Tweet Data Mining: the Cultural Microblog Contextualization Data Set

Yassine Rkha Chaham[1] and Clémentine Scohy[1]
Under the supervision of Sébastien Déjean[2] and Josiane Mothe[3]

[1]INSA Toulouse, Université de Toulouse, Toulouse, France
{chaham,scohy}@etud.insa-toulouse.fr
[2]IMT, UMR CNRS, Université de Toulouse, UPS, Toulouse, France
sebastien.dejean@math.univ-toulouse.fr
[3]IRIT, UMR5505 CNRS, ESPE, Université de Toulouse, UT2J, Toulouse, France
josiane.mothe@irit.fr

Abstract. This paper presents an overview of the data set that was used for the Cultural Microblog Contextualization Workshop at CLEF 2016 and more specifically for the task 1: tweet contextualization. In this paper we first present a descriptive analysis of the data: we consider the variables or features associated with the tweets and analyse them. Then we also analyse the tweet textual content. The results of this work correspond to a first step toward data quality checking. It can also useful in order to understand better the data and its usefulness for some tasks or case studies.

Keywords: Tweet mining ; Cultural Microblog Contextualization ; descriptive analysis

1 Introduction

The objective of this paper is to report a statistical analysis of a database containing 38,686,650 tweets that was used in the Cultural Microblog Contextualization Workshop at CLEF 2016 which is related to festival events [1].

The data was collected by the task organizers using some specific keywords on the Twitter system as “festival”, and some additional terms such as “Cannes” to make a focus on the famous cinema festival that occurs in that town in France.

This study is mainly divided into two stages: first a general data exploration where we analyze different data variables; second an analysis of textual tweet contents.

Initially we set up an exploratory data analysis to observe the distribution and shape of the data. This study was carried out with the R software on different samples. We highlight some features of the data. For example we show a failure in data harvesting over a short period of days; we also show that contrary to what we expected, there is not much difference in the number of tweets between weekdays and weekends. Some tweets have location information. We have therefore studied those tweets more specifically and plot them on a map.

In the second analysis we looked more at the text of the tweets. To start with, we extract the most common words, but also the most frequent hashtags. We analyzed co-occurrences of terms so that it is possible to consider the terms that are the most associated with a given term, a festival for example. We also study the occurrence of terms over time. We report in this paper the results for some terms that correspond to types of festivals and to some cities. This type of visualization could be used in order to detect periods of festivals of certain types or which occur in given cities for example.

The remainder of this paper is organized as follows: In Section 2 we describe the main data set and the three sets of sampled data we built: 500,000 (random), about 132,000 from 3 users, 249,764 geocoded tweets. Section 3 presents the descriptive analysis of the data sets; we consider the main variables and present some features regarding these variables on various data sets. Section 4 reports some detailed results based on the analysis of the contents of the tweets. Section 5 draws some conclusions.

2 Objectives

In the literature, some research focuses on event detection. For example, Sakaki *et al.* [2] detect a target event by using a classifier that uses several tweet features such as the keywords in a tweet, the number of words, and their context. They apply the method on earthquake reporting. Lazard *et al.* [3] analyse tweets in order to extract major themes users express in their posts in relation to a specific event (a diagnosis of Ebola on US). While event detection and tracking is a hot topic for social media and tweet analysis, the objectives of the analysis we conducted is different and is two-folds:

(1) It aims at understanding better the data set the organizers of the CLEF CMC Workshop collected. This knowledge can be useful to decide on the types of tasks that can be drawn from the data set.

(2) The descriptive analysis of the data can be useful to check data quality. For example, it can be useful to know if there are some missing data and eventually to identify the reasons why data is missing. Then, specific processes can be used to handle such data [4].

3 Data Sets

3.1 CMC Tweet Data Set

The CMC collection we used is composed of 38,686,650 tweets (including retweets) and have been harvested using keywords such as "Festival", "Film", "Cannes", "China Festival" on the Twitter system. There are 18,709,732 million of tweets when retweets are excluded, and encoded using UTF8-unicode.

The data is structured using several variables:

- id: tweet identifier (integer)
- from_user: the user name (nickname) -who posted the tweet (string)
- from_user_id: twitter identifier related to from_user (integer)
- iso_langage_code: language used by the user who posted the tweet (string)
- source: variable representing the source of the tweet (e.g. Tweeter for Android, Figaro ...) (string)
- profile_image_url: Link to tweet (string)
- wday: day of the week when the tweet was posted (string)
- created_at: date when the tweet was posted (string yyyy-mm-dd)
- time_s: quantitative variable (integer)
- time_ord: quantitative variable (integer)
- content: content of the tweet (string)
- geo1 and geo2: latitude and longitude

3.2 Building Various Focused Samples of the Data Set

A Large Data Set of 500,000 Tweets (SP500K). After removing retweets, we randomly extracted a set of 500,000 tweets and analysed deeply this data set, that we name SP500K.

