Semantic Mapping: Towards Contextual and Trend Analysis of Behaviours and Practices

Fionn Murtagh

Big Data Lab, Department of Computer Science
and Mathematics, University of Derby, Derby DE22 1GB,
and Department of Computing, Goldsmiths University
of London, London SW14 6NW, UK. Email: fmurtagh@acm.org

Abstract. As a platform for unsupervised data mining and pattern recognition, we use Correspondence Analysis on Twitter content from May to December 2015. The following data characteristics are well addressed: exponentially distributed data properties, and major imbalance between categories. Contextualization is supported. To both focus on informative resolution scale in one's data, and to handle large data sets, the granularity of point clouds offers benefits.

1 Introduction

In the sense of unsupervised classification and exploratory data analysis, our approach is both "The model should follow the data, and not the reverse!" (Jean-Paul Benzécri quotation) and "Let the data speak for themselves" (John Tukey quotation). Our interest is in the data semantics, thus based on interrelatedness, and situatedness, i.e. semantic relative positioning or semantic location. Our fundamental methodology is the Correspondence Analysis analytics platform, used for narrative analytics in [10], and particular themes relating to social narratives and the work of celebrated social scientist, Pierre Bourdieu, in [7].

1.1 Data Selection to Focus Our Analysis

A brief review follows of indirectly related work, that involves analytic focusing, i.e. selecting and recoding relevant or important data.

Summarization of tweets in [14] is through frequent phrases, therefore based on (as expressed in that work) common word usage. A minimum of 100 and a maximum of 1500 tweets are used as a training set. A starting phrase is used, and tweets with this phrase are found and then filtered in regard to some criteria (English language, non-spam). Next a rooted tree is built with that starting phrase as the root, and words before and after it have weights derived from frequency of occurrence of their positions relative to the phrase. In a consensus weighted tree, the greatest weight path from root to terminal provides the summary. Discussion of results follows, related to manual summarization.

Continuing with tweet summarization, from multiple posts, in [5], firstly tweet vectors of terms (i.e., words) that are contained, dissimilarity based on

TF-IDF (term frequency, inverse document frequency) is defined, with however global term frequencies used (and not term frequencies of a tweet). Stop words are removed. Then k topics are determined by k-means clustering, modified in regard to centroid selection, and in regard to bisecting implementation. Other tweet summarizations are discussed and used, as well as manually, for "ground truth". Sub-themes were investigated, of 1500 tweets obtained for 50 different trending topics.

1.2 Objective: Unsupervised Data Mining

While summarizations of Twitter streams are important, in this article, we want to begin to use the potential of Big Data in order to observe and track narratives of behaviours, and of activities. Motivated by [3] (in a very different context), we seek not "content analysis", of the Twitter sources, but rather "map analysis". Alternatively expressed, this is "cognitive mapping, relational analysis, and meaning analysis" among other names. Harcourt [3] notes the need "to visually represent the relationship between structures of social meaning and the contexts and practices within which they are embedded". Related to all of these objectives, having effectiveness and innovative perspectives, and taking care of ethical aspects, we do not seek primarily to have directly predictive mapping (of course, in a secondary sense, we are interested all that is relevant for prediction and forecasting). Rather, our primary interest is in having informative and revealing mapping out of narratives of behaviour and practices, as a foundation for all that follows. This article is a start in this direction, tweets are used but without use, yet, of URLs contained, or of linked Wikipedia or other data sources.

While interest in, and use of, data can require focused, supervised analysis, it is also beneficial and important to allow for unsupervised data mining. This can be the basis for rankings and ratings, and trend following, in all that the data expresses. Here Correspondence Analysis is used. A strength of this analytics platform [9] is in regard to heterogeneous data sources, including exponentially distributed data, and highly imbalanced categorization, as is most typical with textual and behavioural data.

The following are helpful approaches, when using Correspondence Analysis, for handling very large data sets. Firstly, the principle of distributional equivalence, justifying and motivating aggregation; secondly, the basic factor space mapping using aggregated data (cf. below, the use of daily compilations); and compactification or data piling in very high dimensional spaces.

1.3 Twitter Data Used

A set of nearly 12 million (11,952,123) tweets was obtained, from the collection described in [1, 2]. The very first, and the very last, of these tweets were as follows:

```
"600365725350526976", "resist222", "2318838143", "fr", "LeFigaro.fr / Android",
"http://pbs.twimg.com/profile_images/585163960300765184/jyPRAiyf_normal.jpg", "Mon",
```

"2015-05-18", "66117", "1431973317", "GPA : le Défenseur des droits presse Taubira d'agir http://t.co/LKLCqrHZ83 via @Le_Figaro"

"618799838029869056", "TheMBShow", "3357037889", "Select L", "Twitter Web Client", "http://pbs.twimg.com/profile_images/618459240802463744/l7Pzy5-u_normal.jpg", "Wed", "2015-07-08", "54752", "1436368352", "@Kialdn_ Come catch Sneakbo perform at our festival @TheMBShow 29th August • Get tickets here http://t.co/kyah5M49Ho"

The dates are not in sequence. Above these are 2015-05-18 and 2015-07-08. We find, when sorted that the dates extend from 2015-05-11 to the first few days of January 2016, with 2015-12-31 being the final day with a typically large number of tweets. Note may be taken of the fact that various tweets have missing values, and a very small number of dates are thus: 0000-00-00.

Fields are: ids for recipient, author, of the tweet itself; language associated with the author; interface used for posting the tweet; icon given by a URL; day of the week; date; geolocation; and the tweet. There are 75 languages in use, including Japanese, Arabic and so on, with the majority in Roman script.

