Hashtag Sense Clustering Based on Temporal Similarity

Giovanni Stilo*
University of Rome Sapienza

Paola Velardi*
University of Rome Sapienza

Hashtags are creative labels used in micro-blogs to characterize the topic of a message/discussion. Regardless of the use for which they were originally intended, hashtags cannot be used as a means to cluster messages with similar content. First, because hashtags are created in a spontaneous and highly dynamic way by users in multiple languages, the same topic can be associated with different hashtags, and conversely, the same hashtag may refer to different topics in different time periods. Second, contrary to common words, hashtag disambiguation is complicated by the fact that no sense catalogs (e.g., Wikipedia or WordNet) are available; and, furthermore, hashtag labels are difficult to analyze, as they often consist of acronyms, concatenated words, and so forth. A common way to determine the meaning of hashtags has been to analyze their context, but, as we have just pointed out, hashtags can have multiple and variable meanings. In this article, we propose a temporal sense clustering algorithm based on the idea that semantically related hashtags have similar and synchronous usage patterns.

1. Introduction

Hashtags are frequently used in social networks to tag message content, although their intended use may differ in different networks. Unlike other tagging systems (e.g., Flickr[1]), hashtags in Twitter are used more as a symbol of community membership than for organizing a message content, as remarked in Huang, Thornton, and Efthimiadis (2010). Because interest groups in Twitter are very dynamic, hashtag popularity surges and decays; and, furthermore, the same hashtag might refer to different events in different time periods. For example, recently Jawbone tried a *#knowyourself* campaign on Instagram,[2] only to find that the hashtag was already being used generically by thousands of users in all sorts of different contexts. In addition to polysemy, there is also a problem of synonymy: Because new hashtags are freely and continuously introduced by the users, different hashtags, for example

* Dipartimento di Informatica, University of Rome Sapienza, Via Salaria 113, Rome, Italy.
E-mail: {stilo,velardi}@di.uniroma1.it.

1 https://www.flickr.com/.

2 http://tinyurl.com/l5w6pt6.

Submission received: 7 March 2015; revised version received: 18 April 2016; accepted for publication: 14 June 2016.

doi:10.1162/COLI_a_00277

[*#FinaleCDM2014, #WCFinal2014], [#RussiainvadedUkraine, #UkraineUnderAttack*] may share the same meaning, sometimes also because of multilinguality. Another problem regarding hashtags—again, not found in other content tagging policies in social media—is obscurity. Hashtags are often hard to interpret for both humans and machines, as they may consist of acronyms [*#MH17, #GERBRA, #RIP*], concatenated words [*#VamosArgentinaQueSePuede*], neologisms (*#procrastitweet*), or abbreviations (*#tcot*[3]).

Together, these problems reduce the effectiveness of Twitter hashtags as a means both for tracing users' interests (owning to obscurity and sense shifts), and for capturing the worldwide impact of trending topics (because of synonymy and multilinguality). Despite this, however, better methods for analyzing the semantics of hashtags are most definitely needed, since hashtags are readily available, whereas textual analysis techniques, when applied on large and lengthy micro-blog streams, are limited both by complexity constraints and by the very reduced dimension of micro-blog texts (140 characters). What is more, real-time detection of sense-related hashtags could be used to improve the task of hashtag recommendation, thereby facilitating the monitoring of online discussions.

The solution proposed in this article is to use an algorithm for hashtag sense clustering based on temporal co-occurrence and similarity of the related time series, named SAX*. The underlying idea of SAX* is that hashtags with similar temporal behavior are semantically related. The nature of this relatedness is either connected with a specific event as for [*#RussiainvadedUkraine, #UkraineUnderAttack*], or is more systematic and repetitive—for example, when hashtags refer to possibly recurring, cultural, or social phenomena (such as [*#followfriday, #thanksgoditsfriday*])—or, finally, it is a genuine synonymy relation, like for [*#WorldCup, #CopaMundial*].

A conference summary of our research (Stilo and Velardi 2014) has recently been published. With respect to that work, we here introduce the following extensions: 1) we define a cluster splitting technique to reduce the problem of co-clustering synchronous, but unrelated, events; 2) we perform a parameter analysis of the algorithm; 3) we extensively compare against a popular content-based method; and 4) we provide a more systematic quality evaluation, matching SAX* clusters generated during ten days of July 2014 against Google snippets.

The article is organized as follows: Section 2 presents the state of the art on hashtag clustering, Section 3 summarizes our technique for efficiently deriving temporal clusters from large and lengthy micro-blog streams, and Section 4 is dedicated to performance evaluation. Section 5 presents our concluding remarks.

2. Related Work

Hashtag sense analysis is different from word sense analysis, primarily because no meaning catalogs are available as is for common words. Therefore, hashtags' sense similarity and dissimilarity is inferred from the context in which they appear. A number of studies have been specifically concerned with the analysis of usage patterns and their stability in time. In Lisa et al. (2013), and Romero, Meeder, and Kleinberg (2011), the authors analyze the correlation between the semantic category of a hashtag and the diffusion patterns of Twitter messages including it. To this end, in Romero, Meeder, and Kleinberg (2011) a data set of hashtags is defined, which is classified into eight semantic categories, such as *technology* or *sport*. A similar study

3 Means "top conservatives of Twitter." This example is reported in Ferragina, Piccinno, and Santoro (2015).

is conducted in Posch et al. (2013), who analyze a number of hashtags belonging to the same data set used in Romero, Meeder, and Kleinberg (2011). Porche et al.'s study confirms that most semantic categories of hashtags have a fast-changing social structure because users start and stop using such hashtag types frequently, and only a few categories exhibit a more stable behavior, with users adopting hashtags with such categories with fewer changes over time. Finally, in Yang and Leskovec (2011), common temporal shapes of Twitter hashtags are detected using K-Spectral Centroid clustering. Our objective in this article is, however, different: Rather than using a time-invariant measure of shape similarity to detect "generic" patterns of human attention, we cluster temporally co-occurring hashtags with a similar shape in order to induce sense similarity.

Traditional approaches to hashtag sense analysis are based on contextual similarity, without considering the temporal dimension. Several papers use hashtags to help in clustering tweets with similar topics. For example, in Mehrotra et al. (2013) hashtags are used as a pooling schema to improve the quality of topics learned from Twitter using latent models such as latent Dirichlet allocation (LDA) (Jain 2010). In Tsur, Littman, and Rappoport (2013), hashtags are manually associated with the same data set of coarse categories used in Romero, Meeder, and Kleinberg (2011) and Posch et al. (2013). Tweets with hashtags in the same categories are then conflated and a model is learned for each category; finally, the model is used for real-time clustering of new messages.

In Ferragina, Piccinno, and Santoro (2015), the authors address the problem of hashtag relatedness. To overcome the problem of hashtag unintelligibility, they create a bipartite graph, in which nodes are either hashtags or named entities co-occurring with hashtags. Named entities are identified and tagged using the TagMe annotator.[4] A semantic relation between hashtags is inferred via their co-occurring named entities. Because named entities can be related to Wikipedia categories, this information is used to train a support vector machine classifier for assigning semantic categories to hashtags.

