Historical Event Detection of New York Times Articles with DBSCAN

Tracy Stephens
Summer 2020

Abstract
The aim of this project is to build an unsupervised learning algorithm to extract event information from news headlines. I evaluate various implementations of the DBSCAN clustering algorithm on New York Times headlines using both internal cluster validity indices and proposed external cluster validity measures using author byline and subject keyword.

Background
Event detection in NLP is a complex and highly evolving area. Recently much of the research around this topic has been centered on using social media data to identify significant events in real time, like COVID-19 outbreaks[1] or natural and social crises[2]. One of the primary challenges of this area of research is that in most cases no ground truth labels for meaningful events exist. For this reason, the question of how to assess and optimize performance of the algorithms is frequently an issue.

Not all methods of event detection use clustering algorithms. However, when they do the most common one used tends to be DBSCAN and variants of it[3] because it can accommodate variable and irregular cluster shapes and sizes. In my project, I apply event detection techniques to a longer-range dataset of historical news articles. This analysis is intended to provide an organized list of significant events relating to a person, location, or topic (in this example I use a country - 'Brazil'), and corresponding news articles.

Dataset
The underlying data used is New York Times articles between January 2000 and June 2020 obtained through their free and public Archive API. Over that time period, the API has over 1.5 million articles available. However, to limit the scope of the analysis, I limit the articles to those with a particular keyword. For the baseline version of the model, these are only articles that contain the keyword 'Brazil'. In total, there are 2235 articles that contain the keyword 'Brazil' in the dataset, of which 1400 contain a unique headline and abstract.

Approach
Events can differ in size and number, so appropriate algorithms for event extraction should not rely on pre-determined parameters that assume the regular size or number. For this reason,

---

[1] Zong, Baheti, Xu, Ritter. "Extracting COVID-19 Events from Twitter."
[2] Burel, Saif, Fernandez, Alani. "On Semantics and Deep Learning for Event Detection in Crisis Situations."
[3] Capdevila, Cerquides, Nin, Torres. "Tweet-SCAN: An event discovery technique for geo-located tweets"

DBSCAN is an optimal candidate among common clustering algorithms. In contrast to other clustering algorithms, such as K-Nearest Neighbors, DBSCAN aims to partition clusters to optimize their density rather than the distance of each individual point to the center of the cluster [4]. This allows for non-globular shaped clusters. I use cosine similarity as the distance metric to evaluate the similarity between points because it tends to work better than other distance metrics in high-dimensional space[5].

I begin with a baseline model where each headline is represented by its average embedding. For the next iteration, I alter the headline representation by taking into account word frequency in weighting the individual embedding vectors. For the third and final iteration, I remove the first principal component from the weighted average embedding. I evaluate each of these methods across the DBSCAN epsilon parameter with both internal and external cluster validity indices to identify the optimal model iteration and epsilon value.

Headline Embeddings

To summarize the specific characteristics of each headline, I create vector representations using three sentence embedding techniques applied to pre-trained GloVe embeddings of 300 dimensions[6]. A common baseline for these techniques is the simple average of the individual word embeddings, not accounting for word frequency or common components, which I use as my baseline model[7]. With this method, extremely common words such as 'the' and 'and' carry the same influence on the overall headline embedding as uncommon words. But in reality, the uncommon words in a headline likely include more information about its meaning than the common words[8].

To improve upon the baseline version of the model, I use a weighting scheme based on word frequency[9]. More specifically, words that occur more often in the GloVe corpus are given a lower weight than words that occur less often. The final sentence embedding with this method is the sumproduct of these weights and the individual word embeddings.

After accounting for the frequency of individual words, there still remains common elements within the headline embeddings. To remove these common components and potentially improve the model, I strip the first principal component from all the headline embeddings. The result is Smooth Inverse Frequency (or SIF) embeddings[10]. The aim of the extra step of removing the

[4] Lulli, Dell'Amico, Michiardi, Ricci1. "NG-DBSCAN: Scalable Density-Based Clustering for Arbitrary Data."
[5] Mihajlovic, Xiong. "Finding the most similar textual documents using Case-Based Reasoning."
[6] https://nlp.stanford.edu/projects/glove/
[7] Ruckle, Eger†, Peyrard, Gurevych. "Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations."
[8] Piersman, Yves. "Comparing Sentence Similarity Methods."
[9] Arora, Liang, Ma. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings."
[10] Arora, Liang, Ma. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings."

first principal component is the reduce the influence of characteristics common to the entire corpus[11].