2 Days Crossed By Hashtags

We will focus our analysis, in the following way. To have a meaningful aggregation that groups tweets by theme, the following is undertaken: each day's tweets are considered, and the hashtags at issue are obtained. Just as an example, for the very first of our dates, 2015-05-11, there are 23276 rows, that when read as 11 fields this becomes (due to fields with missing values) 23260 tweets. Now, many tweets are retweets, labelled in the tweet by RT. We do not consider retweets so as to remove redundant semantic content. We consider tweets that are not retweets. There are 16837 such not-retweeted tweets. Hashtags are determined for each tweet. It is found that a tweet can have up to 14 hashtags in a tweet.

The very small number of tweets in the first few days of January 2016 were ignored. From the other days, extending from 11 May to 31 December, some days in September had few tweets, while the largest number of tweets was on 2015-05-24, with 293695 tweets. For each day, the following was carried out: non-retweet tweets were selected, the list of hashtags per tweet was assembled. For a given day, repeated use of the same hashtag was not retained, as such. As the next step of the data preprocessing, for each day, the hashtags were collected. Hashtag usage from 1 September to 13 September was small (varying from 31 down to only 2 hashtags, the latter case being the 13 September). For 2015-05-29, there was use made of 17673 hashtags.

The next step was to take the 233 days (from 11 May to 31 December 2015), each with their set of hashtags. As noted above, non-retweet tweets were used, and also used were the unique hashtags used in the day's set of tweets. The hashtag set was listed. This consisted of 70503 hashtags in total. The top ranked hashtags were as displayed in Table 1. The final or bottom ranked hashtags are also listed. The hashtag character, #, was removed, upper case was set to lower case, and for this work, the following was also carried out: punctuation

was removed, and so too were accented characters and non-Roman characters. This first phase of our work did not seek to distinguish between, or separate, languages in use. Thus, `#Reynié.` becomes `reynie` here. With numbers of days that these appear, we have these: `sarkozy` 55, `nanosarkozy` 11, `cesarkozy` 8, and `nicolassarkozy` 2. These are from, respectively, `#Sarkozy` and `#sarkozy`; `#NanoSarkozy`; `#CesarKozy`; `#NicolasSarkozy` and `#nicolassarkozy`. We consider that it is advantageous, in this early stage of the analytics, to simplify hashtag terms by not considering accented characters or punctuation, or numerical characters, and, especially, through considering only the lower case equivalent of upper case characters.

Given that 233 days are at issue here, and that repeated occurrences of a hashtag in a tweet (happening often enough) were not taken into consideration, the following note relates to how, e.g., the hashtag (in lower case, with preceding hash character removed) `cannes` can have 635 occurrences overall. This is due to the use of all the following hashtags: `#Cannes`, `#CANNES`, `#cannes`. Following our preprocessing, these become one hashtag.

Table 1. Top ranked hashtags, coming from the unique hashtag set per day. The # character preceding the hashtag has been removed, and upper case has been set to lower case. A few bottom ranked hashtags are also displayed.

```
cannes 635, avignon 516, france 488, frenchriviera 458, festival 426,
nice 423, cotedazur 419, film 415, paris 414, travel 401, monaco 392,
marketing 381, fashion 379, love 373, news 371, art 365, immobilier 364,
la 362, live 361, hotel 356, photo 356, job 343, culture 339, luxury 338,
provence 338, deal 335, video 332, promos 331, music 327, paca 321,
marseille 317, europe 317, sttropez 312, london 312, emploi 310, cinema 309
...
sorellaterra 2, estilismos 2, provinssirock 2, katzenjammer 2,
electriccastlefestival 2, rapradar 2
```

From the 76098 word list (see section 3 for further discussion), the most frequently occurring 5595 were selected. This selection was from the frequency of occurrence being greater than 30. Such is standard practice in textual data mining, because the analyses are based on commonality of usage, i.e. shared usage, of terms.

The Correspondence Analysis factor space is a latent semantic space, that semantically maps our data, the dual spaces of days and hashtags. Figures 1 and 2 present the first look at this data set, crossing 233 days and 5595 hashtags. All information in the cloud, or point set, of days and of the cloud of hashtags, are covered by 161 factors, i.e. principal axes. More that 50% of the inertia of

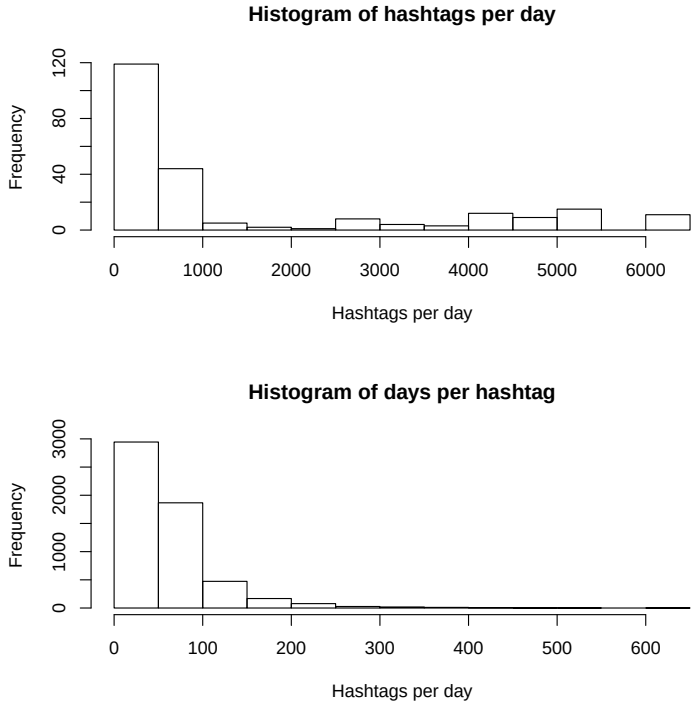

Fig. 1. Illustrating the marginal distributions of the day's tweets crossed by the hashtags used.

the cloud of days (in the hashtag space) and of the cloud of hashtags (in the days space), is explained by the principal 11 factors. From Figure 2, displaying the relative importance of inertia explained, our main focus will be on the first factor and also on the first four factors.