Only a few papers deal with hashtag clustering, which is the topic of our article. In Muntean, Morar, and Moldovan (2012), the authors represent a hashtag h by the set of words in those messages including h, and then use a MapReduce implementation of K-means to create clusters. In Ozdikis, Senkul, and Oguztuzun (2012), the authors cluster hashtags on the basis of their contextual similarity and then use this information to expand context vectors associated with tweets including these hashtags. In Carter, Tsagkias, and Weerkamp (2011), hashtags from different languages are clustered using a machine translation tool, MOSES. In Antenucci et al. (2011), a combination of co-occurrence frequency, graph clustering, and textual similarity is proposed. Finally, in Feng et al. (2015), hashtags are clustered in real-time along three dimensions: time, space, and contextual similarity. Hashtags are represented as context vectors, including both co-occurring words and hashtags. A nearest-neighbor algorithm is used to create clusters of similar hashtags. The temporal dimension is analyzed in a simple way, by selecting only hashtags whose frequency is above an empirically determined threshold; and the spatial dimension is considered using geo-tagged tweets (we note, however, that geo-tagged tweets are only a small percentage of the total traffic[5]). Feng et al. are the only ones to consider the temporal variability of clusters; however, clusters are generated on the basis of lexical similarity.

4 tagme.di.unipi.it/.

5 http://dfreelon.org/2013/05/12/twitter-geolocation-and-its-limitations/.

3. Clustering Hashtag with Symbolic Aggregate ApproXimation

The SAX* algorithm was first presented in Stilo and Velardi (2016), where it was applied to the task of event detection. In this section we shortly summarize the algorithm[6] and we introduce a new cluster splitting step. The phases of SAX* are as follows:

1. The temporal series associated with hashtags are sliced into sliding windows of length W, normalized and converted into symbolic strings using Symbolic Aggregate ApproXimation (Lin et al. 2003). The parameters of this step are the dimension of the alphabet $|\Sigma|$ and the number $\frac{W}{\Delta}$ of partitions of equal length Δ.
2. Using a set of seed keywords related to known events, we convert their temporal series into symbolic strings and automatically learn regular expressions representing in a compact way common usage patterns. For example, with an alphabet of three symbols, we learn the following expression:
$(a + [bc]?[bc][bc]?a+)?(a + [bc]?[bc][bc]a*)?$
which captures all the temporal series with one or two peaks and/or plateaus in the analyzed window. Only hashtags with frequency higher than a threshold f and matching one of the learned regular expressions are considered in the subsequent steps. These are hereafter denoted as **active hashtags**.
3. Hashtags are analyzed in sliding windows W_i and the detected active hashtags are clustered in each W_i using a bottom–up hierarchical clustering algorithm with **complete linkage** (Jain 2010) and similarity threshold δ.

An example of SAX* cluster is shown in Figure 1 (Paris attacks on November 2015).[7] With $|\Sigma| = 3$, $\Delta = 1day$, and $W = 10$, the cluster centroid is represented by the string: *aabccaaaaa*.

In a minor number of cases, clustering hashtags with similar and synchronous temporal shapes may accidentally merge different, though concurrent, events. Intuitively, this phenomenon is less likely to occur with larger temporal spans and world-wide events: for example, if W = 10 days and $\Delta = 1$ day, it is unlikely that two or more events started, peaked, and ended precisely on the same dates, though it is possible. An example (discussed later in this article) is the buzz around the Argentina–Holland match during the World Cup on 9 July 2014 and the Gaza war news peaking during the very same day. Instead, collisions are more likely to occur as we reduce the length of W, Δ, and the frequency threshold f.

In order to cope with **temporal collision** we have introduced an additional cluster splitting step, which refines a clustering result C^W in a window W, as described in Algorithm 1. First, we build a graph $G = (V, E)$ for each cluster $c \in C^W$ previously detected by SAX*. As shown in Figure 2, a graph G is built associating each vertex $v \in V$ with a hashtag h_i and adding an edge (h_i,h_j) if hashtags h_i and h_j co-occur in a number

6 The interested reader is referred to that paper for additional details.

7 The figure shows only an excerpt of the detected cluster, which was as follows: [*PrayForParis, TousUnis, France, Hollande, ISIS, Muslim, Muslims, Paris, ParisAttack, ParisAttacks, Bataclan, Courage, NotAfraid, Islam, religion, world, parisattacks, refugees, terrorism, Lebanon, Syria, Siria, Palestine, pray*].

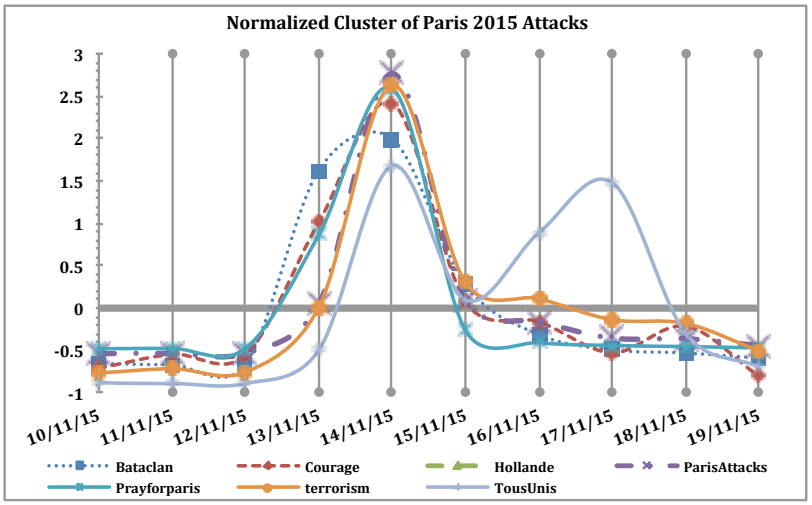

Figure 1
Normalized time series for ISIS attacks in Paris on November 2015.

of documents greater than a threshold τ. Second, we detect connected components in G. Extracting connected components from a graph is a well-established problem (Hopcroft and Tarjan 1973) and does not heavily impact on the computational cost of the entire algorithm (Reingold 2008), especially because the size of the graphs, thanks to the previous temporal pruning step, is very small.

The intuition behind cluster splitting is the following: if two hashtags do not sporadically co-occur in tweets (and they are therefore connected in the related graph), then the underlying senses must be related. We note that geolocalization could also help in

Algorithm 1 Cluster Splitting Algorithm

Input: documents of W, D^W, Clustering C^W
Output: Clustering C'^W

$C'^W \leftarrow \{\emptyset\}$
$g \leftarrow new \quad G(V,E)$
for each c $\in C^W$ **do**
  $V \leftarrow \{\emptyset\}$
  $E \leftarrow \{\emptyset\}$
  for each term $h_1 \in$ c **do**
    for each term $h_2 \in$ c **do**
      if ($h_1 \neq h_2$ **and** $cooccur(h_1,h_2,D^W) > \tau$) **then**
        V $\leftarrow$ V $\cup \{h_1,h_2\}$
        E $\leftarrow$ E $\cup$ $edge(h_1,h_2)$
      end if
    end for
  end for
  for each component $cc \in connectedComponents$(g) **do**
    $C'^W \leftarrow C'^W \cup toCluster(V(cc))$
  end for
end for
return C'^W

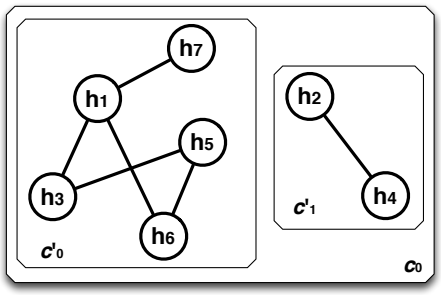

Figure 2
Sub-clusters c'_0 and $c'_1 \in C'^W$ extracted by Algorithm 1 from the original cluster $c_0 \in C^W$.

separating synchronous events (if not world-wide); however, as we already pointed out in Section 2, only a small percentage of tweets can be reliably geolocated.

