem by using some computation that will filter the highest label name. in this situation, the output might be more reasonable as it known a little about the user and the system.
Our work will focus more on this second aspect of labelling because we have not been in contact with any user. What are the metrics to a meaningful cluster label? Labelling a cluster require to understand what the vocabulary in the cluster of word. Under a document clustering analysis, each document is transformed into a vector of words ……

[5] https://en.wikipedia.org/wiki/Elbow_method_(clustering)

A number **N** of most relevant (most appearing) keyword from a cluster can be selected and study the topics they have in common. This can be done manually reading through the cluster or done mathematically using TF-IDF.
This labelling is not a step as its own because human comprehension of the topic influence the choice in the stop words and thus the remaining words in the cluster. On advantage is that in this process is that it can lead to a comprehensive storage of the documents since they will be group. The ambiguity is that this process is not solving all the problems, if the term is too specific.
After we have clustered the documents, we model the topic. This process in powerful in order to when it comes to identify latent topic from a dataset.

Labelling the document can either be performed explicitly or automatically. In the process of automatically generating a cluster label. Considering the scenario where a user enters one query term to search for a document, retrieving the information from the system require knowing which clusters cover those that word. Therefore, the return result is not only from one cluster.

3.4 Similarity measure

A persistent problem in the domain of unsupervised pattern recognition is the similarity between documents. In order to see in which category a document should be placed an analyse of the content in a word document could be done. What does it consist of? Let's say that we have a set of documents, the similarity measures are ways to check out how similar or different two documents are.
Generally existing there are techniques are not said to be just for one type of text document.
Such measures are quite important in the blogs data mining knowing that certain blogs do not have a certain originality, their work is just a replication of what is already there.

Density based clustering algorithm , with this clustering algorithm dense point shows that there is a cluster therefore rely this method on the similarity measures [4]. With this method, clusters are located in dense area. A dense point reflects close of a data point to neighbouring. The similarity of the two blogs content can be evaluated by comparing their vectors. A trivial propriety of document similarity is two documents d1 and d2 has a cosine similarity equal to one 1 are identical.

- Euclidian distance

This measure (Figure 5) of the distance between two points is in this context a metric while evaluating the similarity between two documents. As a geometrical solution can measure the distance between points in a two or three-dimensional space.
It is trying to bring solution to one of the main problem in clustering which is the classification. K-means apply it for the measure of the distance between a centroid a document.
It is considered as a reliable metric due

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Figure 5 [6]Euclidian distance computation

- Cosine similarity

This measure is applied on vectorised documents. The cosine similarity is the measure that compute the cosine of the angle between two vector (directions) on the space.
It will take into consideration the directions of the vectors. Here, the angle between two vectors (blog a) and (blog b) in figure 6 inform about the similarity. The short the angle between then the higher is their similarity. With the computation perform using the following equation, 0 tells that the vectors are very dissimilar where 1 tells that the two items are very similar.

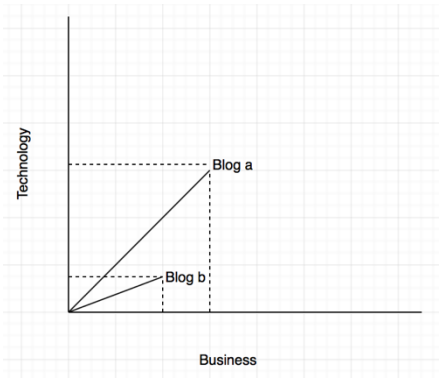


Figure 6 Cosine similarity of 2 blogs a and b

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Equation 1 Similarity measure eaquation

3.5 Topic modeling

Topic modelling has recently attracted particular interest. It is a statistical model for generating the words in a document based on the latent variable[9]. Generally assumed that a document dataset has a small number of topics with different word frequencies[24]. Other researches have been performed in "Evaluating Tag Quality for Blogger Modelling via Topic Models" to detect spam using LDA [33] and determine if a tag assigned to a blog is spam or not.
A semantic analysis of the blogs can lead us to the necessity to see if there are similar blogs within the same clusters. A hypothesis that we can experiments is that documents from same clusters should have similarity and documents from different clusters should have low similarity.
Other approaches in this context is the measure of similarity between topics [11].
An approach named Wisclus [34]have been proposed by Agarwal et al where they cluster blogs taking into consideration the link strengths and labels hierarchy. By comparing Wisclus and SVD they came to

[6] http://www.saedsayad.com/k_nearest_neighbors.htm

the conclusion that future works in the field need to consider multiple source of information in the blogosphere like tags, and labels. This approach might not be relevant as the document is long enough to reveal context.

- Latent Dirichelet analysis (LDA)

LDA is a powerful topic modelling tool in the analysis of textual data in the context of that is help to discover topics in group of documents. It has been introduced by David
Blei in 2003, is a probabilistic topic model based upon the hypothesis that documents are mixtures of topics, where a topic is a probability distribution over words [5]
Pushman and Scheler illustrates distinctly some strengths and weaknesses of LDA [35]. This method is an evolution of the latent semantic model and offers the possibility to offer a probabilistic clustering of documents. As a probabilistic method, it computes the similarity measure considering the probability of the text having the same topic. LDA takes into consideration the fact that in a document it can happen that more than one topic is related specially in blogs topics can overlap.
Not just the concept within a document, one word can belong to more than one category. "LDA does not output interpretable labels of the learned topics. Usually, topics are represented by the top 10 most probable terms generated by each topic "[35]. One main challenge that have been mentioned is the is lack of large scale training data. However, they underlined that the problem can be solved using the by using external data.

Chapter 4. Analysis

In this section, we analyse the two approaches used for the evaluation of our clusters that will be studied in our experimental work. This analysis aims to review the system and motivate our choice of algorithm and tools. We examine the problem that will help us later on the choice of for our experiments. This analysis will help to understand the requirements for a cluster analysis process.

4.1 Cluster analysis

"*Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).*"[7]
At the beginning, we do not know which blog belong to which group even we do not know how many categories we can group them. The process of cluster analysis figure 7 start from the cleaning of the data to the clusters.We take this analysis as in the basis that more terms give us a data sample that is long enough to conduct a semantic analysis.

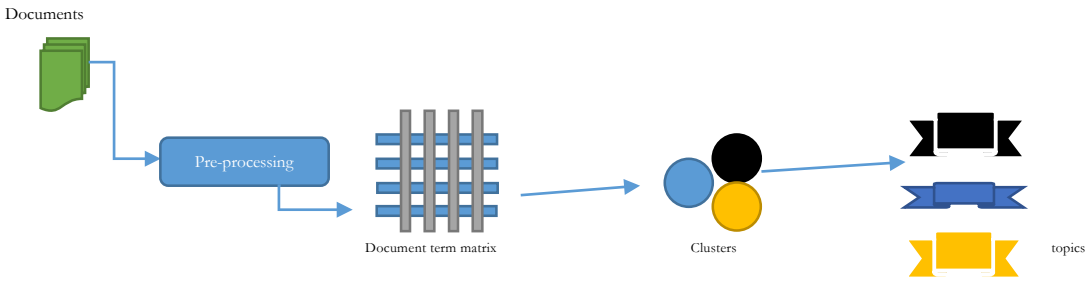


Figure 7 Clustering analysis process

- SVD analysis

The dataset is run through a semantic analysis, where we get a documents terms matrix. I order to analyse the matrix content, a SVD singular value decomposition is performed to bring the matrix in a structure that can lead a lower dimension of our matrix. The square matrix obtained is then decomposed into a product three other matrices. All the eigenvectors that are very small do not almost contains any information therefore some of them can be ignored during the process.
Finally, the documents that we started with are reduced to smaller number of factor that facilitate the computations. There from, the similarity between to document can be studied with a geometrical representation. Thus, the result can be represented geometrically by a spatial configuration in which the dot product or cosine between vectors representing two documents corresponds to their estimated similarity.

[7] https://en.wikipedia.org/wiki/Cluster_analysis

Methods	Applications	Performance & comments
LSA	Search & retrieval	
	• (Herdiyeni 2009) [24]	Automatic image annotation.
	• (Chen et al. 2008) [14]	Medical image retrieval
	Automatic essay grading	
	• (Kakkonen et al. 2006) [29]	LSA > PLSA
	• (Kakkonen et al. 2008) [30]	LSA > PLSA, LDA
	Spam filtering	
	• (Gansterer et al. 2008) [21]	VSM > LSA
	• (Sun et al. 2008) [50]	LSA + SHA > KNN[1)]
	Topic detection	
	• (Sidorova et al. 2008) [47]	Abstracts data of journals

Models	Characteristics / Limitations
Latent Semantic Analysis (LSA) [18, 34]	Characteristics • Reduces dimensionality of *tf-idf* using Singular Value Decomposition. • Captures synonyms of words. • Not robust statistical background. Limitations • Difficult to determine the number of topics. • Difficult to interpret loading values with probability meaning. • Difficult to label a topic in some cases using words in the topic.

Table 1 LSA performance and limitations [4]

The table above is from an evaluation of some clustering algorithm and the general finding that can be concluded is LSA is a great tool but not enough to visualize clusters.

Other observations are that CA itself is a better in understanding the data and if we are expectation better knowledge of our corpus CA is not enough. It is recommended to have a combination of algorithms that analyse the clusters that have been found.

- Comparing two blogs posts' similarity

In order to investigate the prevalence of how much two blogs are similar in topics, it will be conducted in this paper a document similarity analysis. The dataset is collected without the title to avoid a biased labelling. The experiments that have been first applied is the LSA clustering. Each document is represented is represented as vector.
In order to know if the clusters that have been drawn were relevant enough, we decide to apply a given number of cluster using K-means.

- Cluster evaluation

Assessing clustering results is said to be a difficult [8]. However, among other ways to evaluate the cluster, an external evaluation is done with using data from outside that have not been used in the clustering process. A benchmark set which is a pre-grouped class and a comparison is done to see how this one is close to the one we actually finished as an example there is the purity measure. By looking into figure 8 a supervised learning can be performed to evaluate the clusters. New content can be classified by looking to its nearest neighbours KNN.
The coherence of a clustering can be measured with the **PURITY and ENTROPY** metrics.

[8] https://en.wikipedia.org/wiki/Cluster_analysis

Purity

The entropy does not only consider the number of elements in a cluster. Generally smaller value of the entropy result related good clustering result. The quality of a clustering result can be evaluated using purity and entropy. With **entropy,** the cluster labelling is done looking into terms that dominate inside the cluster and the same category labelled can be assigned to other clusters that present the same dominant words. The general assumption with **purity** evaluation is that all samples are predicted to from the same cluster and entropy measures the distribution category in a same cluster

Figure 8 Twitter blog page interface June 2017

4.2 Labelling analysis

This could be done manually or automatically by looking inside the cluster. Clustering as indicated earlier in the beginning, groups documents that share the same concept.

The cluster labelling analysis follows that cluster analysis step

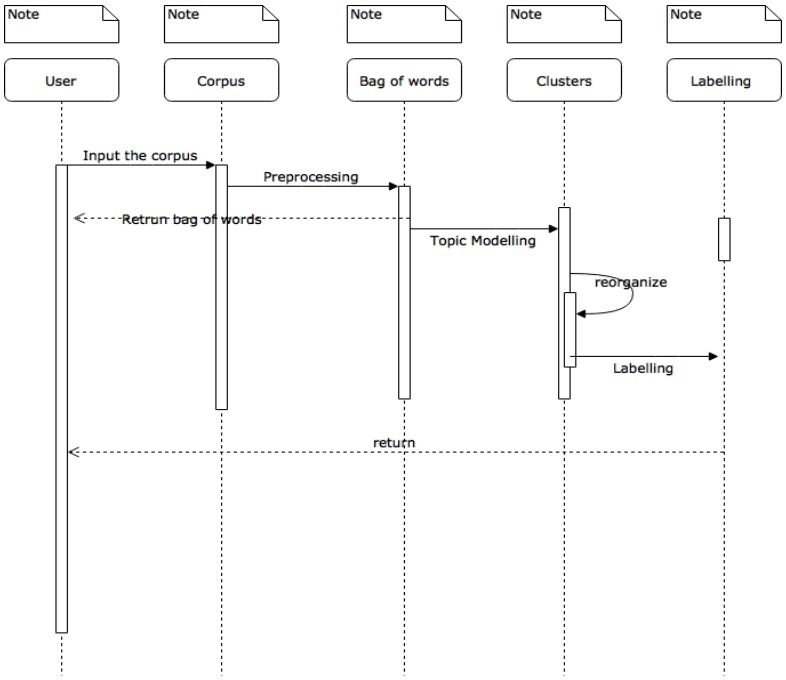


Figure 9 Cluster labelling sequence diagram

In the process of labelling cluster, the main challenge remains to find title that are both meaningful in the context of IR and also for any USER.

4.3 Hypothesis

Hypothesis 1: The main hypothesis in this research is that two blogs belonging to the same clusters must have a higher cosine similarity than two blogs from different clusters.

The quality of a clustering result was evaluated using two evaluation measures—purity and entropy, which are widely used to evaluate the performance of unsupervised learning algorithms

Hypothesis 2: By computing similarity between two document belong to the same tag, the results should be close to 1 Different clusters should not return the same label.

Chapter 5. Experiments

5.1 Dataset

Our corpus has been collected from the Twitter blog https://blog.twitter.com. The collection's process has been described in the methodology part (Chapter 2). At the beginning of the research the first challenge was the choice of the dataset. The idea started with the aim of laying our research in the context of machine learning and document clustering of blogs. Our first choice was turned on tweets, however some literature consulted have depicted them as short dataset that are not so convenient to work with in semantic analysis.
Other investigation on the web led us to the twitter blog website where we could see that weblogs have been emerging. Because of their accessibility and their impact on the society blogs are exploding and are great source of data analysis. Therefore, our aim of visualising the challenges related to blogs clustering has been done using the following dataset.