The principal plane is shown in Figure 3. Figure 4 displays the hashtag projections (most but not all in the plot window that is displayed), not labelled but just as a dot at the location of the hashtag's projection.

Due to the confused display of a large number of hashtag terms here, consider, firstly, the highest contributing hashtags to factor 1, listed in Table 2. Table 3 lists the greatest and least projected values on factor 1. For Factor 2, these results are listed in Tables 4 and 5.

3 Days Crossed by Words Used

In order to investigate semantic content, words are extracted, as strings of adjacent letters. The semantic relationships and patterns among words is the main

Table 2. Factors 1 and 2 projections of the hashtags with the highest contributions to factor 1.

	[,1]	[,2]
cannes	1.185046	0.1157641
avignon	1.341975	0.3057062
france	1.259990	0.1704328
paris	1.199826	0.1786310
travel	1.161375	0.1893705
monaco	1.181179	0.1488680
immobilier	1.283816	0.2323571
promos	1.321608	0.4335699
appartement	1.563006	0.4364088
carhaix	1.941503	1.0825339

Table 3. Factors 1 and 2 projections of the hashtags with the greatest and lowest projections on factor 1.

	[,1]	[,2]
planscul	2.339680	0.9930509
offresemplois	2.436154	1.9741365
intemp	2.392753	1.0840280
aya	2.308942	0.8265798
intemperies	2.366595	1.0561119
eglgbti	2.343183	1.1127613
villaincannes	2.339515	1.0218940
festivaldedanse	2.289062	0.9611641
lourdesleon	2.435461	1.0791402
picassomania	2.333926	0.9841345

	[,1]	[,2]
powerofcricket	-0.6980843	0.09094040
futurenseine	-0.7001860	0.08547853
summergathering	-0.7074827	0.09807065
vabeach	-0.7077370	0.09236909
isleofwhite	-0.7245547	0.10680055
atxtvs	-0.7289408	0.10617401
sfjazz	-0.7292949	0.10577881
announo	-0.7221056	0.08919958
outsidein	-0.7217502	0.11530522
stepneygreenpark	-0.7422039	0.11113410

Table 4. Factors 1 and 2 projections of the hashtags with the highest contributions to factor 2.

	[,1]	[,2]
carhaix	1.94150283	1.082534
finist	1.86502290	1.347036
viproom	0.49887231	-1.430724
ubercopter	0.01191271	-1.526498
lorient	0.10540259	1.606021
offresemplois	2.43615356	1.974137
quimper	0.87351563	1.958572
ampav	0.16830070	-2.132900
carhaixplouguer	1.74397047	2.504840
arch	1.22985351	2.362014

Table 5. Factors 1 and 2 projections of the hashtags with the greatest and lowest projections on factor 2.

	[,1]	[,2]
lorient	0.1054026	1.606021
offresemplois	2.4361536	1.974137
quimper	0.8735156	1.958572
eau	2.2184275	1.884511
carhaixplouguer	1.7439705	2.504840
arch	1.2298535	2.362014

	[,1]	[,2]
oscardelarenta	0.1860428	-1.988590
ampav	0.1683007	-2.132900
festivaldicannes	0.2161622	-1.954455
seaoftrees	0.2456611	-1.983685
marchedufilm	0.3141766	-2.027599
womeninmotion	0.1671224	-2.115963
galacroisette	0.1583530	-2.064233
theseaoftrees	0.1479140	-2.077604
nespressocannes	0.1249056	-1.967695
semainedelacritique	0.1626134	-2.126001

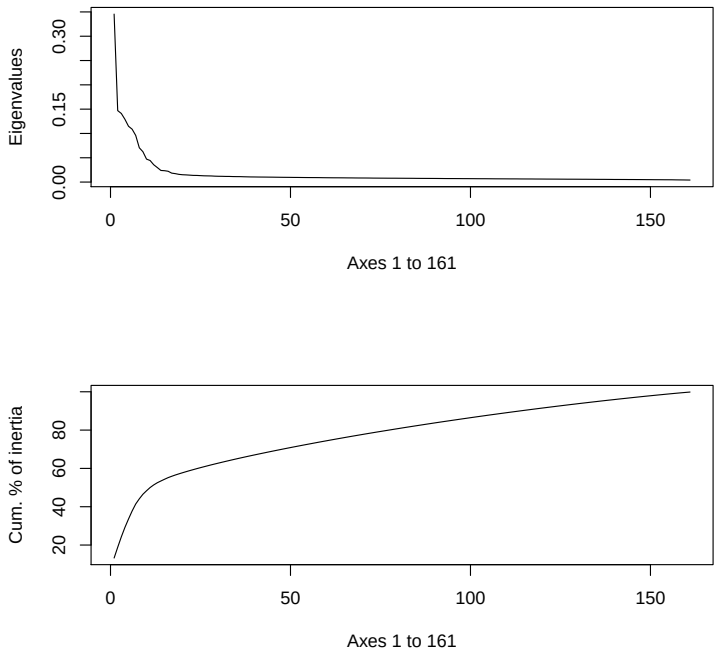

Fig. 2. For the clouds of days and of hashtags, we see the relatively great importance of the first axis, and the cumulative importance of the first four axes.