Performance Metrics

Evaluating performance of unsupervised algorithms is not as straightforward as it is for supervised algorithms. Metrics for evaluating clustering algorithms are commonly referred to as "Cluster Validity Indices"[12]. These metrics generally fall into one of two categories: those that require an external ground-truth measure of similarity (External Cluster Validity Indices), such as class labels, and those that do not (Internal Cluster Validity Indices). The relative performance of each clustering algorithm differed substantially depending on the metric applied.

Internal Cluster Validity Indices

The Silhouette Index evaluates cluster validity based on a comparison between the average distance of the point to all other points in its cluster to the average distance of the point to each point outside the cluster. The Calinski Harabasz Index compares the average sum of squares within each cluster to the average sum of squares between each cluster. For both of these measures, higher values indicate a more optimal partition of the dataset.

Figure 1: Silhouette Score

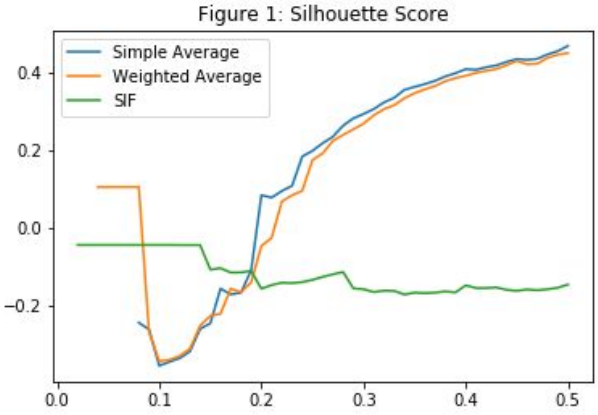
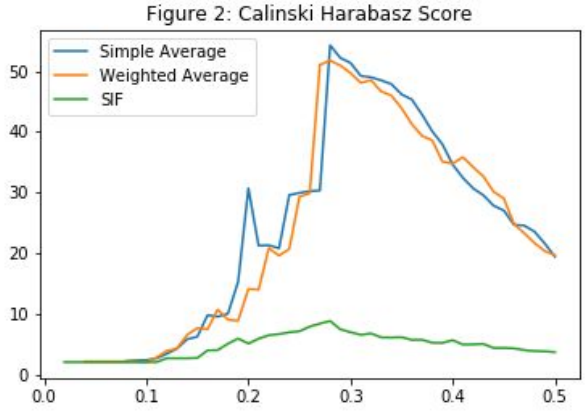


Figure 2: Calinski Harabasz Score



A drawback of both the Silhouette Index and the Calinski Harasz Index are designed for convex clusters. Since DBSCAN does not produce convex clusters, it makes more sense to use a density-based cluster validity measure[13]. In other words, because DBSCAN aims to achieve the highest level of intra-cluster density, rather than the highest level of closeness to a central point, cluster validity indices that account for density are more appropriate than those that only account for the intra-cluster distance between all points in a cluster. Two examples of density-based validity indices are the SD Index and S-Dbw Score.

The SD index compares the variance of each cluster to the variance of the dataset as a whole. The S-Dbw Score compares cluster compactness, measured by intra-cluster variance, vs cluster separation, measured by inter-cluster density. In contrast to the Silhouette Index and the

[11] Voleti, Liss, Berisha. "Investigating the Effects of Word Substitution Errors on Sentence Embeddings."
[12] Liu et all. "Understanding of Internal Clustering Validation Measures."
[13] Legany et all. "Cluster Validity Measurement Techniques."

Calinski Harabasz Index, more valid partitions of the dataset with these metrics have lower SD and S-Dbw scores.

As Figure 5 illustrates, of the two density-based internal cluster validity metrics, the S-Dbw score proved to prove more clear differentiations between the three versions of the model.

