Mapping Satirical Articles to Source Event Article

Sho Ohata
sho.ohata@berkeley.edu

Abstract

Satire is a genre of literature that features irony and sarcasm and is usually aimed at critiquing social, cultural and current event issues. There are many satirical news outlets, the most prominent US one being "The Onion". While much of The Onion's (and other satirical news outlets) articles are satirizing social and cultural issues, many of the articles are triggered and are responses to current events. The goal of this paper is to outline and explore document similarity techniques of the titles to investigate the most effective ways to map satirical articles headline to a "source" article headline. Since traditional techniques such as tf-idf have difficulty grasping semantic meaning tf-idf was used as a baseline. Word Mover's Distance is one such technique that is able to capture semantic meaning better and was used in this paper to compare against the baseline tf-idf approach.

1 Introduction

Satire is a genre of literature that features irony and sarcasm and is usually aimed at critiquing social, cultural and current event issues. There are many satirical news outlets, the most prominent US outlet being The Onion. While much of The Onion's (and other satirical news outlets) articles are satirizing social and cultural issues, many of the articles are triggered by current events. What is meant by this is that some satirical articles were written in response to a recent event while many others are articles that are simply absurd, ironic or immaterial news. For example, The Onion published an article titled "Black Man Given Nation's Worst Job", after Barack Obama was elected in 2008, a case where a satirical article was written based on recent events. On the other hand, The Onion routinely publishes articles that have nothing to do with current events such as, "7-Year-Old Unable To Maintain Single Cohesive Storyline While Playing With Action Figures".

Just like many other news websites satirical news websites rely on advertisement revenues and wants as many clicks to their websites. It can be argued that headlines are even more critical to satirical news websites since they are relying on the humor of the headline instead of the subjects and contents which traditional news websites rely on. While much human creativity goes into creating these headlines any assistance to speed up the process may be useful. Currently, there are no public datasets that comprehensively labels and maps a source article - an article written about a specific event - to a satirical news article. However, there are many services offered that applies natural language processing on published articles and serves that data. Event Registry is one such service that processes articles from thousands of news sources and generates concpets, associates an article to an event and generates other useful scores such as trending scores.

Once this sort of clustering and association are accomplished, a satirical headline generator can potentially be created with actual news and current articles used as inputs. While this paper will not discuss creating models that will generate satirical headlines given a source article input, this paper will discuss, outline and evaluate the steps taken to map satirical articles headline to a "source" article headline.

2 Methods

The main source of the data was from the Event Registry service. Event Registry provides an API that can be easily queried to obtain articles from specific sources. In addition to The Onion, The New Yorker's Borowitz report (the satire section

of the New Yorker) was also queried for additional data. Event Registry stored had articles dating back to January of 2014.

One of the advantages of using the Event Registry API is the rich features they provide for a given article. As mentioned in the Introduction section, Event Registry provides details such as concepts associated with the article. Event Registry also provides a powerful feature that tags an event to the article. The Event Registry API can then also be queried with the event to obtain articles that covered the event. These related articles can be thought of as some of the candidate articles that map to the satirical articles. While this feature is useful, out of the 3,500 articles collected from Event Registry only a third of them had associated event tags. In addition to the lack of candidate articles from related events, some of the candidate articles seemed to also be simply incorrect. Because of this lack of data supplemental data was collected to add to the candidate articles list by collecting article headlines from The New York times and subsetting article headlines that have published dates in close proximity to the satirical article. Since we are looking for source articles that triggered the writing and publishing of the satirical article, time proximity will narrow down the list of candidates.

Once candidate articles were determined, document similarity algorithms were used against the article headlines. Cosine similarity with tf-idf was used as a base case scenario. The second method explored was Word Mover's Distance with word embeddings from word2vec. Word mover's distance algorithm are able to utilize word embeddings to capture document similarity even if the documents have no word in common. This was also an attractive method to experiment on since these algorithms were going to be applied to documents that are very short.

Several literatures point to the potential benefits of using word embeddings and algorithms such as Word Mover's Distance. Kustner, et. al improved upon k-nearest neighbor document classification error rates. Further De Boom, et. al, 2016 have found word embedding aggregation performing well for short texts. As such, utilizing Word Mover's distance has some potential.

2.1 Cleaning and Filtering the Data

Prior to procesing the data, some steps were needed to be taken to clean and filter the data.