interest, so numerical characters are not considered, nor is punctuation, and upper case is reset to lower case. Since we can best analyze the following in their own terms, these are removed from the tweets: tweeter name, preceded by the @ character; hashtag preceded by the hash character, #; and URL. For the 233 days (as above, from 11 May to 31 December 2015), 160503 words were extracted. The list of words with 1000 or more occurrences was retained. This provided a list of 6407 words. To filter this set of words, the following were deleted from this list: prepositions, parts of verbs, Cyrillic, Chinese, Japanese, Greek, Turkish script, fairly evident abbreviations or parts of words, many 2-letter words. Often, abbreviations were left in the list, also left were some words like days of the week, also some salutations or other informal words (including expletives!) since they may indicate emotional context. Also what could be part or complete personal names, or names of places. Portuguese, Spanish, French, English, etc. were treated in the same way. The word list became 5820 words. Table 6 lists the top ranked and the bottom ranked words that are thus retained.

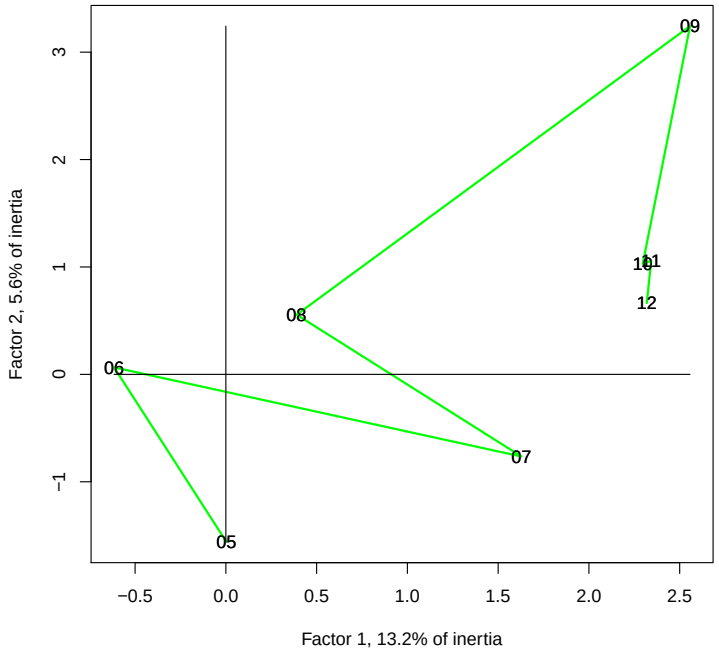

Fig. 3. Months May, denoted as 05, to December, denoted as 12, 2015, as supplementary elements.

This 5820 word set is now our corpus. The total word set, i.e. use of words in this corpus, is 32507477 words.

Figure 5 illustrates the imbalance of words, which is very much as expected. Per day the total word use, from the specified word corpus, was minimum, median and maximum: 73, 13497 and 658683. For each word in the corpus, the minimum, median and maximum of usage in the set of days at issue here: 1000, 2208.5 and 4052390.

Figure 6 shows the relative importance of the factor space axes, determined from the inertia of the clouds of days and of words, and endowed with the Euclidean metric. The eigenvalues of the first four axes are 0.457, 0.245, 0.235 and 0.214.

The principal plane, Figure 7, shows `festival` and `cannes` near the origin, therefore very central and average here in their semantic relevance and importance. Based on the May to December 2015 Twitter set here, there is centrality of the Twitter exchanges that related to the Yulin Festival, and the Cannes Film

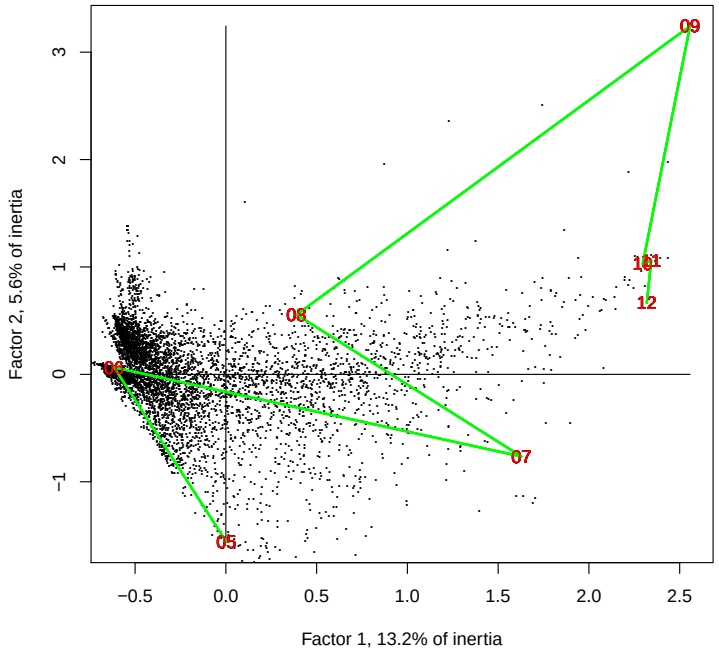

Fig. 4. Months May, 05, to December, 12, 2015, as supplementary elements. The cloud of dots are the locations of the hashtag projections.

Festival. Figure 8 displays the highest ranked coordinates in the principal plane. Then for factors 3, 4, Figure 9 just a somewhat more differentiated perspective.

Table 7 explains further just how the first factor is defined. We find dominance of the words `cannes` and `festival`. Next, but a lot less central semantically, are the words `yulin`, `doc`, `meat`. If we were to pursue further the undercurrents and trends in semantics, we would carry out our analytics with these two most dominant terms considered as supplementary elements in the analysis.