Figure 5: Internal Validity Metrics by Epsilon

	Silhouette Score			**Calinski Harabasz Score**			**SD**			**S-Dbw**		
Epsilon	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF
0.05		0.10	-0.04		2.17	2.06		2.15	2.06		0.37	0.20
0.1	-0.36	-0.34	-0.04	2.38	2.02	2.06	1.68	1.75	2.06	0.45	0.44	0.20
0.15	-0.25	-0.23	-0.11	6.25	7.66	2.75	1.49	1.52	1.90	0.59	0.58	0.33
0.2	0.08	-0.05	-0.16	30.69	14.12	5.12	1.75	1.69	1.59	0.74	0.63	0.47
0.25	0.20	0.17	-0.13	29.93	29.31	7.15	1.70	1.84	1.75	0.78	0.81	0.53
0.3	0.29	0.27	-0.16	51.34	49.68	6.98	2.12	2.15	1.88	1.14	1.14	0.55
0.35	0.36	0.35	-0.17	46.15	43.87	6.16	2.05	2.10	1.96	1.22	1.22	0.59
0.4	0.41	0.39	-0.15	34.58	34.78	5.67	2.08	2.24	1.91	1.33	1.32	0.64
0.45	0.43	0.43	-0.16	27.04	29.02	4.40	2.07	2.23	2.23	1.31	1.29	0.65
0.5	0.47	0.45	-0.15	19.45	19.68	3.71	2.10	2.27	1.91	1.35	1.40	0.68

External Validation Metrics

External Validity Metrics are measures that match cluster partitions to external sources information, like class labels. Since there is no ground truth partition of historical articles into meaningful historical events, I needed to find a suitable proxy in order to calculate external validity. This method could be called ‘semi-supervised’[14] in that labels are used, but they are known to be weak and noisy. I explore two options that were available in the dataset: byline and subject keyword. Within each metric, I analyze homogeneity (if the elements within each cluster come from the same class), completeness (if the elements within each class fall in the same

[14] Ruder, Sebastan. "An overview of proxy-label approaches for semi-supervised learning."

cluster), and V-measure Score (the harmonic mean between homogeneity and completeness), which is essentially equivalent to Normalized Mutual Information[15].

Partitioning articles by byline, or author, is one proxy. Of course using byline as a proxy for ground truth operates on the assumption that the same reporters will cover the same events. It also assumes that the model is able to separate similarity based on content from similarity based on some other author-dependent characteristic, like writing style or word choice. Theoretically, if the same authors use similar words but report on different events, the clustering algorithm could pick up similarities in their writing styles, rather than the content. In this case using byline as a proxy would make the clusters look valid when they are actually not.

Figure 5: External Validity Metrics vs. Author

	Homogeneity Score			**Completeness Score**			**V-Measure Score**		
Epsilon	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF
0.05	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
0.1	0.58	0.57	0.00	0.72	0.65	1.00	0.64	0.61	0.00
0.15	0.10	0.12	0.18	0.42	0.43	0.48	0.16	0.19	0.26
0.2	0.01	0.01	0.42	0.34	0.35	0.48	0.01	0.03	0.45
0.25	0.00	0.00	0.41	0.37	0.31	0.55	0.00	0.01	0.47
0.3	0.00	0.00	0.35	1.00	1.00	0.45	0.00	0.00	0.40
0.35	0.00	0.00	0.44	1.00	1.00	0.50	0.00	0.00	0.47
0.4	0.00	0.00	0.30	1.00	1.00	0.53	0.00	0.00	0.39
0.45	0.00	0.00	0.26	1.00	1.00	0.45	0.00	0.00	0.33
0.5	0.00	0.00	0.17	1.00	1.00	0.42	0.00	0.00	0.25

Using byline to evaluate cluster validity has several flaws. For example, authors tend to use unique, distinguishable writing styles and word choice[16]. These differences could have a significant impact on the initial cluster partition, therefore making this method of evaluation redundant. The fact that the texts being analyzed are very short, with the average headline in the dataset being only 8.6 words, may reduce the impact of author-specific distinctions in writing style[17].

Subject keywords proves to be a more promising ground-truth proxy. Keywords are tags manually added to each article at publication, and they are conveniently available in the Archive API. Each keyword is given a type, such as ‘geolocation’, ‘person’, ‘organization’, and ‘subject’.

[15] Rosenberg and Hirschberg. "V-Measure: A conditional entropy-based external cluster evaluation measure."
[16] Qian et all. "Deep Learning based Authorship Identification."
[17] Sanderson, Conrad. "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation."

