Final Paper:

RSS News Feed Identification

University of Amsterdam

Graduate School of Communication Research

Research Master

Course: Big Data and Automated Content Analysis

Lecturer: Prof. Damian Trilling

Student: Janice Butler

Student number: 12356093

31.May 2020

Wordcount: 4496

Introduction

The main aim of the project and this resulting paper is to be able to automatically identify which, of a selection of articles, is based on essentially the same news story. News articles which are published on a common topic, for example "USA pulls out of WHO" can be reported on in different ways. How exactly a news story is presented - for instance, which aspects of the story are highlighted in a positive or perhaps negative sense - have the ability to influence how the consumer feels about the topic.

This is clearly shown by Tian and Stewart (2005), who write: "News reporting is more like telling a story about the world than presenting information, even though there are factual elements in the stories".

In this project news stories from around 50 different online sources from across a wide political spectrum as identified by https://mediabiasfactcheck.com/ are analysed and automatically clustered with a view to identifying groups of articles reporting the same underlying story.

Framing Theory

The conceptual communication framework underlying the analysis is that of framing theory. Framing theory proposes, in a journalistic sense, that the way a news story is presented to an audience influences ("frames") the choices that are made by people when processing the information presented to them (Davie, n.d.). In order to conduct (automated) analysis of large numbers of articles, there is thus a strong motivation – in a first step – to achieving an automatized mechanism for detection of common stories. The applications of this non-trivial task would be found in the analysis of media-bias, the detection of fake news or in the role of news-agency copy and the degree of its proliferation.

Semetko and Valkenburg (2000) explain that there are two approaches when analysing frames: the inductive or the deductive approach. The inductive approach begins by analysing a news story with loosely defined presuppositions, the aim being to reveal possible frames. The deductive approach is to clearly define frames as variables, meaning that frames that are not specified a priori can be overlooked. The paper presented here analyses frames according to the more labour-intensive inductive approach, only the dates on which the data is collected are identified.

A frame that is particularly relevant in this context is the "human impact frame", as defined by Semetko and Valkenburg (2000). They hypothesize that in order to keep the audience interested in news, an effort is made to personalize and emotionalize the story. The authors argue that this occurs due to a highly competitive market. Since the paper was published in 2000 it can only be said that the news market has grown even more competitive, especially considering that social media has risen as an alternative source of information. This is also reflected by the increased analysis regarding framing in context of social media. Ahmed, Cho and Jaidka (2018) being only one example.

It only follows that the emotionalization of the news due to competitiveness has also potentially been impacted. It can be very important to analyse news stories from a sentiment analysis perspective.

Sentiment Analysis

A way of framing a particular issue can be through evoking a particular sentiment. Research into the processing and analysis that aims to find meaning, underlying themes and attitudes behind texts is called sentiment analysis. Sentiment analysis uses various text analysis techniques to study the emotion that an author attempts to convey. Often machine learning is used to classify if sentiment is positive, negative or neutral (i.e. the polarity) and the strength of the sentiment expressed (i.e. the valence).

That the media has been able to influence opinions is inherent in the field of communication science and the influence of the media has been especially been highlighted by the framing theory. That how the media constructs a story is influenced also by the political alignment of the journalist and the news organisation is made clear from the example of DellaVigna and Kaplan (2007). They were able to show that through the entry of Fox News (politically more right on the spectrum, as identified by Media Bias / Fact Check) into the media landscape that the politically wide-ranging choice of voting has been impacted. Fox News had successfully convinced somewhere between 3 and 28% of its viewers to vote for Republicans between the analysed years of 1996 and 2000.

Diversity of media is important to analyse, as Garcia Pires (2014) writes, it ensures free enterprise and by extension also freedom and democracy. By extension the convergence of opinion as an issue also depends on the way it is presented. Frames are able to simplify a complex topic by lending greater weight to certain arguments over others and thus explaining why an issue could give rise to problems, who is responsible and what mitigations may exist. Through this, framing can provide points of reference and meaning between key actors. Successful framing is able to link two concepts making people understand the connection and the underlying message (Nisbet & Scheufele, 2009).

Given the theoretical background and problems stated above, the following research question is asked:

RQ: Is it possible to identify "identical" news stories emanating from disparate media-sources through automated content analysis?