While most of the Onion headlines are satirical there are sections of the websites where the satire does not appear in the headline but only in the article itself. For example, the section "American Voices" features fictional characters that speak about ongoing current events in a satirical manner. The title of these section ends up being just a description of the current event where these fictional characters speak about, so there are no inherent satirical features in them. Other sections were also removed due to similar issues.

Another issue stems from the practice of capitalizing leading letters of the words in titles. Because word embeddings vector distinguishes between the word "new" and "New", the title had to be cleaned up so words where the leading letters should be capitalized, e.g. Apple (the company) vs apple (the fruit), remained intact while making the other words lowercase. Further, in word2vec, words like San Francisco appears as "San_Francisco" so these words had to be identified. This was accomplished by identifying the named entities in the full text of the articles and then replacing and modifying each words to lower case when it was not a named entity.

3 Results and Discussion

A full evaluation of the results was not feasible given the unsupervised nature of the topic. However, even from a few examples it can be easily seen that the cosine similarity with tf-idf seems to perform much better. For example, below are results returned from the cosine similarity measure with tf-idf with an input satire article from the Borowitz Report, "Jeb Bush Resigns as George W. Bush's Brother". The article contained related articles from Event Registry.

- Jeb Bush Resigns From Board Seats, Possibly Edging Closer to Presidential Run
- No. Oklahoma Shuts Down George Mason,
- Jeb Bush Wont Attend Immigration Critics Event in Iowa
- Jeb Bush resigns from all boards
- Jeb Bush resigns from board memberships

The tf-idf seems to be picking up the key word "Jeb Bush" as well as the name "George". Overall for this example tf-idf seems to be doing quite well. On the otherhand, with Word Mover's Distance the results seems quite all over the place.

- Edith Pearlmans Honeydew
- A Simple Gift
- Soy on the Lower East Side
- The Rise of Evgeny Lebedev
- Can Writers Still Make It New?

None of these titles seem even remotely close to the satirical title from the Borowitz Report. Another example, again with cosine similarity with tf-idf with an input satire article from The Onion, "Secret Service Adds Emotional Protection Division To Safeguard Trumps Psyche", which had no related articles in Event Registry:

- Trump Seems to Side With Russia in Comments on Ukraine
- How Trump Chose His Supreme Court Nominee
- How Attorneys General Became Democrats' Bulwark Against Trump
- A Quiet Giant of Investing Weighs In on Trump

Once again tf-idf seems to be putting a lot of weight to the appearance of the word "Trump" and the results seems to capture some semblance of the satirical article. On the otherhand, with Word Mover's Distance is once again all over the place.

- Today in History: score: 1.86
- Amnesty: Up to Hanged in Syria's 'Slaughterhouse': score: 0.90
- Well, Then, Would You Like to Dance?: score: 0.88
- A Gravity-Defying Champion at Rest: score: 1.85
- Norman Rockwell's : score: 0.89

Additional examples can be found on the Jupyter notebook.

The dissappointing results seems to stem from the fact that the candidate articles were rather small. The number of candidate articles ranged from 300-500 since it was only covering related articles as well as NY Times articles that were published in very close proximity of the satirical article. Because it wasn't a large pool of candidates tf-idf was able to simply pick up the keywords found in the satirical articles that showed up in the list of candidate articles. While there were many articles covering Trump as a candidate and Trump as a president it is still a small percentage of overall article and tf-idf were able to take advantage of this. Another potential issue is that titles are created to ensure that it captures and summarizes the essence of the article and usually does not add subtlety, humor or roundabout ways especially for articles that are not opinionated. Another issue is that the word2vec embeddings were created in a time where entities that are showing up in the news headline. An updated word embedding could potentially improve the results a bit more.

4 Next Steps and Conclusion

The initial hypothesis was that Word Mover's distance would at least map similar article headlines with some interesting results. As discussed in the previous section the results were not as satisfying and cosine similarity with tf-idf seems to generally be a better approach. Applying this similar approach to the entire document will likely not improve the results by much. A better approach would likely be focusing on building word embeddings that include more entities of recent news and events.

References

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin and Kilian Q. Weinberger. 2015. *From Word Embeddings To Document Distances*. Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, Bart Dhoedt 2016. *Representation learning for very short texts using weighted word embedding aggregation*. Pattern Recognition Letters.