Figures 10 and 11 consider the time evolution by projecting the monthly aggregate location into the semantic factor space, as supplementary elements in the factor mapping. The following may be considered in the interpretation: the 68th annual Cannes Film Festival was held on 13–24 May 2015, and the annual Yulin Dot Meat Festival was held on 21–30 June 2015.

Table 6. Top ranked words, and bottom ranked words, with the frequency of use in the tweet set.

---

festival 4052390, cannes 1615550, film 388301, music 299792, yulin 204177, dog 172423, meat 147538, china 134485, day 119675, new 111196, stop 103054, avignon 102499, people 95039, all 91626, rock 91521, lions 88681, edinburgh 88424
...
opener 1001, tiap 1001, pire 1001, calgary 1001, revés 1001, gbp 1000, harbor 1000, stalls 1000, francesco 1000, thinks 1000

Table 7. The words with the highest contributions to the first axis. Highest of all are `cannes` and `festival`. Firstly here are shown the contributions to axes 1 and 2. Secondly, here, there are the coordinates.

Contribs.	Dim 1	Dim 2
festival	5.0486425	0.058077104
cannes	12.4044960	0.111256026
film	0.8462713	0.003516824
music	0.6005026	0.001852411
yulin	0.6770539	0.027186154
dog	0.4839718	0.013704251
meat	0.4657401	0.016922335
salma	0.8294958	0.735496378
hayek	0.8182728	0.683801901
franco	0.5409874	1.629173172
michel	0.5086740	1.502823560
portman	0.4568916	0.088364783
actriz	0.4802585	0.302854474

Coords.	Dim 1	Dim 2
festival	-0.4303220	-0.03375638
cannes	1.0682951	0.07399639
film	0.5691574	-0.02683490
music	-0.5456439	-0.02216499
yulin	-0.7020533	-0.10289148
dog	-0.6459134	-0.07949485
meat	-0.6849860	-0.09549643
salma	1.7655692	-1.21594797
hayek	1.7573206	-1.17493594
franco	1.4854169	1.88532190
michel	1.5314790	1.92527323
portman	1.8212840	-0.58581139
actriz	1.8676415	-1.08472605

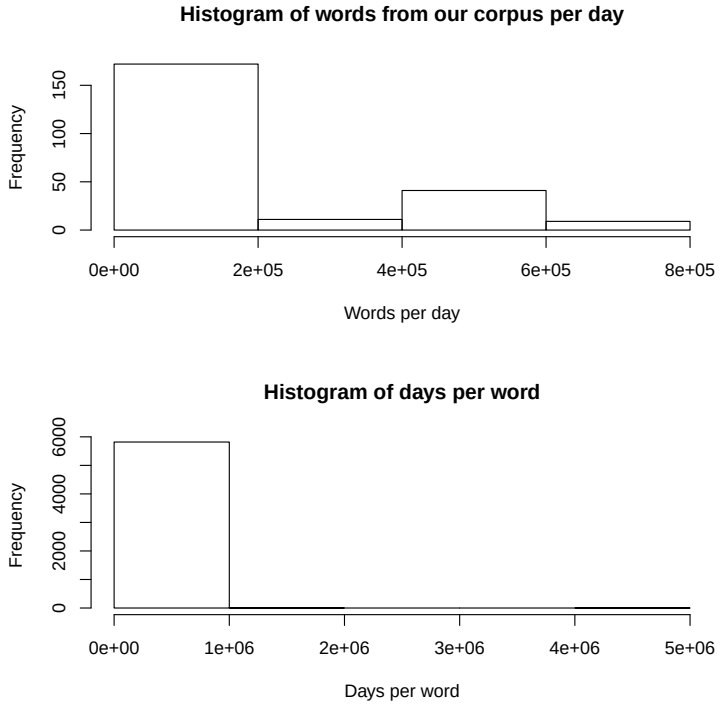

Fig. 5. Illustrating the marginal distributions of the day's set of tweets crossed by the frequency of word use, from the word corpus.

4 From Twitter: Relationships Between Festivals

The following festivals are selected: Cannes Film Festival (13–24 May 2015); Fèis Ìle, Islay (Scotland) Festival (23–31 May 2015); Berlin Film Festival (19–21 May 2015); CMA, Country Music Association (Nov. 2015); Yulin Dog (June 2015); and Avignon Theatre Festival (4–25 July 2015).

Figure 12 displays the planes of axes 1, 2 and of axes 3, 4, exactly as previously. We do see here how the principal factor plane is especially a contrasting engagement with the Cannes Film Festival for axis 1, and the Avignon Theatre Festival for axis 2. Meanwhile, both axes 3 and 4 can be said to be especially relevant for the Avignon Theatre Festival.

Further analysis could well concentrate on the festivals. Furthermore, analytics could well be pursued about the particular aspects of various festivals, including their content. A longer term aim is to support social science research, by having validation for, and contextualization of, specific and directed studies. Contextualization can be most straightforwardly addressed by having supplementary elements (input data array rows or columns) projected into the latent

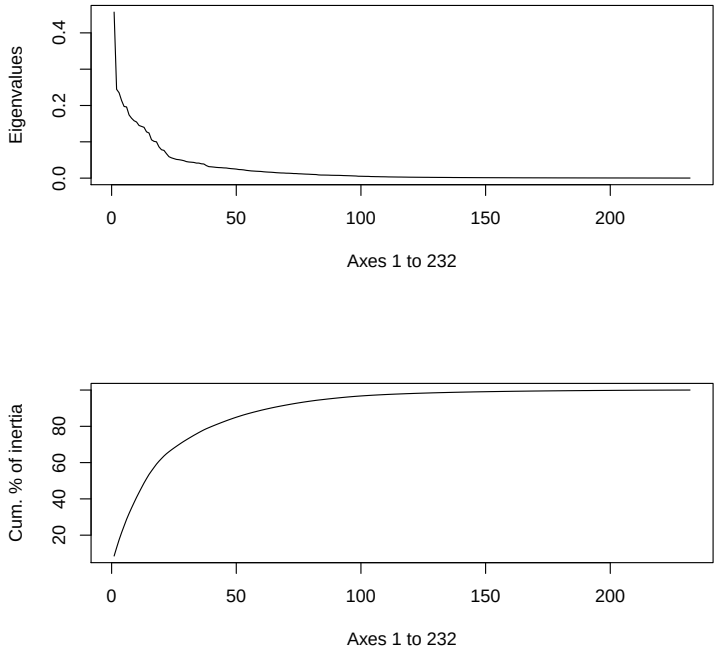

Fig. 6. Quite similarly to Figure 2, we note the relatively great importance of the first axis, and the cumulative importance of the first four axes.

semantic, or factor, spaces. Issues such as the following are relevant here: public protest expressed in Twitter, in particular from the United States in regard to the Yulin Dog Meat festival in China; or Mexican celebratory Twitter microblogs for the Cannes film festival. Context such as geolocation information can be directly used in the analyses or, where there is limited data availability, such information can be projected into the analyses as supplementary elements. Some consideration of very negative sentiment, e.g. related to the Yulin festival, or other particular emotationally-laden content, may also need to be taken into account.

In the concluding section to follow, it will be noted that the primary objective of this paper has been to carry out an exploratory analysis of the data. Further motivation is as follows, as we express in our comments in [6]: "The bridge between the data that are analysed and the calibrating 'big data' is well addressed by the geometry and topology of data."

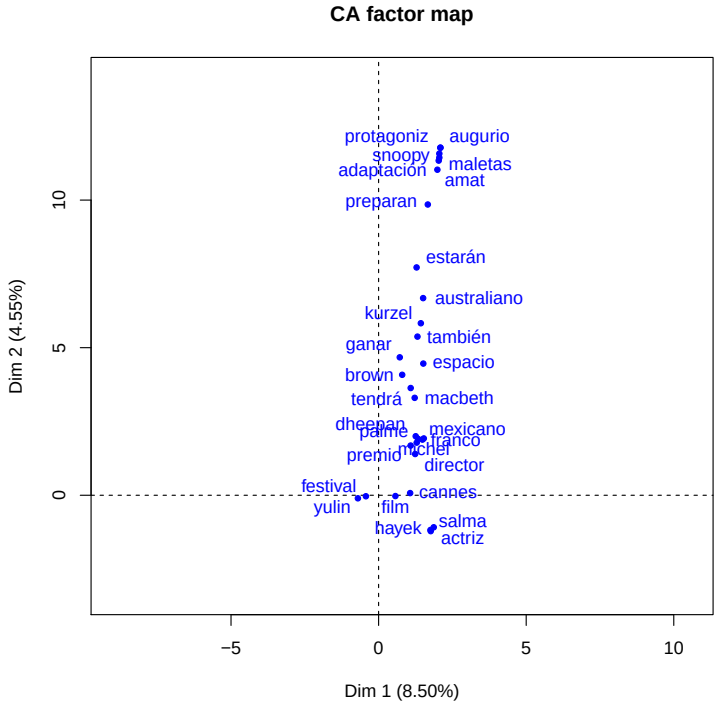

Fig. 7. Projections of 30 words with the highest ranked contributions to the principal plane.

5 Discussion and Conclusion

This work sets out a platform for data mining, and knowledge discovery, in large data volume analytics. The focus is on unsupervised pattern finding, including trend finding, in the data. Data analysis of Twitter flows using Correspondence Analysis are feasible, cf. [11], notwithstanding the need to address the misspellings, abbreviations, and so on, in many tweets.

A short summary of procedural aspects follow. While these were driven by computational reasons, they were also motivated and justified in this early phase of our analytics, by the focus or orientation of the exploratory analysis.

While in this work, web addresses (URLs) were removed, both they and the not infrequent use of smileys, are to be considered when extending this work. It is to be noted that in the near 12 million tweets used in this work, there were references 1702980 times to URLs. The number of unique URLs was 474709. The number of unique hashtags in this data collection was 74384. The number of tweeters, with name preceded by the at sign, was 123953, that is, the number of unique named tweeters. While a range of languages were accepted for work

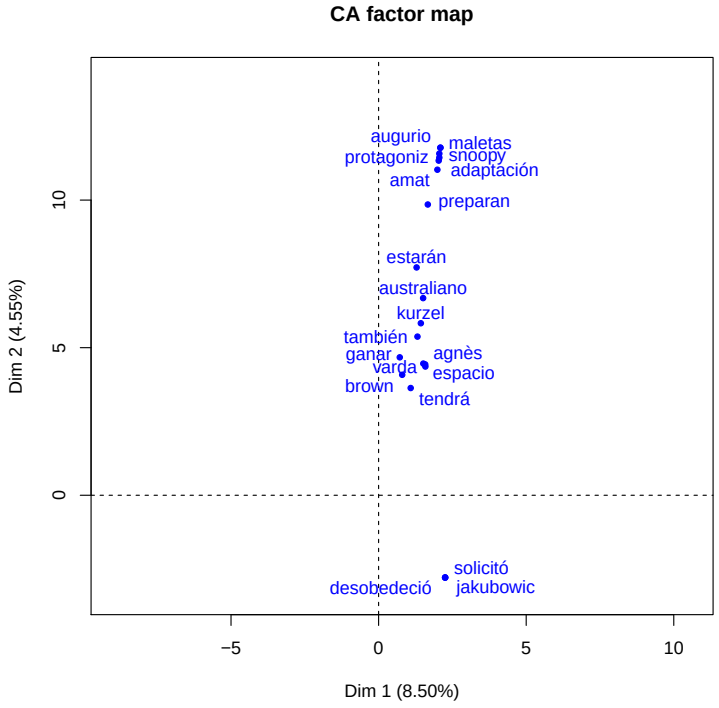

Fig. 8. Projections of 20 words with the highest ranked coordinates in the principal plane.

that is reported on here, it has also been noted how Cyrillic, Chinese and other languages were not included here.