A User Focused Dataset (3USERS). We select the 50 most frequent users and pick-up three of them. We then extracted the tweets these 3 users had posted in the entire dataset. This process makes the 3USERS dataset. It is composed of 123,546 from userA (the one who posted the most posts in the SP500K), 5,444 from usersB, and 3,253 from userC.

Geo-localised Dataset (GEOSET). We select tweets that contain geo-localisation data and make the GEOSET composed of 249,764 tweets. We focused on three main variables: id, geo1 (latitude) and geo2 (longitude)

4 Descriptive Analysis

In this section, we consider each tweet variable or feature individually and analyse some of them over the three datasets we built and which are presented in section 2.

4.1 From SP500K

Even after retweet removal, we found out that there are many very similar tweets. It could be automatic tweets from “sharing” functions some internet sites have. These tweets are not completely identical since they use some tools to shorten URLs with different functions.

In this analysis, we focused more on the following variables: id, iso_language_code, source, created_at and content.

Users. We found out that one user was more frequent than other in this sub-set of tweets: he posted 3,491 tweets while the 34 other authors who posted the most sent

245 tweets on average (14 times less). Figure 1 shows the frequency of each user in the SP500K data set.

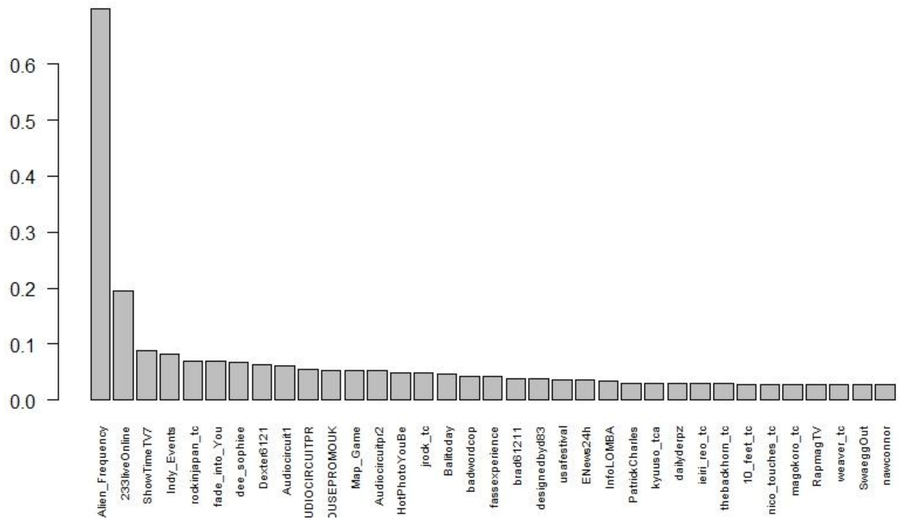


Fig. 1. Frequency of users (in percentage) who posted the tweets – SP500K data set

Language. In the data set, there are 66 different values for the *iso_language_code* variable. Some values are redundant. For example, en / en-AU / en-gb / en-GB / en-IN all indicate tweets in English. We decided to fuse the different values by keeping the two first language identification letters only. We also deleted the 选择语 which means « choose the language ». Not surprisingly, the main language used (for more than 50% of the tweets) is English, followed by Spanish, Portuguese and French. Figure 2 provides more details on the distribution of the languages used in tweets.