4. Data Analysis and Evaluation

To tune and evaluate our approach we collected 1% of Twitter traffic, the maximum freely allowed traffic stream, from January 2012 to November 2015, using the standard Twitter API.[8] Our data set, hereafter referred to as the **1% Twitter stream**, is about 5,107 million tweets (of which about 2,750 million are in English). This is considerably larger than the 250 million tweets of the Twitter 2013 collection,[9] which was, to the best of our knowledge, hitherto the largest available collection used in micro-blog analysis. What is more, the 2013 data set spanned only two months and is not multilingual. We considered that a longer time span was indispensable in order to track a sufficiently wide variety of events and to demonstrate the computational efficiency of the proposed method.
In this section, we extensively evaluate SAX* clusters in the following ways:

1. First, under different parameter settings, we evaluate the quality of generated clusters with reference to a number of selected world-wide events that captured the attention of media during Summer 2014. We show that under appropriate parameter settings, all the selected events are captured; and, furthermore, different events in the same temporal slot can be separated effectively.
2. Second, we evaluate systematically all the clusters generated in a randomly selected 10-day window with the best parameter setting determined in the previous step. Cluster quality is evaluated generating Google queries, including the hashtags in each cluster and the cluster peak date. We show that the majority of generated clusters matches at least one Google snippet and has a clear explanation.
3. Third, we evaluate the quality of clusters in a fully automated—yet coarser—way, using two available data sets of semantically tagged

8 https://dev.twitter.com/docs/streaming-apis.
9 https://sites.google.com/site/microblogtrack/.

hashtags. We show that SAX* is able to create almost "pure" clusters, that is, clusters whose members belong to the same semantic category (purity is also qualitatively assessed in the two previous evaluation steps).

4. Fourth, we use two standard cluster internal validity measures, *Inter* and *Intra* cluster similarity, along with a new measure, the *Random* similarity. Our purpose is to show that tweets are more similar to each other when they include hashtags co-occurring in the same temporal window and same cluster than when they include the same hashtags but in different temporal windows.

Evaluation results in all experiments are compared against a popular lexically based hashtag clustering method, K-means lexical clustering (Muntean, Morar, and Moldovan 2012). We implemented the algorithm as described by the authors: For every hashtag h_j we build a virtual document D_j with all the tweets mentioning h, using the same lexical filters described by the authors, and use the Jaccard distance between document pairs to assign hashtags to clusters at each iteration. For every experiment E, we run K-means in sliding windows with the same length W and same frequency threshold f as for SAX*, using a K-value equal to the average number of SAX*-generated clusters for that experiment. The K-values we used are in the range of those reported in Muntean, Morar, and Moldovan (2012), who experimented with K = (20, 100, 500). The results of all our experiments are available as supplementary material in Additional-Materials.zip,[10] along with the full list of SAX* clusters derived in summer 2014.

An additional evaluation experiment was performed in Stilo and Velardi (2016), where we analyzed the computational complexity of the SAX* algorithm, showing that SAX* is one order of magnitude more efficient than other temporal mining methods (Weng et al. 2011; Xie et al. 2013), and two orders of magnitude more efficient than lexical clustering methods such as LDA (Mehrotra et al. 2013) and K-means (Muntean, Morar, and Moldovan 2012). This is a notable advantage when processing large streams of data.

4.1 Tuning of SAX* Parameters and Qualitative Analysis

According to the algorithm description summarized in Section 3, SAX* has several parameters: W, Δ are the discretization parameters; the alphabet Σ determines the granularity of the discretization process; δ is the clustering parameter; f is the minimum frequency for observed hashtags; and finally τ is the co-occurrence threshold used during the cluster splitting phase. To systematically analyze the effect of variable parameters on the cluster quality, we considered a smaller subset of our 1% Twitter stream, from 1 July 2014 to 30 August 2014, referred to hereafter as the **Summer2014** stream. This subset includes 276,057,840 multilingual tweets, of which 114,797,700 are in English.

Manually evaluating the quality of generated clusters in a systematic way, with different combinations of parameters settings, is impossible, both because of the large number of generated clusters and also because hashtags are difficult for external evaluators to interpret, either because they have some local or specific meaning (e.g., *ViernesDeGanarSeguidore*) or they are abbreviations (e.g., *RIP*, *GERBRA*). We thus

10 http://acl.stilo.di.uniroma1.it/ACL-Additional-Materials.zip.

decided to identify and analyze, among the clusters generated during Summer 2014, those related to seven world-wide events (with 11 sub-events), namely: *Independence day (July 4th), three final matches of the Football World Cup (on 8, 9, and 12 July, respectively), the Israel-Gaza conflict (capturing the attention of media during mid-July 2014), the Malaysia airplane crash in Ukraine and subsequent Russia-Ukraine crisis (the crash was on 17 July, and the Russian invasion started later in August), the suicide of Robin Williams (who died on 11 August and was subsequently honored during the Emmy Award later in August), the shooting of Michael Brown and the Ferguson protest (the protest peaked around 14 August), and ISIS killings in Iraq (James Foley was executed by ISIS on 20 August)*.

These events were very diverse in nature (sport, disasters, military, politics, media) and length, and they all had an impact on social media, therefore they are a good test of the ability of SAX* at clustering the related hashtags. In our experiments we selected 10 possible parameter combinations and we manually inspected the derived clusters to study the influence of each parameter on the quality of derived clusters. In some experiments we did not apply character normalization in order to "simulate" synonymy (e.g., in the absence of normalization *WorldCup, WORLDCUP, worldcup* may be considered different tokens); we refer to these hashtags as **pseudo-synonyms**. Table 1 summarizes the effect of each parameter on purity, which is the percentage of cluster members belonging to the same event, and recall, which is the percentage of events captured by at least one SAX* cluster.

We now analyze in more detail one of the best-quality experiments, an excerpt of which is shown in Appendix A1. The experiment refers to the following SAX* setting: $W = 10$, $\Delta = 1day$, $|\Sigma| = 2$, $\delta = 0.35$, $f_{engl} > 49$, $\tau = 2$, *Languages*: only English and Multilingual, no character normalization (to detect pseudo-synonyms). With this setting, the "only English" SAX* experiment includes 3,045 different *active* hashtags (see Section 3), an average number of 74.07 clusters in any of the 60 sliding windows W_i and an average cluster size of 3.44 with standard deviation SD = 0.80. For the multilingual experiment, the values are 3976, 91.15, and 3.73, respectively. With K-means, when setting k = 74 as previously explained, 6,095 different hashtags are clustered, of which 27% are in common with SAX*. The average cluster size is 26.74 and the SD is 0.08. We found that SAX* successfully captures several clusters related to all the listed events and sub-events (therefore the recall is 100%), while the best K-means clusters include at most two event-related hashtags. Furthermore, even though the average SAX* cluster size

Table 1
Effect of parameter setting on performances.