Most articles contain at least one, and likely more than one keyword. Therefore this method of creating proxy class labels comes with an added layer of complexity in that the classes are not mutually exclusive. As an example, the four most common subject keywords are ‘Politics and Government,’ ‘Economic Conditions and Trends,’ ‘Corruption (Institutional),’ and ‘Elections.’

In order to calculate external validity across multiple subjects, I first calculate the validity score for each of the top 50 subjects individually as a binary class, and then take the weighted average of each, with the weights being proportional to the frequency that the subject keyword occurs.

Figure 6: External Validity Metrics vs. Byline

	Homogeneity Score			**Completeness Score**			**V-Measure Score**		
Epsilon	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF	Baseline	Weighted Average	SIF
0.05	1.00	0.86	0.86	1.00	1.00	1.00	1.00	0.86	0.86
0.1	0.48	0.58	0.86	0.11	0.09	1.00	0.18	0.15	0.86
0.15	0.05	0.06	0.57	0.03	0.04	0.18	0.04	0.04	0.19
0.2	0.00	0.01	0.29	0.02	0.01	0.06	0.00	0.01	0.10
0.25	0.00	0.00	0.26	0.02	0.02	0.05	0.00	0.00	0.08
0.3	0.00	0.00	0.18	1.00	1.00	0.04	0.00	0.00	0.07
0.35	0.00	0.00	0.22	1.00	1.00	0.04	0.00	0.00	0.06
0.4	0.00	0.00	0.14	1.00	1.00	0.03	0.00	0.00	0.05
0.45	0.00	0.00	0.12	1.00	1.00	0.03	0.00	0.00	0.05
0.5	0.00	0.00	0.05	1.00	1.00	0.02	0.00	0.00	0.03

Results

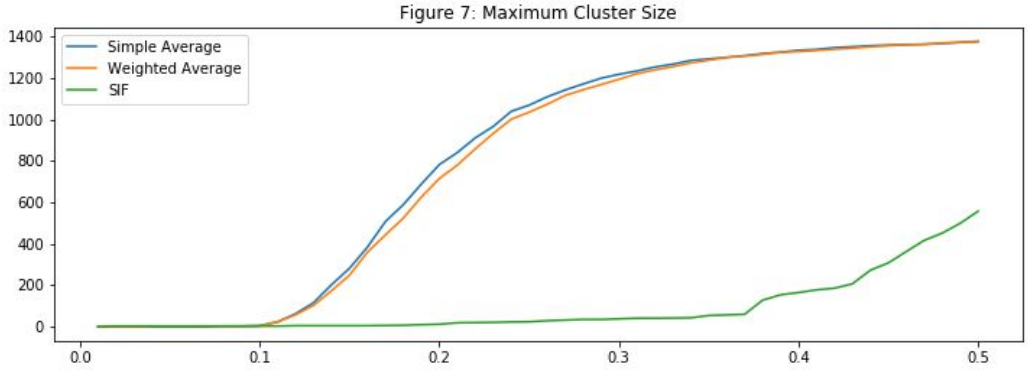
Of the internal cluster validity indices, since the density-based metrics are more appropriate for this particular application, it makes sense to place more importance on those in evaluating performance. SD score does not give clear differentiation in performance between the three versions of the model. However, S-Dbw does give a clear differentiation in performance across a wide range of values of epsilon. It is clear that the SIF version of the model outperforms both the baseline and the frequency-weighted version on this metric.

The SIF-based model outperforms on the external cluster validity measures based on both byline and subject keyword as well in terms of homogeneity and V-measure score. However, in terms of completeness, none of the three versions of the model are able to achieve decent scores with more than one cluster.

Overall, the cluster partitions proved successful but still left room for improvement. The size of the clusters was one weakness. As would be expected, the size of the largest cluster tended to increase with higher values of epsilon. However, the extent to which this occurred in most

iterations was too much for the model to be practical. In the simple average and weighted average versions of the model suffered from this problem much more than the SIF version, as the largest cluster would contain more than half the dataset when epsilon > 0.2. In the SIF version, the max cluster size remains relatively constant until epsilon reaches levels near 0.4.