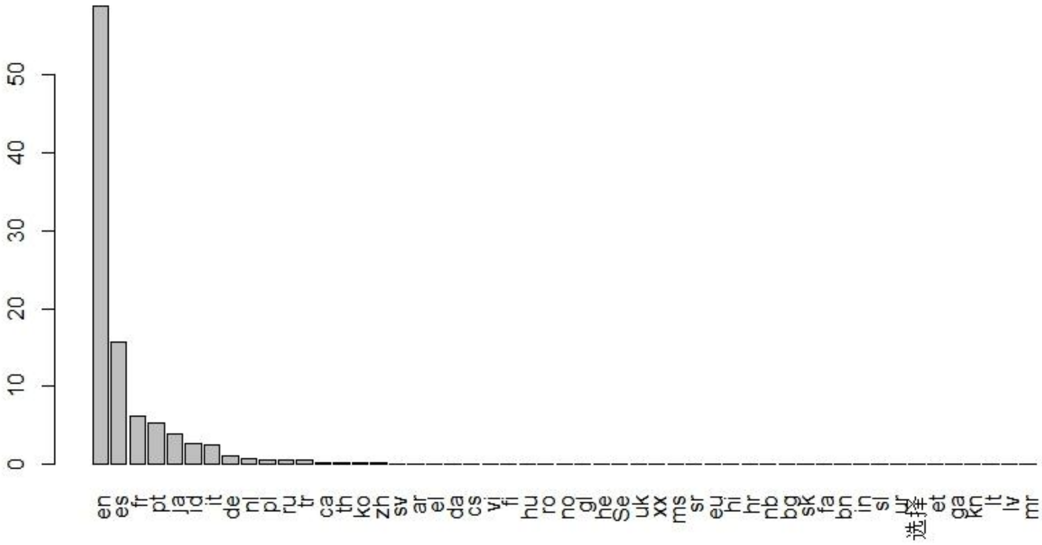


Fig. 2. Frequency of languages(in percentage) of the tweets – SP500K data set

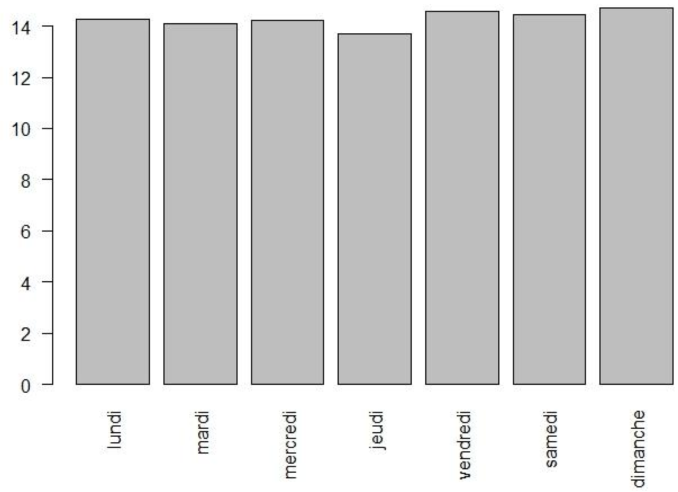


Fig. 3. Frequency of the tweets according to the week days

In the rest of the analysis, we kept the tweets in English only. There are 294,213 remaining tweets.

Even if we filtered out the tweets according to the l*anguage* variable, there are still some tweets that are not in English. It could be that the language indicator is the main language of the tweeter/author.

Wday. The distribution of the tweets over the week days is balanced as it can be seen in Figure 3.

Created_at. Tweets are from May 11th 2015 to January 10th 2016, that is to say on a 244 days period. Moreover,

- ¼ of the dates (*1st Quartile*) have been posted from before June 28th 2015,
- ½ of the dates (*Median*) have been posted before August 28th 2015,
- ¾ of the dates (*3rd Quartile*) have been posted before October 11th 2015.

Figure 4 displays the number of tweets per date when ordered in the chronological order.

We can see that there is period end of August -- early September where there are very few tweets. Apart from this short period, there is a peak of posts in June and another in September, while in contrast the period from November to January seems to be quieter. Figure 5 provides the frequency of tweets per month.

However it is worth recalling that we are analyzing a random sample of only a very small part of the entire collection. Thus, these observations might be due to drawing at random.