Parameter	Role of parameter	Effect on performance
W	window size	larger windows best suited to capture world-wide events
Δ	discretization size	larger discretization best suited to capture world-wide events
$\|\Sigma\|$	dimension of alphabet	larger alphabets positively affect purity, negatively affects recall
f	frequency threshold	higher frequency positively affects purity, negatively affects recall
δ	clustering parameter	higher values positively affect purity, negatively affect recall
τ	splitting parameter	higher values positively affect purity, negatively affect recall

is 3.44, clusters related to the selected events are much larger: *this is a desired feature*, because important, world-wide events are represented by much larger clusters than the average. Instead, K-means clusters have more or less the same dimension (as also shown by the standard deviation, which is one order of magnitude lower than for SAX*). Furthermore, the number of clusters K with SAX* is not a parameter: both the number of clusters and their size primarily depend on the number of occurred events in a time span and on the intensity of the social buzz, rather than on the clustering algorithm.

In Appendix 1, column (1) shows, for five of the selected events, the corresponding SAX* cluster[11] when considering only English tweets, and column (2) shows Multilingual clusters,[12] column (3) is the date of the slot Δ_j in which each cluster peaked, column (4) shows the most related K-means English clusters extracted in corresponding windows, and column (5) is a short text describing the event. Furthermore, in columns (1) and (2), unrelated hashtags in SAX* clusters are in bold; in column 4, event-related K-means hashtags are underlined. We adopt a different highlighting policy because the majority of clustered SAX* hashtags are related to the target events, whereas the majority of clustered K-means hashtags are unrelated, as can be easily noted inspecting the clusters in column (4).

We remark that whereas SAX* clusters are tagged with the peak date of the related temporal series in each window W_i, K-means clusters are only implicitly tagged with the start and end date of each W_i in which the algorithm was run, therefore manually identifying the clusters of interest was more difficult. In practice, our search was guided by the hashtag labels of previously identified SAX* clusters, because at least the most frequent hashtags in the same W_i are likely to be found also in K-means clusters.[13] As shown in column (3), SAX* clusters peak in the same day or shortly around the peak day for the related event, as compared with Google Trends.

Concerning *cluster purity*, only six clusters[14] out of 31 (19%) have a *single* unrelated hashtag, and one cluster has two unrelated hashtags. Unrelated hashtags are in bold in Appendix 1. With higher values of τ even this small amount of noise disappears, but recall is also reduced. Note also that SAX* perfectly separates two synchronous events on 9 July: the Holland–Argentina match and the Gaza war. Similarly (not shown in Appendix A), on 14 July, the clusters related to the Ferguson protest are perfectly separated from those still commenting on the final match between Germany and Argentina. With $\tau = 0$ (no splitting), mixed clusters were generated.

As far as *multilinguality* is concerned (see column (2)), the most frequent languages are English and Spanish, but we also captured other languages: *Palestineveincra* [Palestine will win], a French tag in clusters of the Gaza war; *MalaysiaBerkabung* [Malaysia mourning], a Malaysian hashtag in clusters of Malaysia airplane crash; and *FinaleCDM*2014, *FinaleMondiale*2014 [WC Final 2014, world final 2014], two Italian hashtags in the cluster of Germany–Argentina. Multilingual clusters have a similar purity performance to that of column (1), with one single exception: the Malaysia airplane crash cluster C88, shown in Appendix A, is fully mixed with a local event in Mexico, "Los mejores vestidos en Premios Juventud 2014," peaking on the same day

11 We show only one cluster and five events for the sake of space. See supplementary material.

12 Note that even when considering only English tweets we can capture a few hashtags in other languages.

13 Remember that SAX* clusters and K-means clusters both include hashtags exceeding the same frequency threshold f, but SAX* hashtags in a cluster must exhibit a similar temporal shape in the considered window, whereas K-means hashtags in a cluster must show a contextual similarity.

14 Only two of which are shown in the excerpt in Appendix A.

(17 July), during which the best looks of participating celebrities were awarded prizes, as we found out.

To conclude, SAX* detects all types of similarity relations between hashtags: event-related hashtags (e.g., *[Putin, Russia, MH17, Ukraine, Malaysia, Malaysia-Airlines, prayforMH17]*) pseudo-synonyms (e.g., *[MichaelBrown, michaelbrown])*, and multilingual synonyms (e.g., *[Germany, Alemania]*, *[FinalMundial2014,worldcupfinal2014, FinaleMondiale2014]*, *[FreePalestine, Palestinalibre]*.

4.2 Qualitative Analysis and Comparison with Lexical Similarity Methods

To conduct a more systematic quality evaluation, we inspected the full set of clusters generated during the randomly selected window W_i, starting on 2 July through 11 July, 2014. We used for SAX* and K-means the same settings as in columns (1) and (4) of Appendix A, respectively. To help manual evaluation, we proceeded as follows: For any generated cluster, we consider the first five most frequent hashtags of the cluster (or fewer, if the cluster is smaller) and we then generate a Google query concatenating these hashtags with the peak date of the cluster, $\pm$ 1 day. A similar approach is used to validate K-means clusters. Next, we inspect the first page returned by Google in the search of a matching snippet. Surprisingly, Google does a good job at handling hashtag queries with acronyms and concatenated words, therefore "good" clusters do return good snippets, including all the query terms, whereas queries including unrelated hashtags produce no results or snippets with partial matches.

In Appendix B we show the first (in temporal order) 10 SAX* clusters in the selected window W_i, with a link to the related best hit page and an explanation of the cluster, if any. Overall, SAX* generated 70 clusters in the considered window. Out of these 70 clusters, 51 (73%) include remarkably related hashtags for which at least one matching page and an explanation was found. For seven clusters (10%) no hits were returned, although the cluster members are lexically similar, for example *[LittleMixLiveStream, LittleMixOnEllen]*. In two cases, matching snippets and explanations were found, but the cluster merges unrelated synchronous events, for example, as in the cluster: *[ARGBEL, ARGvsBEL, MCCvsROW]* that merges the World Football Cup match Argentine vs. Belgium and the Marylebone Cricket Club (MCC) vs. Rest of the World (ROW) cricket match, both occurring on 5 July 2014. Finally, there were 11 clusters (15%) with no Google hits and lexically unrelated or partly related hashtags, for which no explanation was found. K-means clustering produced 73 clusters (available in the supplementary material), of which only three include some related hashtags for which at least one matching page and an explanation was found; for the remaining 71 clusters no related hits were returned or hits are completely missed.

All in all, our results appear to be extremely good in comparison not only with K-means clusters, but with the (few) examples of clusters provided by the authors of all related papers, for example, Ozdikis, Senkul, and Oguztuzun (2012), Muntean, Morar, and Moldovan (2012), and Feng et al. (2015).

4.3 Quantitative Cluster Evaluation

This section describes two quantitative evaluation experiments: The first is an "objective" evaluation based on two available reference classifications; the second provides internal validity measures of cluster cohesion. Both are standard validation approaches adopted in clustering literature and in previous work on hashtag sense analysis (Carter,

Tsagkias, and Weerkamp 2011; Muntean, Morar, and Moldovan 2012; Ozdikis, Senkul, and Oguztuzun 2012; Mehrotra et al. 2013; Tsur, Littman, and Rappoport 2013).

The purpose of the first experiment is to verify that cluster members generated by SAX* belong to the same semantic category (according to a reference categorization), which is weaker than assessing their synonymy or relatedness. The previous section already provided a qualitative manual analysis of a reduced number of clusters. The purpose of the second experiment is to show that tweets whose hashtags belong to the same cluster in the same temporal window W_i are more similar to each other than tweets in the same W_i with hashtags belonging to different clusters, and even than tweets with the same hashtags but extracted in different non-overlapping windows W_{i1} and W_{i2} (as a consequence of sense shifts).

The two experiments are conducted on the previously introduced Summer2014 stream (where we compare with K-means clustering, as before) and on a larger one-year subset of our stream, from June 2012 to May 2013. In both cases, we apply character normalization (e.g., *argentina* and *ARGENTINA* are now merged) to avoid an artificial increase in purity. Furthermore, for the one-year stream we used a higher frequency threshold $f > 99$. In the one-year stream we clustered a set H of 124,345 hashtags in 365 sliding windows. The average number of clusters per window was 33.24 (with standard deviation SD = 6.76) and the average cluster dimension was 10.29 (SD = 3.29). Note that because of the extremely large processing time of K-means clustering,[15] we do not compare with K-means on this larger Twitter stream, but with a simpler baseline, obtained by considering the N most frequent *active* hashtags in each slot Δ with N equal to the average SAX* cluster dimension in the considered window.

4.3.1 Experiment 1: Gold Standard Evaluation. In order to evaluate the purity of extracted clusters, we used two available classifications: (1) The hashtag classification in Tsur, Littman, and Rappoport (2013):[16] this data set (named hereafter TSUR) includes 1,000 highly frequent hashtags manually assigned to nine categories: *Music, Movies, Celebrity, Technology, Games, Sports, Idioms, Political*, and *Other* and has been commonly used to validate hashtag clusters by other authors (Romero, Meeder, and Kleinberg 2011; Posch et al. 2013). (2) A user-populated hashtag taxonomy on the TWUBS on-line hashtag directory:[17] this taxonomy has three top categories (*Event, Organization*, and *Topic*) and 32 sub-categories. For example, *Topic* has the following categories: *Art, Education, Entertainment, Gaming, Health&Beauty*, and so forth. Note that in TWUBS a hashtag may belong to more than one class. We crawled TWUBS and we downloaded about 40,000 hashtags with related classifications.

We remark that only a subset of the TSUR and TWUBS hashtags are *active* in at least one sliding window W_i in our 1% Twitter streams. Overall, in the Summer2014 stream, 102 active hashtags matched those in TSUR and 201 with TWUBS. In the one-year stream, 243 hashtags matched the TSUR data set and 617 from the TWUBS data set (recall that in the one-year stream we used a higher frequency threshold).

The purpose of the evaluation presented here is to quantitatively measure the purity of SAX* clusters, for example, computing the percentage of their members belonging to the same category. Let $H(W_i)$ be the set of active hashtags detected in a window W_i and

15 In their paper, the authors use Map Reduce on Hadoop; however their experiment was conducted on a 3-day stream.

16 We thank the authors for providing the data set.

17 `http://twubs.com/p/hashtag-directory/`.

Table 2
Precision and information gain of SAX* in the hashtag clustering task (Summer2014).

Golden Classifications:	SAX* τ=0		SAX* τ=2		SAX* τ=4		K-Means	
	TSUR (|Cat.=9|)	TWUBS (|Cat.|=32)	TSUR (|Cat.|=9)	TWUBS (|Cat.|=32)	TSUR (|Cat.|=9)	TWUBS (|Cat.|=32)	TSUR (|Cat.|=9)	TWUBS (|Cat.|=32)
Average NIG	0.9679	0.7447	**0.9819**	**0.9269**	0.9804	0.9337	0.6259	0.3451
Stand Dev (NIG)	0.0398	0.1354	0.0282	0.0828	0.0301	0.0825	0.0174	0.0691
Average Precision	0.8484	0.6984	**0.9208**	**0.9069**	0.8569	0.8924	0.2904	0.4948
Stand Dev (Precision)	0.1568	0.1512	0.1345	0.1903	0.2322	0.2130	0.0136	0.0338
Average # of clusters with $\|c_i\| > 1$ in W_i	4.5172	7.0	2.1463	1.4833	1.6486	1.2667	62.1667	63.6667

let $C^{W_i} = \{c_1..c_N\}$ be the clusters generated by SAX* in W_i. Further, let $C^T : \{t_1, t_2..t_K\}$ be the correspondent "ground-truth" classification obtained by grouping $H(W_i)$ in clusters, such that each cluster t_m includes hashtags belonging to just one category.[18] Because TSUR has 9 categories and TWUBS 32, the value of $|C^T| = K$ is upper bounded by one of these two integers, depending upon the adopted reference classification. We compute the **Information Gain (*IG*)** as follows:

$$IG(C^T, C^{W_i}) = \sum_{j=1..K} \left(\frac{|t_k|}{\sum_{j=1..k} |t_j|} \right) log\left(\frac{|t_k|}{\sum_{j=1..k} |t_j|} \right) - \sum_{n=1..N} \left(\frac{|c_n|}{\sum_{j=1..N} |c_j|} \right) \cdot \sum_{k=1..K} \left(\frac{|c_n \cap t_k|}{|c_n|} \right) log\left(\frac{|c_n \cap t_k|}{|c_n|} \right) \quad (1)$$

In the equation, with reference to a set of clusters C^{W_i} of $H(W_i)$ *active* hashtags in window W_i, $K = |C^T|$ is the number of categories of the reference classification (either TSUR or TWUBS) having at least one member in $H(W_i)$; N is the number of clusters generated by SAX* in C^{W_i}; the minuend is the initial entropy of the set $H(W_i)$, namely, the initial impurity of the examples; and the subtrahend is the weighted sum of entropies of each cluster $c_n \in C^{W_i}$. The *IG* then provides a measure of the improvement of SAX* over a baseline classifier assigning a category to a hashtag based on the a priori probability distribution of the various categories in $H(W_t)$. We actually compute the normalized *IG* (*NIG*), because K may vary in each W_i.

Table 2 shows average and standard deviation (*SD*) of *NIG* and Precision over the 60 clusterings C^{W_i} of Summer2014. IG values are computed with $\tau = 0$ (no cluster splitting) and $\tau = 2$ and $\tau = 4$. The *NIG* is also computed for K-means clustering, with the same settings as for column (4) of Appendix A. Table 3 shows the same measures computed over the 365 clusterings C^{W_i} derived in one year. Again, we evaluate purity with and without cluster splitting. In this case, comparison is performed against the simpler daily frequency baseline, as previously explained.

The tables show that SAX*-induced clusters are indeed highly semantically related, at least as far as the reference semantic categories are concerned. The average NIG is close to the maximum value of one bit for TSUR and slightly lower for TWUBS, which also has a lower precision. This is consistent with the fact that the number of available

18 Note that K, the number of categories in W_i, is in general lower than the total number of available categories in the two classifications. Also note that TWUBS allows for multiple classifications; therefore some hashtags may belong to more than one category.

Table 3
Precision and information gain of SAX* in the hashtag clustering task (one-year stream).

Golden Classifications:	SAX* τ=0		SAX* τ=2		Baseline Clusters	
	TSUR (|Cat.|=9)	TWUBS (|Cat.|=32)	TSUR (|Cat.|=9)	TWUBS (|Cat.|=32)	TSUR (|Cat.|=9)	TWUBS (|Cat.|=32)
Average NIG	0.967	0.778	**0.98**	**0.82**	0.77	0.6005
Stand Dev (NIG)	0.042	0.1002	0.043	0.1066	0.25	0.27
Average Precision	0.88	**0.77**	**0.97**	0.76	0.73	0.73
Stand Dev (Precision)	0.127	0.128	0.085	0.26	0.27	0.29
Average # of clusters with $\|c_i\| > 1$ in W_i	4.85	7.86	2.45	2.94	6.2	4.5

categories is more than three times higher for TWUBS (32 vs. 9) and, in addition, in TWUBS some hashtags have multiple classifications. In general, clusters are very pure (e.g., members belong to a unique category), as shown in the following cluster, in which hashtags have been replaced by their semantic labels in TWUBS: On: <Jun 25, 2012>: {[SPORTS] [MOVIES, MOVIES] [MOVIES] [SPORTS, SPORTS] [IDIOMS, IDIOMS, IDIOMS, IDIOMS, IDIOMS] [POLITICAL, POLITICAL] [MOVIES, MOVIES] [IDIOMS, IDIOMS, IDIOMS, IDIOMS, IDIOMS]} (**NIG=1.00**)

Both tables also show that best results are obtained with $\tau = 2$, which is consistent with the qualitative evaluation of Section 4.1. As far as K-means and the daily frequency baseline are concerned, the tables show significantly lower performance for both *NIG* and Precision. Note that the daily frequency baseline is slightly better than K-means, which is probably due to the fact that to compute the baseline we consider active hashtags, whereas K-means only uses a frequency threshold. Finally, the last row in Tables 2 and 3 show the average number of generated clusters per window: These numbers are lower than those reported in Section 4.1, because the set of hashtags with a corresponding TSUR or TWUBS category are much less than the total number of active hashtags, according to SAX* and K-means filters, respectively.

4.3.2 Experiment 2: Internal Cluster Validity Measures. The objective of this experiment is to analyze the temporal shift of hashtags' meaning by comparing the similarity of messages sharing the same hashtag in the same window against that of messages with the same hashtag but in non-overlapping windows.

Similarly to other papers (Muntean, Morar, and Moldovan 2012; Ozdikis, Senkul, and Oguztuzun 2012; Tsur, Littman, and Rappoport 2013), we represent each hashtag with a *tfidf* word vector of the virtual document D^i_h created by conflating all tweets including a given hashtag h, but we also add the constraint that tweets must co-occur in the same window W_i. We introduce three metrics. The first two are well known measures commonly used to evaluate the lexical similarity of two documents D^i_k, D^i_j belonging either to the same or to different clusters in a window W_i. The third one computes the similarity between co-clustered hashtags, but now documents D^{i1}_k, D^{i2}_j are created by collecting tweets occurring in two different randomly chosen and non-overlapping windows W_{i1}, W_{i2}. The objective of this third measure is to verify our hypothesis of a temporal shift of hashtag meaning. Note that lexical similarity is used in this experiment only for evaluation purposes: SAX* does not exploit terms in messages other than hashtags.

For each hashtag pair $h_k, h_j \in c_n$ and for all clusters C^{W_i} detected in window W_i we compute the average intra-clusters similarity IntraSym(C^{Wi}), based on the cosine similarity[19] $sym(D^i_k, D^i_j)$:

$$IntraSym(C^{W_i}) = \frac{1}{|C^{W_i}|} \cdot \sum_{c_n \in C^{W_i}} \left[\frac{1}{|c_n|(|c_n|+1)} \sum_{h_k, h_j \in c_n, k \neq j} sym(D^i_k, D^i_j) \right] \quad (2)$$

Next, for each hashtag pair $h_k \in c_n$, $h_j \in c_{n'}$ and all clusters C^{W_i} detected in window W_i we compute the average inter-clusters similarity InterSym(W_i) based on the cosine similarity sym(D^i_k, D^i_j):

$$InterSym(C^{W_i}) = \frac{1}{|C^{W_i}|(|C^{W_i}|-1)} \sum_{c_n, c_{n'} \in C^{W_i}, n \neq n'} \left[\frac{1}{|c_k||c_{k'}|} \cdot \sum_{h_k \in c_n, h_j \in c_{n'}} sym(D^i_k, D^i_j) \right] \quad (3)$$

Finally, for each pair of co-clustered hashtag $h_k, h_j \in c_n$ and for all clusters C^{W_i} detected in window W_i we compute the average random clusters similarity RandSym(C^{W_i}) based on the cosine similarity of the related virtual documents sym(D^{i1}_k, D^{i2}_j) when h_k, h_j occur in two non-overlapping randomly selected windows W_{i1}, W_{i2} where $i \neq i_1 \neq i_2$ and $i_1, i_2 \in random(W)$. In other terms, given that h_k, h_j are found to be related in W_i, we measure the stability of this relation when the very same hashtags occur in two different temporal windows W_{i1} and W_{i2}.

$$RandSym(W_i, W) = \frac{1}{|C^{W_i}|} \sum_{c_n \in C^{W_i}} \left[\frac{1}{|c_n|(|c_n|+1)} \cdot \sum_{k \neq j, i \neq i_1 \neq i_2 \in random(W)} sym(D^{i1}_k, D^{i2}_j) \right] \quad (4)$$

Inspired by the information gain, we compute the similarity gain by the following formula:

$$SymGain(W_i) = \frac{IntraSym(W_i) - RandSym(W_i, W)}{RandSym(W_i, W)} \quad (5)$$

Tables 4 and 5 show the values and standard deviation SD of the ratios $R1 = \frac{InterSym(W_i)}{IntraSym(W_i)}$, $R2 = \frac{RandomSym(W_i, W)}{IntraSym(W_i)}$ and $SymGain(W_i)$, averaged over the total number of sliding windows (i.e., 60 for Summer2014 and 365 for the one-year stream). Results are also compared with K-means clusters in Summer2014 and with baseline clusters in the one-year stream.

Note that both for $R1$ and $R2$ best values are close to zero (the desired result is that $IntraSym(W_i) \gg InterSym(W_i)$ and the expected result in the case of temporal sense shift is that $IntraSym(W_i) \gg RandSym(W_i, W)$).

As in previous experiments, SAX* clusters outperform K-means and baseline clusters, and $\tau = 2$ obtains the best values, with the exception of the similarity gain value in Summer2014, where $\tau = 0$ performs slightly better. Note, in all columns, that the similarity gain and the difference between $R1$ and $R2$ are higher when results are averaged on a one-year stream, as expected, since hashtag sense shifts are more likely to be observed in larger time spans.

19 http://en.wikipedia.org/wiki/Cosine_similarity.

Table 4
Cluster similarity measures (Summer2014).

	SAX* τ=0			SAX* τ=2			SAX* τ=4			K-Means		
	R1	R2	SymGain	R1	R2	SymGain	R1	R2	SymGain	R1	R2	SymGain
Average	0.0320	0.0770	**16.2472**	**0.0131**	**0.0878**	12.0377	0.0140	0.1002	12.3644	0.9419	0.3667	1.7669
St. Deviation	0.0194	0.0445	9.2044	0.0074	0.0330	4.9338	0.0078	0.0479	10.2259	0.1775	0.0479	0.3596

Table 5
Cluster similarity measures (one-year stream).