Analytic Strategy

A large number of articles from more than 60 different RSS feed sources on the left-right spectrum of politics were accessed and parsed using the Python package Universal Feed Parser. It is important to note here that the data obtained was not homogenous. Not all of the feed articles initially tested contained useful content. Thus, the corpus analysed was reduced down to articles that did contain content or a significant summary text. This step was intended as a simplification, as of course one could have followed the link of the RSS feeds back to the original article and then collected the data

separately. Additionally, articles with a publication date before the start of the project were discarded. For instance, the Independent had several articles dating back further than other feeds. To answer the research question, different articles discussing a story would be analysed. If for instance all the older Independent articles had been kept, this would have led to a lot of noise and no comparison would have been possible for those articles, due to the limited amount of articles dating this far back. Another specific case that had to be addressed in data cleaning was the presence of New York Times Briefings (of the 8'000 Times articles 28 were daily briefings). These are daily summaries where all important news of the day is presented at once. This data had to be excluded since a topic separation at article level was being pursued. Article containing multiple stories dilute any distinction between topics, leading to extra noise.

Data from the RSS feeds had to be collected and archived continuously throughout the project to avoid any gaps in the timeline. Nonetheless, with this simplified approach enough data was able to be gathered. Around 11700 articles were ultimately collected and analysed. Effectively a semi-automatic RSS-aggregator was implemented for the data-gathering part of the project.

In persisting the collected data to disk, an initial attempt using json serialization was abandoned, since on re-reading the json data was deserialized into objects in a different data format requiring additional code to traverse the object structure. Instead the, to avoid losing structure, pickle was used rendering the data into binary format therefore being able to preserve its structure as well as being more efficient and compact. One slight disadvantage is, of course, that the binary file are not in human-readable form.

The articles after collection underwent pre-processing steps before they were further procesing. Pre-processing involved several steps. First articles were cleansed of stop-words. A collection of stop-words was loaded in from the NLTK (Natural Language Toolkit) Python package as described in Chapter 7: Automated Content Analysis of the course book and in Brid, Loper and Klein (2009). Though the predefined stop words did take out unnecessary data, another list of custom stop words and phrases was compiled to get rid of some phrases with little value such as "view entire post" and "like story share", which has nothing to do with the story content of the feed articles. In a next step both the content of the articles and the titles with teasers were in the process of lemmatization (and stemming) reduced down to the stem of the words. Thus elect, elects, elected, electing and election for instance were reduced to elect as all of these words have essentially the same meaning as the root word "elect". Lemmatization was done by importing the Snowball Stemmer that originates, as the stop-word package, from the NLTK Python package discussed in Chapter 7 of the course book and in Brid, Loper and Klein (2009).

A final step in the pre-processing was the normalisation of synonyms and acronyms. Words like European Union and EU or NATO and North Atlantic Treaty Organisation have the same meaning and should not be separately interpreted by the program.

After these aforementioned pre-processing steps, the news stories were then clustered. News stories talking about the same topics were identified and grouped into clusters and displayed using various visualisation techniques such as box plots, bar plots and cluster plots visualising the evolution of the subtopics over time. Clustering is able to provide the basis for many other types of comparative analysis.

Topic modelling is running statistics over content to pick out groups of words which can be collected into topics. Documents are represented as random mixtures over latent topics, a topic in this sense is characterized by a distribution over words. A topic, in the context of topic modelling, is a list of words occurring in statistically significant methods (Jelodar, Wang, Yuan, Feng, Jiang, Li and Zhao, 2019).

The different algorithms using for deriving topics are Latent Semantic Analysis (LSA), probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Algorithms (LDA). As Blei, Ng and Jordan (2003) write, both pLSA and LSA are based on a Bag of Words (BOW) approach, neglecting the order of the words in a given text. Due to this limitation and considering that LDA accounts for the exchangeability of words and documents the choice was made for using LDA in topic modelling. As there are a significant amount of documents which are analysed a tool named LDAvis is used to interactively visualise the many dimensions. The original R version has been adapted to Python, making it applicable for usage in this project. Terms within the topics are ranked according to relevance. One can interactively select a topic revealing the most relevant terms. Additionally, by moving the mouse over a term the conditional distribution in the topics is visualised (Sievert & Shirley, 2014).