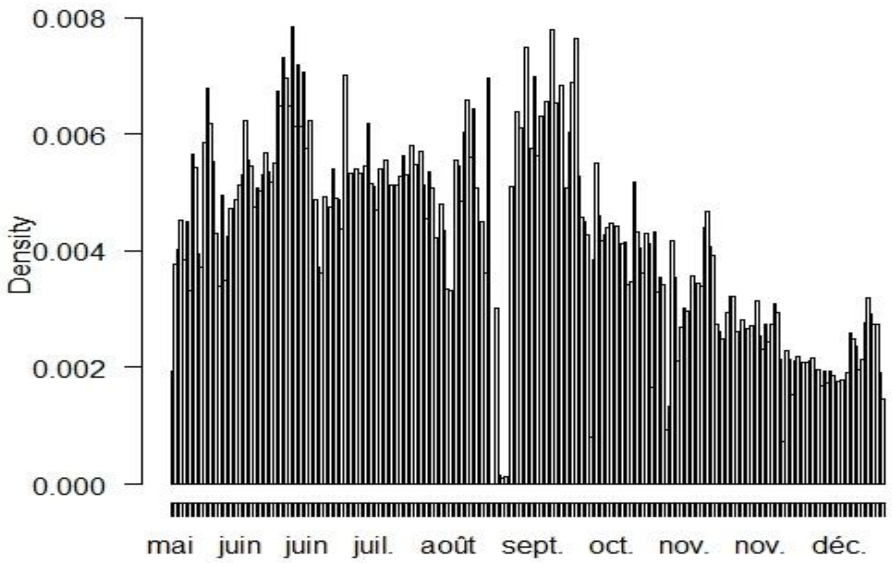


Fig. 4. Frequency of tweets per day

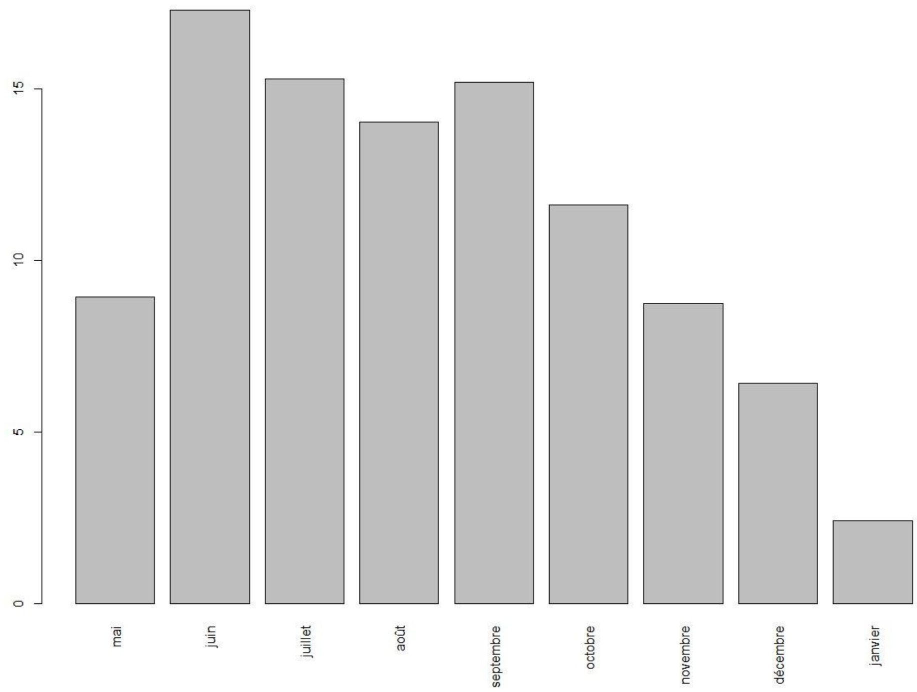
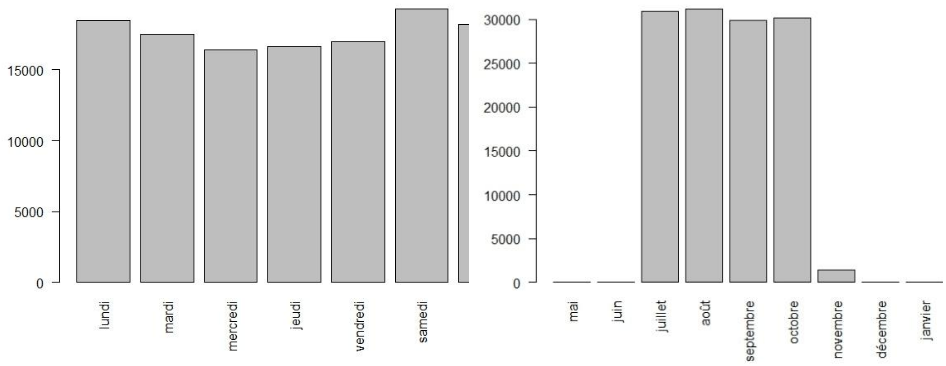


Fig. 5. Frequency of tweets per month

4.2 From 3USERS

Figure 6 provides the number of tweets posted by one of the users (userA) who posted the most tweets in the SP500K sample. We report the number of posts a) per week day and b) per month. In total in the entire collection, there are 123,546 tweets this user posted.