	SAX* τ=0			SAX* τ=2			Baseline Clusters		
	R1	R2	SymGain	R1	R2	SymGain	R1	R2	SymGain
Average	0.0245	0.1093	12.9158	**0.0094**	**0.0646**	**28.9702**	7.3427	0.5331	2.6003
St. Deviation	0.0161	0.0766	10.4426	0.0101	0.0903	26.6849	25.6822	0.5987	2.9237

Comparably low $R2$ values (observed in SAX*, K-means, and baseline clusters) demonstrate one of the main claims of our methodology: hashtag similarity is time-related. As a consequence, hashtag clustering models based only on content features cannot cope with sense shifts. As an example, consider two hashtags, *CNN* and *America*, co-occurring in a cluster starting on 22 October, 2012. Two examples of tweets in this window are:

#America #CNN

23 Oct 2012: *Final presidential debate is tonight tune in #America!!!*
24 Oct 2012: *Final Debate, Tune in on #CNN*
As to be expected, the content of the tweets is very similar. However, the same two hashtags may be used in very different contexts when found in different temporal windows, as for example:

#CNN:

29 Oct 2012: *Might watch a bit of #CNN to follow #Sandy*

#America:

14 Dec 2012: *Very sad day in #America. Pray for the families in Connecticut.*

To conclude, both qualitative and quantitative evaluation shows that lexical similarity, even when adopting the virtual document heuristics in order to obtain larger contexts, does quite a bad job at detecting hashtag similarity in micro-blogs.[20] In marked contrast, temporal similarity allows for very meaningful clusters (both large and precise) to be created.

20 We incidentally note that the few cluster examples reported in Muntean, Morar, and Moldovan (2012) are not very good, either.

5. Concluding Remarks

In this article we introduced a hashtag sense clustering algorithm, named SAX*, based on the novel notion of temporal similarity. Our algorithm converts temporal series into a sequence of symbols using Symbolic Aggregate ApproXimation (Lin et al. 2007), and then clusters hashtags with similar and co-occurring sequences. Our work was inspired by the observation that hashtags suffer from polysemy, ambiguity, and most notably, sense shifts, a problem that, though neglected in current literature, affects purely content-based models. Clustering hashtags in the same temporal slot reduces all these problems and allows the detection of clusters with a very high contextual similarity.

SAX* was evaluated in a variety of ways, on a very lengthy Twitter stream, showing a number of advantages over previously used methods:

- First of all, our qualitative evaluation in a large number of experiments (available both in the Appendix of this paper and as supplementary material) showed that the detected clusters are very pure, because all hashtags in clusters appear to be related with very few exceptions. Furthermore, the number of generated clusters and their size depends primarily on the number and impact of the events occurring in the related time span, rather than on clustering parameters. This is a notable advantage over clustering methods previously used in literature.
- The quantitative evaluation, based on available reference categorizations and on internal validity measures, confirmed empirical findings on clusters' quality.
- Both qualitative and quantitative evaluations show that lexical methods perform very poorly even in small temporal windows (i.e., in the absence of sense shifts), because context sparsity in Twitter messages cannot be adequately coped with, even when merging all tweets including the same hashtag.

An additional, but not minor, advantage of our algorithm is its reduced computational complexity, as extensively demonstrated in Stilo and Velardi (2016).

We conclude by noting that, with SAX*, the coverage of hashtag usage patterns depends on the density of the analyzed stream: With a poorly dense stream (such as our 1% flow), world-wide events are very precisely captured, but minor events may either be missed or conflated into the same cluster. We believe that precision and recall on local and minor events would significantly improve with denser streams, possibly using geolocalization to better separate synchronous events, though we could not test our hypothesis in the absence of an adequate data set (i.e., a local or denser Twitter stream).

Appendix A.
Excerpt of hashtag clusters related to 5 Summer2014 events (full data in additional material).

SAX* parameters: $W = 10$, $\Delta = 1day$, $\tau = 2$, $\delta = 0, 35$, $|\Sigma| = 2$, $f_{english} > 49$
K-means parameters: as in Muntean, Morar, and Moldovan (2012), with K = 73 (average # of clusters in SAX* when $f_{english} > 49$)

SAX* clusters, only English tweets	**SAX* clusters, multilingual tweets**	**SAX* cluster peak-date**	**Best k-means cluster**	**Event description**
C27 [Freedom, America, GodBlessAmerica, Happy4thofJuly, merica, Happy4th, Merica, HappyIndependenceDay, Murica, fireworks, happy4th, indepenceday, murica, freedom, america]	**C14** [indepenceday, merica, Happy4th, America, Merica, Happy4thofJuly, Murica, america, fireworks, freedom, murica, happy4th]	July 4^{th}	[4thOfJuly, 5SOSCANADIANTOUR, 5sosAmnesiaMusicVideo, BestFeelingKapag, Boston, FIR365, GazaUnderAttack, GermanyBattleReady, HappyIndependenceDay, **+more unrelated hashtags**]	Independence day
C6 [hollandvsargentina, netherlandsvsargentina, NEDvsARG, NEDARG, ARGvsHOL, ARGvsNED]	**C53** [*VamosArgentina*, arg, ARG, Arg, Argentina, Messi, Romero, *VAMOSARGENTINAQUESEPUEDE*, *VamosArgentinaQueSePuede*, *VamosMessi*, *argentina*, messi, Higuain, WM2014, *Penales*]	July 9^{th}	[AMAs, AsianAnal, BRA, CFL, CallMeCam, CoblosNomor1_PrabowoHatta, Friends, HNYTrailerOn15thAugust, HupHollandHup, Knicks **+more unrelated hashtags**]	Holland Argentina (World Cup)
C19 [prayforgaza, PrayForPalestina, FreePalestine, PrayForGaza, PrayForPalestine, SaveGaza, SavePalestine, **HappyHeechulDay**]	**C52** [SavePalestine, FreePalestine, *Palestinalibre*, PrayForGaza, PrayForPalestina, PrayForPalestine, SaveGaza, *PalestineVaincra*, **HappyHeechulDay**]	July 9^{th}	[5sosLikeSocks, Battleground, CALLMECAM, EUI4LTU, FebruaryWish, FreePalestine, Gemini, IsraelUnderAttack **+more unrelated hashtags**]	Gaza war
C54 [Putin, Russia, MH17, Ukraine, Malaysia, MalaysiaAirlines, prayforMH17, 5DaysFor1D, PrayForMH17, **LouisDeservesTheWorld**, **WeLoveYouLouis**, MH370, MalaysianAirlines, PrayforMH17]	**C88**[Putin, Russia, MH17, Ukraine, Malaysia, MalaysiaAirlines, PrayForMH17, MH370, MalaysianAirlines, PrayforMH17, prayforMH17, ukraine, mh17, **LouisDeservesTheWorld**, **WeLoveYouLouis**, **SmileLouis**, **reinasdeluiscoronel**, **LuisCoronelPJ**, **luiscoronelpj**, **mejorvestidoluiscoronel**]	July 17^{th}	[49ers, 5sosRockOutWithYourSocksOutTour, ASG2014, BGC12, DepressingDisney, EminemWembley, GirlProblems, HappyEngagementElounor, MalaysiaAirlines, NYC, NiallsNeymarTattoo, PopAsiaGOT7, PrayForMH17**+more unrelated hashtags**]	Malaysia airplane crash and Ukraine crisis
C58 [RobinWilliams, sosad, DeadPoetsSociety, MrsDoubtfire, Jumanji, Legend, Robin, legend, RIPRobin, RIPRobinWillams, Rip, riprobinwilliams, RIPRobinWilliams, robinwilliams, RIP, rip, sad]	**C53** [RIP, Legend, MrsDoubtfire, RIPRobinWilliams, RobinWilliams, legend, robinwilliams, Jumanji, makeuptransformation, rip, RIPRobinWillams, sad, QEPD, RipRobinWilliams, riprobinwilliams, DEP]	August 11^{th}	[19MillionSalmaniacsOnFB, 1MonthOf1DInPortugal, DownloadDirtyDancerNow, EngvInd, GetToKnowYouOnItunes, JAKEBOYS, JustWaitOnIt, LLWS2014, LetsBeCops, Monday, ProphetMuhammad, RIProbinwilliams**+more unrelated hashtags**]	Robin Williams suicide