We started by going on the website and collect few content to start with. Moreover, beautiful we have scrapped figure 11 the data from twitter website as we did not find API to pull twitter blogs content.

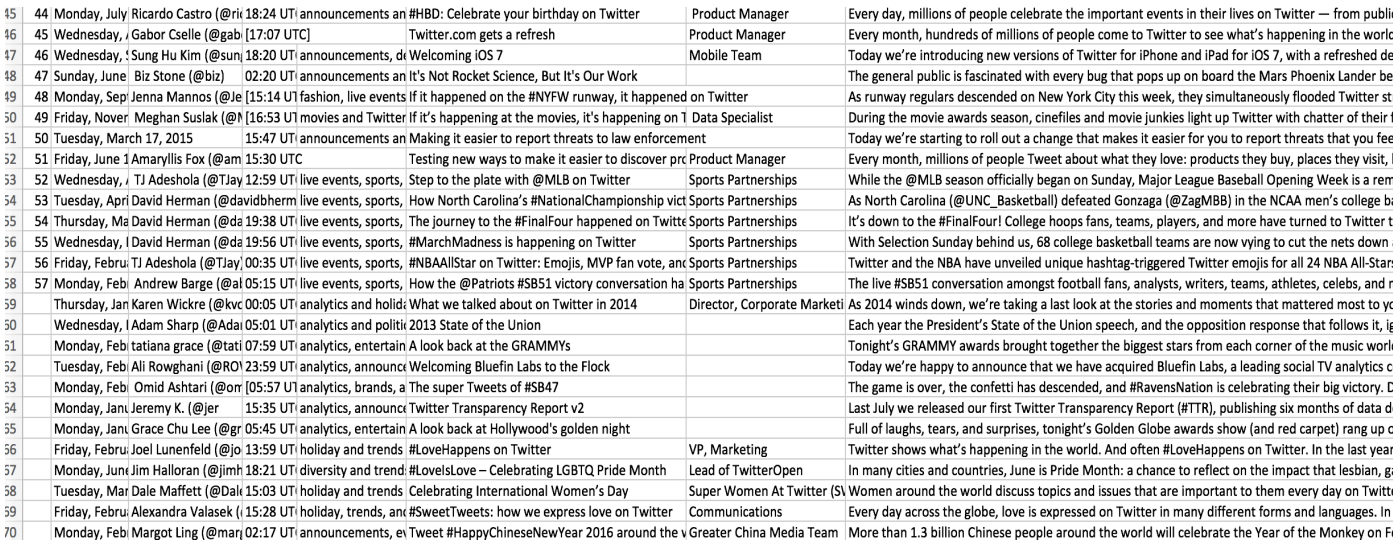
We have been able to access blog posts from different tags. Note that from recent update it has been grouped into four categories twitter company, insights and products and events "Twitter interface" Using the python library Beautiful soup we have been scrapping blogs post.

The page has been downloaded using the GET request to the link https://blog.twitter.com/official/en_us/topics/events/2017/Eid2017.html we could access the server without any need for authentication.

```
import requests
from bs4 import BeautifulSoup

page = requests.get("https://blog.twitter.com")
#page = requests.get("https://blog.twitter.com/official/en_us/topics/e

soup = BeautifulSoup(page.content, 'html.parser')
#list(soup.children)
html = list(soup.children)[2]
#list(html.children)
body = list(html.children)[3]

list(body.children)

p = list(body.children)[1]

soup = BeautifulSoup(page.content, 'html.parser')
soup.find_all('p')
soup.find_all('p')[0].get_text()
```

Figure 10 Twitter blogs web scrapping

Blogs paragraphs are fetched by looking through the html web contents and its children. We will only collect tags that contains block of text, date posted, tag name authors and current position of the author.



45	44	Monday, July	Ricardo Castro (@ric	18:24 UT	announcements an	#HBD: Celebrate your birthday on Twitter	Product Manager	Every day, millions of people celebrate the important events in their lives on Twitter — from publi
46	45	Wednesday, /	Gabor Cselle (@gab	[17:07 UTC]		Twitter.com gets a refresh	Product Manager	Every month, hundreds of millions of people come to Twitter to see what's happening in the worlc
47	46	Wednesday, !	Sung Hu Kim (@sun	18:20 UT	announcements, d	Welcoming iOS 7	Mobile Team	Today we're introducing new versions of Twitter for iPhone and iPad for iOS 7, with a refreshed de
48	47	Sunday, June	Biz Stone (@biz)	02:20 UT	announcements an	It's Not Rocket Science, But It's Our Work		The general public is fascinated with every bug that pops up on board the Mars Phoenix Lander be
49	48	Monday, Sep	Jenna Mannos (@Je	[15:14 UT	fashion, live events	If it happened on the #NYFW runway, it happened	on Twitter	As runway regulars descended on New York City this week, they simultaneously flooded Twitter st
50	49	Friday, Nover	Meghan Suslak (@M	[16:53 UT	movies and Twitter	If it's happening at the movies, it's happening on T	Data Specialist	During the movie awards season, cinefiles and movie junkies light up Twitter with chatter of their f
51	50	Tuesday, March 17, 2015		15:47 UT	announcements an	Making it easier to report threats to law enforcement		Today we're starting to roll out a change that makes it easier for you to report threats that you fee
52	51	Friday, June 1	Amaryllis Fox (@am	15:30 UTC		Testing new ways to make it easier to discover pro	Product Manager	Every month, millions of people Tweet about what they love: products they buy, places they visit, l
53	52	Wednesday, /	TJ Adeshola (@TJay	12:59 UT	live events, sports,	Step to the plate with @MLB on Twitter	Sports Partnerships	While the @MLB season officially began on Sunday, Major League Baseball Opening Week is a rem
54	53	Tuesday, Apri	David Herman (@davidbherm	live events, sports,		How North Carolina's #NationalChampionship vict	Sports Partnerships	As North Carolina (@UNC_Basketball) defeated Gonzaga (@ZagMBB) in the NCAA men's college ba
55	54	Thursday, Ma	David Herman (@da	19:38 UT	live events, sports,	The journey to the #FinalFour happened on Twitte	Sports Partnerships	It's down to the #FinalFour! College hoops fans, teams, players, and more have turned to Twitter t
56	55	Wednesday, I	David Herman (@da	19:56 UT	live events, sports,	#MarchMadness is happening on Twitter	Sports Partnerships	With Selection Sunday behind us, 68 college basketball teams are now vying to cut the nets down
57	56	Friday, Februa	TJ Adeshola (@TJay	00:35 UT	live events, sports,	#NBAAllStar on Twitter: Emojis, MVP fan vote, and	Sports Partnerships	Twitter and the NBA have unveiled unique hashtag-triggered Twitter emojis for all 24 NBA All-Stars
58	57	Monday, Feb	Andrew Barge (@al	05:15 UT	live events, sports,	How the @Patriots #SB51 victory conversation ha	Sports Partnerships	The live #SB51 conversation amongst football fans, analysts, writers, teams, athletes, celebs, and r
59		Thursday, Jan	Karen Wickre (@kvc	00:05 UT	analytics and holida	What we talked about on Twitter in 2014	Director, Corporate Marketi	As 2014 winds down, we're taking a last look at the stories and moments that mattered most to yc
60		Wednesday, I	Adam Sharp (@Ada	05:01 UT	analytics and politi	2013 State of the Union		Each year the President's State of the Union speech, and the opposition response that follows it, ig
61		Monday, Feb	tatiana grace (@tati	07:59 UT	analytics, entertain	A look back at the GRAMMYs		Tonight's GRAMMY awards brought together the biggest stars from each corner of the music worl
62		Tuesday, Feb	Ali Rowghani (@RO	23:59 UT	analytics, announc	Welcoming Bluefin Labs to the Flock		Today we're happy to announce that we have acquired Bluefin Labs, a leading social TV analytics c
63		Monday, Feb	Omid Ashtari (@om	[05:57 UT	analytics, brands, a	The super Tweets of #SB47		The game is over, the confetti has descended, and #RavensNation is celebrating their big victory. D
64		Monday, Jan	Jeremy K. (@jer	15:35 UT	analytics, announc	Twitter Transparency Report v2		Last July we released our first Twitter Transparency Report (#TTR), publishing six months of data d
65		Monday, Jan	Grace Chu Lee (@gr	05:45 UT	analytics, entertain	A look back at Hollywood's golden night		Full of laughs, tears, and surprises, tonight's Golden Globe awards show (and red carpet) rang up c
66		Friday, Februa	Joel Lunenfeld (@jo	13:59 UT	holiday and trends	#LoveHappens on Twitter	VP, Marketing	Twitter shows what's happening in the world. And often #LoveHappens on Twitter. In the last year
67		Monday, June	Jim Halloran (@jimh	18:21 UT	diversity and trend	#LoveIsLove – Celebrating LGBTQ Pride Month	Lead of TwitterOpen	In many cities and countries, June is Pride Month: a chance to reflect on the impact that lesbian, g
68		Tuesday, Mar	Dale Maffett (@Dal	15:03 UT	holiday and trends	Celebrating International Women's Day	Super Women At Twitter (S	Women around the world discuss topics and issues that are important to them every day on Twitt
69		Friday, Februa	Alexandra Valasek (	15:28 UT	holiday, trends, an	#SweetTweets: how we express love on Twitter	Communications	Every day across the globe, love is expressed on Twitter in many different forms and languages. In
70		Monday, Feb	Margot Ling (@mar	02:17 UT	announcements, e	Tweet #HappyChineseNewYear 2016 around the v	Greater China Media Team	More than 1.3 billion Chinese people around the world will celebrate the Year of the Monkey on F

Figure 11 Dataset structure

Blogs csv is the dataset that have been directly collected from twitter blogs. At the time it was collected, it was grouped in 6 different groups security (6), design (7), policy (18, announcements (206), analytics (41), mobile (33). AS they already propose a grouping, we in the perspective of trying to find more categories and compare their grouping with what we propose, we will investigate and discuss the challenges in text clustering and topic modelling. This dataset is not too huge we could decide to save it in a database but this is no relevant at this level as we are not performing any query to directly to the system. As shown on figure 11 above the list the blog text has at minimum 1000 words. Still in their previous interface, twitter blog had the tags we see above "announcements", "security, "best practices" and so one. this dataset could be sampled from any other platform where a set of documents could be collected. Not that main page where we had all the tags does not exist anymore https://blog.twitter.com/official/en_us/tags/ and the information architecture have been updated since august 2017.

5.2 Tools

- **Matlab**

MATLAB is a multi-paradigm and fourth generation program language. It allows matrix manipulations, plotting of functions and data, implementation of algorithms creation of user interfaces. It can be used others programming languages like c, java and python[9].
This tool is one Mathwork[10] products to solve mathematical and engineering problems. It contains a number of sharp toolbox for image processing programming and so forth. Is has also rich libraries for computing and data modelling. Our choice of Matlab in this process have been motivated by its packages.

- **Anaconda Python distribution**

This freemium is a python distribution, it can also be used with R programming language. The main motivation behind this product is to simplify package distribution. This data science ecosystem is large of 4.5 million users[11].

In this project, python has been used with beautiful soup parser in web scrapping process. Working with this environment. In the clustering process the following python libraries have been used:

nltk: this tool is a platform that provides the corpora resources such as WordNet , it also include a number of text processing libraries for example the stopword and stemming tokenizing processing.

numpy:" NumPy is the fundamental package for scientific computing with Python. It contains among other things a powerful N-dimensional array object sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, and random number capabilities"

matplotlib: "*Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits".* [36]".

sckit-learn: we are interested in the whole package. It contains the tfidfvectorizer and this class transforms our raw document collection into (TF-IDF features) statistical numbers making it ready and useful for the k-means step.

5.3 Experimental Design and Process

In the first experiment, we need to define our clusters.
The clustering has been done firstly using the latent semantic analysis model LSA and afterwards, we did the clustering using K-means. On a second part, the top words resulting from the clustering will serve as input in the topic modelling (Figure 10). Thereby, in the evaluation part if the user has been including, then a number of clusters could be presented to them and they will be asked to label them. For example, they could be asked to give the maximum number of topics could be extracted. For each of the clusters the topic are widely presented in the second part of the experiment. Human evaluation

[9] https://en.wikipedia.org/wiki/MATLAB
[10] https://se.mathworks.com/products/matlab.html
[11] https://www.continuum.io/what-is-anaconda

or judgement on the formed clusters may not be highly reliable. We proceed by manually collecting the data from the clusters and visualizing the top words in tag crowd.

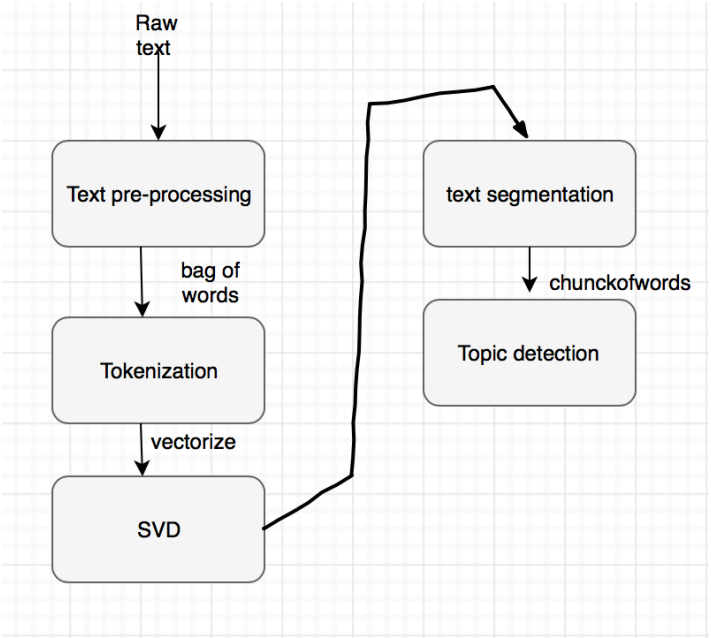


Figure 12 Experimental Design

We try to find latent structures in the blog posts dataset. In this process of clustering the will try to identify the challenges and how these two methods LSA and K-means perform.
While classifying the blogs, we will measure to which extend documents from same cluster can be similar using similarity measures. The hypothesis derived from the analysis in chapter 4 are the basis of our experiments.