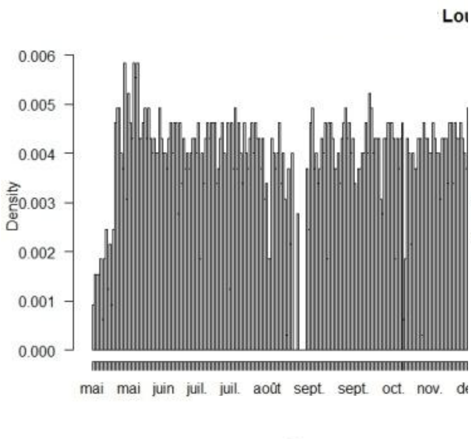


a) Per week day for the main twitter b) Per month for the main twitter

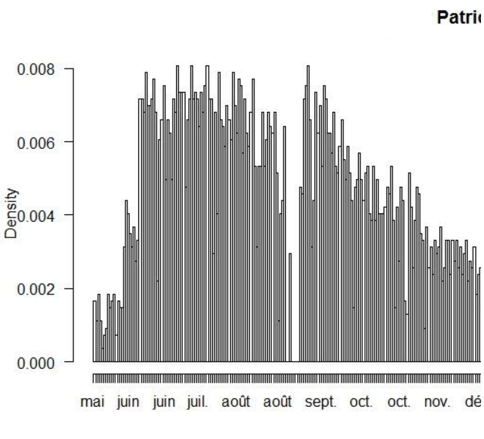
Fig. 6. Number of tweets for user A

There is no significant difference between the week days. May and February are not complete; that explains the difference we can observe in these two months Figure 6b. This user may be a spammer or an automatic system since he has a quite atypical behaviour. The other users who post lots of tweets have a stable level of posting across months. This could be checked going back to Twitter. After checking, this user seems to be an automate.

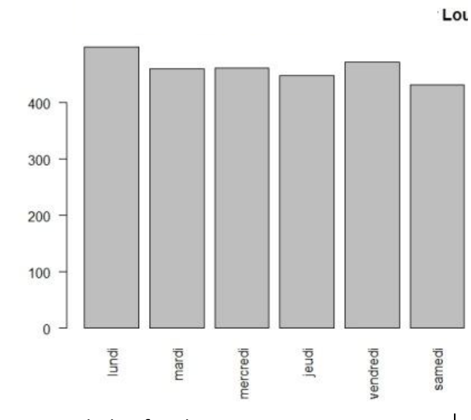
Figure 7 is similar to Figure 6 but for two other users we selected. There are 5,444 tweets in the entire data collection for userB and 3,253 for userC.



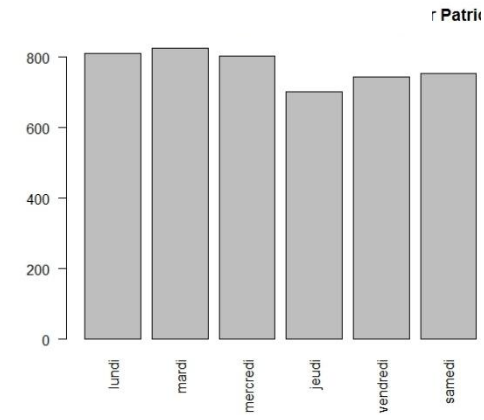
Per day



Per day



Per week day for the main twitter



Per week day for the main twitter

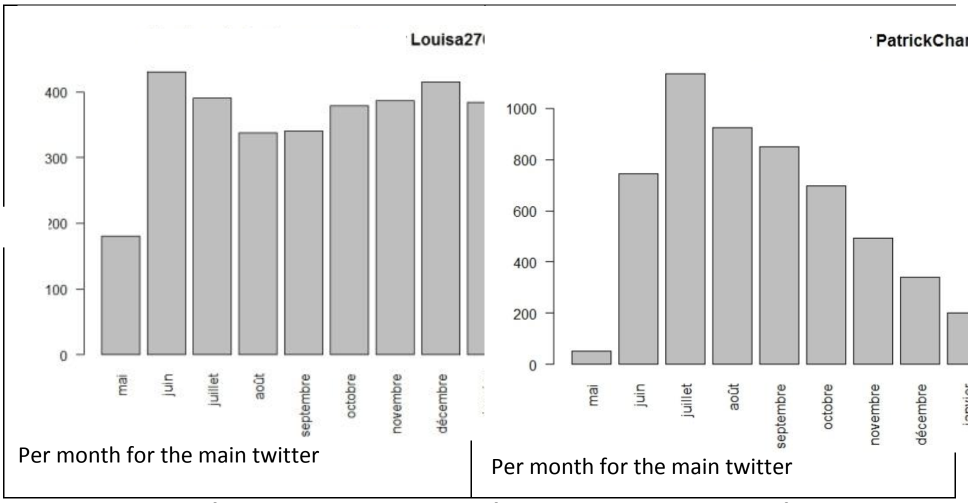


Fig. 7. Number of tweets in the entire data set for two users among the most frequent users

While the three users have a similar profile with regard to the week days (there is no large difference depending on the week day); they have very different profiles with regard to the months when they posted the tweets (and of course regarding the day).

In this paper, we just extract three users without any intention related to the choice of these users. A deeper analysis could focus on some specific users: the most actives for in a given event for example, or in the all data set, or in a given period in order to get some trends on those users. From this type of analysis, we could also extract the users who have similar behaviour. In addition, by crossing these results with tweet content analysis we could detect whether the users are more ordinary people or specialists of some types of festival for example.

4.3 From GEOSET

In Figure 8, we plot the tweets on a world map (rgdal library from R). We only plot the tweets for which we had the latitude and longitude values. We can see that most of the tweets were posted from West of Europe and from Asia.

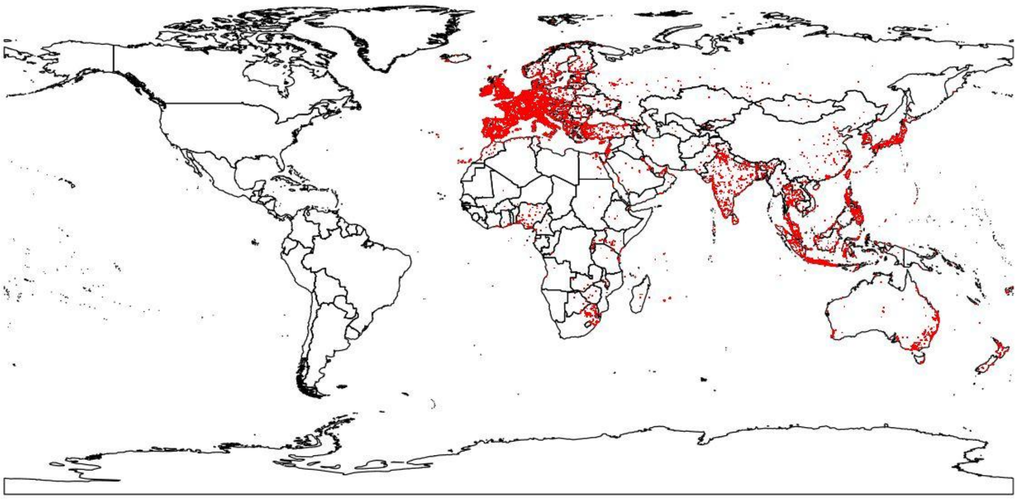


Fig. 8. Localisation of the tweets

We have no explanation why there is no tweet from Americas. One could be that the users do not allow geo-localisation. This hypothesis could be partially checked going back to the public user profiles for example.

5 Tweet Content Analysis

In this section, we analyse the textual content of the tweets; we focused on two topics: the type of festival and the town where festivals occur.

Types of Festivals

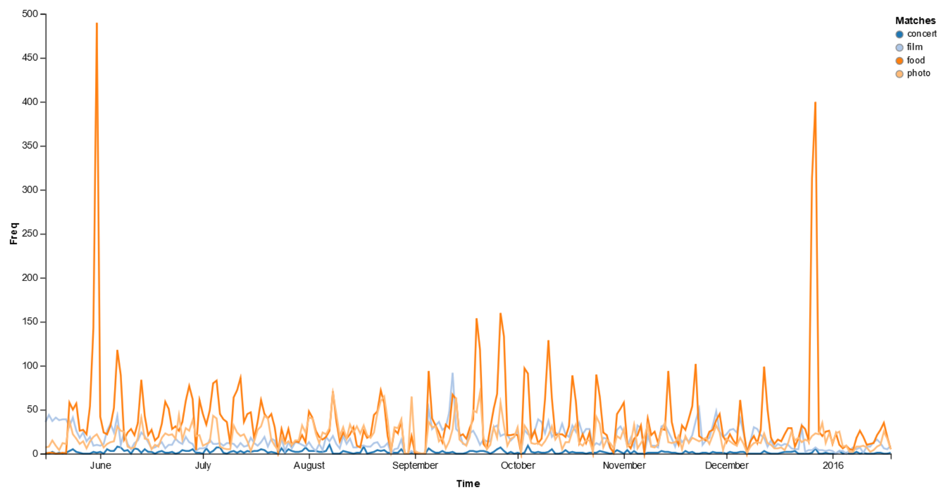


Fig. 9. Evolution of terms related to the type of festival

We first consider different types of festival and analyse the frequency of the associated terms. Figure 9 displays the frequency over time for the words ”concert”, “film”, “food” and “photo”.

We can observe some peaks in June, October, and January for the “food” term for example (See Figure 9 line on the top). We can also observe that this term is much more frequent than the other terms we choose. For a clearer comparison and analysis, it would be mandatory to consider also synonyms of the terms rather than just the terms as we did.

5.1 Towns

We did a similar visualization for some town where we know or find there are festivals and for which some data was in the data set.

In Figures 10 to 12, we select “kuala” (Kuala Lumpur), “Avignon”, “Phuket”, “cairo”, “Jakarta”, “rennes” and “dubai” and display the evolution of frequency of these words in the tweets, by day (Figure 10), by week (Figure 11) and by month (Figure 12).

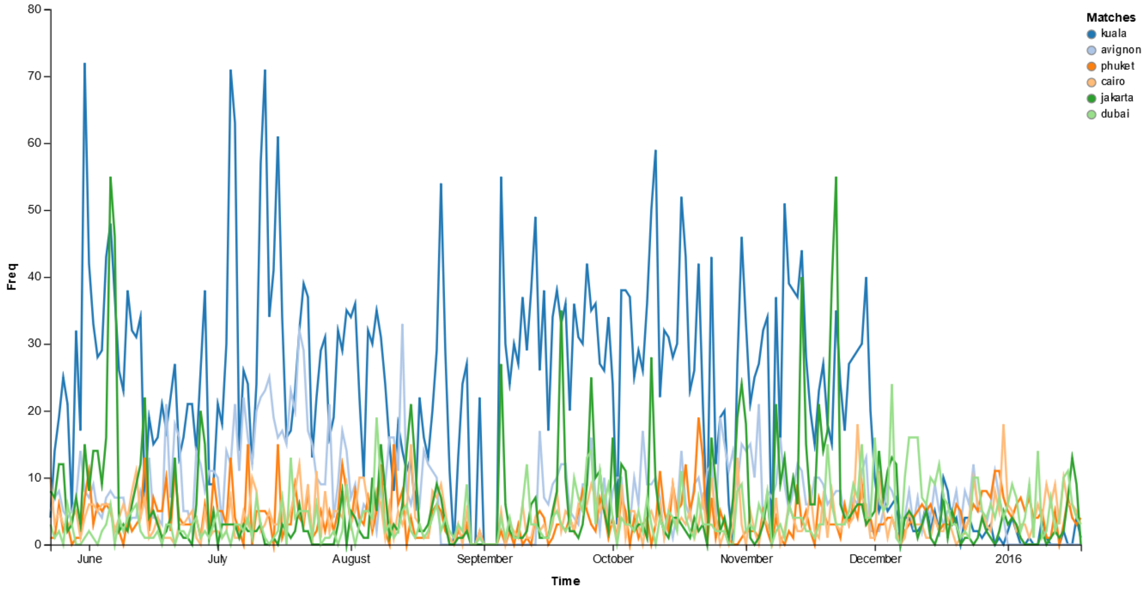


Fig. 10. Evolution of the frequency of some town name references in the data set by day

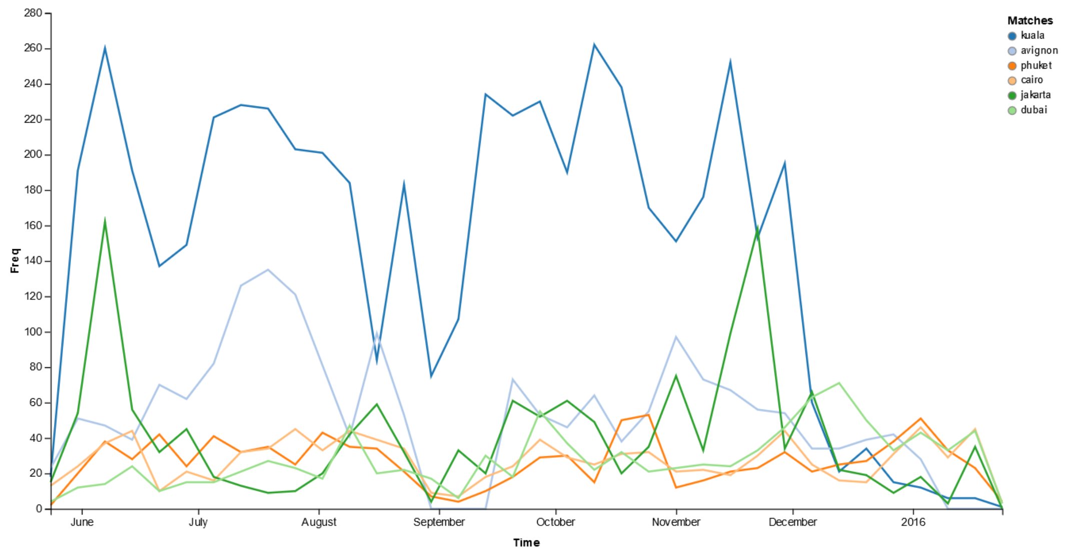


Fig. 11. Evolution of the frequency of some town name references in the data set by week

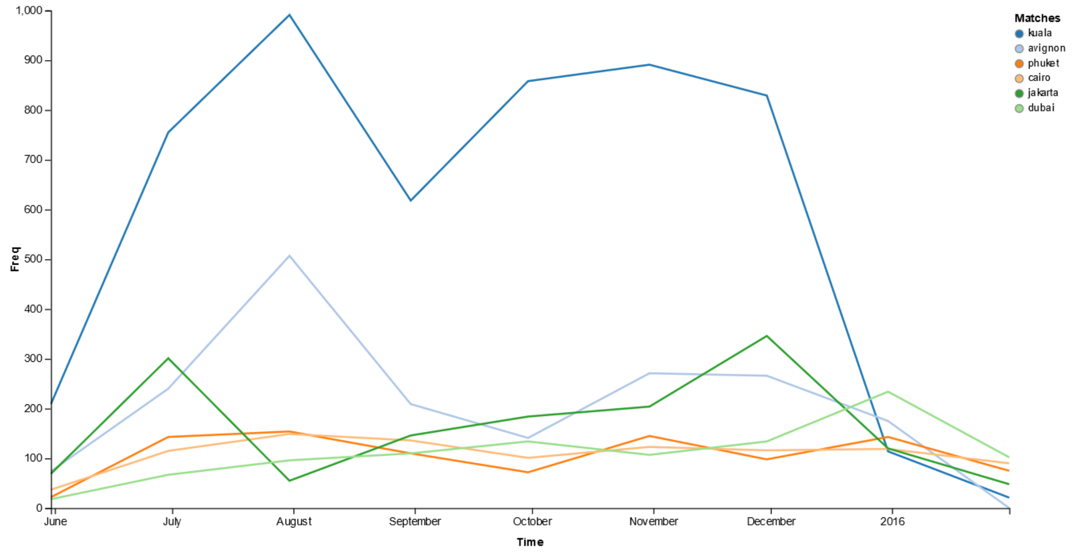


Fig. 12. Evolution of the frequency of some town name references in the data set by month

While it can be difficult to extract any information for a human from a visualisation by day, humans can extract more interesting information from the two other Figures.

If we consider "Jakarta" for example, one can observe two peaks in the frequency: one in July and the other in December. Not surprisingly, there is a very important festival in December in this town (international film festival) and other festivals in July.

The analysis of peaks in social media is a quite reasonable means to use to detect festivals in the world and when they occur. However, we could use more sophisticated means to extract locations such as the one presented in [5] or [6].

6 Discussion and Conclusions

In this paper, we did an analysis of the data set which is provided to participants of the CMC tasks of CLEF 2016. We built some sub-collections for focusing on different sub-problems or sub-analysis. We first conducted a descriptive analysis considering some of the features. We then analyse the tweet contents. This type of analysis could be useful to check data quality of the data set (e.g. detecting missing data) and to get an idea of the possible tasks that could be associated to the data set.

In the future, we would like to work on sentiment analysis of tweets. We think such study could be useful for example for festival organizers to get the flavour of what is said, positively or negatively about the festival they organize. Such methods are very popular for analysing events such as presidential elections [7] or other types of events. We would also like to work on data visualization in order to help discovering trends in the data [8] [9].

Although we did not participate in any of the tasks proposed in the CMC workshop, we think that the analysis we provide in this paper can be useful for understanding the data set.

References

1. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., Sanjuan, E.: Overview of the CLEF 2016 Cultural Microblog Contextualization Workshop, Experimental IR Meets Multilinguality, Multimodality, and Interaction. In: Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016). LNCS, vol. 9822. Springer, Heidelberg (2016).
2. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering, Vol 25, no 4, pp. 919-931 (2013).
3. Lazard, A.J., Scheinfeld, E., Bernhardt, J.M., et al.: Detecting themes of public concern: A text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. American journal of infection control, Vol 43, no 10, pp. 1109-1111 (2015).
4. Graham, J.W., Missing data analysis: Making it work in the real world. Annual review of psychology, Vol 60, pp. 549-576 (2009).
5. Ritter, A., Clark, S., Etzioni, O.: Named entity recognition in tweets: an experimental study. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 1524-1534, (2011).
6. Hoang, T.B.N., Mothe, J.: Building a Knowledge Base using Microblogs: the Case of Cultural MicroBlog Contextualization Collection. In: CLEF 2016 Working Notes, Balog, K.

and Cappellato, L. and Ferro, N. and Macdonald, C. Editors, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1609/ (2016).
7. Wang, H., Can, D., Kazemzadeh A., et al.: A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, pp. 115-120 (2012).
8. Murtagh, F., Taskaya, T., Contreras, P., Mothe, J., Englmeier K.: Interactive visual user interfaces: a survey. In: Artificial Intelligence Review, Vol 19, no 4, pp 263-283 (2003)
9. Dkaki; T., Mothe, J.: An energy-based model to optimize cluster visualization. In Research Challenges in Information Science (IEEE RCIS), (2014).