.5ex 2em

ASSOCIATION ANALYSIS AND TWEETS

Project for the course Algorithms for massive datasets (DSE)

Martina Corsini
Marzio De Corato

Facts are stubborn things, but statistics are pliable
Mark Twain

"We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study."

ABSTRACT

The aim of this research is to perform an association analysis on several textual datasets, each of which is made up of tweets that contain a specific hashtag, so to understand which words or groups of words are linked with the hashtag itself and the relevance of these connections. The datasets are extracted from Twitter exploiting the Twitter APIs, that make it possible to access public Tweets by searching for specific hashtags. After a pre-processing phase of data clean-up, in which all characters are converted to lowercase and stop words and punctuation marks are removed, and tokenization, two algorithm, implemented in Spark [20] and MLxtend [14] , are used in order to reach the goal, the FP-growth and the A-priori algorithms. Depending on the technicality and specificity levels of the hashtags used to retrieve the dataset, the words and the groups of words associated to a dataset can belong to only one or several fields. Such behaviour is well-captured by the shape of the associated support distribution: therefore, it could be possible to use this characteristic to gain a first assessment of the quality of information retrieved with a particular hashtag.

CONTENTS

1 INTRODUCTION

This paper is organized according to the following scheme. A first section will provide a rapid theoretical overview, based on cited textbooks, of the concepts and the techniques used in this work. In particular, we will take into account the many steps involved in the process of preparing text data for use, such as the cleaning and the tokenization; subsequently, we will consider the theoretical framework of the association analysis and the algorithms involved, with a particular focus on its connection with the information theory and on the possibility to improve memory management. A second section will describe how our datasets were generated, describing the process of extraction of tweets with common hashtags from Twitter through the Twitter APIs and the ratio behind the choice of the hashtags, namely to have several levels of specificity and technicality and to have datasets concerning different topics. The last paragraph will contain a discussion of the results obtained for each dataset. Finally, we will propose a possible way to obtain a first approximation about the quality of the information obtained from a hashtag.

2 THEORETICAL FRAMEWORK

2.1 TEXTUAL DOCUMENTS MINING

The kind of data one gets from social media is usually unstructured. It contains, for example, a vast number of stop words and symbols, but also typos and abbreviations that need to be cleaned up so that a model can grasp the truly interesting associations between the words in the texts. The reliability of the model is highly dependent upon the quality of the data. The process of preparing text data for the analysis is called text pre-processing. Many steps can be involved in this task. First of all, it may be helpful to get rid of unhelpful parts of the data, or noise, by converting all characters to lowercase, removing stop words and punctuation marks. In particular, removing stop words (the most common words like articles, prepositions, conjunctions etc.) allows the model to consider only key features because these words typically do not contain much information [8]. Many variations of those words do not bring any new information and create redundancy, ultimately bringing ambiguity when analyzing the critical information. Moreover, it is crucial to remove stop words since this process reduces the dataset size. Concerning the Python framework, NLTK (Natural Language Toolkit) [4] has a list of stop words stored in 16 different languages. It is also possible to add other words in the list according to the task we perform and the goal we want to achieve. Finally, text must be tokenized. Tokenization is the process of splitting a text document into smaller units, such as words, numbers or punctuation marks, that are called tokens. Finally, in order to make our analysis, we must vectorize the text, for example using the bag of words model, in which each tweet is represented as a word-count vector, with a size equal to the number of elements in your vocabulary. It is important to note that the resulting matrix will be sparse since most of its elements will be zero.

2.2 ASSOCIATION ANALYSIS

In this section, we are going to recap the main concepts and algorithms involved with the market basket analysis. After a brief introduction about the main indices involved, in which we will follow the Leskovec et al. [10] as well the Tan et al. approaches [16], we will report how such indices can be connected with the information theory. For this purpose, we will follow the Yao paper [19]; furthermore, concerning the rapid exposition about the main concepts of the information theory, we will consider the approaches of the two standard textbooks about this topic, written by D.MacKay [11], and T.Cover [7]. Finally, following the Tan et al. approaches [16], we will analyze and compare the two algorithms used in this research for the association analysis: the A-priori and the FP-growth.

2.2.1 Basic concepts

Given a large dataset, a first-order approach to mine useful information from it is to perform an association analysis: the underlying idea of this approach is that itemsets that appear together may[1] reveal an interesting relationship. Such analysis is commonly referred to as basket analysis, since one of the first and common uses of this technique is to retrieve itemsets that are bought together; however, such technique can be used in a number of field, such as for the analysis of the earth science data or even for astronomical purposes [6]. It is worth noting that the patterns discovered with this technique can contain spurious correlations. Therefore, the conclusions obtained can be seriously affected by the Simpson paradox [15, 5]. In essence, the Simpson paradox states that the existence and the order of causality can depend by the variables considered: the removal or the inclusion of them can seriously alter the inference obtained. Thus, the conclusions obtained from an association analysis should be validated by a model and not taken recklessly for granted. With this important premise in mind, let us move to consider how this analysis works. A market basket dataset can be represented with a binary table in which each row i represent a case to be analyzed (transaction) and each column j an element that may be present in the transaction. The matrix element m_{ij} is set to 1 if the object j is present in the case

1 But not necessary

(transaction) i, otherwise it is equal to o (perhaps a boolean TRUE/FALSE can also be considered). Once defined how the transaction can be described into a compact form, we can move to the description of indices that highlight the important associations. For this purpose, we label with $I = \{i_1, i_2, ..., i_d\}$ the set of all itemset and with $T = \{t_1, t_2, ..., t_d\}$ the set of all transactions: in this way, the support count $\sigma(X)$ can be defined as follow [16]:

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}| \tag{2.1}$$

Such quantity can be normalized with the overall number of transaction N; the resulting value [16]:

$$s(X) = \frac{\sigma(X)}{N} \tag{2.2}$$

is called support index. If this value for an item X is greater than a certain threshold arbitrarily chosen called *minsup*, item X is said to be frequent. By now, we focused on single items, but our propose is to find a connection between them: in particular, we are interested in finding associations, or implications, $X \rightarrow Y$ with the condition that $X \cap Y = \varnothing$. These can be measured with the *support* index [16]:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{2.3}$$

as the formula shows, such an index measures how often a rule can be applied for the chosen dataset. In addition, one can replace the denominator of the following expression with the support index of X, thus giving [16]:

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2.4}$$

this index is called *confidence* and measures the frequency of item Y in the transactions that contain item X. The meaning of these two indices will become more evident within the information theory framework provided in the following paragraph: here we would anticipate that, as pointed out by Tan et al.[16] a low support signals an association that is likely to happen by chance, while the second one provides an estimate of the conditional probability of Y given X. Finally, another helpful index for the assessment of an association rule is the *lift* defined as [17]:

$$lift = \frac{P(A \cap B)}{P(A) \cdot P(B)} \tag{2.5}$$

If the lift has an absolute value vastly different from 1, we have an association rule that largely prevails on the product between the ratios of the two separate events; on the contrary, if the lift value is close to one, the value of the association is poor. As we will see in the paragraph dedicated to the information theory, the lift index is connected to the point mutual information concept. On this basis, following the Tan et al. approach, we can state the Association Rule discovery problem as [16]:

> **Association Rule Discovery**: Given a set of transactions T, find all the rules having support $\geqslant$ *minsup* and confidence $\geqslant$ minconf where *minsup* and *minconf* are the corresponding support and confidence thresholds

What is the computational effort required for this task ? It can be proven that with a dataset that contain d items the possible association rules are [16]:

$$R = 3^d - 2^{d+1} + 1 \tag{2.6}$$

As one can point out, this number grows fast with d; thus it is crucial to prune the rules that will never have sufficient support and confidence values in order the make the association rule discovery task affordable. This aim can be achieved with the algorithms that will be reviewed at the end of this section.

MEMORY MANAGEMENT In order to store the pair counters, one could build a matrix. However, in this way there is a waste of memory that amount to half the space, because we are not interested in the order of the association. A possible solution can be the linearisation of the matrix, so a one-dimensional representation of the association with no wasted space. Furthermore, the efficiency in finding pairs is maintained using the formula k = (i-1) (n-1/2) + j- i, which makes us able to know the correct position of the counter. If the matrix is sparse, another possible and more space-efficient alternative is to represent counts of pairs as triples (i, j, c). An index structure such as a hash table allows us to efficiently find the triple for (i, j). It is important to note that we can have collisions, so using the linearised matrix is preferable if we have no space problems.

2.2.2 Information theory and its connections with association analysis

This subsection was added by Marzio De Corato in order to connect his background with the theoretical framework of this work

As stated by Cover textbook, [7], the information theory comes out from the issue of how much and at which rate data can be compressed. Indeed, its initial purpose was to quantify how much a signal can be compressed and send via a finite communication channel. However, as its founder, Claude Shannon, discovered, such problem that was initially conceived within the communication theory involves many fields of science such as statistical physics, computer science, statistical inference, and economics: a bird's-eye view about the formidable ability of this theory to combine different disciplines is given in Fig. 2.1. The fundamental concept of information theory is entropy. In this context, such quantity, which was originally conceived in physics to explain the lack of symmetry in time-reversed transformation for physical processes [2], measures the degree of uncertainty of a random variable. Formally, given a random variable X with the accessible states (alphabet) $\mathcal{X}$ and probability distribution $p(x) = \Pr\{X = x\}, x \in \mathcal{X}$ [3], the entropy of this variable is defined as [7]:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2.7}$$

where the logarithm is in base 2. With this definition, the entropy, which is usually referred to as Shannon entropy, is measured in bits. Once defined this quantity for a single variable, we can move to two variables. The joint entropy $H(X, Y)$ of a discrete random variables with joint distribution $p(x, y)$ is given by the following expression [7]:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{2.8}$$

Furthermore we can define the conditional entropy as [7]:

$$H(Y|X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \tag{2.9}$$

2 like the expansion of gas

3 It is worth noting that this is the standard definition of a stochastic variable [18]

The joint entropy and the conditional entropy are mutually bounded by the following relation [7]:

$$H(X,Y) = H(X) + H(Y|X) \tag{2.10}$$

Finally we can measure the mutual information of two random variables as

$$I(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2.11}$$

Note that the entropy of the single variable are bounded to the mutual information according to the following expression [7]:

$$I(X,Y) = H(x) + H(y) - H(X,Y) \tag{2.12}$$

As can be pointed out such quantity is an average between the two alphabet, in our case where we are interested in the single associations, the point-wise mutual information should be considered [7]:

$$i(X,Y) = \log\frac{p(x,y)}{p(x)p(y)} \tag{2.13}$$

It is worth mentioning that the mutual information can be recast in a more elegant way with the Kullback-Leibler distance. Lets focus for while on this crucial concept. Given two probability distribution p(x) and q(x) the Kullback-Leibler distance or the relative entropy is defined as [7]:

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \tag{2.14}$$

The meaning of this quantity, not by chance called distance, is that given a distribution probability p, it would describe it with another distribution q we would have an excess of entropy equal to $D(p||q)$ concerning the entropy $H(p)$. In this way, the KL distance is an inefficiency measure connected with the distribution p. Moreover, as stated by the Gibbs inequality [11]:

$$D(p||q) \geqslant 0 \tag{2.15}$$

the inefficiency can always be positive or at least null[4]. This is the formalization, and more importantly, the quantification,

4 This is the equivalent, in the information theory framework, of the second principle of thermodynamics. This is particularly clear looking to the M.Plank statement: *Every process occurring in nature proceeds in the sense in*

of the argument that given a random variable (the weather of tomorrow) we cannot reach with a forecast an entropy lower concerning the phenomenon itself. We would remark that these measures, which were initially used for the inference of a transmitted signal, can be applied basically for every inference problem, including the object of this work: the inference of the association pattern within a dataset. Once defined and clarified the fundamental concept of DL distance, we could recast the mutual information expression into the elegant form [11]:

$$I(X,Y) = D(p(x,y)||p(x)p(y)) \tag{2.16}$$

Based on this theoretical framework, following the Yao paper [19], we can recognize in the logarithm (with base 2) of lift index, described before, the point-wise mutual information described before. This allows, as one expects, to use the formalism and the results achieved by the information theory in data mining. We say as one expects because we can think of the dataset as a communication channel in which nature encodes the association rules: the data mining process can be seen as the decoding of these association rules given a dataset (a remarkable analysis on how the information theory concepts can be used in the data mining context can be found in [21]). A remarkable description of the connections between learning algorithms and the information theory is provided by D.MacKay [11] and by E.T.Jaynes [9] .

2.2.3 Algorithms

Before reviewing the algorithms used in this work for the frequent item generation, we would introduce the itemset lattice: in each n-level (line) the possible combination between n elements is given (starting from 0, the avoid ensemble). Furthermore, each cell is connected with the cell that contains a subset of itself. The advantage of such representation will become evident during the A-priori description. There are three way by which the computational costs can be reduced: the first possibility is to reduce the number of candidates (this is the underlying idea of A-priori principle); the second option is to reduce the number of comparisons by exploiting an intelligent data structure (this is the idea of the FP growth algorithm); finally, one

which the sum of the entropies of all bodies taking part in the process is increased. In the limit, i.e. for reversible processes, the sum of the entropies remains unchanged [13]

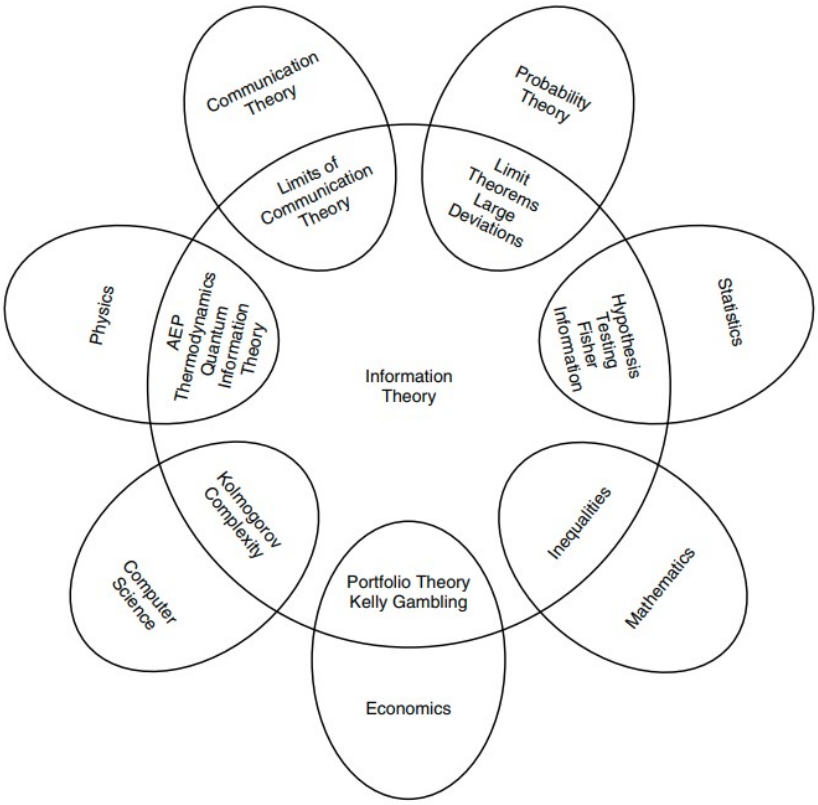


Figure 2.1: Connections of information theory with different science fields. Image taken from [7]

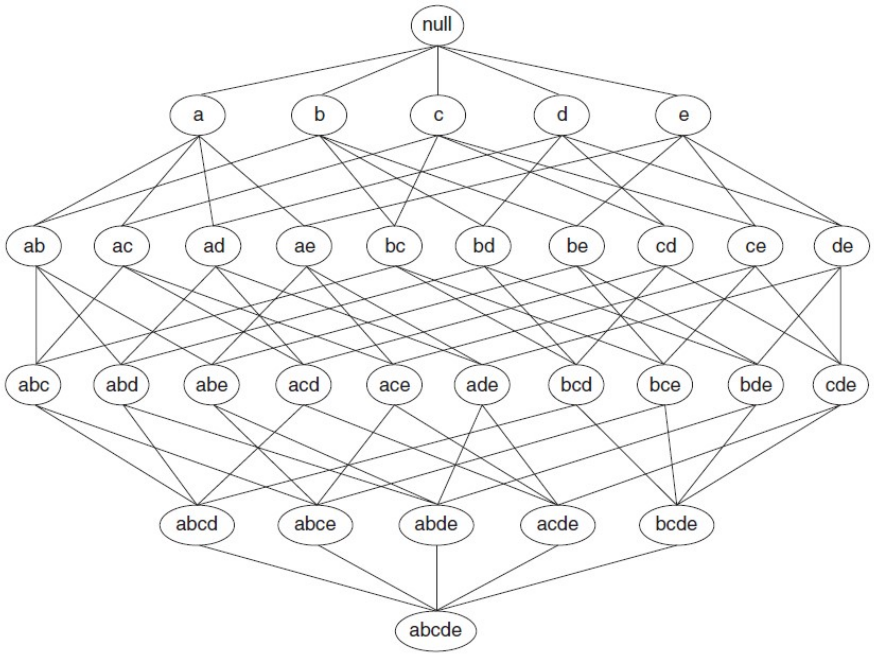


Figure 2.2: An example of itemset lattice for $I = \{a, b, c, d, e\}$. Image taken from [16]

can reduce the number of transactions. In this work, we will consider the algorithms that use the first two approaches; thus, we would limit our theoretical analysis to these two only.

A PRIORI The A-priori algorithm is based on the following principle, as stated by Tan et. al [16]:

> **Apriori principle**: If an itemset is frequent, then all of its subsets must also be frequent

The support benefits of the Anti-monotone property [16]:

> **Anti-monotone property**: A measure f possesses the antimonotonone property if for every itemset X tha is a proper subset of itemset Y, i.e $X \subset Y$ we have $f(Y) \leqslant f(X)$

Therefore, we can prune the candidates by fixing a support threshold and removing those whose subsets are not frequent enough. The crucial point, as shown in Fig. 2.3 is that the algorithm starts from the smallest subset, and then, if an infrequent item is found, all the larger subset that include it are removed

from the candidates. In this way the overall number of candidates is vastly reduced, thus lowering the computational cost required for the association analysis.

FP GROWTH The Frequent Pattern (FP) algorithm represents an alternative approach to the A-priori: here, the key concept for reducing computational cost is played by how the data are stored, instead of the candidates' pruning. For this purpose, the dataset is stored in to the FP-tree: this is a diagram, whose construction is explained in Fig. 2.4 , in which each node represent a label of an item accompanied by a counter that reports the number of transactions mapped on the given path. The key point is that different transactions may have elements in common: thus, since some nodes in the diagram may overlap, the overall number of nodes is reduced (more precisely, we have an overlap between two paths if they have a common prefix). As we will see, the computation cost for the association analysis, given a FP-tree, depends on the number of distinct nodes: if this number is reduced, the computational effort is reduced. Now that we described the core concept for the data compression, we move on to how the frequent itemset are generated. This procedure is performed through a bottom-up approach in which the algorithm first evaluates the support for a single suffix, then evaluates the support of two element suffix, then for three element suffix, and so on. The evaluation of the support can be quickly obtained by evaluating the paths that contain a chosen node (this is why also the dashed lines were included in the diagram reported in Fig. 2.4). As illustrated in Fig 2.5, once a particular suffix is fixed, the algorithm generates the conditional trees that are nothing more than the part of the FP tree that verifies a particular condition i.e., to end with a particular suffix. The key point that represents one of the significant advantages of the FP algorithm is that once a particular suffix is chosen, the conditional tree that is generated provides a new subset of problems (see Fig. 2.5) in this way, the complex problem of the association analysis is attacked with a divide and conquer strategy. At this point is clear how the run-time of the FP algorithm is strictly dependent on the number of distinct nodes: if the overlap is low, the performance of this approach degrades. On the contrary, if the compression of the transaction dataset into the FP-tree is high, this method can outperform the A-priori algorithm by a different order of magnitude [16]

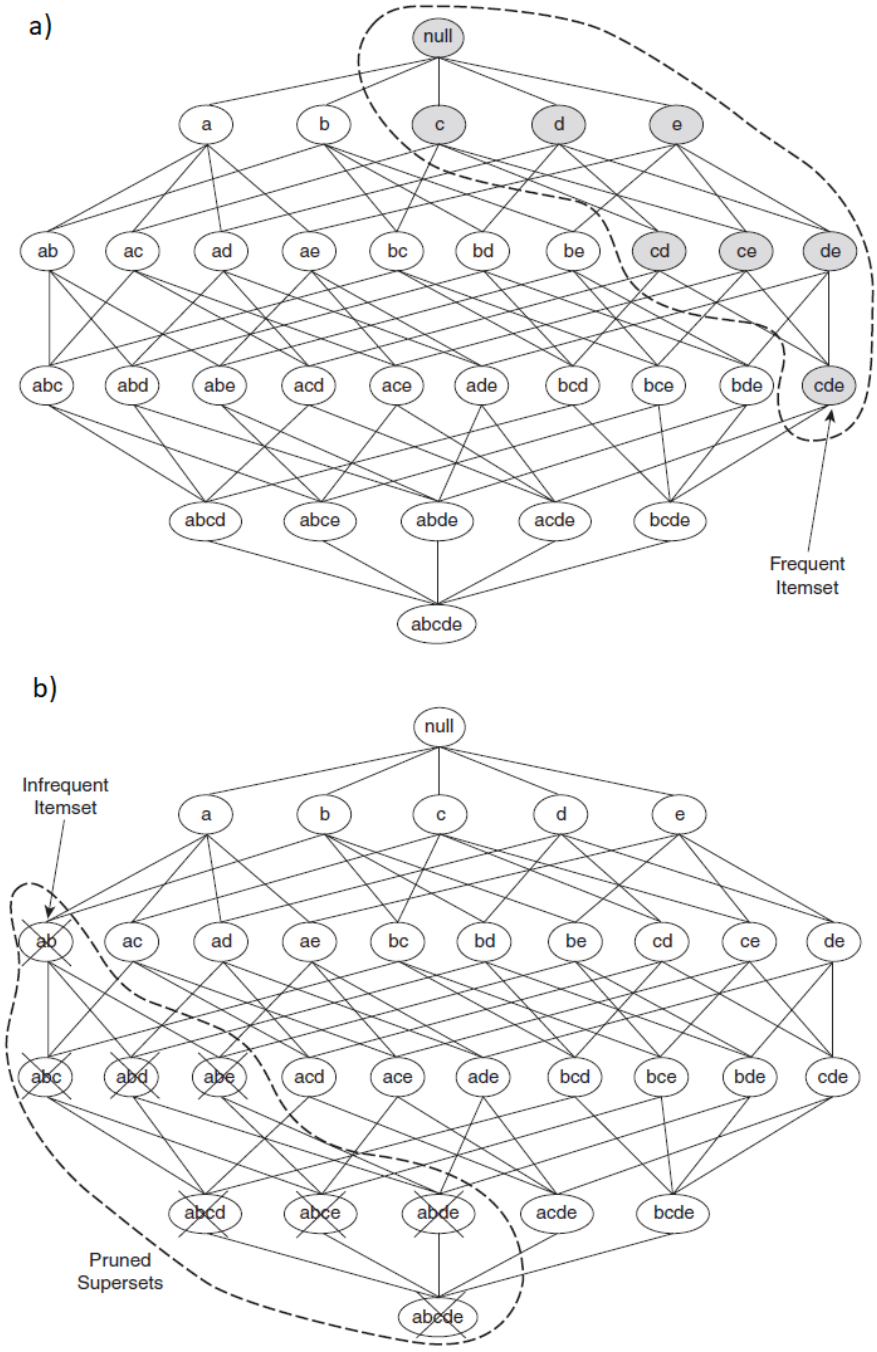


Figure 2.3: The mechanism of A-priori algorithm: in the upper panel the fact that the $\{cde\}$ is frequent implies that the subset by wich it is composed are also frequent. On this basis one can implement an algortihm that prunes the larger itemset by investing its subset: if a subset is infrequent all the set that include it are also infrequent (lower panel). This approach considerably reduce the candidates and thus the computational cost of the association analysis. Images taken from [16]

Transaction Data Set

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

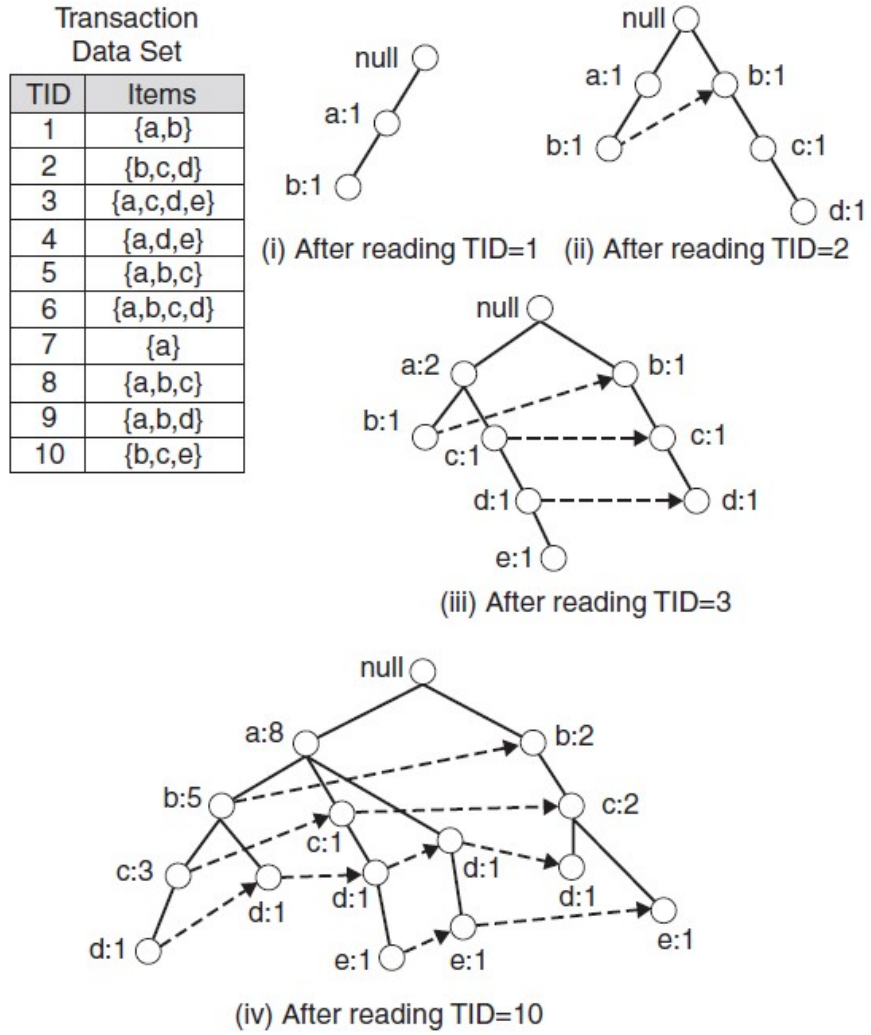


Figure 2.4: The implementation of an FP-tree given a transaction. Note the dashed lines which provide the paths that connect identical elements in the diagram, but are not in the same path due to the fact that there is no common prefix. Image taken from [16]

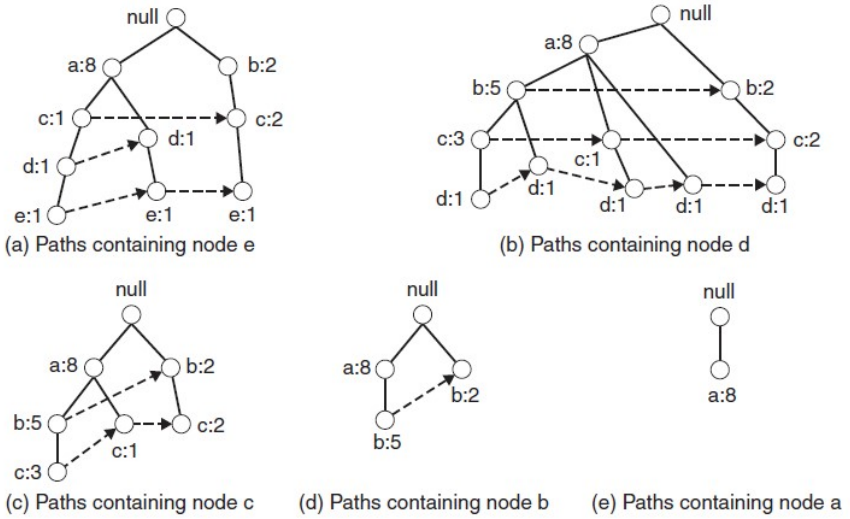


Figure 2.5: The problem of basket analysis is decomposed into different sub-problem for each suffix. Image taken from [16]

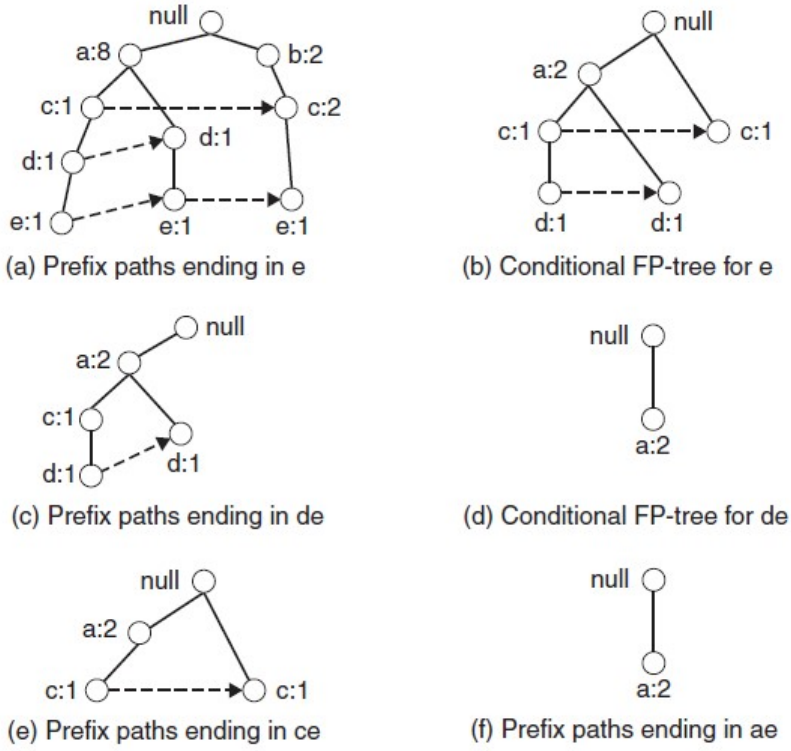


Figure 2.6: The conditional FP-trees generated as a condition is imposed. Image taken from [16]

2.2.4 PCY variant

The a priori algorithm is characterized by a waste of memory in the first pass of the process: in fact, a considerable portion of RAM is not used, namely the amount of RAM used in the second pass to count the pairs. The PCY variant of the a priori algorithm uses the remaining part of the memory in the first pass to reduce the candidates. In fact, it constructs a hash table on the first pass, using all main-memory space that is not needed to count the items. Pairs of items are hashed to be mapped into buckets; the buckets are used as counters of the number of times a pair hashed to that bucket. On the second pass, we only have to count pairs of frequent items that hashed to a frequent bucket (one whose count is at least the support threshold) on the first pass.

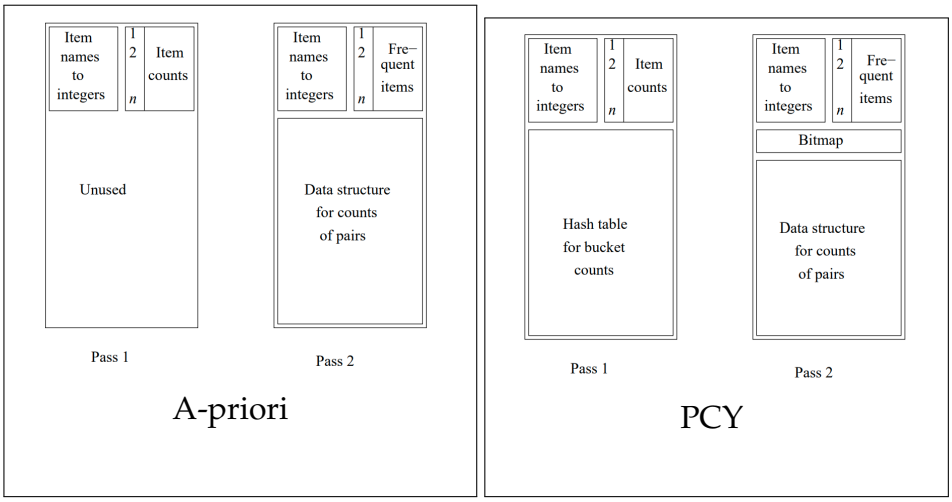


Figure 2.7: Use of memory for A-priori algorithm and for PCY variant. Image taken from [10]

3 DATASET GENERATION

3.1 TWEET RETRIEVAL

Twitter and other social media platforms represent an important and updated source of information with regard to the development of our society, since social media platforms generate a vast amount of data daily on a variety of topics. Therefore, it is important to be able to retrieve tweets in order to be able to analyze them. In order to reach this goal, Twitter offers companies, developers, and users programmatic access to Twitter data through API (Application Programming Interfaces). Using the Twitter APIs, developers can access public Tweets by searching for specific keywords [1]. To retrieve all tweets with a specific hashtag, it is possible to use the search/tweets GET endpoint. By default, the results return tweets sent in the previous seven days. Tweets are delivered in reverse-chronological order, starting with the most recent. Therefore, the sample extracted can not be considered, in principle, a random one. However, under the condition that there is no particular event from the first to the last tweet extracted that alters the association,[1] we can consider the sample as randomly extracted. Furthermore, it is also possible to select the language of the tweets. The resulting csv file is available in the following format: hashtag | timestamp | tweet_text | user_name | language . For the purpose of our project, only the tweet_text is necessary, since we will consider each tweet as a basket and the words contained in them as the items.

3.2 CHOICE OF THE DATASET

We chose to work on five datasets composed of 10000 tweets that contain respectively the following hashtag: *freedom, Afghanistan, rare earth, Covid-19 and Pfizer*. The ratio behind the choice of the hashtag has been to have tweeted with several levels of specificity and technicality, and concerning different topical ques-

1 We verified that this condition was true from different on-line press agencies

tions, in particular the Afghanistan crisis and the pandemic situation. The hashtag *freedom* is a generic hashtag that can be linked both with the rise to power of the Taliban regime and with the possibility to make vaccination compulsory. *delta*, *Pfizer* and *vaccines* are related to the pandemic, but they show a different level of technicality, since vaccines is a common word while the others two refer to a more technical field, even if they have become very popular in the public debate. *Afghanistan* and *rare earth* refer to the Afghanistan situation: in this case, the latter shows a high level of specificity since it is about an issue that is not widely known, namely the interest of some countries to exploit Afghanistan's rare earth metals, some crucial elements in modern technology.

4 RESULTS AND DISCUSSION

For each term, we conducted a preliminary analysis on its cleaned document-term matrix: as it is possible to point out from Fig. 4.1, a sparse matrix was obtained, as expected, for each hashtag. We then applied the FP-growth algorithm, as implemented in Spark suite of codes [20] , and the A-priori algorithm, as implemented in MLxtend suite of codes [14]. With the first one, we retrieved the frequency for each itemset, the association rule for each basket, and the predictions of the items that are most likely to appear given a basket. A support threshold equal to 0.03 was chosen. A preview of the results is given in the Tab. 4.1, 4.2, 4.3, 4.4,4.5 and 4.6 for the 5 most popular items, while the associations (the five with the higher lifts) are given in 4.7, 4.8, 4.9, 4.10 and 4.11 (note that the table for hashtag freedom is not present since there is no association with a support equal to 0.03). It is worth noting that in the pre-processing phase the hashtag was removed, thus the singleton represents, in this perspective, a binary combination between the word of the hashtag and the singleton itself. For this reason, in our analysis we will consider as multiple association a combination of two or more words, since the hashtag is implicitly included in the association. The full results (in the form of an Excel table) obtained for each hashtag are provided here [2], and a good visual representation of them is provided in Fig. 4.2 and Fig. 4.3. Let us start from the Rare earth hashtag, that present a large number of high frequency peaks given by the words *minerals, china, Afghanistan, metals* and *taliban* (see Tab. 4.1): the presence of the words China, Afghanistan and talibans is associated with this hashtag due to the interest of some countries, such as China, to exploit Afghanistan's rare earth metals. Furthermore, also the multiple association table Tab. 4.1, in which we see a high lift value for the association *lithium* and *copper*, provide interesting informations: in fact, this association is due to the fact that these two metals are largely present in Afghanistan together with rare earth [1]. In the case of the Afghanistan hashtag,

1 The reader must be aware that lithium and copper are not rare-earth. A full discussion about the chemistry and the industrial uses of these materials can be found into a general chemistry textbook such as [12]

we have a support distribution with a very sharp peak connected to *Taliban* and *Kabul*: while the first one is connected because it is the Afghanistan capital (see Tab. 4.2), the second one is an effect of Taliban conquer of Afghanistan happened at the end of August 2021, after US troops withdraw. The importance of the US in the political scenario of the Afghanistan is also captured by the words *us*, *biden*, while the association between Afghanistan and *kabulairport* is given both by the importance of this place for the evacuation of the US troops and by the terroristic attack that took place during the evacuation. These words and their associations provide the first peak and the second broader one. Looking to the multiple associations table (Tab. 4.8), we can see that the association with larger lift is given by *nowazd* and *usarmy*; this connection is due to the Nowzad Dogs charity, whose supporters provided financial support for the evacuation of the association staff and their animals [3]. The following association are basically connected with the capture of Kabul by Taliban and the civilian evacuation. Concerning the hashtag delta, we can see that we have an overall sharp peak made firstly by the words *covid19*,*covid*, *variant*, *amp* and *vaccine* (see Tab. 4.3). This findings clearly signal the extensive use of word delta within the pandemic framework, while other fields in which this word appears, such as, for instance, the delta airlines, does not contribute to the spread of this hashtag in the social media. An inspection of the multiple associations between the word delta and other words, reported in Tab 4.9, highlights the words *vaccine*, *covid19* and *variant*: however, in contrast to the previous case (hashtag Afghanistan), we can note that the maximum lift is very close to 1, thus the interest of these associations is poor. Moving to the hashtag pfizer, we see that we have a very sharp peak connected with the term *vaccine*, followed by *covid19*, *fda* (food and drug administration) and by the association *covid19*,*vaccine*. In this case, as we can note looking to Tab 4.10, we have an interesting lift value for *full*,*vaccine*: this is the effect of the public debate about the vaccination campaign, also connected with the discussion on the possibility of a third dose, as one can see from the lift value of the association *full, approval*. Concerning the hashtag vaccines, we point out the presence of large peak at the origin connected with the word *covid19*, followed by *covid* and *amp*. With regards to the multiple association, we can see that these has a lift value close to 1. Finally, the distribution connected with freedom is almost horizontal, highlighting that this hashtag is associated

Table 4.1: Most popular items for hashtag *RareEarth*

	Support
Minerals	3215
China	2948
Afghanistan	2400
Metals	1804
Taliban	1551
Us	1217

with several fields, and therefore a low information content can be retrieved with it. The lack of interesting associations is also captured by the fact that no multiple association with a support equal or larger with respect to 0.03 is present, thus in this case not table is provided. In order to quantify this contrast, we propose to evaluate the area under the curve (remembering that the curves here reported referring only to the itemsets whose support is larger to the chosen threshold): the idea is that if this area is lower, the specificity of the hashtag searched is very low; on the contrary, if the area is high we have a fine specificity (i.e. only a few words are usually combined with a hashtag with the high area). From an information theory point of view, the first case corresponds to a large entropy (every state is equally probable), the second one to low entropy (few probable states). The results of this evaluation for the hashtag here considered is provided in Tab. 4.12: the RareEarth is associated with the higher value, followed by Afghanistan and Pfizer; on the other side, freedom has low specificity since it can be associated with a cumbersome amount of words. In the author's view, the area can be considered as an assessment of the information content obtained by mining tweets with this word: if a word with a high area (above the support threshold) is used, it is associated almost with specific terms. Thus the tweets retrieved used with this hashtag were will more likely to provide useful information. On the other side, a hashtag connected with a low area will provide many spurious tweets, and thus, in the end, no useful information will be retrieved.

Table 4.2: Most popular items for hashtag *Afghanistan*

	Support
Taliban	1803
Kabul	1675
Us	1598
Biden	1135
Amp	1016
Afghan	914

Table 4.3: Most popular items for hashtag *Delta*

	Support
Covid19	2728
Covid	1774
Variant	1721
Amp	1189
Vaccine	968
Cases	836

Table 4.4: Most popular items for hashtag *Pfizer*

	Support
Vaccine	3821
Covid19	2607
Fda	1847
Covid19	1467
Approval	1120
Moderna	1086

Table 4.5: Most popular items for hashtag *Vaccines*

	Support
Covid19	4007
Covid	1596
Amp	1080
Get	965
People	755
Vaccinated	740

Table 4.6: Most popular items for hashtag *Freedom*

	Support
Amp	872
Get	802
People	709
Free	636
Us	532
Dont	506

Table 4.7: Association rule for the hashtag *RareEarth*

Antecedent	**Consequent**	**Lift**	**Confidence**	**Support**
lithium	copper	6.07	0.27	0.03
deposits	lithium	4.56	0.53	0.03
russia	china	2.99	0.88	0.04
china	afghanistan	2.62	0.13	0.04
afghanistan	biden	2.44	0.14	0.03

Table 4.8: Association rule for the hashtag *Afghanistan*

Antecedent	**Consequent**	**Lift**	**Confidence**	**Support**
nowzad	usarmy	21.16	0.68	0.03
nowzad	penfarthing	16.80	0.72	0.03
kabulattack	kabulairport	9.01	0.70	0.03
airport	kabul	4.60	0.77	0.05
kabulairport	kabul	3.49	0.27	0.05

Table 4.9: Association rule for the hashtag *Delta*

Antecedent	**Consequent**	**Lift**	**Confidence**	**Support**
vaccine	covid19	1.65	0.45	0.04
covid19	cases	1.40	0.12	0.03
covid19	variant	1.38	0.24	0.06
covid19	covid	1.15	0.20	0.06
amp	covid19	0.95	0.26	0.03

Table 4.10: Association rule for the hashtag *Pfizer*

Antecedent	**Consequent**	**Lift**	**Confidence**	**Support**
full,vaccine	approval	8.19	0.92	0.04
full	approval	7.08	0.79	0.05
approved,vaccine	fda	3.97	0.73	0.04
approval,covid19	vaccine	2.18	0.83	0.03
full, approval	vaccine	2.04	0.78	0.04

Table 4.11: Association rule for the hashtag *Vaccine*

Antecedent	**Consequent**	**Lift**	**Confidence**	**Support**
covid19	please	1.96	0.08	0.03
news	covid19	1.91	0.76	0.03
covid19	well	1.84	0.08	0.03
covid19	pandemic	1.64	0.09	0.03
covid19	need	1.56	0.10	0.04

Table 4.12: Area under the support distribution curve normalized by the lowest value (Freedom). Note that the items with support lower with respect to the threshold 0.03 were discarded

	Area
RareEarth	6.14
Afghanistan	4.60
Pfizer	2.86
Vaccines	2.85
Delta	2.63
Freedom	1.00

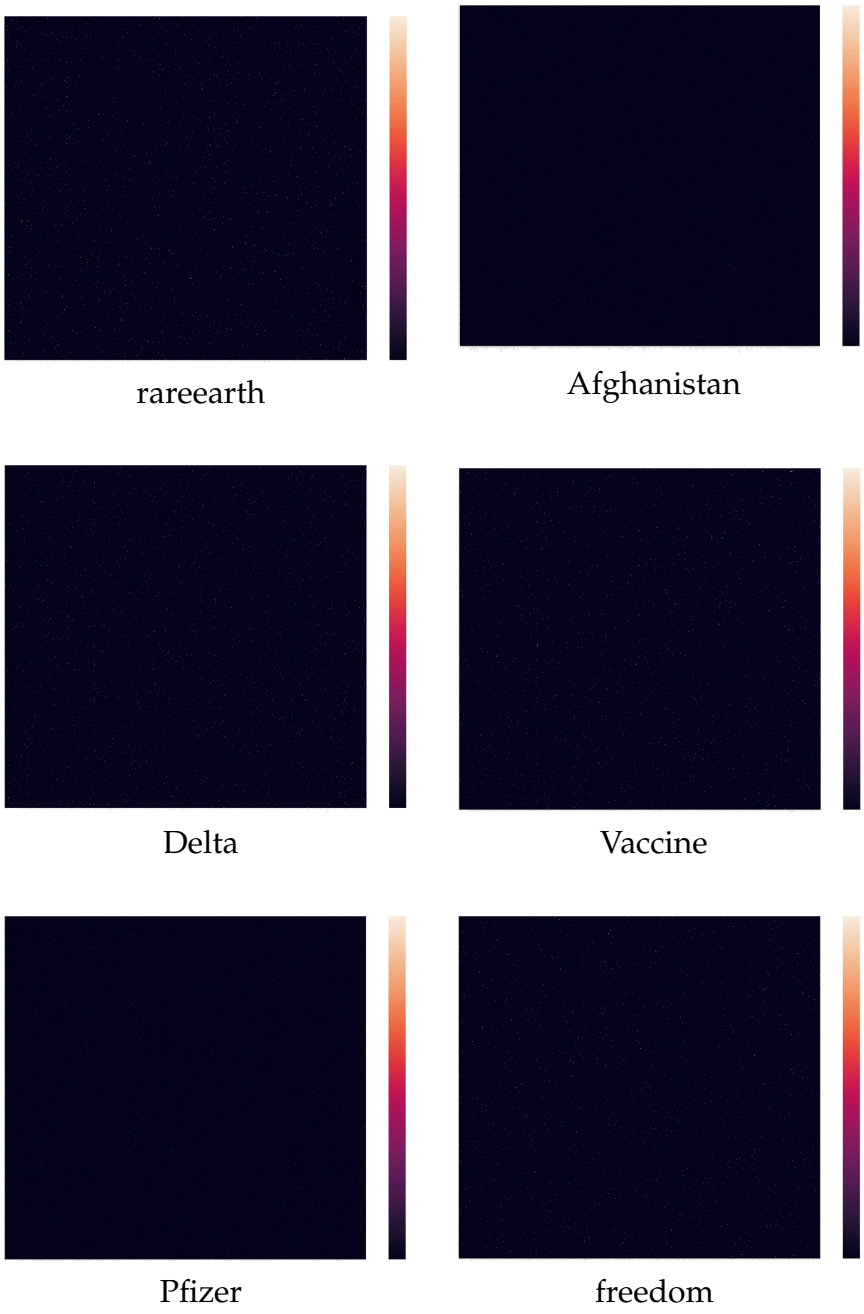


Figure 4.1: Document-term matrix for the different hashtag. The high quality ones are available at [2]

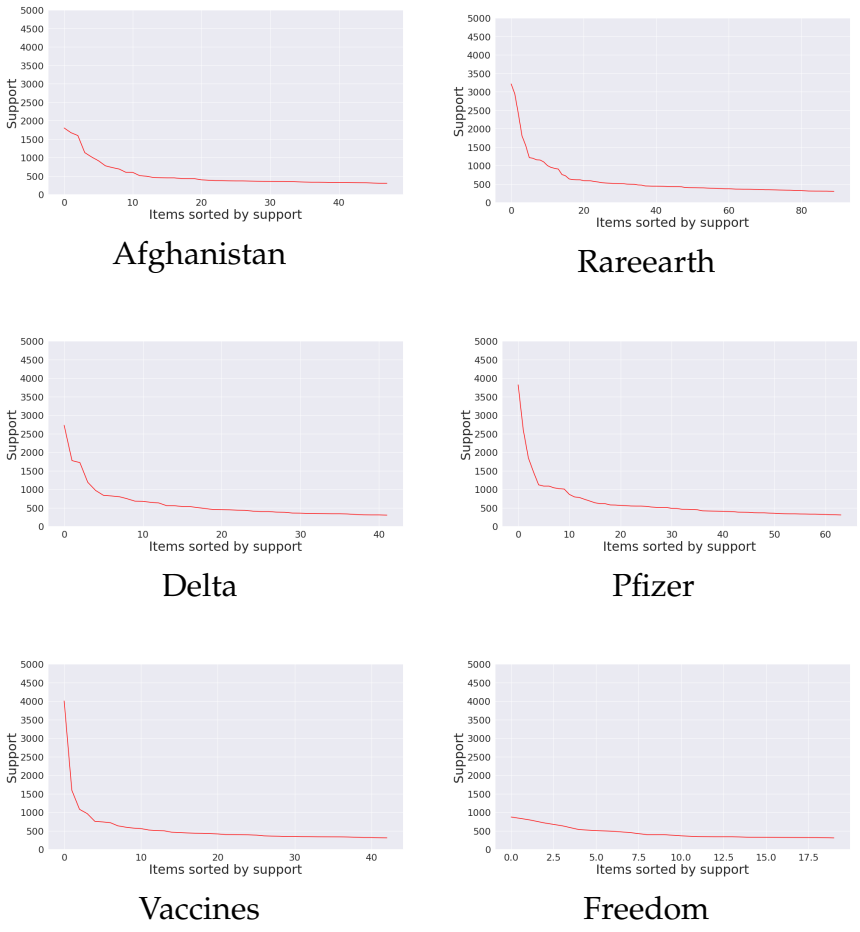
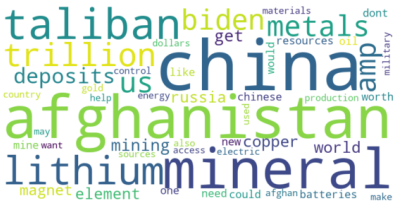
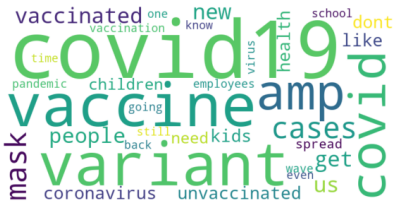


Figure 4.2: Support distribution for the different hashtag

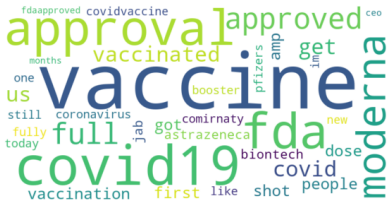
Afghanistan



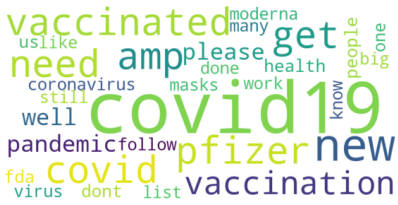
Rareearth



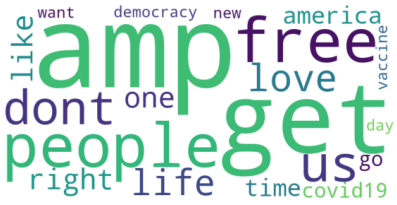
Delta



Pfizer



Vaccines



Freedom

Figure 4.3: WordCloud for the different hashtag

5 CONCLUSION

From the basket analysis that we conducted on the datasets that we generated from a set of hashtags through the Twitter APIs, we saw that a given hashtag can involve words that can span from a single very specific area of interest to many fields. The first case is what happens with the *RareEarth* hashtag, the second one with the *Freedom* one. The shape of the support distribution captures such behaviour: a horizontal one (high entropy) highlights a word that is used in an extensive set of context, while a distribution with a high peak at the origin (low entropy) highlight a word that is used into a very reduced context. Since the quality of information is linked to the specificity of words used with it, we propose to use the overall area under the support distribution (once the area below the support is removed) to assets numerically the information quality connected with a particular hashtag: in this framework, the behaviour of *RareEarth* hashtag is emblematic since it is more specific and technical with respect to *freedom* that can be used almost in every context.

BIBLIOGRAPHY

[1] https://help.twitter.com/it/rules-and-policies/twitter-api.

[2] https://www.dropbox.com/sh/2whpk4r2ompd42e/AACHTKOBwx2WaBgp9KMNNX4ga?dl=0.

[3] https://www.independent.co.uk/asia/central-asia/pen-farthing-nowzad-afghanistan-kabul-b1908359.html.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[5] Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

[6] Anindita Borah and Bhabesh Nath. Rare pattern mining: challenges and future perspectives. *Complex & Intelligent Systems*, 5(1):1–23, 2019.

[7] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. A Wiley-Interscience publication. Wiley, 2006.

[8] Wael Etaiwi and Ghazi Naymat. The impact of applying different preprocessing steps on review spam detection. *Procedia computer science*, 113:273–279, 2017.

[9] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[10] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[11] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[12] Ralph H Petrucci, F Geoffrey Herring, and Jeffry D Madura. *General chemistry: principles and modern applications*. Pearson Prentice Hall, 2010.

[13] Max Planck. The theory of heat radiation. *Entropie*, 144(190):164, 1900.

[14] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of open source software*, 3(24):638, 2018.

[15] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[16] P.N. Tan, M. Steinbach, V. Kumar, and A. Karpatne. *Introduction to Data Mining eBook: Global Edition*. Pearson Education, 2019.

[17] Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

[18] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[19] YY Yao. Information-theoretic measures for knowledge discovery and data mining. *Entropy measures, maximum entropy principle and emerging applications*, pages 115–136, 2003.

[20] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

[21] Mohammed J Zaki, Wagner Meira Jr, and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.