As a second and alternative method of clustering of the documents the soft cosine similarity is used as a measure of similarity irrespective of length of the documents or frequency of single words. Cosine similarity uses the term frequency – inverse document frequency (TF-IDF) which down weights frequently occurring words in a document. Mathematically speaking, cosine determines the angle between two vectors. In this case the vectors describe words. The cosine similarity measure (from 0 to 1) is especially relevant, since it measures the distance irrespective of the length of an article (document) the words appear in, thus not artificially enhancing the importance of repetitiveness. With the high cosine similarity of 1, as Boumans, Trilling, Vliegenhart and Boomgaarden (2018) write, two documents are thought to be identical (though the words are not necessarily in the right order) and with the low score of 0 the documents use completely different words. For the purposes of this project the soft cosine similarity measure was taken, as it is able to further consider the semantic meaning of a word and therefore words such as, “hi” and “hello”, though not exactly alike, are prescribed a similar meaning in vector space and are thus closer together. As a first step, it does not use a bag of words approach like in the cosine similarity but word embeddings converting words into high-dimensional vectors. This facilitates clustering of articles allowing for diverging vocabularies or even misspellings

in those articles, but with recognition of similarities due to the vector representation of an independent and very large model of words (the word embedding model).

In order to compare the measure of soft cosine similarity it was decided to include two alternative methods, with the same aim of discovering groups of cluster topics of the articles, namely Levenshtein distance. This distance is commonly used in the detection of plagiarism. The distance looks at the amounts of single-character edits needed to be undertaken in order to change one string of text into another. This can be a deletion, substitution or addition of a character. The more edits needed, the more different the texts are from each other and the higher the Levenshtein distance (Boumans et al., 2018).

The measures referred to extract the topics from the articles. As a complement to this approach a fuzzy string comparison was also implemented. For this purpose the fuzzywuzzy python package was used. Originally the code was created to find a way to identify if, for instance, two labels of a sporting or event ticket are for the same event. This approach used the Levenshtein distance, as described previously, additionally to the exclusion of "common" words (this is different than the previous exclusion of stop words as the common words are not stop words per se but occur nonetheless often enough to provide no additional value) (SeatGeek, 2011).

In order to visualise the high dimension data in a meaningful way, it had to undergo dimension reduction. This was done using principal component analysis (PCA). As, due to the high dimensionality of the data, visualisation cannot otherwise be understood properly. PCA aims at getting meaningful geometric coordinates extracting a low dimensional set of features through maximising the informational content and minimising the loss of information, by taking a projection of irrelevant dimensions (Analytics, 2016).

The visualization of group separation was on the one hand achieved with dimension reduction via PCA. There are several other dimension reduction techniques though, which can also be used when aiming to visualise multi-dimensional data. As these techniques might lead to different outcomes it is planned to also test these approaches, which can in successive steps be applied in combination with oneanother. The first alternative is to use truncated singular value decomposition (SVD) the other alternative is TSNE. Truncated SVD uses as a basis latent semantic analysis as discussed above. As opposed to the initial dimensionality reduction technique PCA, this measure does not attempt to center the data before the singular value decomposition is calculated. This also has the additional value that the technique works also with scipy.sparse matrices efficiently, where most values in the matrix are rendered to a null value (Sklearn Truncated SVD, n.d.). The other alternative to PCA implemented is the t-Distributed Stochastic Neighbour Embedding (t-SNE). This technique was developed by Laurens van der Maaten and is a dimension reduction technique especially adapted for high-dimensional datasets (van der Maaten, n.d.). As PCA aims at maximizing variance this can lead to problems when attempting to visualise data with large differences. T-SNE aims at preserving small pairwise

differences in contrast to PCA preserving large ones. The algorithm behind t-SNE "calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space. It then tries to optimize these two similarity measures using a cost function" (Violante, 2018).

To further accentuate the optical separation, spectral clustering was applied to colorize the "bubbles" in the 3d plot Colouring of the topic's groups used the spectral clustering package from sklearn. The expected number of clusters can be specified in advance. Regions are defined by choosing cuts in the graph, the gradient ratio along the cut is minimised thus the volume of the region weight of the edges is cut small compared to the weight of the edges inside each cluster (Clustering: 2.3.5. Spectral Clustering, n.d.).

