MSDS453_Sec56_Natural Language Processing

Assignment 3 Report

1. Abstract

AG is a collection of more than one million news articles which have been gathered from more than 2000 news sources by ComeToMyHead, an academic news search engine, in more than one year of activity. The dataset is provided by the academic community for research purposes in data mining, information retrieval, data compression, data streaming, and any other non-commercial activity. The AG's news topic classification dataset is constructed by Xiang Zhang from the previous dataset. The goal of the second assignment is to preprocess text documents, create document vectors and classify documents based on topics.

2. Introduction

Vast availability of news needs its effective organization and retrieval. One of the ways to easily access these documents is to classify them into topics using text classification techniques. The AG's news topic classification dataset is divided into 4 classes: World, Sports, Business, and Sci/Tech. Each class has 30,000 training samples and 1,900 testing samples. The total number of training samples is 120,000 and testing 7,600. My goal throughout this project is to apply different unsupervised learning methods on the TD-IDF vectors to understand the domain of content represented by the documents, to identify common themes and to organize documents into groups. I will use cluster analysis, t-SNE multidimensional scaling, topic modeling, and biclustering in this assignment.

The sections of this paper include Literature Review which reviews similar work done by others, Methods which discusses how I will be conducting my research, Results which lays out what I have learned from this research, and Conclusions which summarizes the findings and discusses the next step.

3. Literature review

With the rapid growth of network information, the Internet has become the most popular source of data. As the textual presentation of data continues to be the most widely used form of communication, it is important to use the computing power of the machines to uncover the insights from the text. One of the first steps of exploratory analysis is to use topic modelling to identify themes or topics within a corpus of many documents, or to develop or test topic modelling methods. The use of topic modelling cannot provide the complete meaning of the text but it does provide a good overview of the themes, which is hard to be obtained otherwise. In research from Lancichinetti et al. (2015), Latent Dirichlet Allocation (LDA) is considered to be the most used and state-of-the-art method. The framework can even be fully automated. An inspiration for this can be found in the article by Brocke et al. (2017) where they find that topic modelling can be automated and argue that the use of a good tool can easily present good results, but the method relies on the ability of people to find the right data and to interpret the results.

4. Method

For this project, I focused on five main parts. The first part was to preprocess documents and create a TF-IDF vector. Since the original dataset of 120,000 rows took a long time to finish, I extracted 1000 articles from each class, which totaled 4000 items for this project. The second part was to perform cluster analysis (K-means) and hierarchical clustering with documents as objects. I specifically used sklearn KMeans and sklearn Agglomerative Clustering for these. For each type of clustering, I used t-SNE for multidimensional scaling to visualize the solution in two-dimensional space. The third part was to develop a topic modeling solution using Latent Dirichlet Allocation. The fourth part was to employ Spectral Biclustering. The chosen numbers of clusters for all algorithms are four. For each method, to understand the results further, I

summarized the top ten words, predicted possible topics, and counted the number of documents of each cluster. The last part was to develop an ontology that is broad enough to provide a framework for further research and analysis. The figures and result tables can be found in the appendix.

5. Result

Comparing K-Means clustering and hierarchical clustering, we can see that K-Means' top words are more descriptive of the classes, and we can easily tell which topic they belong to. For hierarchical clustering, though I already use up to twenty words, it is not very clear what the main topic is about for each class. There is also not a clear distinction of how each cluster differs from the others. Besides, the distribution of documents using K-Means seems to be more even than that of hierarchical clustering. In the multidimensional scaling map for K-Means, we can find clusters 0, 1, and 2 centering around their centroids while cluster 3 scattering across the map. In the one for hierarchical clustering, clusters 0 and 3 are dominant while cluster 2 is hardly visible. Overall, K-Means clustering provides more clear-cut and interpretable results between the two methods.