For the Correspondence Analysis, and much of the interactive analysis, both the FactoMineR package, cf. [4], was used, and when, not infrequently, there were problems of unbounded computational and memory requirements, then the essential processing steps were, both effectively and efficiently, used. See [8] for this. R was used for this processing. For data selection, filtering, extraction and reformatting, prior to the analytics in R, scripts were written and then sourced into R.

Data compactification, or data piling, are known properties of very high dimensional spaces. In the dual spaces relating to the large word cloud, and the daily tweets cloud, the latter – the daily tweet set – is embedded in a very high dimensional space, i.e. the space of words. There is very large correlation, close to 1, between all of the following: the total inertias of cloud points; random projections of cloud points on random axes, with uniformly distributed coordinates in [0,1] (cf. [12]); and the marginal distribution, or masses of the cloud points.

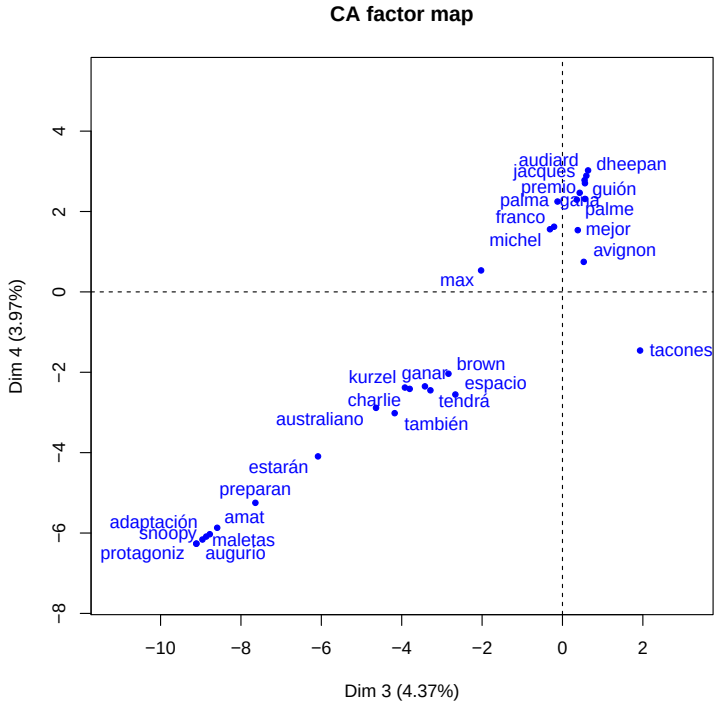

Fig. 9. Projections of 30 words with the highest ranked contributions to the plane of factors 3, 4.

Data mining, as such, has been a primary objective. Future objectives will include profiles and reputations of classes of festival, and both continuity and change of classes of festival over time.

References

1. Goeuriot, L., Linarès, G., Mothe, J., Mulhem, P., SanJuan, E.: Building evaluation datasets for cultural microblog retrieval, LREC preprint, 5 pp. (2015)
2. 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, Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016), Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany (2016)
3. Harcourt, B.E.: Measured interpretation: introducing the method of Correspondence Analysis to legal studies. University of Illinois Law Review, pp. 979–1017 (2002)

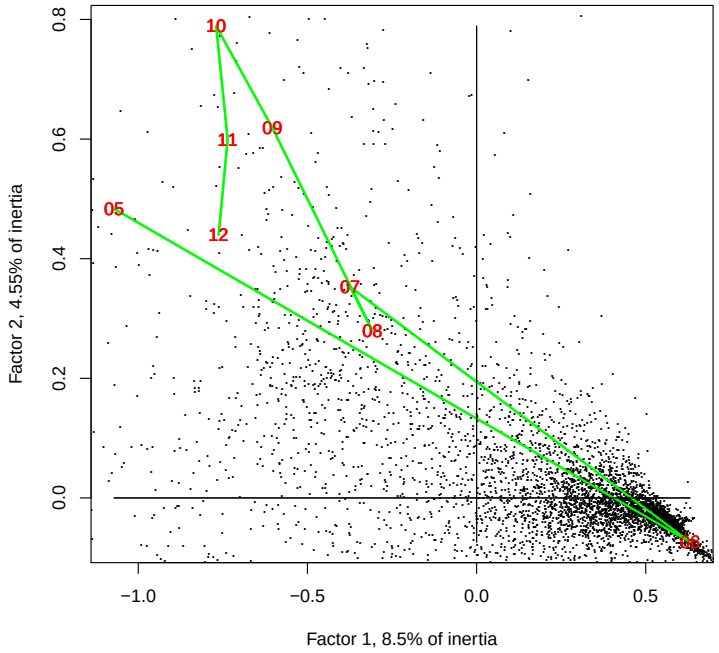

Fig. 10. Evolution by month. Compared to Figures 7 and 8, here the months are displayed as supplementary elements. The month of June, label 06, is quite far apart semantically. Dots are at the locations of words.