Appendix B.
The first 10 SAX* events in the window starting on 2 July through 11 July, with the related best hit page and an explanation.

C. #	Date	Cluster	Eval.	Web Link	Event
1	*03-07-14*	[4thofJuly2014, 4thofJuly, fourthofJuly]	**OK**	https://www.youtube.com/watch?v=RinK52M7XfM	Independence day
2	*03-07-14*	[catfish, catfishmtv]	**OK**	http://heavy.com/entertainment/2014/07/jeff-megan-catfish-mtv-episodes-watch-season-3/	Catfish episodes on MTV
3	*03-07-14*	[callmecam, CalIMeCam, CallMeCam, Callmecam, CALLMECAM, CaIIMeCam, CallmeCam]	**OK**	https://www.youtube.com/watch?v=4-VWameDPBc	Cameron Dallas (Internet personality and actor) tweets
4	*03-07-14*	[Bring1DToSerbia, SerbiaNeedsOneDirection]	**OK**	https://www.facebook.com/Where-We-Are-Tour-One-Direction-2014-Croatia-Bosnia-and-Serbia-519922548074435/	Where We Are tour of One Direction in Serbia
5	*03-07-14*	[HurricaneArthur, IndependenceEve]	**OK**	http://nypost.com/2014/07/03/rainy-weather-expected-to-clear-by-4th-of-/07/y-eve/	Rainy weather expected on Independence day
6	*03-07-14*	[independenceday, BornInAmericaVideo]	**OK**	https://www.facebook.com/officialmagicmack/videos/346354782180581/	Independence day video
7	*03-07-14*	[3AM, 3am]	**nothing found**		
8	*03-07-14*	[JANOSKIANStour, JANOSKIANSTour, JANOSKIANtour]	**OK**	http://www.songkick.com/artists/6692944-janoskians	The Janoskian's tour in July 2014
9	*03-07-14*	[TelehitAustinMahone, TelehitCD9, OneDirectionEsMejorQueTheBeatles, TelehitJB, TelehitOneDirectionRetoKit, RetoTelehitJary, RetoTelehit, RetoTelehitJB, TelehitJonasBrothers2, TelehitJonasBrothersGuitarra, retotelehit, JonasBrothers, Retotelehit, RetoTelehitCD9, TelehitJonasBrothersGuitarra2, TelehitOneDirectionRetoKid, RetoTelehitKitOneDirection, TelehitPerfumeJustinBieber, Telehitonedirectionretokit]	**OK**	https://www.facebook.com/telehit/photos/a.146482682947.112844.131292887947/10152576091587948/	Telehit prizes on 3 July 2014
10	*03-07-14*	[SaveTwitterFromBOT, HBDCancerian]	**nothing found**		

References

Antenucci, Dolan, Gregory Handy, Akshay Modi, and Miller Tinkerhess. 2011. Classification of tweets via clustering of hashtags. Technical report, EECS 545 Final Project.

Carter, S., M. Tsagkias, and W. Weerkamp. 2011. Twitter hashtags: Joint translation and clustering. In *Proceedings of Web Science 2011*, Koblenz.

Feng, Wei, Chao Zhang, Wei Zhang, Jiawei Han, Jianyong Wang, C. Aggarwal, and Jianbin Huang. 2015. Streamcube: Hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *Proceedings of the IEEE 31st International Conference on Data Engineering (ICDE)*, pages 1561–1572, Seoul.

Ferragina, Paolo, Francesco Piccinno, and Roberto Santoro. 2015. On analyzing hashtags in Twitter. In *Proceedings of AAII Conference*, Oxford.

Hopcroft, John and Robert Tarjan. 1973. Efficient algorithms for graph manipulation. *Communications ACM*, 16(6):372–378.

Huang, Jeff, Katherine M. Thornton, and Efthimis N. Efthimiadis. 2010. Conversational tagging in Twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 173–178, New York, NY.

Jain, Anil K. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Lin, Jessica, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11, New York, NY.

Lin, Jessica, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing sax: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144.

Lisa, Posch, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2013. Meaning as collective use: Predicting semantic hashtag categories on Twitter. In *22nd International World Wide Web Conference*, Rio de Janeiro.

Mehrotra, Rishabh, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 889–892, New York, NY.

Muntean, Cristina Ioana, Gabriela Andreea Morar, and Darie Moldovan. 2012. Exploring the meaning behind Twitter hashtags through clustering. In Witold Abramowicz, John Domingue, and Krzysztof Wecel, editors, *BIS (Workshops)*, volume 127 of *Lecture Notes in Business Information Processing*. Springer, pages 231–242.

Ozdikis, Ozer, Pinar Senkul, and Halit Oguztuzun. 2012. Semantic expansion of hashtags for enhanced event detection in Twitter. In *Proceedings of VLDB 2012 Workshop on Online Social Systems*.

Posch, Lisa, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2013. Meaning as collective use: Predicting semantic hashtag categories on Twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 621–628, Geneva.

Reingold, Omer. 2008. Undirected connectivity in log-space. *Journal of the ACM*, 55(4): 17:1–17:24.

Romero, Daniel M., Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 695–704, New York, NY.

Stilo, Giovanni and Paola Velardi. 2014. Temporal semantics: Time-varying hashtag sense clustering. In Krzysztof Janowicz, Stefan Schlobach, Patrick Lambrix, and Eero Hyvaanen, editors, *Knowledge Engineering and Knowledge Management*, volume 8876 of *Lecture Notes in Computer Science*. Springer International Publishing, pages 563–578.

Stilo, Giovanni and Paola Velardi. 2016. Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Mining and Knowledge Discovery*, 30(2):372–402.

Tsur, Oren, Adi Littman, and Ari Rappoport. 2013. Efficient clustering of short messages into general domains. In *ICWSM*, Ann Arbor, MI.

Weng, Jianshu, Yuxia Yao, Erwin Leonardi, and Francis Lee. 2011. Event detection in Twitter. In *International AAAI Conference on Weblogs and Social Media*, Seattle, WA.

Xie, Wei, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. 2013. Topicsketch: Real-time bursty topic detection from twitter. In *IEEE 13th International Conference on Data Mining (ICDM)*, pages 837–846, Dallas, TX.

Yang, Jaewon and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 177–186, New York, NY.