Figure 7: Maximum Cluster Size



Clusters also tended to center around niche events with specific word choice. For example, one cluster includes these three articles from 2002:

> *A Leftist Surges in Brazil's Turbulent Presidential Election*
> *LEFTIST CANDIDATE TAKES A FIRM LEAD IN BRAZIL ELECTION*
> *Relations With U.S. a Challenge for Leftist Elected in Brazil*

It is clear that these articles were clustered together because they include the uncommon word “Leftist.” This cluster is likely homogenous, but not complete, as these were not the only New York Times articles written about the 2002 Brazil elections. This example to some extent undermines the use of bylines as a ground-truth proxy in the above section.

Conclusions

This project served to show both the capabilities and limitations of context-free word embedding. The SIF method was able to achieve decent performance across the two most appropriate metrics, S-Dbw and external validity with subject keyword.

The frequency-weighted average and simple average embeddings performed similarly across the metrics, which suggests that accounting for word frequency in headlines does little to improve performance. However, the added transformation from the SIF method, which involved stripping a common component from the headline embeddings did improve performance both on S-Dbw and on the external metrics. One potential reason for this could be that headlines contain similarities in word choice that are not accounted for by simply using word frequency, but are represented in the first principal component.

The limitations of context-free word embeddings are shown here as well. None of the three versions of the model were able to achieve high values of completeness in the external validation metrics. This could be a result of the clusters being too dependent on specific words (such as ‘Leftist’ in the example above). For further research, comparing these methods to more complex headline embedding methodologies and clustering methods could likely improve performance on the metrics shown here.

Appendix

Code - https://github.com/tracy-stephens/headlines

Code References

- SIF: https://github.com/PrincetonML/SIF
- S_Dbw: https://github.com/fanfanda/S_Dbw
- GloVe: https://github.com/datasci-w266/2020-summer-main

Paper References

Zong, Baheti, Xu, Ritter. “Extracting COVID-19 Events from Twitter.” *Ohio State University*. Web. <https://arxiv.org/pdf/2006.02567.pdf>

Burel, Saif, Fernandez, Alani. “On Semantics and Deep Learning for Event Detection in Crisis Situations.” *Stanford*. Web. <http://web.stanford.edu/~vinayc/olearning/literature_review/EventDetectonUsingCNN.pdf>

Capdevila, Cerquides, Nin, Torres. “Tweet-SCAN: An event discovery technique for geo-located tweets.”

Lulli, Dell’Amico, Michiardi, Ricci1. "NG-DBSCAN: Scalable Density-Based Clustering for Arbitrary Data." Web. http://www.vldb.org/pvldb/vol10/p157-lulli.pdf

Mihajlovic, Xiong. "Finding the most similar textual documents using Case-Based Reasoning." ETH Zurich. Web. <https://arxiv.org/pdf/1911.00262.pdf>

Ruckle, Eger†, Peyrard, Gurevych. "Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations." Ubiquitous Knowledge Processing Lab (UKP). Web. <https://arxiv.org/pdf/1803.01400.pdf>

Piersman, Yves. "Comparing Sentence Similarity Methods." NLP Town. Web. <https://nlp.town/blog/sentence-similarity/>

Arora, Liang, Ma. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings." Priceton. Web. <https://openreview.net/pdf?id=SyK00v5xx>

Voleti, Liss, Berisha. "Investigating the Effects of Word Substitution Errors on Sentence Embeddings." Arizona State University. Web. <https://arxiv.org/pdf/1811.07021.pdf>

Liu et all. "Understanding of Internal Clustering Validation Measures." 2010 IEEE International Conference on Data Mining. Web. <http://datamining.rutgers.edu/publication/internalmeasures.pdf>

Legany et all. "Cluster Validity Measurement Techniques." Web. <https://pdfs.semanticscholar.org/581c/71da74bd3baa06693cc6d0751e7c60f81bb3.pdf>

Ruder, Sebastan. "An overview of proxy-label approaches for semi-supervised learning." Web. <https://ruder.io/semi-supervised/>

Rosenberg and Hirschberg. "V-Measure: A conditional entropy-based external cluster evaluation measure." Columbia University.

Qian et all. "Deep Learning based Authorship Identification." Stanford. Web. <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760185.pdf>
Sanderson, Conrad. "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation."