Topic modeling and summarization of live TV shows via Twitter

Dimitris Spathis[1], Mirsini Demi[2]

Aristotle University of Thessaloniki, School of Informatics, Thessaloniki, Greece

Abstract. Watching TV is usually accompanied with comments about the content. We tend to addresss these comments to nearby people or online friends. In this empirical study we retrieved 30k Twitter status updates during popular TV talk shows. Topic modeling analysis allows us to separate themes and, eventually, summarize long TV broadcasts automatically.

Keywords: Topic modeling, TV summarization, Twitter.

1 Introduction

Real time social networks, such as Twitter, enable people to share short updates and thoughts that reflect their current view. Watching TV might not be considered the most social activity, but the comments about the content could be regarded as a meta-commentary online. These tweets are ephemeral by means that you have to tune to the show in order to get the gist. While ephemeral would mean useless in this context, aggregation, computational linguistics and statistics could help us use the sheer volume of tweets during a TV show as metadata. Video editors and related professionals could exploit topic modeling of tweets during a TV show in order to decide when to cut and which sections to keep. Also, automatic topics could help archivers of multimedia content to store it efficiently.

Topic modeling is gaining increasing attention in different text mining communities. Latent Dirichlet Allocation (LDA) [Blei et al., 2003], a probabilistic model, is becoming a standard tool. Due to the 140 character restraint in Twitter, while many researchers wish to use standard text mining tools to understand messages (such as tf-idf), the restricted length of those messages prevents them from being employed to their full potential. In a thorough study [Hong & Davison., 2010] showed that by training a topic model on aggregated messages we can obtain a higher quality of learned model which results in significantly better performance. In event summarization, [Shen et al., 2013] proposed a participant-based event summarization approach which "zooms-in" the Twitter event streams to the participant level, detects the important

1 Research, data analysis and visualization.

2 Data retrieval.

sub-events associated with each participant using a novel mixture model that incorporates both the "burstiness" and "cohesiveness" of tweets, and generates the event summaries progressively. We incorporate a much simpler method of choosing the "official" hashtag of the TV show and rely on these tweets in order to find sub-events by volume peaks and topics.

We collected around 30k tweets for 6 night talk shows in Greek TV. These talk shows average about 3 hours of runtime. We performed LDA and computed peak volumes. Our work is described in the following sections that cover the data retrieval, cleaning, analysis and visualization. We provide the source code, datasets and Jupyter notebooks, contributing to the reproducible science movement.

Table 1. The analyzed TV shows.

TV show	Broadcast date	Tweets volume
Anatropi	5/4/2016	483
Anatropi	12/4/2016	3109
Enikos	4/4/2016	10866
Enikos	1/4/2016	6459
Enikos	18/4/2016	8494
Ellinofrenia	5/4/2016	410

2 Methodology

In this section we describe our algorithm and the challenges we faced during implementation.

2.1 Retrieval

Twitter provides an API that enables us to gather tweets by hashtags. Second author, who was responsible for data collection, downloaded the real time tweets using the Python library tweepy[3].

2.2 Preprocessing

Our approach was to extract the tweet and the timestamp from the Twitter API output, which was in JSON format. We used the Jupyter Python[4] along with Pandas [5]framework due to better support of Greek language and unicode. After extracting the relevant fields, we transformed the timestamp timezones into local time, in order to be

3 http://www.tweepy.org/

4 http://jupyter.org/

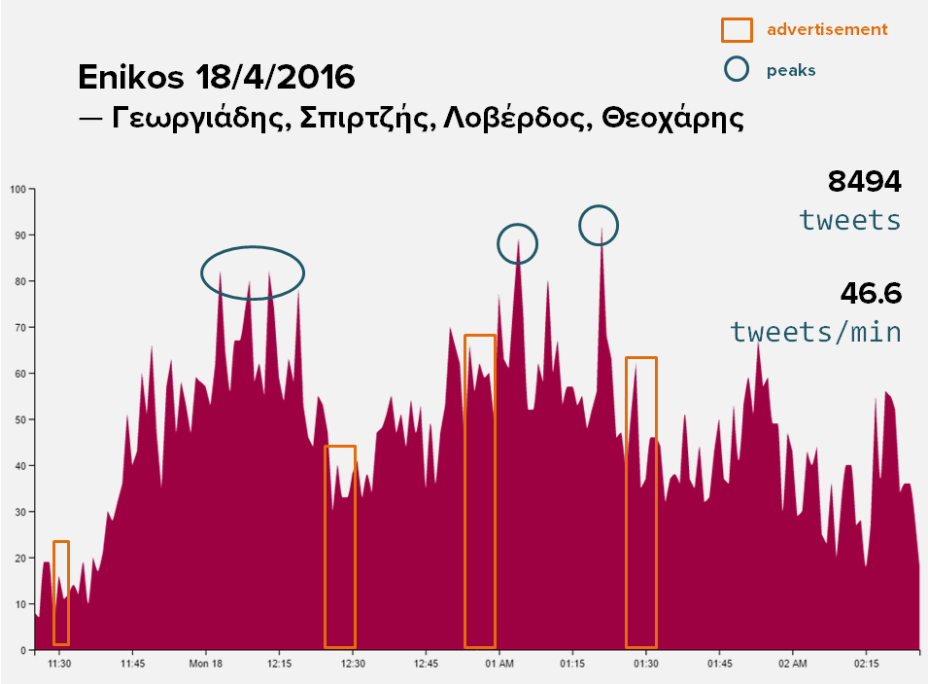
5 http://pandas.pydata.org/

able to match them with real time broadcast video.