The Latent Dirichlet Allocation method results in a good overview of the corpus. Looking at the top ten words with the highest probability in each class, we can certainly recognize the main theme. The word list is also quite similar to that of K-Means clustering. The Spectral Biclustering method does give a good sense of what the corpus is about. However, from one bicluster to another, there is still not a clear cut. Some biclusters appear to focus on certain themes while others can be a mixture of several different topics. This method does a better job at extracting meaningful words and topics than hierarchical clustering but does not generalize as well as K-Means clustering and Latent Dirichlet Allocation.

6. Conclusion

From this project, I learned a lot about topic modelling and how to use different unsupervised learning methods for this purpose. Data normalization and vectorization was very essential to the whole process. Minor changes made at this level can affect everything else, so the choices being made must be thoughtful. All topic modelling methods provide some insights into the corpus. For this project specifically, K-Means and Latent Dirichlet Allocation work better than hierarchical clustering and spectral biclustering in terms of extracting meaningful words and topics. However, these methods are also subjective to personal judgement and are only as good as the data collected. Overall, this assignment got me thinking out of the box of what I could accomplish from a large collection of documents.

References

Brocke, J., Muller, O., & Debortoli, S. (2017). *Class Notes: Power of Text-mining in BPM*. BPTrends. https://www.bptrends.com/class-notes-power-of-text-mining-in-bpm-2/

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Kording, K., & Amaral, L. A. N. (2015, January 29). *High-Reproducibility and High-Accuracy Method for Automated Topic Classification*. Physical Review X. https://journals.aps.org/prx/abstract/10.1103/PhysRevX.5.011007