4. Husson, F., Lê, S., Pagès, J.: Exploratory Multivariate Analysis by Example Using R, Chapman and Hall/CRC (2011)
5. Inouye, D., Kalita, J.K.: Comparing Twitter summarization algorithms for multiple post summaries. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 298–306 (2011)
6. Keiding, N., Louis, T.A.: Perils and potentials of self-selected entry to epidemiological studies and surveys, Journal of the Royal Statistical Society, Series A, 179, Part 2, pp. 319–376 (2016)
7. Le Roux, B., Rouanet, H.: Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis, Kluwer (Springer), Dordrecht (2004)
8. Murtagh, F.: Correspondence Analysis and Data Coding with R and Java, Chapman and Hall/CRC Press (2005)
9. Murtagh, F.: The Correspondence Analysis platform for uncovering deep structure in data and information, Sixth Boole Lecture. Computer Journal, 53 (3), 304–315 (2010)
10. Murtagh, F., Ganz, A., McKie, S.: The structure of narrative: The case of film scripts. Pattern Recognition, 42, 302–312 (2009)

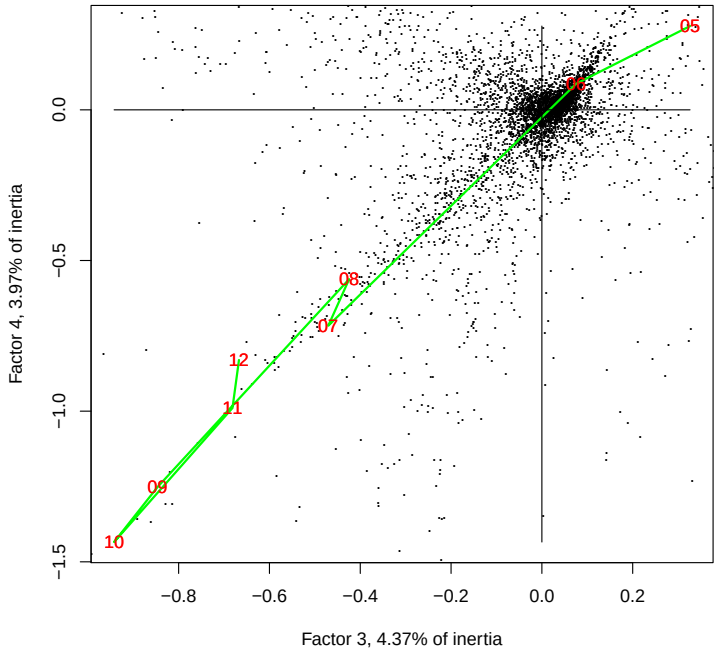

Fig. 11. Evolution by month in the plane of factors 3, 4. Dots are at the locations of words.

11. Murtagh, F., Pianosi, M., Bull, R.: Semantic mapping of discourse and activity, using Habermas's Theory of Communicative Action to analyze process, Quality and Quantity, 50(4), 1675–1694 (2016)
12. Murtagh, F. and Contreras, P.: Random projection towards the Baire metric for high dimensional clustering. In A. Gammerman, V. Vovk and H. Papadopoulos, Eds, Statistical Learning and Data Sciences, Springer Lecture Notes in Artificial Intelligence (LNAI) Volume 9047, 424–431 (2015)
13. Séguéla, J., Saporta, G.: A comparison between latent semantic analysis and correspondence analysis. CARME Conference presentation, (2001) http//carme2011.agrocampus-ouest.fr/slides/Seguela_Saporta.pdf
14. Sharifi, B., Hutton, M.-A., Kalita, J.: Automatic summarization of Twitter topics. In National Workshop on Design and Analysis of Algorithms, Tezpur, India, pp. 121–128 (2010)

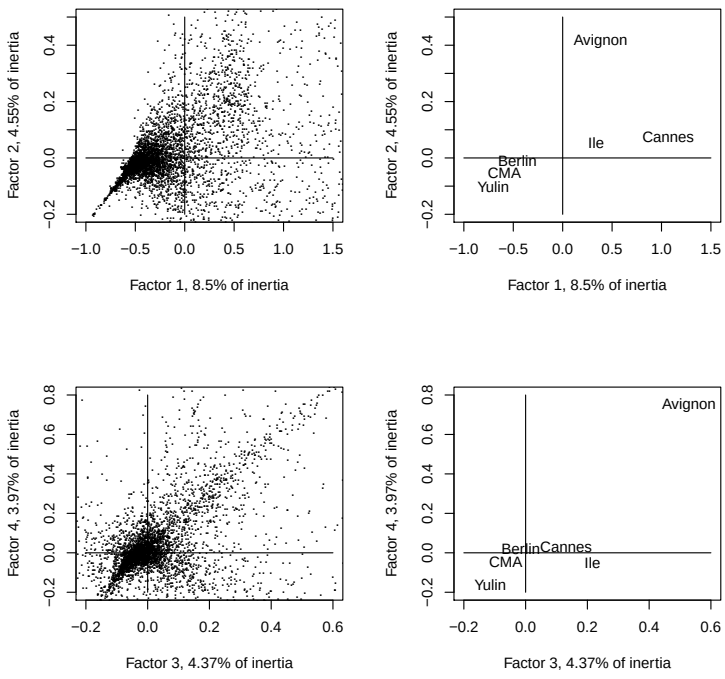

Fig. 12. The principal factor plane, in the top two panels, and the plane of factors 3,4 in the bottom two panels. The left panels display all words, with a dot at each word location. The right panels display the selected festivals.