5.4 Clustering with LSA

As we already have collected our blogs, we want to find groups of blogs that are semantically connected. This is motivated by the aim to find patterns by take into consideration synonyms. LSA is an algorithm to discover hidden concept in documents.

In the latent semantic analysis, documents can be represented as a vector which gives the possibility to compare the similarity between them. Each document is characterized by a set of terms that tells about topics approached in the document. Therefore, a document to document search and also a term by term evaluation could be conducted. This could be helpful while doing document classification. What about word association in this case to better know what the search is really about. The function stop words ae irrelevant words that occur lot of time. As mentioned in section Chapter 3 where we have described what we collated as data. The myLSA class has the main steps needed for LSA available on Matlab webpage [37]. this class is the starting point

- Pre-processing

In this step, we prepare our data in order to keep only words that will have for the document, general English words as "and", "if" and so forth that leads to no information are cut off.

$D = \{d_1, d_2, ..., d_N\}$ is our corpus also the corpus and $W = \{w_1, w_2, ..., w_n\}$ be the set of word or dictionary obtained after applying natural language process NLP such as tokenization , stop words and stemming [7]

1. **Tokenized function**: this function is taking out the stop words we have predefined, set the rest of the words into lowercase, and stemming, remove numbers from [0 9], drop all the empty cells. It splits our blog text into word tokens using white space.

```
/Users/seynabou/Documents/MATLAB/masterproject/myLSA.m
function tokenized = tokenizer(self,strarray)
    % Tokenize the text in a cell array of strings using
    % delimiters, stemmer, and stopwords defined in the properties

    % if it is a single line of text, place it in a cell
    if ischar(strarray) && size(strarray,1) == 1
        strarray = {strarray};
    end
    % standardize to lowercase
    strarray = lower(strarray);
    % remove numbers
    strarray = regexprep(strarray, '[0-9]+','');
    % tokenize text by delimiters
    tokenize = @(x) textscan(x, '%s', 'Delimiter', self.delimiters);
    tokenized = cellfun(tokenize, strarray);
    % drop empty cells
    dropEmpty = @(x) x(~cellfun(@isempty, x));
    tokenized = cellfun(dropEmpty, tokenized,'UniformOutput', false);
    for i = 1:length(tokenized)
        % remove stopwords
        tokenized{i}(ismember(tokenized{i},self.stopwords)) = [];
        % stem words
        tokenized{i} = cellfun(self.stemmer, tokenized{i}, 'UniformOutput', false);
    end
```

Figure 13 Tokenizer function

2. The stopwords function: the choice of the stopwords have a non-negligible influence on our final docterm. At a first step , it will also remove most frequent words from the document.

```
function stopwords = get_stopwords(url)
    % Download stopwords from the web
    if exist('stopwords.txt', 'file') ~= 2
        websave('stopwords.txt',url)
    end
    stopwords = fileread('stopwords.txt');
    stopwords = strsplit(stopwords,',');
    fid = fopen('stopwords.txt');
    stopwords = textscan(fid,'%s','Delimiter',',');
    fclose(fid);
    stopwords = stopwords{:}';
end
```

Figure 14 Stop word function

3. The stemmer function: with this algorithm strimmer, the suffix. Still after applying the process the output need to be careful studies otherwise that are relevant will remain there filling our corpus. After applying

the standardisation, words to be stemmed are put in an array of characters. In this process, it is not all the words that are process, we fix the of words going inside the strimmer at two (2).

```
function stemmer = get_porterstemmer(url)
    % Download Porter Stemmer from the web
    if exist('porterStemmer.m', 'file') ~= 2
        websave('porterStemmer.txt',url);
        movefile('porterStemmer.txt','porterStemmer.m')
    end
    stemmer = @porterStemmer;
end
```

Figure 15 Stemmer function

To tokenize the documents as we already have our dataset and in initiated our myLSA class, then we can tokenize by call the function tokenizer.

tokenized = LSA.tokenizer(documents)

DOCUMENT-TERM MATRIX: Each document dj is represented as a vector in an-dimensional vector space[7] where each term t_k of the document d_j is weighted w_{kj} . This step has several issues, one is to define which term is relevant for the document and not only for the whole corpus. It has been assumed with IDF that terms that are present but rare in the document are not irrelevant. Words that appear a lot in a document are not exactly more important than word that appear few times in the document.

```
[word_lists,word_counts] = LSA.indexer(tokenized);
```

```
function docterm = docterm(self,words,counts,minFreq)
    % Create a document-term frequency matrix from word lists and
    % word count vectors

    % if minFreq is not given use all the words
    if nargin == 3
        minFreq = 1;
    end
    % if the vocab property is empty
    if isempty(self.vocab)
        % set it to the unique words from all documents
        self.vocab = unique([words{:}]);
    end
    % initialize docterm where rows = doc, cols = words
    docterm = zeros(length(words),length(self.vocab));
    % populate docterm with word count for each doc
    for i = 1:length(words)
        cols = ismember(self.vocab,words{i});
        docterm(i,cols) = counts{i};

    end
    % drop any words that didn't meet minFreq criterion
    self.vocab(sum(docterm) < minFreq) = [];
    docterm(:,sum(docterm) < minFreq) = [];
end
```

Figure 16 Docterm function

From the word list, we can see what is remaining from the processing we did earlier. The indexer function in the myLSA is what is utilised for creating the vectors word_lists and word_counts.
At this level we can loop back to the stopwords, by just adding some words that are present in words list and we judge not meaningful to our clustering. For example the document "6" has 173 words , if we inspect the list of words "avail". "before", "best", "bring", "end", "find", "g".

```
docterm = LSA.docterm(wordl_ists, word_counts,2);
```

- TF-IDF weight

High number of occurrence between one document and other documents

```
function [tfidf,varargout] = tfidf(self,docterm)
    % Convert raw term frequency values to weighted values using
    % TF-IDF method

    % compute the term frequencies for each doc
    tf = self.termfreq(docterm);
    % if idf property is empty, compute it
    if isempty(self.idf)
        self.idf = log(size(docterm,1) ./ sum(docterm));
    end
    % tfidf is a product of tf and idf
    tfidf = bsxfun(@times, tf, self.idf);
    varargout = {tf};
end
```

Figure 17 Tf-IDF function

Our new matrix created with the TF-IDF takes into consideration two metrics: words that appears a lot in the document but are really meaningful to it and words that rarely appear in the document but are very important for that document. in this approach, a stopword and stemming have been performed to clean the document to its irrelevant content. The computation of the TF and the IDF gives the weights that is mainly used in such a context of can be related as the weighting. this weight look into how unique are words in the document by counting their frequency. Rare term have also important to the document they are weighted using IDF

On figure 16 , we can see that the function takes the "docterm " matrix , compute the frequency of each term appearing in the document. As it is a product of TF and IDF, it creates a balance between frequents and rare words appearing in the document and words.

Each word is weighted based on its relevance to the document. however, one can discuss about the relevance of a word to a document. mathematically the TFIDF look for the term frequency and the inverse term frequency where TF looks into the how many times a word appears in a document an IDF looks into the inverse of the number of time the word appear in the document.

The main implication of this is that word that appear a lot in the document are weighted less.

- SVD Dimension reduction

Singular value decomposition SVD allows the dimension reduction of our matrix and to automatically derive semantic "concepts" in a low-dimensional matrix it is also used as the basis of latent-semantic analysis (Ricci et al., n.d.p 68).

One of the major issue related with the svd is finding the low dimension space and the features representing our reduced dataset. Even though its computation is said to be complex, the matrix from the tf-idf is clearly decomposed into three other matrices (Ricci et al., n.d. p 67).

With LSA, it is assumed that words in the document are somehow related, terms are statistically measured in order to present the latent relation in a document. "*The latent semantic analysis is to use the singular value decomposition, and to find out the relationship between the word and the word in the document.*" [38].

$$svd(M) = [U,S,V] ;$$

M= tfidf is the matrix (m x n) resulting from the previous computation
U (m x r) is the Blogs matrix (an orthogonal matrix)
V (r x r) is the term matrix (diagonal, matrix)
S (r x n) is an orthogonal matrix which contains the eigenvalues or singular values.

```
explained = cumsum(S.^2/sum(S.^2)); figure
plot(1:size(S,1),explained)
xlim([1 30]);ylim([0 1]); line([5 5],[0 explained(5)],'Color','r')
line([0 5],[explained(5) explained(5)],'Color','r')
title('Cumulative sum of S^2 divided by sum of S^2')
xlabel('Column')
ylabel('% variance explained')
```

Figure 18 Variance

After the SVD decomposition of the matrix M, the eigenvalues are on the diagonal of the (r x r) matrix. They are represented following a decreasing magnitude. Thus, the largest singular values from the decomposition are namely the first on the diagonal. The reduction of the initial matrix is happening cutting the data from a given k.

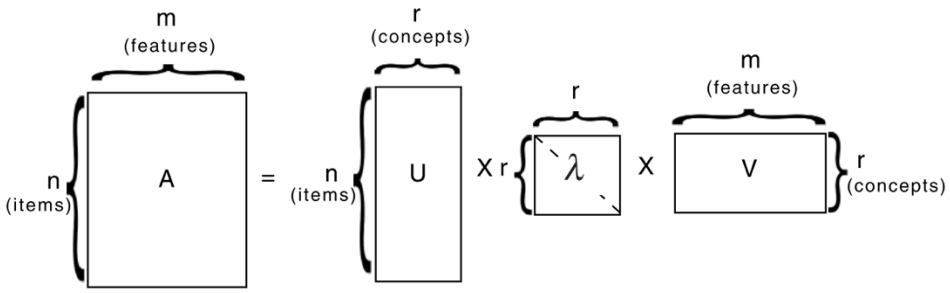


Figure 19 SVD factorization

The reduced uk v_k s_k is a approximation of the matrix M at a rank k where Ak is supposed very close to the initial matrix where the new space is reduced to k –dimension and still the maximum of the initial data.

"It is always possible to decompose a given matrix A into A =U VT . Given the n×m matrix data A λ (n items, m features), we can obtain an n×r matrixU (n items, r concepts), an r×r diagonal matrix λ (strength of each concept), and an m×r matrix V (m features, r concepts). Figure 2.3 illustrates this idea. The λ diagonal matrix contains the singular values, which will always be positive and sorted in decreasing order. The U matrix is interpreted as the "item-to-concept" similarity matrix, while the V matrix is the "term-to-concept" similarity matrix" [7]

```
% a low rank matrix that retains 60% variance 0.6
[Uk,Sk,Vk] = LSA.lowrank(tfidf,0.7);

figure()
scatter(U(:,1), U(:,2),'filled')
title('Twitter Blogs LSA clusters')
xlabel('Dimension 1')
ylabel('Dimension 2')
%xlim([-.3 -.03]); ylim([-.2 .45])

% moving the document in the middle for better visualize
xlim([0.00 .3]); ylim([-.35 .3])
maxtfifd = max(tfidf);

for i =1:length(LSA.vocab)
   %if(maxtfidf(i)>3)
     text(V(i,1).*3, V(i,2).*3, LSA.vocab(i))

   %end
end
```

Figure 20 Blogs and terms in a 2 dimensions space

- Cosine similarity

The purpose of using the cosine similarity here is to define for example a group of documents that we are interred in. this is relevant when subdividing the data. Later on, this similarity measure is used to define similar user and documents.
As the documents are vectorise the angle between them is computed to see the similarity. Both in k-means and Lsa they can easily be compared. However, k-means include as default in Euclidean distance as distance measure for grouping similar documents.

>> doc_norm = LSA.normalize(U(:,1:9));

```
>> LSA.score(doc_norm(1:9,:),doc_norm(1,:))

ans =

    1.0000
    0.9318
    0.6025
    0.7337
    0.7104
    0.3892
    0.4151
    0.5039
    0.3637
```

As the comparison is done between first 9 document row and first ones , the similarity is at its highest level = 1 because the document are the same.

5.5 Clustering with K-means

We examine to what extent the cluster labels can show groups. The experiments with K-means are designed to address to issues:

- One is the ambiguity in the choice of the cluster label, is the distance similarity measure between document should be a similarity measure or a cosine similarity measure. other possibilities exist however we will put our focus on this on those e two methods. For that purpose, in our experiment, we will study the output from the k-means method.
- Which of the 2 clustering methods performs the best, is it LSA or k-Means

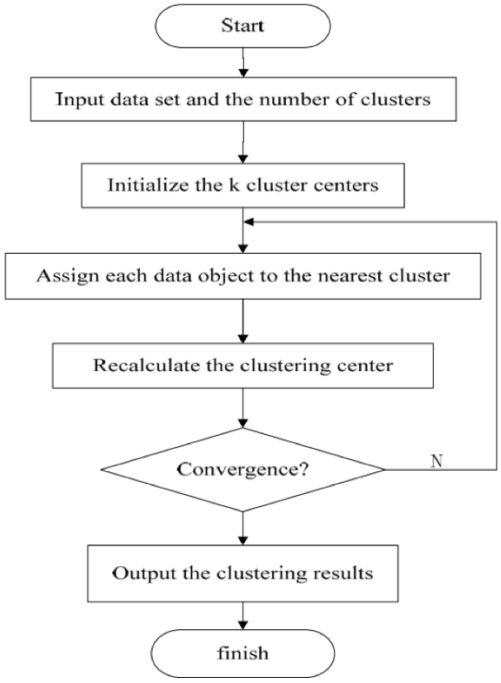


Figure 21 K-means flow chart [38]

The design is the following: We start by specifying a number **k** clusters for our chosen set of data. Firstly, we specify to potentials k clusters by choosing k random data points will work as initial centroids.
On a table, we record all the remaining data points and their distance to each of those points. When we data point, it is assigned to for example cluster 2 if its distance to that point is the smallest compared with distance the three other clusters.
Next, new computations are performed for the newly formed clusters, documents belonging to centroids are revaluated looking into the new position of the centroids. This process is done again and again until the documents still remain din the same after we have updated the centroid.
However apart from having difficulties in choosing the initial number of clusters to start with, the choice of the initial centroid is also challenging. Knowing that different seed lead to different grouping, the choice on the initial centroid for a cluster is important. Arthur et al have tried to improve the basic k-means clustering by proposing an k-means++ method that focus on the choice on the centroids to all clusters [39]
Distance measure is performance several times to find closer data objects specially distance within elements of the same cluster. Similar documents are those that are closer to each other.

K-means will at any case find clusters for our data, however, as we are applying this algorithm and we do have the following assumptions:

1. The choice of the number of clusters k affects the results?
2. Even though this algorithm is reliable it can happen that a cluster is not meaningful?
3. As described in chap 3, our choice for the clustering methods and the number of clusters have an influence on the performance of our system. The following experiments are conducted in order to visualize the impact of the choice number of clusters on the results.
4. Each blog will belong to the nearest mean. Such computation is similar to the Gaussian distribution as they both use centroids goes through iterations to refine the results.
5. How do similarity between documents can impact on our clusters?

The data collected from the blogs have been through a SVD where we end out that the first 9 dimension can capture up to 70% of the original data. The reduced matrix includes the first 9 dimensions meaning that there will be less computation. The computation leading to the choice of the number of dimension have been already computed in figure 18.

```
filename = 'datasetblogs.csv'
mydata = pd.read_csv(filename, sep='|', encoding='latin-1')

>>> stopwords= nltk.download("stopwords")
```

The words that we want to remove from the dataset are from the English dictionary and correspond to conjunction, article, and so one.
The words are broken down and reduced to their root word for example “stems”, “stemming” and “stemmer” are reduced to root stem.

- Finding the centroids

After we have our data we need to setup up four random centroids. The means for each of the clusters (because we have chosen k= 4) are randomly calculated using the Euclidian distance. From the first

round, the centroids are just guest and they randomly appear on the space, they are distinct to each other

The number of time we do the centroid guessing is also important as the it can result to wrong clusters if initial centroid have been chosen between to probable clusters.
While trying to define the position of the centroid s several iterations are performed. It starts randomly, but the through the iterations,

- Distances between documents and centroids

Recalculating the position of the centroid is done until is there is no more moving. While using the k-means python library then the details of the computation are not seen. However, the more iterations, the closer is the document to the cluster mean and is finally assigned to that cluster.
In this experiment, choosing the number of iteration helps to tell the system when to stop.

- Cluster documents according to minimum distance to the centroid

Here we compute all the distances from the centroids means to all the others points. Still, from the results we can see that even though k-means is efficient as a cluster method, it is not capable of drawing clusters that could be said meaningful. Therefore, we need to make use of supplement tools that can help in labelling our clusters.

Genism is a free python library [12] that can help in analysing and retrieving semantically similar document.

```
vectorizer = TfidfVectorizer(stop_words='english')
X = vectorizer.fit_transform(documents)

true_k = 2
model = KMeans(n_clusters=true_k, init='k-means++', max_iter=100, n_init=1)
model.fit(X)

print("Top terms per cluster:")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names()
for i in range(true_k):
    print("Cluster %d:" % i),
    for ind in order_centroids[i, :10]:
        print(' %s' % terms[ind]),
    print


print("\n")
print("Prediction")
```

Figure 22 Top terms in the 4 clusters

5.6 Labelling/ Topics modelling

The aim of this part of the experiment is to find the topic that have been raised in each of the clusters topic modelling is to pull out the topics that are developed in the blogs. With the help of our clustering and modelling methods we will create groups and assign them topic.

[12] https://radimrehurek.com/gensim/

A trivial way to label our clusters would be to look at the top term from each group and look at dominant terms that mostly describe the cluster.
Labelling the clusters can be performed with other technique, in this experiment the LDA method is applied. As a probabilistic method, it is capable of expressing certain values even if they are uncertain in the corpus and to assign words to topics.
Clusters labelling is not only the process of finding topic the topic from each cluster but also to show the limit of each cluster.
Main limitation of certain labelling technique is that they highly base the classification on feature statistical representation in the document, for example, computation of the most frequent words

- Manually labelling with tag crowd

These following screenshots are the results of the visualization of the top words each of the clusters we got from the k-means.

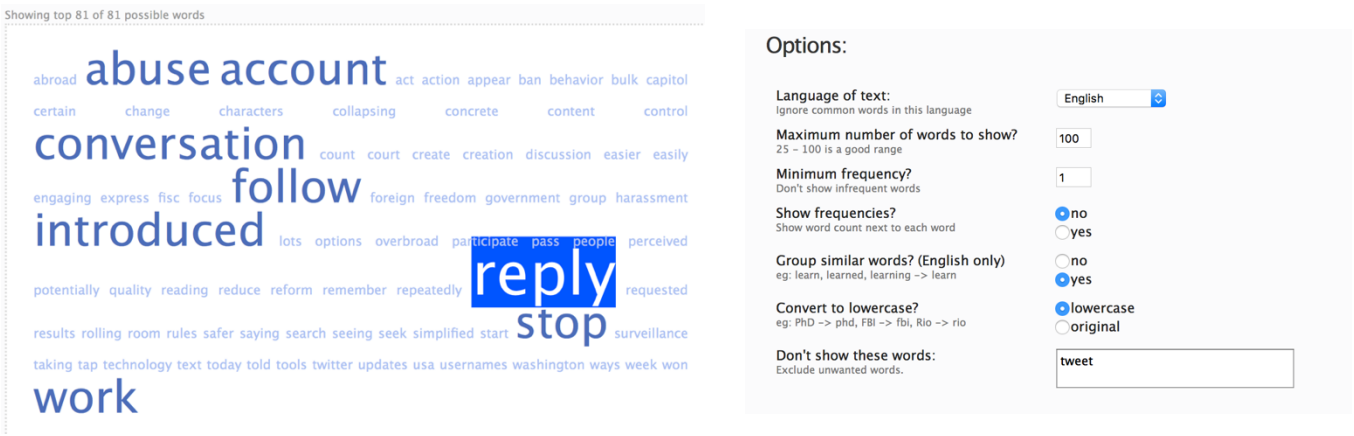


Figure 23 cluster 0 K-means with top 100 words

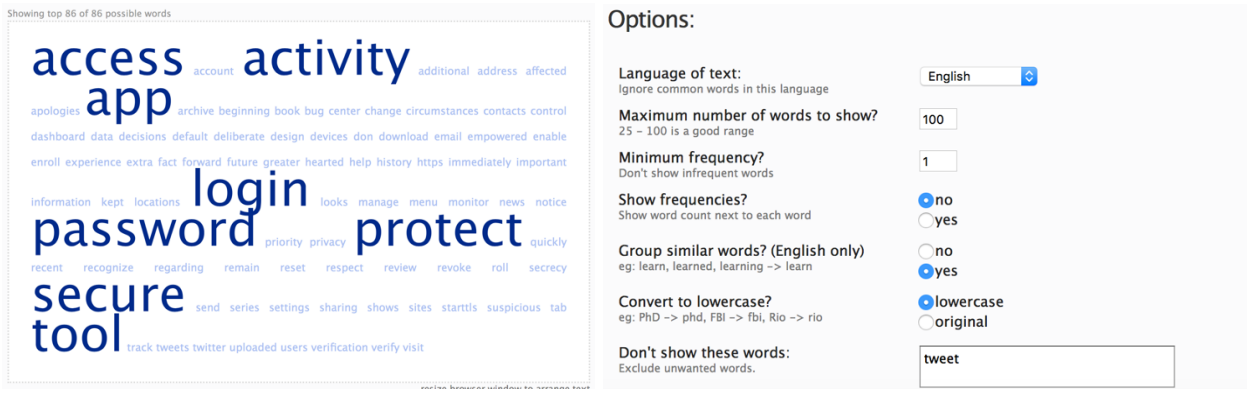


Figure 24 Cluster 1 with top 100 words

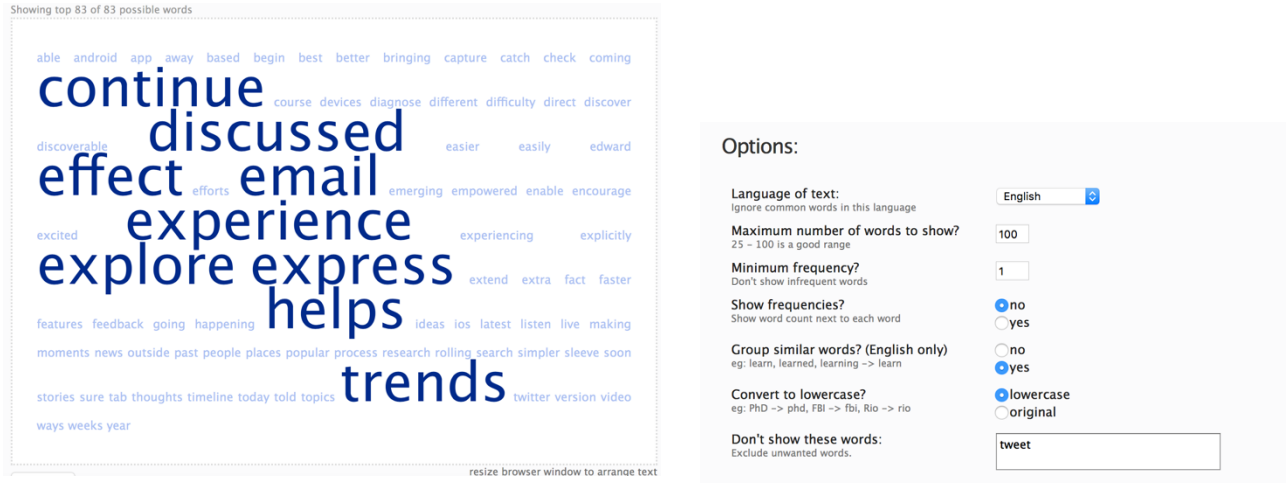


Figure 25 Cluster 2 with top 100 words

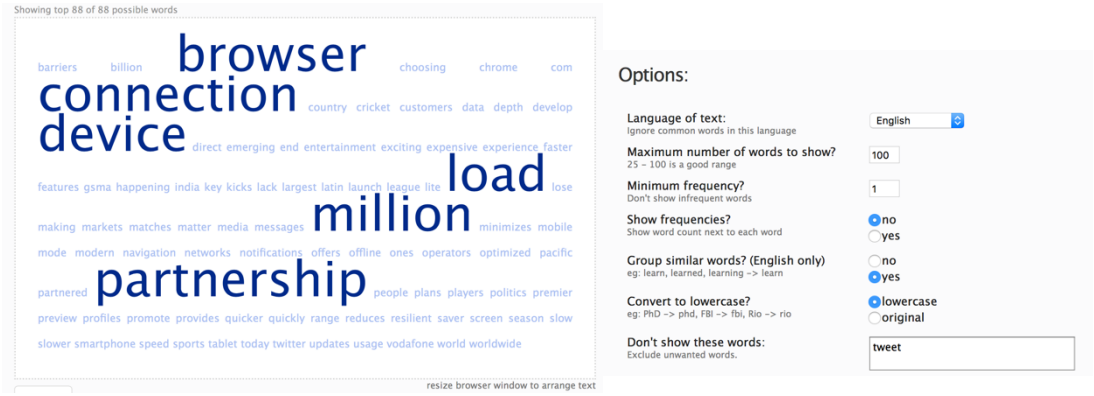


Figure 26 Cluster 3 from k-means

Here are the clusters we end up. Two possibilities can be drawn to label. One could be to use "Human visualization" or the topic labelling method LDA to find out developed topics.
In a first sub conclusion

Cluster 0: Business, social
Cluster 1: Technology, IT
Cluster 2: Security,
Cluster 3: Actuality, events

- Topic modelling with LDA

This method has been already introduced earlier in chapter 3. First proposed by David Blei in 2003 this probabilistic topic modelling method starts with the assumptions that topics present in a document are a probabilistic distribution of words.
Blog posts are filled with words that create a mixture of latent topics that point to several directions. Here we ca say that we do have a bag of words for each document. The topic modelling can be approached in two ways. One way is to find the topics the are discussed in each cluster and the assign the labels. If this process is adapted then in the IR blogs can be retrieved by looking into the term entered the context in which it is utilized and the cluster it belongs to. Another way to apply this topic modelling could be to directly apply it on the documents (blogs). The resulting is that we find a number topics that are in majority touched in the blog post.
Looking into the relationship between words, here, two models that have different words but in the same context they will be considered as similar because they fall in the same latent topic. That

Preparing documents

- Cleaning and Pre-processing: as previously tokenisation, stemming
- Preparing document term matrix

```
# list for tokenized documents in loop
texts = []

# loop through document list
for i in doc_set:

    # clean and tokenize document string
    raw = i.lower()
    tokens = tokenizer.tokenize(raw)

    # remove stop words from tokens
    stopped_tokens = [i for i in tokens if not i in en_stop]

    # stem tokens
    stemmed_tokens = [p_stemmer.stem(i) for i in stopped_tokens]

    # add tokens to list
    texts.append(stemmed_tokens)

# turn our tokenized documents into a id <-> term dictionary
dictionary = corpora.Dictionary(texts)

# convert tokenized documents into a document-term matrix
corpus = [dictionary.doc2bow(text) for text in texts]

# generate LDA model
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics=2, id2word = dictionary, passes=20)
```

Figure 27 Lda topic modelling on blogs dataset

>>> print(ldamodel.print_topics(num_topics=2, num_words=4))

[(0, '0.026*"twitter" + 0.016*"tweet" + 0.011*"re" + 0.010*"account"'), (1, '0.033*"twitter" + 0.017*"s" + 0.010*"live" + 0.010*"world"')]

The result is showing here to number of topics and we can see that all deal with twitter. we can even apply human judgment to see if the clusters are topics are relevant.

.

Chapter 6. Results & Findings

In this section, we present results of our experiments in relation with the hypotheses fixed earlier in the analysis part.

Throughout the clustering experiments we have defined the principal components for our reduced matrix in the SVD decomposition. Looking into the S matrix with its Eigen values, we derive the component for the reduced matrix.in the choice of the first column that represents the variance are better represented with only 7 dimensions instead for 100, it is possible to keep 80% of the information. Both the matrices representing the blogs and the word are displayed on the same space. Four main clusters seem to derive from this clustering model however we still need to label them looking into the document contents.
From the experiments, several findings could be drawn but we emphasis on few of them:

- Difficult to segment the number of clusters:
- Select a given number of clusters bias the result
- After we have set our number of cluster to 4, the top terms have been used to label the cluster. However, the diversity is not really represented

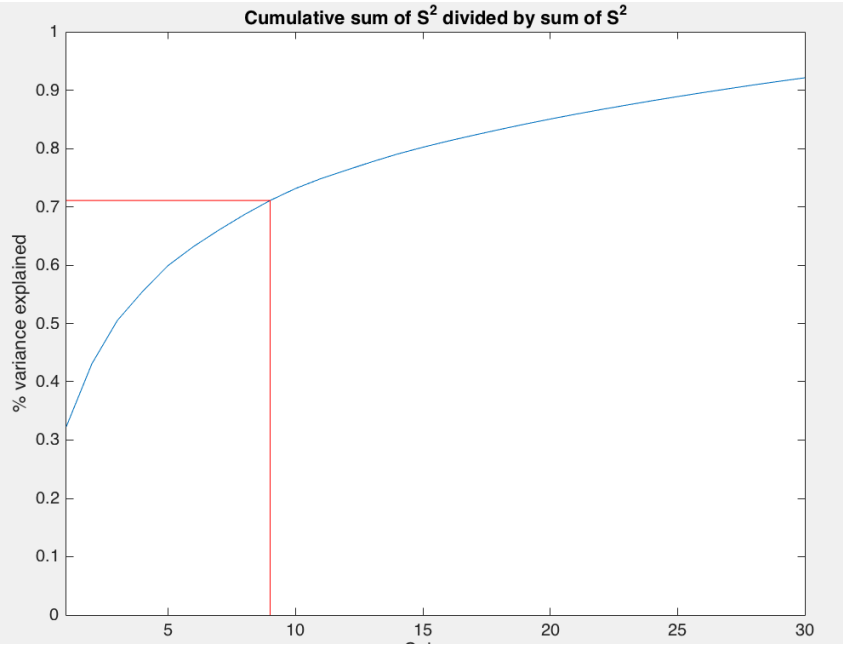


Figure 28 Plotting the Eigenvalues

Defining which components can represented the deducted matrix is approximately measured using the diagonal matrix containing the Eigen values. The cumulative values of the of the percentage variance plotted in figure 28 shows that the low rank matrix can be represented with 70 pour cent of the previous data using only the first 9 columns. Looking to the Eigen values, our rank-9 approximation is retaining 70% of the information of the original matrix. The 100 blogs with be represented in a 9-dimension space.

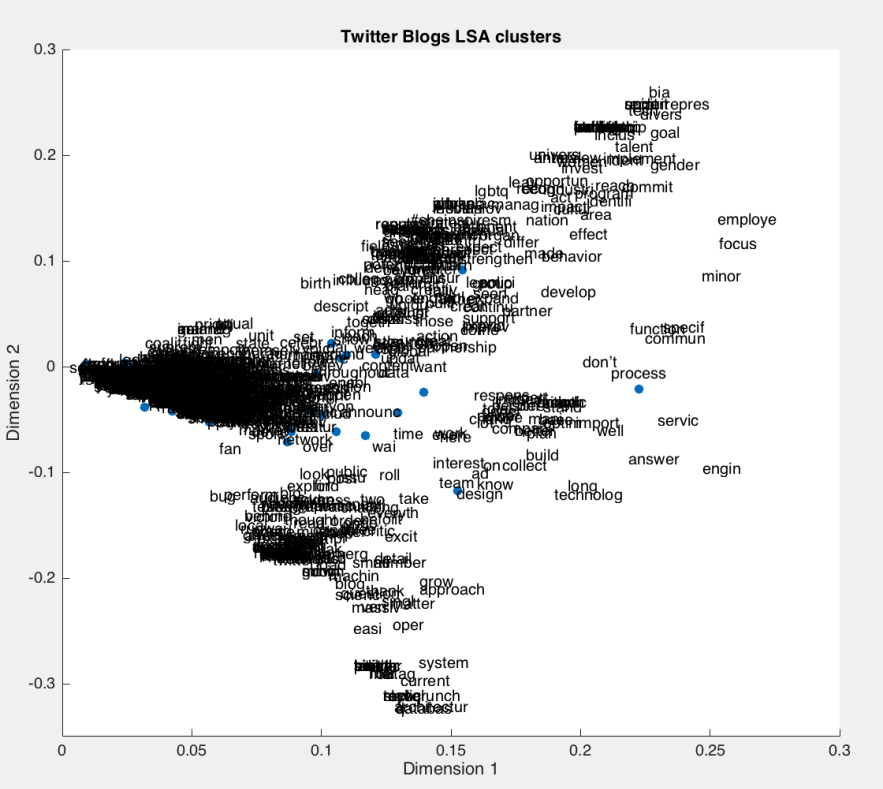


Figure 29 Clusters found with LSA

Figure 20 shows the cluster output from the latent semantic analysis. Clusters and the documents they contain seems to be quite relevant even though the filtering with the key word is not quite extensive. The algorithms are of high quality as they draw
With regards to the automatic labelling the generated topics could still be discussed, the topic modelling could give a more relevant output if we apply a strong stopword filter..

The following table presents the results obtained using k-means. Here are listed the first 10 most important word in using tfidf. The number of computation to find the centroids have been fixed to 100 iterations to give more reliability to the output.

Cluster 0:	Cluster 1:	Cluster 2:	Cluster 3:
characters tweet 140 conversation conversations text group usernames replying abusive	lite twitter mobile data smartphone networks connections devices updates way	account twitter bug surveillance security login password access information help	search explore abusive moments results potentially make accounts new trends
Business, social	**Technology, IT**	**Security,**	**Actuality, events**

Table 2 K-means clusters with top ten words

From the table 2, we have selected the first 10words of the in each of the four clusters selected. This could be quite selective and sometime not taking account of the whole diversity in the document or in a cluster. it works here as a filter and only word that are weighted very important remain in the op words. Another annotation where we do not only look at the first 10 but the first 100 words is done. The purpose is here to if the topic manually proposed here could be maintained.
As human, we have tried to propose topics by looking into the interconnection between words. The class topic that we have derived could be more but we have decided to narrow it down and resume it to a number of four.
Based on this analysis from the result, we can say that the results from both LSA and k-mean were inconsistent. Some outliner was clearly visible whereas other clusters were quite straightforward. From LSA even if we are not sure about the delimitation of the clusters, places where clusters, there are two clear clusters. At this level of the work, the result of the grouping has not been evaluated by other potential users. The next level will be to involve users and evaluate the credibility of our clusters.

6.1 Evaluation of the results

- Cluster evaluation:

The first think we have tried here is preliminary see if two blogs that have been placed in the same group are similar as document. We took to blogs text "Web scrapping" from our dataset respectively 1 and 10 just picked randomly, the result shows a cosine of 0.45.

Another way to assess our results would be to apply the Elbow method to see if effectively the number of clusters we have been work is good enough

- Labels assessment:

This could be approached in two different ways collecting users need through a requirement specification document of mathematically solving it thanks to the use of algorithm. Our primer goal

here is to give equal chance to all the twitter blogs while a user is searching for a specific topic. Therefore, we will need to group all similar content in the same cluster where is the and then automatically label them based its top words. An external evaluation can be applied in a future work.

6.2 Comparing LSA and K-means

From the experiments, several findings could be drawn but we emphasis on few of them:

- Difficult to segment the number of clusters:
- The final result can be biased by the selected number k
- After we have set our number of cluster to 4, the top terms have been used to label the cluster. However, do really know if the words diversity is fairly represented?

	K-means	LSA
Advantages	• It addresses polysemy • Simple algorithm	• Simple algorithm • TF-IDF • SVD • Sparsity of the data
Disadvantages	• Choosing the k • The centroid is difficult to	• Can no handle polysemy • Depend on svd

Table 3 K-means and LSA comparison

As a sub conclusion in the comparison of this two approaches we can end up saying that k-means and both LSA are great tool in the document clustering. they may seem competitive technologies however each of them has its advantages and disadvantages.

Chapter 7. Discussion

In this chapter, we bring some reflexion to clustering as such using the two methods LSA and K-means. What we firstly bring in the discussion is the relevance and challenges with the clustering methods we have chosen. We could also discuss the fact that when we select a given number of clusters bias the result. Moreover, after we have set our number of cluster to 4, the top terms have been used to label the cluster. Another challenge in the top terms of the cluster is that the diversity is not really represented

- **Discussing the nature of found clusters:**

Even though our experiments are solely focused on ways to groups blog post and categorize them, we believe that some of the discussions can be also applied to other scenarios that deal with group segmentation.
In many cases, we have tried to apply our personal knowledge to supply the difficulty in finding cluster. For example, in the choice of the stopwords. Even if we go beyond the list of English terms and make our own word filter, the pre-processing choice deeply shape the finals results. Stopword removal can be repeated across topics leading to a slightly or fully different topic. In this work, words like "twitter", "follow" are not giving any extra information and redirect to same concept, therefore the need to filter them. Choice of the stopwords could be done looking into words that occur < 0:2% and > 15% of the documents are removed. [26].
From the latent semantic analysis, it was difficult to visualize the number of clusters and the number of Main criticism on cluster analysis have been pointing the difficulty of finding relevant clusters. Results from the first experiment using LSA raise the following discussions.
To which extent we can rely on the semantic relation in LSA? If we look at the results from its cosine similarity, is shows a high similarity between the first 9 document and the first row. Even though the first and the second document are developing different topics they have been estimated as similar with high cosine similarity of 0.9.

Cosine similarity has the property to be independent of the document length. Blogs with same content but are not the same length will be treated as identical.

Based on a study of similarity preference to certain words, with a combination of human similarity rating. Based on the results of study 1, LSA cosine values seem sensitive to differences in similarity between items, at least when these differences are quite large. However, the data of McRae and Boisvert (1998) suggested that LSA was more predictive at the lower end of the similarity range. To investigate this, further, we used an expanded item set and collected participants' ratings of the similarity of a variety of word pairs. The purpose was to sample word pairs across a wide range of LSA scores, so as to examine whether their predictive validity varies across the similarity range.[4].

Strictly speaking, LSA better performs with large number of items. But this could be relative as in the scikit learn interface, clustering can be performed with a dataset of less than 10 k . The size of the dataset is also quite important as LSA has slightly lower performance with small dataset [40].
The polysemy problem is still here, and most of the time the due to the query's polysemy, documents that are retrieved tends to several topics. In such situation is up to front the necessity of labelling the categories in order to avoid long time processing and help the user while searching for specific topics. If the system has indexed the documents which is in most the case, the word used to characterize a document might be different from the one entered by the user while searching for that document.

Several discussions have been raised concerning the labelling of the clusters. Firstly, using an automatic annotation thanks to human evaluation. This action may be relevant if the "USER" have been involved in a categorisation process. We can put all the top words on sheet of papers and take the categories we have found or we want to end up with and ask them grouping. Such a process can be quite useful but can also end up with a huge number of combination. The semantic organization can better make sense the two options "humans" and "algorithm" are combined.

- **Cluster labelling**

Several techniques exist in the labelling process. After we have grouped our data, the next step has to give title to the formed categories. From our experiments in the LSA and K-means we could already derive some group title. However, if we take the example cluster 0 in k-means some of the words in the top ten are not really mirroring the diversity that we have inside the clusters.
In one hand, we can trust the system and let it chose the topics for us, but we are still setting up the parameters. Documents of a given topic tends to have a number of words that inform more or less on its content. There are numbers of techniques to the in the process of labelling clusters. We have tried to utilize human judgment to give title to the formed categories and also to apply the LDA model to find the topics. From our experiments in the LSA and K-means we could already derive some group title. However, if we take the example cluster 0 in k-means some of the words in the top ten are not really mirroring the diversity that we have inside the clusters.
In one hand, we can trust the system and let it chose the topics for us, but we are still setting up the parameters. As Puschmann et al mention that "Only the human analyst can make sense of the topics that have been learned" [35].
Again, when defining the topic to our clusters, blogs that contains words "mother", and "brother" I dealing probably with "family" and word like "swim", "match", "competition" belong to a topic of "sport". A word like "kids" can belong to both of the topics. In such case the semantic analysis is relevant to supply in details.
Moreover, this is also this problem of category label that might be different from one system to another. One solution be an implication of standard labelling as it already exist but apply it in a way to reduce confusion. Other prosed solution have been the use of cross media problem that make use existing data to train a classifier that will be applied to our data.

- **Defining the position of the centroids**

Regarding the choice of the centroid in k-means the true position of the centroid can be wrong. Let 's take the situation where the initial centroids are placed between two clusters then this this can lead to error in the formed groups. Even though this method is considered to be very efficient choosing the initial points is not an easy task. If you choose the random centroid wrong then the rest will be also wrong. How to ensure that we are choosing a good starting point?
The process will up to a certain level before stopping. In order to ensure good results, it will be wise to try the algorithm several times. We set the number of iteration and the number of k clusters we want. Nevertheless, the more iterations the more processing time is needed. Moreover, this is not the only problem in k-means, the number of k is also a guess game at the beginning. Applying the Elbow method at an early stage could help to see exact number of cluster.
In particular, to the problem of finding the centroid, several researches have been conducted. As an example, K-means ++ is one of the improvement of the method. [39] . as any other clustering method, it seeks to minimize distance between elements from the same cluster. As main downside of this algorithm, they point out the low-level guaranty specially in the choice of the centroids.

They have been proposing a randomized seeding technique. With such a set up they have been able propose a more accurate positioning of the centroids. The canter is chosen looking into the shortest distance between a data point that be been chosen closer to it. The results are that k-means++ perform quite faster.

Chapter 8. Conclusion

Clustering blogs or just any other documents text is a quite challenging task. We could see that this field is so complex and have been source of many researches and applications.
Throughout this report, we have mainly tried to relate the challenges and walkthrough of two well-known algorithms that are LSA and K-means and ways to tag formed groups. One advantage of taking the blogs as dataset is that it provides rich content.
At the beginning, we fixed some objectives that in order to answer the research questions:

How to find latent topic-group and label them in a blogging environment?

- How do we organize a document collection clustering into semantically connected keywords?

We have been applying of the LSA and K-means to visualize latent structure in our dataset. This objective could be said to be partially fulfilled. In fact, clusters have been drawn but the quality could be improved by a sharper selection of key words and similarity measures. How to find latent topic-group and label them in a micro-blogging environment? The question of semantic properties has been addressed in LSA. With K-means clustering, the problem of polysemy has to be taken into account. Moreover, for that purpose several Blogging environments have the advantage of being rich in content, tags, diversity and topics. By diving in that space, we have been able to collect relevant data for semantic analysis purpose. Blogs can be similar in the way present the content or just in defend by the proposed information's.

- How can we evaluate similarity between blogs?

We have proposed an analysis of two clustering methods and present challenges of their applications. In an evaluation, the similarity of documents that have been placed in the same tag, and the similarity between our 9 principal components and the first line of documents.

- If we apply semantic analysis strategy do we get similar grouping as on twitter blogs?

The similarity measure has its importance in several contexts. When it comes to classify a new content, when want to see how similar two blogs. The question of similarity in clustering have been subject of many research from spam detection to health studies, search features on mac computers and so forth. K-means and LSA have shown their effectiveness and number improvement have been made on those two methods. More work should be done in particular on ways for evaluating the relevance of their output.

- What are the implications of clustering documents in different ways?

This question has not been answered, however we could argue that different metrics for example different number of k in k-means leads to different results.

- What is the relation between topics and cluster labelling?

For sure there is need for addressing the cluster labelling in documents clustering. In the context of IR, the clustering of the blogs can be useful for both the user and the business. For the system, the can recommend personalized content or promote novel or new content. For the user who come on twitter blog to search for blogs that might cross their interest, clustering can provide the grouping and even at another level a hierarchical distribution of the topics.

.

- While applying K-means and LSA which one perform better?

Both of these two models have their advantages and disadvantages as listed earlier. In our experiments chapter the clustering challenges have been approached from different angle. Matlab and Anaconda python environment are both providing relevant packages and libraries process, vectorise and plot the data matrices.
Finally, the relevance to this clustering and topic modelling are closely related specially if we want to give meaning to the formed groups, give greater user experience and lighten the system.

Chapter 9. Future Work

As an extension of this work if at the first level we have the blogs clustered and labelled, we will be considered sublabels and recommendation of blogs to visualize. Here are some issues that in a future research we would be considering:

- **User involvement**:

Naturally users demand on to get search results that matches their expectations. More they are curious to see other contents that is different from what they are to see like most recent and most popular. We have tried to mathematically define the category in which the user blogs can be categorised, however this could be done by implicating the user in the choice of the categorisation. F. Ricci et al report Card sorting as "*It is used to create taxonomies of items based upon the naturalistic mental models users hold for the relationships between content or concepts. The method basically consists of asking users to sort a collection of cards, each which depicts a content item or sub classification, into groups based on similarity*" (Ricci et al., p 366.) the resulting groups formed can be analysed using results from our cluster analysis.

Other testing procedure that need to be addressed in the future is the evaluation of the number of k. the Elbow method have been applied sometimes applied before clustering to know what is the number of clusters.

- **Recommend blogs to read or to follow**

Apply hierarchical clustering in order to cover subcategories and clusters that overlap
Moreover, applying filtering based on the author of the blog, or the similarity between document could be a novel way to recommend content to users. In the context of SEO and recommender system, the search features and items categorisation is actually largely implemented by online services like amazon, movies services, online books stores, to increase the user experience. So, this is a recurrent and recurrent problem of trying to collect organize store and query for big data. Information retrieval and RS are therefore tightly related.
From the LDA the text mining can reveals topics that might interests a given user by looking to similar author, or similar topic or just bring a novel way of presenting the content to the user.

- **Improvement of LSA**

This could be done with a focus on three main levels, that are the choice of the stopwords, the document similarity evaluation and the segmentation of the clusters.
A future work can improve subspace formation for the Lsa. Another perspective in the blogging environment could be in the semantic and labelling compare and evaluate the

Chapter 10. Bibliography

[1] G. Wilkes and B. Traynor, “Folk Classification of Social Media Platforms : Preliminary Findings.”

[2] R. Nugroho, J. Yang, W. Zhao, and C. Paris, “What and With Whom ? Identifying Topics in Twitter Through Both Interactions and Text,” vol. 14, no. 8, 2017.

[3] G. Hope, T. G. Wang, and S. Barkataki, “Convergence of Web 2 . 0 and Semantic Web : A Semantic Tagging and Searching System for Creating and Searching Blogs,” pp. 201–208, 2007.

[4] A. Huang, “Similarity measures for text document clustering,” *Proc. Sixth New Zeal.*, no. April, pp. 49–56, 2008.

[5] M. F. alhamid Majdi rawashdeh, “BOOSTING TAG-BASED SEARCH IN SOCIAL MEDIA SITES,” pp. 1–6.

[6] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, “Hybrid Linear Modeling via Local Best-Fit Flats,” no. June, pp. 217–240, 2012.

[7] F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, and F. Ricci, *Recommender Systems Handbook 123*. .

[8] G. Xue, C. Lin, Q. Yang, W. Xi, H. Zeng, Y. Yu, and Z. Chen, “Scalable Collaborative Filtering Using Cluster-based Smoothing *,” pp. 114–121, 2005.

[9] F. Wikipedia, “Unsupervised learning,” pp. 4–7, 2017.

[10] A. C. Billings, *Sports Media: Transformation, Integration, Consumption*. Taylor & Francis, 2012.

[11] L. Hong and B. Davison, “Empirical study of topic modeling in twitter,” *Proc. First Work. Soc. …*, pp. 80–88, 2010.

[12] T. H. Parisa, B. Razeghi, N. Okati, D. M. Souran, and M. Mir, “Collective wisdom based blog clustering,” *6th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2015*, pp. 1–6, 2016.

[13] Ritzau, “FAKTA : Her er det globale fænomen Twitter,” *TV2 Samf.*, no. november 2012, pp. 2012–2013, 2017.

[14] K. J. Geras, “Prediction Markets for Machine Learning,” *Artif. Intell.*, no. 1, p. 76, 2011.

[15] D. Ramage, S. Dumais, and D. Liebling, “Characterizing Microblogs with Topic Models.”

[16] R. Videos, P. Focus, and O. Resources, “Predictive Modeling for Healthcare,” pp. 24–26, 2017.

[17] “Birkbak, Carlsen_2016.pdf.” .

[18] S. Deerwester, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis.”

[19] D. G. Dastidar, “Top 3 Best ‘ Word Cloud ’ Generators,” pp. 2016–2018, 2017.

[20] C. Torniai, J. Jovanović, D. Gašević, S. Bateman, and M. Hatala, “E-learning meets the Social Semantic Web,” *Proc. - 8th IEEE Int. Conf. Adv. Learn. Technol. ICALT 2008*, pp. 389–393, 2008.

[21] D. A. Gómez-Aguilar, M. A. Conde-González, R. Therón, and F. J. García-Peñalvo, “Reveling the evolution of semantic content through visual analysis,” *Proc. 2011 11th IEEE Int. Conf. Adv. Learn. Technol. ICALT 2011*, pp. 450–454, 2011.

[22] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, “A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval,” *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM ’14*, pp. 101–110, 2014.

[23] C. H. Brooks and N. Montanez, “Improved annotation of the blogosphere via autotagging and hierarchical clustering,” *Proc. 15th Int. Conf. World Wide Web - WWW ’06*, p. 625, 2006.

[24] S. W. Cho, M. S. Cha, and K. A. Sohn, “Topic category analysis on twitter via cross-media strategy,” *Multimed. Tools Appl.*, vol. 75, no. 20, pp. 12879–12899, 2016.

[25] T. P. Ping, N. Kulathuramaiyer, and A. A. Julaihi, “Semantic Characterisation: Knowledge Discovery for Training Set,” *Exchange*, vol. 312, no. 535, p. 852, 2013.

[26] I. S. Dhillon “Co-clustering documents and words using Bipartite Spectral Graph Partitioning,” 2001.

[27] P. Lynggaard, “ICTE course : Development of ICT and Media Services,” no. Cmi, p. 58, 2017.

[28] C. Fry, “Can we Group Similar Amazon Reviews : A Case Study with Different Clustering Algorithms,” pp. 374–377, 2016.

[29] V. S. Reddy, P. Kinnicutt, and R. Lee, “Text Document Clustering : The Application of Cluster Analysis to Textual Document,” 2016.

[30] R. Xu, S. Member, and D. W. Ii, “Survey of Clustering Algorithms,” vol. 16, no. 3, pp. 645–678, 2005.

[31] K. N. Stevens, T. M. Cover, and P. E. Hart, “Nearest Neighbor,” vol. I, 1967.

[32] M. R. Brett, “Topic Modeling: A Basic Introduction,” *J. Digit. Humanit. 2.1*, pp. 1–5, 2012.

[33] L. Shan, C. Sun, L. Lin, M. Liu, X. Wang, and B. Liu, “Evaluating Tag Quality for Blogger

Modelling via Topic Models," no. 61300114, pp. 1770–1776, 2015.

[34] N. Agarwal, M. Galan, H. Liu, S. Subramanya, N. Agarwal, M. Galan, H. Liu, and S. Subramanya, "Clustering Blogs with Collective Wisdom ∗," pp. 2–5.

[35] C. Puschmann and T. Scheffler, "Topic modeling for media and communication research : A short primer," 2016.

[36] S. Numfocus, "Toolkits Citing Matplotlib," pp. 11–13, 2017.

[37] L. Shure, "Can You Find Love through Text Analytics ?," pp. 4–9, 2017.

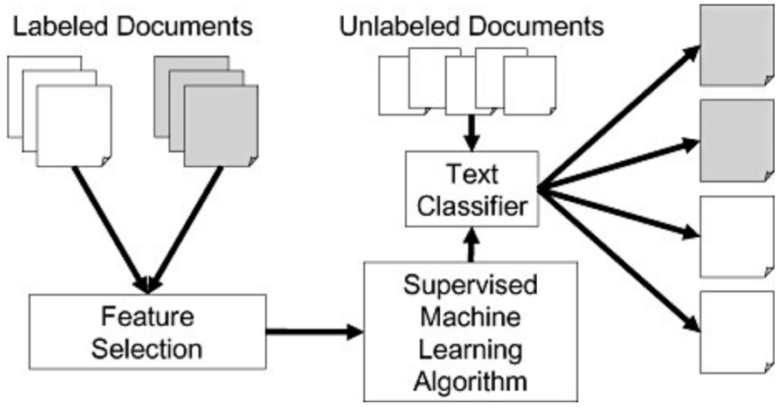
[38] C. Wenli, "Application Research on Latent Semantic Analysis for Information Retrieval," 2016.

[39] D. Arthur and S. Vassilvitskii, "k-means ++ : The Advantages of Careful Seeding," vol. 8, pp. 1–11.

[40] S. Simmons and C. Cv, "Using latent semantic analysis to estimate similarity LSA in Application Overview of LSA," 1997.

Chapter 11. Appendix

1. Labelled and unlabelled documents clustering



2. Web scrapping

```
>>> soup.find_all('p')
```

[Twitter is the best and fastest place for people to <a
href="https://twitter.com/search?q=%23SeeEverySide&src=tyah">#SeeEverySide of what's
happening around the world, and for the past month, it's been the place where Muslims around the world shared
and discussed their #Ramadan
experiences in real time. This year, Tweets around <a
href="https://twitter.com/search?q=%23ramadan&src=typd">#Ramadan and <a
href="https://twitter.com/search?q=%23eid&src=typd">#Eid increased to 118 million worldwide.
Click on the heat map below to see how people used Twitter to share their Ramadan experiences during <a
href="https://twitter.com/search?q=%23eid&src=typd">#Eid.
, The top English-language Tweet around <a
href="https://twitter.com/search?q=%23ramadan&src=typd">#Ramadan this year was by the
French footballer <a
href="https://twitter.com/paulpogba?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor">
@PaulPogba.
, We released an Eid-themed emoji that embodies the spiritual aspect of <a
href="https://twitter.com/search?q=%23eid&src=typd">#Eid. The following hashtags unlock a
visual of a mosque, which will be live until the end of the week:
, The top hashtags used globally for the Ramadan and Eid emojis were: , , In the
Middle East, we launched the first Ramadan bot on Twitter that functions as a TV guide in
partnership with Arab entertainment news portal @FilFan.
, , More than 200 Ramadan-themed Moments and the first-ever 360 videos on Twitter
were produced by regional broadcasters and publishers such as <a
href="https://twitter.com/CBCEgypt">@CBCEgypt, <a
href="https://twitter.com/dubaitv">@DubaiTV, <a
href="https://twitter.com/i/moments/871338351605022725">@Fatafeat, <a
href="https://twitter.com/akhbarak">@Akhbarak and <a
href="https://twitter.com/AlBayanNews">@AlBayanNews over a variety of topics.
, Sheikh @MohamadAlarefe posted videos
on Periscope throughout the holy month to engage his followers in live conversations over the Muslim faith,
with broadcasts that exceeded 45,000 views in total.
, Saudi Broadcasting Corporation's <a
href="https://www.periscope.tv/qurantvsa/1kvJpQebYgOxE">@qurantvsa has been broadcasting the
Holy Mosque in Makkah live through its Periscope channel.
, The dual screen experience between television and Twitter couldn't be more evident than during
Ramadan. Again this year, the sequel ٣ سيلفي has become the most Tweeted about show this Ramadan with more
than 2.5 million Tweets around the show. The top talked about shows on Twitter this year were: الزيبق ، هذا
المساء، كفر دلهاب، غرابيب سود، رامز تحت الأرض and 30 يوم.
, Exclusive Twitter Q&As took place with celebrities across the region which gave fans the
opportunity to directly engage in live video conversations. These engagements totalled over 2.5 million video
views with celebrities such as Mohamed Henedy (<a
href="https://twitter.com/OfficialHenedy">@OfficialHenedy), Ahmed Malek (<a
href="https://twitter.com/jrMalek">@jrMalek) and Ahmed Fahmy (<a
href="https://twitter.com/AFahmyOfficial">@AFahmyOfficial).
, A number of celebrities joined Twitter during <a
href="https://twitter.com/search?q=%23Ramadan&src=typd">#Ramadan to build stronger
relationships with their fans during and after the holy month, including: <a
href="https://twitter.com/Mohamed_Emam">@Mohamed_Emam, <a
href="https://twitter.com/EngyKhattab">@EngyKhattab, <a
href="https://twitter.com/RAbdelGhafour">@RAbdelGhafour, <a
href="https://twitter.com/midoadel">@MidoAdel and <a
href="https://twitter.com/Owisses">@Owisses. , In addition to the above, viewers got the
opportunity to access timely and tailored Ramadan related videos right on their Twitter feeds from key
publishers in the region such as @Fatafeat, Rotana's <a
href="https://twitter.com/Khalejiatv">@KhalejiaTV and <a

href="https://twitter.com/layalina">@Layalina. The content was sponsored by brands that include Nestle, Unilever Personal Care and Magnum respectively, enabling viewers to access premium food, lifestyle and documentary near live video highlights.
, In Indonesia, as the largest Muslim-populated country, people celebrated #Ramadan by sharing unique local experiences on Twitter - from sharing photos of their #suhoor meal, videos of unique activities while waiting for breaking the fast, to Periscope broadcasts of their homecoming journey.
, We partnered with @netmediatama TV station who integrated Twitter within its Sohoor programming for one of the most popular variety shows Ini Sahur (@Ini_Talkshow). Celebrity guests shared special Ramadan message on Twitter for their fans from the sets of Ini Talkshow. People could also send a Direct Message to @netmediatama to get daily prayer time, iftar time, and video of short sermon in their DMs.
, The TV station also actively used video to promote their Ramadan programs. There were at least 2 million views generated from the short videos Tweeted by @netmediatama during Ramadan.
, In collaboration with the @travellerkaskus community, Indonesians also actively took part in various Tweet-powered competitions during the past month. People flocked to Twitter to share photos, videos and Periscopes using different hashtags: #PilihYgSegar, #CeritaRamadan, #MudikGan, #BukberDmn, #AyoNgababutrip, #GuyonSahur, and more.
, , The day before Eid and the day of Eid itself became memorable moments where Indonesians shared local unique #Eid experiences from their hometowns on Twitter.
, We wish all Muslims a blessed #EidMubarak. May you have a wonderful celebration with your family and loved ones.
, Kinda Ibrahim ,
@kindaibrahim
, Media Partnerships Director, Middle East and North Africa, Twitter , Your official source for what's happening.

Need a hand? Visit https://t.co/jTMg7YsLw5 , © 2017 Twitter, Inc.]

1. People basis of grouping of social media in General [1]:

A study realised by Wilkes and Taynor that examined how active users were grouping social media platforms. 59 respondents completed an open card sort activity where they categorized 19 social media applications according to their own preferences. Data was also collected on frequency of use of Social Media Platforms as well as perceived use in comparison with peers.

TABLE 1. HYPOTHETICAL RATIONAL CHOICE SORTS.

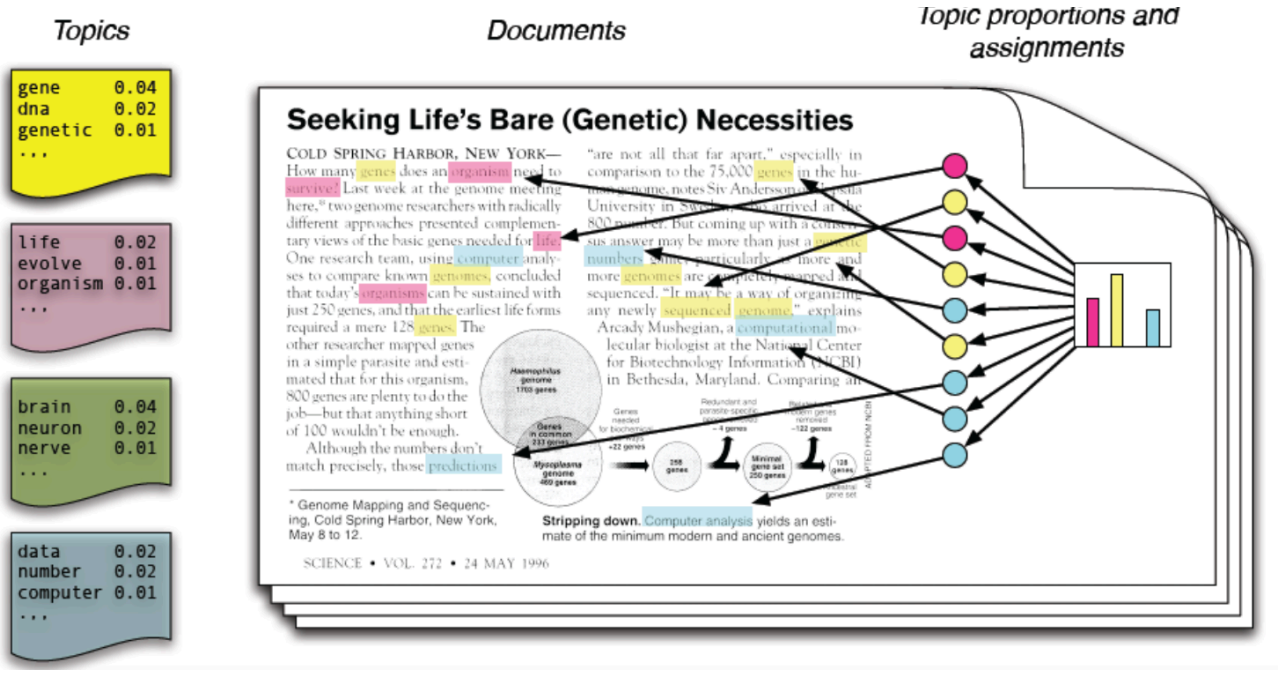
Platform sort	Where users group platforms (services) based on whether accessed through smartphones, tablets, laptops, or desktops.
Content sort	Where users group platforms based on kinds of content, e.g. Youtube.ca for videos, Reddit.com for memes or humor, Facebook.com for family images.
Propagation sort	Active twitter feeds, for example, can propagate content more quickly than other services, which can prompt users to return more often to view the passing show or participate.
Affordance feature sort	Tumblr.com for example, is a blogging platform that automates how content is posted and allows browsing by means of a dashboard, features that support engagement.
Affinity group sort	A platform may host a group or groups of interest or of value to the participant. Special interest group sort—A service may host a group or groups who share and therefore make available specialized content
Special interest group sort	A platform may host a group or groups who share and therefore make available specialized content not easily found or aggregated elsewhere.
Personal - Professional sort	A participant may organize online activities according to boundaries among work, family or in-group social personae to protect the integrity of work, family, and friend relationships.
Diurnal organization sort	The time of day may be grounds for selecting among services according to divisions of work and leisure or time-zones.

TABLE 2. HYPOTHETICAL SOCIAL INFLUENCE SORTS.

Liking sort	Here we refer to liking as a principle of social influence developed by Cialdini and Goldstein [4]. What is liked is trusted. Where an affinity or special-interest group sort has a functional character, "I post from Instagram to Tumblr.com because I like Ahmed and his sister who also hang out there and we exchange stuff" would be a sort based on liking.
Authority sort	Authority as grounds for grouping services can realize itself as direction, e.g. "My professor had us use twitter," or discretion "My boss is a connection on Linked.com," or avoidance "I stay away from Facebook.com because my mom is there."
Conformity - compliance sort	Where the values of a group expressed in its collective activities become grounds for categories: "All of my classmates share information on a Facebook.com page."
Reciprocity	Where categories derive from content or opportunities for contact offered by others "My girlfriend and her friends would send me memes from Reddit.com."

2. Labelled and unlabelled documents clustering

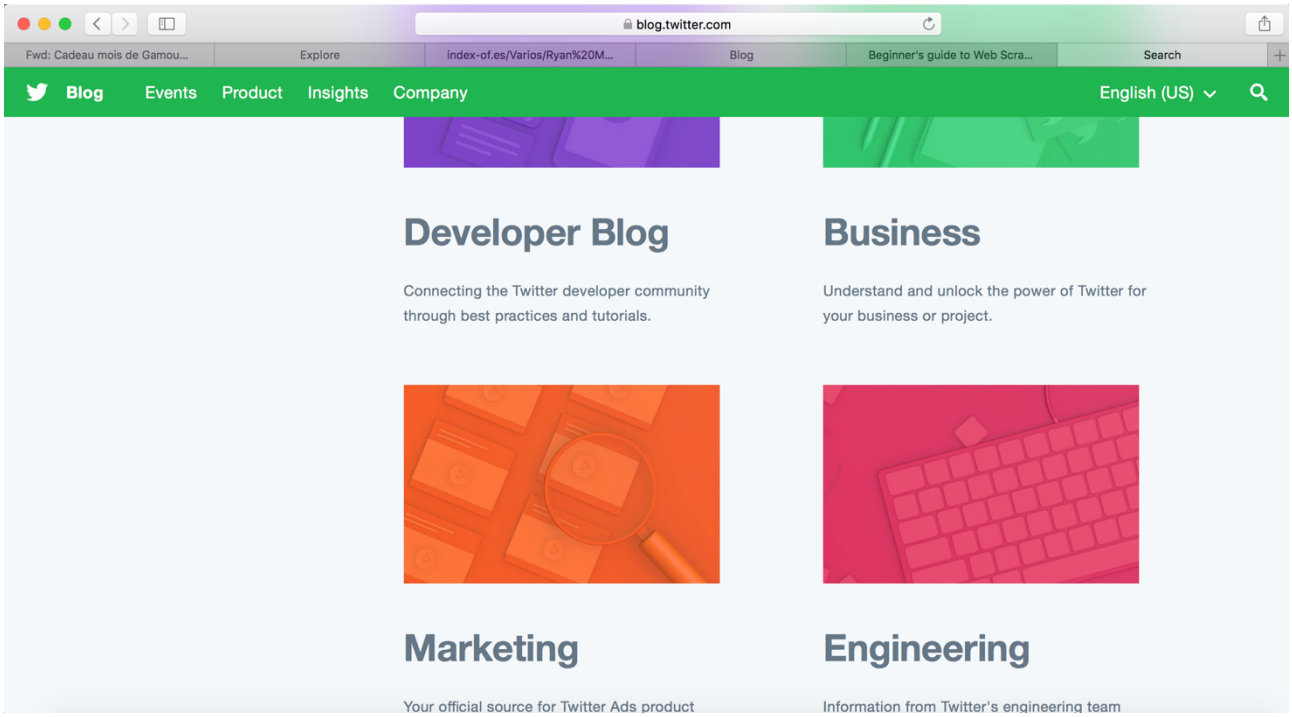
David blei's LDA model [13]



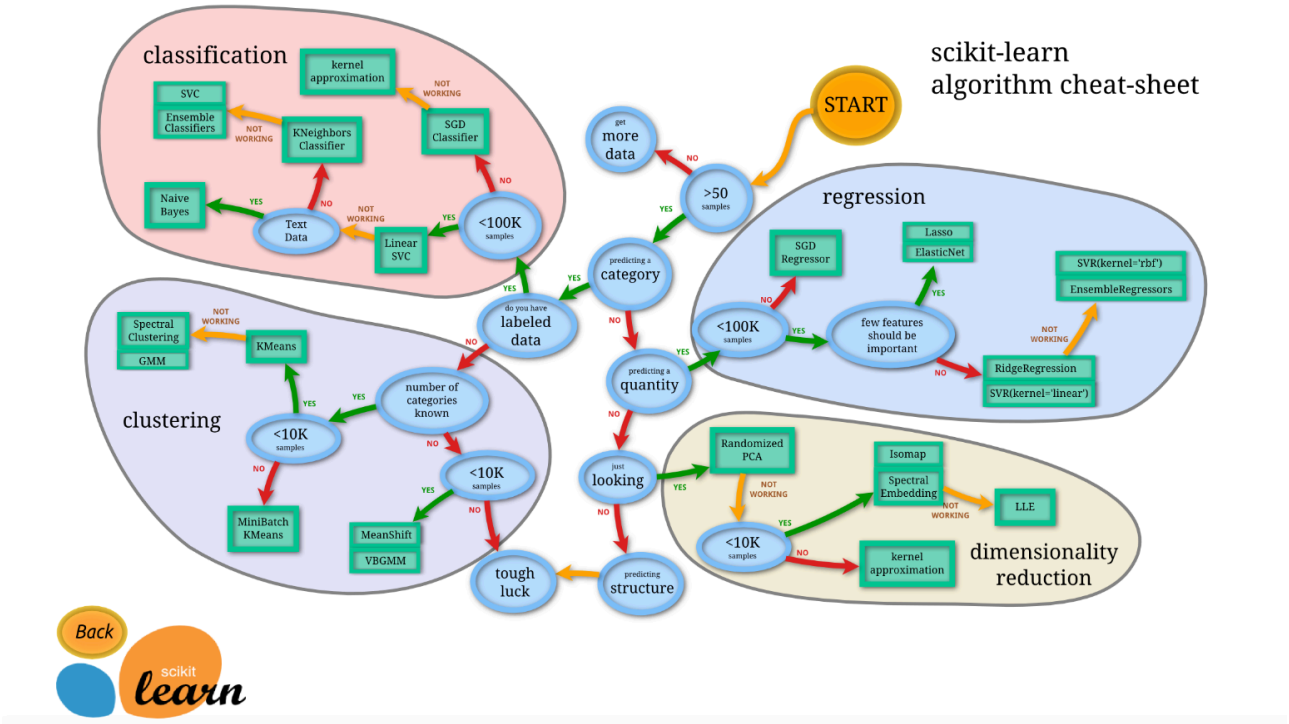
3. Twitter blog interface : twitter blogs have been updated the 20 july 2017 with a new interface and a new grouping of the

blogs.

[13] http://www.scottbot.net/HIAL/index.html@p=221.html

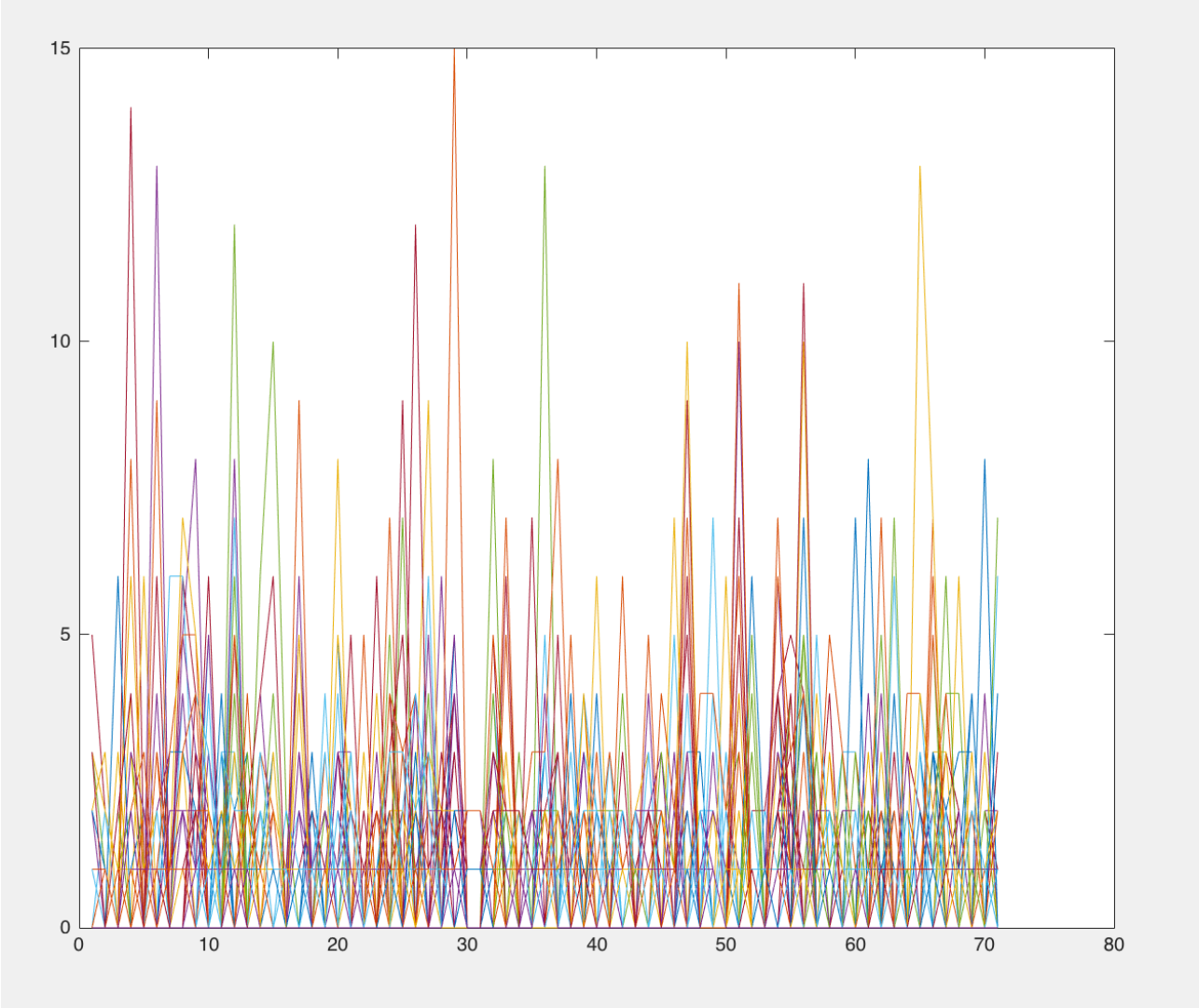

blog.twitter.com
Fwd: Cadeau mois de Gamou...
Explore
index-of.es/Varios/Ryan%20M...
Blog
Beginner's guide to Web Scra...
Search
Blog
Events
Product
Insights
Company
English (US)
Developer Blog
Connecting the Twitter developer community through best practices and tutorials.
Business
Understand and unlock the power of Twitter for your business or project.
Marketing
Your official source for Twitter Ads product
Engineering
Information from Twitter's engineering team

4. Often the hardest part of solving a machine learning problem can be finding the right estimator for the job.

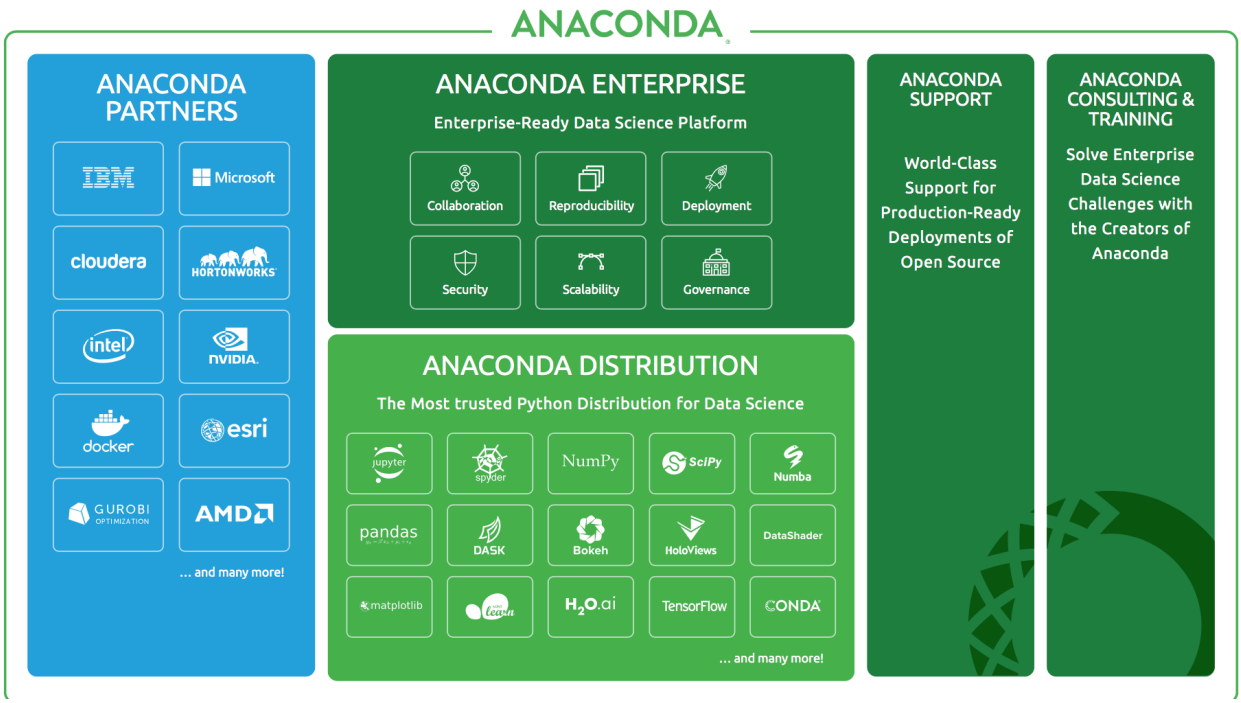


http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

7. Plotting the document term matrix:



8 Anaconda python distribution



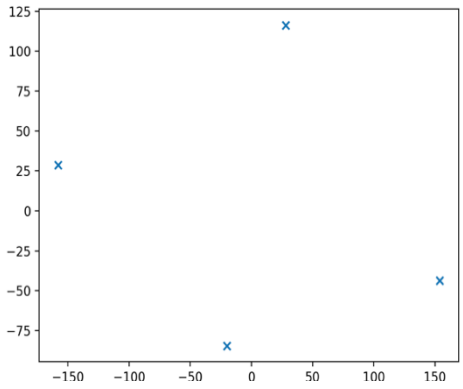
9 cosine similarity of 2 blog posts belong to the same tags: announcements.

blogtext1 = 'Today, we're continuing to roll out to all users the Twitter data dashboard — a new tool to help you monitor and manage your account. From the beginning, Twitter has empowered people to share information with the world. To put you in control of your information, we've made a series of deliberate design decisions that help protect your privacy and security. For example, you don't need to use your real name on Twitter. Your privacy settings let you control whether your Tweets are kept public, and you can enable login verification for greater account security. We respect Do Not Track, and we secure your Twitter experience with HTTPS by default, StartTLS and forward secrecy. Now, your Twitter data dashboard — which you can access from the settings menu on twitter.com — shows your account activation details, the devices that have accessed your account and your recent login history. With this information, you can quickly review your account activity and verify that everything looks the way it should.",
"If you see login activity from an app that you don't recognize, you can go to the apps tab in your settings to revoke its access to your Twitter account. If you notice logins from suspicious locations, you can change your password immediately, and you can enroll in login verification for extra security. From your dashboard, you can also manage your uploaded address book contacts, download your Twitter archive, and more. Visit our Help Center for additional information. Your privacy and account security remain a priority for us and we look forward to sharing news regarding additional tools in the future.",
"In 2013, following the revelations by Edward Snowden about the scope of national security surveillance both domestically and abroad, Twitter joined with a number of other technology companies to seek concrete reform in Washington, D.C. of our surveillance laws and practices. The Reform Government Surveillance Coalition has been fighting on Capitol Hill to pass the USA Freedom Act.The bill was introduced with clear objectives: explicitly ban bulk collection of telephony and Internet metadata; create a Public Advocate in the Foreign Intelligence Surveillance Court (FISC) — the court that reviews and authorizes government surveillance — to argue against the government when the requested surveillance is perceived to be overbroad or otherwise in conflict with the law;'

blogtext43 = There are many ways to see what's happening on Twitter. Outside of your timeline, trends show you what topics are being discussed right now, Moments capture the most popular stories so you can catch up, and search helps you find anything and everything. Until today, you had to go to a few different places to find each of these experiences. As part of our continued efforts to make it easier to see what's happening, we're bringing all these together. Very soon, you'll be able to find trends, Moments, search, and the best of live video, all within the new Explore tab. Over the past year, we've been exploring different ways to make it simpler for people to find and use trends, Moments, and search. During our research process, people told us that the new Explore tab helped them easily find news, what's trending, and what's popular right now. Nothing is going away – we're just making it easier to find what you want. Explore will begin rolling out today on Twitter for iOS, and in the coming weeks on Twitter for Android. Make sure you have the latest version of your app to check it out. And of course, we will continue to listen to your feedback to make Explore even better, based on your thoughts and some ideas we have up our sleeve!.'

10 Defining the centroids in kmeans

```
vect = TfidfVectorizer()
X = vect.fit_transform(documents)

clf = KMeans(n_clusters=4)
clf.fit(X)
centroids = clf.cluster_centers_


tsne_init = 'pca'  # could also be 'random'
tsne_perplexity = 20.0
tsne_early_exaggeration = 4.0
tsne_learning_rate = 1000
random_state = 1
model = TSNE(n_components=2, random_state=random_state,
             init=tsne_init, perplexity=tsne_perplexity,
            early_exaggeration=tsne_early_exaggeration,
            learning_rate=tsne_learning_rate)

transformed_centroids = model.fit_transform(centroids)
print (transformed_centroids)
plt.scatter(transformed_centroids[:, 0], transformed_centroids[:, 1], marker='x')
plt.show()
```



11 Project flow with supervisions as vertical orange line