Appendix

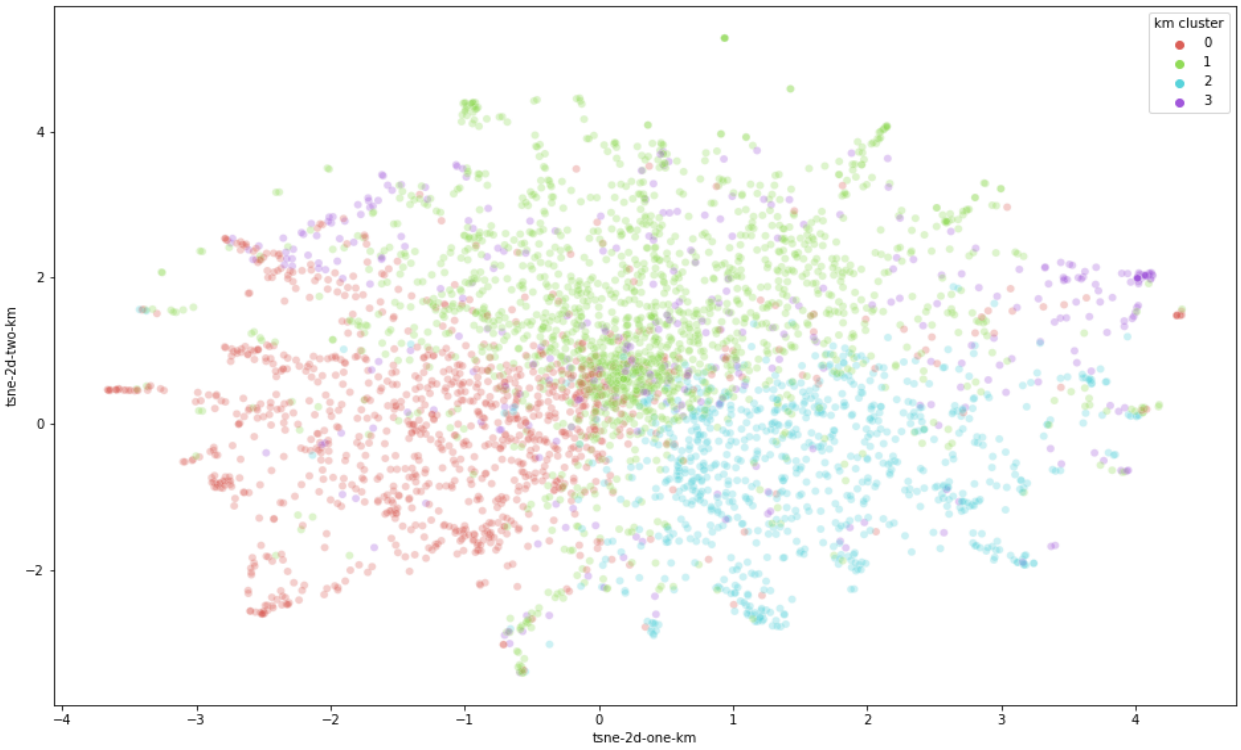

Figure 1. Multidimensional scaling map for K-Means clustering method

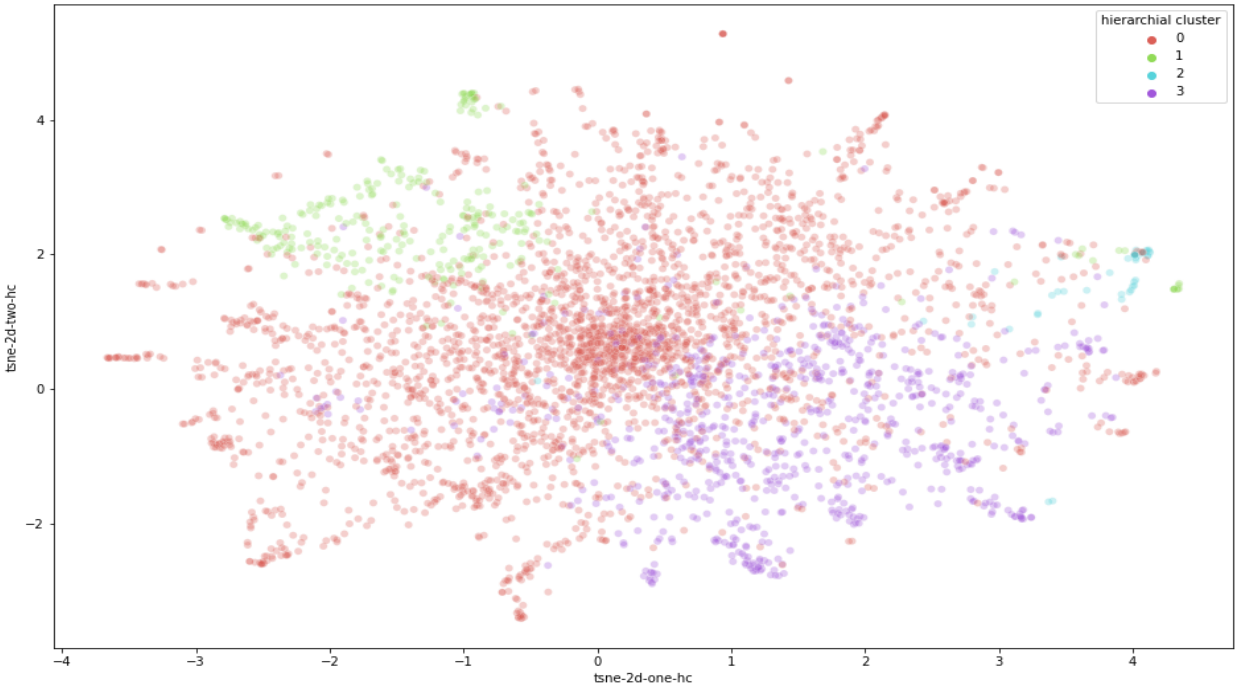

Figure 2. Multidimensional scaling map for hierarchical clustering method

Class	Top 10 Words	Possible Topic	Number of Documents
0	wednesday, killed, united, friday, iraq, minister, people, president, reuters, said	World	1086
1	business,service, internet, computer, said, corp, software, microsoft, million, company	Sci/Tech	1752
2	world, games, coach, win, victory, cup, team, night, season, game	Sports	801
3	corp, street, friday, tuesday, prices, oil, stocks, reuters, york, new	Business	361

Table 1. K-Means clustering results

Class	Top 20 Words	Possible Topic	Number of Documents
0	minister,time,years,sunday,night,season,company,world,friday,president,tuesday,people,million,thursday,york,monday,wednesday,reuters,new,said	World or Business	1945
1	game,monday,security,year,announced,thursday,software,quot,united,friday,today,wednesday, yesterday,tuesday,york,world,company,reuters,	Sci/Tech or Sports or Business	1926

	said,new		
2	announced,airline,service,yesterday,step,company,said,market,years,maker,new,wants,file,scientists,atlanta,airways,round,chief,executive,wireless	Sci/Tech or Business	92
3	ruling,set,food,free,mission,early,movies,flight,following,hurricane,russian,security,giant,satellite,new,international,users,station,nasa,space	Business or World or Sci/Tech	37

Table 2. Hierarchical clustering results

Class	Top 10 Words	Possible Topic
0	new, york, reuters, said, percent, corp, shares, tuesday, wednesday, prices, company	Business
1	software, company, microsoft, computer, new, internet, service, online, technology, million	Sci/Tech
2	season, team, game, night, win, victory, cup, world, final, games	Sports
3	said, people, reuters, minister, iraq, president, oil, killed, government, officials	World

Table 3. Latent Dirichlet Allocation result

Bicluster	Numbers of Docs and Words	Top 10 Words	Possible Topic
0	610 documents, 126 words	militant, strip, wounded, arafat, blair, fallujah, gaza, israeli, usled, yasser	World
1	1567 documents, 682 words	exports, airways, singapore, century, supply, winter, drugs, flight, euro, ban	Business or World
2	1800 documents, 682 words	browser, opensource, firefox, net, sony, hard, product, ipod, warner, application	Sci/Tech
3	1800 documents, 135 words	game, season, san, night, victory, network, sports, points, league, games	Sports or Sci/Tech
4	1760 documents, 60 words	quote, google, wireless, windows, mobile, ibm, music, version, profile, phone	Sci/Tech

Table 4. Spectral biclustering results

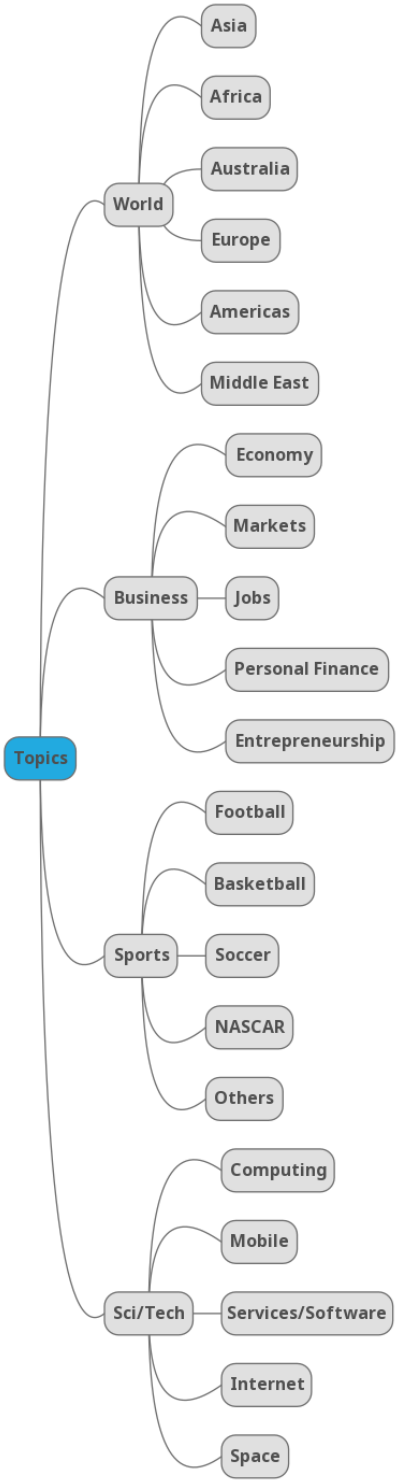

Figure 3. Ontology