2.3 Timeseries visualization

The next step is to create a time series in a per minute minute format. We resampled the timestamps at 1 minute format using the built in *resample* function of Pandas. The average tweet count is plotted in the figure below. We note the advertisement time from the broadcast.

Figure 1. Tweets volume for the TV show Enikos. The guests' names are indicated in Greek.



2.4 Frequency analysis

In order to find the most frequent words for each TV show we used the NLTK Python library along with a greek words stemmer vocabulary that we open source with this paper. Thus, we found the top-k frequent words that occur during a TV show. The most common words such as "the", "and" etc are eliminated as stopwords.

2.5 LDA Topic Modeling

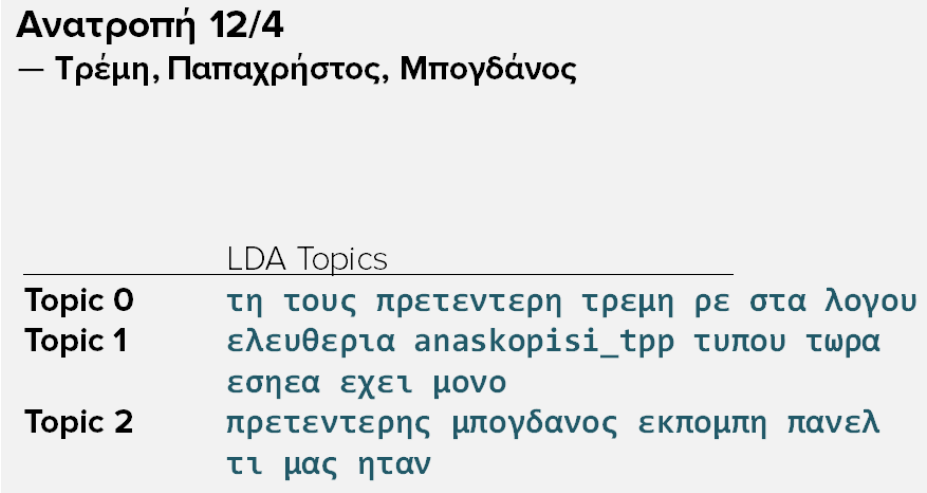
Latent Dirichlet Allocation is an unsupervised machine learning technique which

identifies *latent* topic information in large textual datasets. It uses a "*bag of words*" approach, which treats each document as a *vector* of word counts. Each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. LDA defines the following generative process for each document in the collection:

- For each document, pick a topic from its distribution over topics.
- Sample a word from the distribution over the words associated with the chosen topic.
- The process is repeated for all the words in the document.

Our implementation used the LDA Python library[6] along with sklearn for vectorization. First, we strip all accents from words because Greek contain many variations of the same letter. Then, we applied a few common heuristics: words occurring in *only one* document or in *at least 95%* of the documents are removed. We arbitrarily choose the number of topics we want to find in the documents. In our case we limit it to 3 topics. Below, we can see an example topic model.

Figure 2. Topics from the TV show Anatropi. We observe that topic 0 is about the two hosts, topic 1 is about free speech which was the main topic of the show and topic 2 contains the guest.



3 Conclusion

We analyzed a large dataset of real time tweets regarding live TV shows and attempted to find sub-events via topic modeling, peak volumes and word frequencies. The main

6 https://pypi.python.org/pypi/lda

findings of our work are grouped below:

- **Inadequate NLP resources and tools for non-English languages.** While for English it should be quite easy to perform stemming analysis, in Greek one shoud find a specific vocabulary that is not included in the popular libraries.

- **Inherent difficulty to detect actual time in recorded TV broadcasts.** Most TV shows are uploaded online after a couple of days, but the advertisements are cut off. Thus, we cannot estimate the actual time of the broadcast moment, resorting us to record the broadcast ourselves.

- **Single-topic hashtags are hard to apply topic modeling.** The *designated* hashtags of the TV shows are *doomed* to talk about the same thing. So, we cannot really distinguish completely different topics, only slight sub-topics.

- **Frequent words are the people who spoke more**. The frequency analysis shows that the most repeated words are those of the hosts, the guests and some users.

- **Anonymous users gain popularity during TV broadcasts**. Some anonymous Twitter users express quite popular opinions using the hashtag of the TV show. Most of the cases are humorous comments or remarks. The TV shows could incorporate these tweets as feedback.

References

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993-1022.

2. Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics* (SOMA '10). ACM

3. Shen, C., Liu, F., Weng, F., & Li, T. (2013). A Participant-based Approach for Event Summarization Using Twitter Streams. In *HLT-NAACL* (pp. 1152-1162).