For each of the individual subtopics a sentiment analysis using VADER (Valence Aware Dictionary and sEntiment Reasoner) was conducted in order to evaluate each article on the balance of coverage. VADER is a lexicon based analysis tool, it has been specifically created with social media in mind but has also previously been used to analyse movie reviews and New York Times editorials. Most importantly it doesn't require any training data thus being directly applicable it has also been originally trained using the human gold standard with people from Mechanical Turk evaluating the ratings. Additionally, rather than just giving a positivity score, it gives a range of how positive or negative a given string is (Pandey, 2018). A sentiment analysis is further able to provide a comparison between the authors of the texts and the media outlets. Sentiment, being a range, then was displayed on a spectrum on single articles according to media outlet. The project implemented not only the sentiment analysis, but also a 4-dimensionalal (3 dimensions plus color) visualization of the sentiment.

Conclusion

Conclusions are drawn on the balance of coverage across the multiple online news sources. The results accompanied with a substantial number of comments have been written into the Jupyter Notebook. Thus, this section will just discuss the main results, draws conclusions upon them and will go into further detail of what could have been done differently or additionally in this project including recommendations for future research.

The articles that are analysed are current ones, collected during the time of the project. It should be noted that the timeframe of the project is in the middle of the worldwide corona virus pandemic. This has the effect that very many articles, currently being written (almost all, it seems) have some relation to the corona virus, thus being harder to distinguish from each other than they would be if articles had been analysed under "normal" news conditions. The feature of "timeliness" is also a staple of why an article would be deemed newsworthy (Galtung & Ruge, 1965). Therefore, a lot of the articles feature updates on the situation regarding corona virus. Regarding the distinction of clusters, it is more a case of distinguishing a topic of corona virus in combination with "economics" or in combination with

"health", therefore making it much harder for various approaches to find any clear lines between groups of topics. The results invariably show a very large "corona" cluster in the 3d plots – however, the separated clusters do have common themes such as business, corona virus stories from SE Asia or sport. This demonstrates that the techniques being employed do work to some significant degree.

The sentiment analysis and visualization does seem to produce excellent and reliable results. The 3d visualization is quite fascinating in that plotting the positive, negative and neutral sentiment components on orthogonal axes always results in a triangular plane of data. The colorization according to the compound sentiment gives rise to no surprises (it is a function of the other sentiment values) but does help in interpreting the results.

A simple parameter that could be changed are the upper and lower N-gram bounds. At the moment the sweet spot for news articles seems to be the tri-gram. Septa-gram and octa-grams were also experimented with, producing surprisingly clear-cut results, but producing groupings of somewhat strange stories away from mainstream news whilst not detecting the main stories of the day. It is suspected that these were stories taken 1:1 from a single source and syndicated to other news outlets with no individualized editorial input.

Further Application

Already, it can be surmised that some components of this research could find application elsewhere. Using ranking to identify the "interestingness" of articles as described by Pon, Cárdenas, Buttler and Critchlow (2011) and Karimi, Jannach, and Jugovac (2018). Karimi and his colleagues explain that the cosine similarity can be subsequently used to compare the profile to a potentially relevant news article.

The Levenshtein Distance measure could be a useful tool for instance in chatbot applications to support helpdesks. Assuming the pre-condition that you have lots of FAQs with set answers as a type of knowledgebase, customers or agents could enter questions as clear text to a chatbot, behind the scenes the best match of the entered question to one of the FAQs is searched for. The answer is the ultimately supplied as recommendation of the best fitting FAQ.

Another application of the matching would be in detecting plagiarism. The aforementioned matching would be used to detect similar passages in the analysed text.

Possible Improvements

To improve the results of this project there are two main areas to approach: tuning of the existing implementation (including enhanced data cleansing) and implementing additional/alternative algorithms for key sections of the processing.

As regards tuning, there are many variables, (adaptable) parameters and techniques used in the project. The next phase would be to go through the same analysis with a limited amount of data (i.e. articles

from a limited window in time) in order to tune the "system" and increase the accuracy of the analysis. This could be done with historical documents, as news (as demonstrated by the covid-19 pandemic), due to its high adaptiveness and low saturation rate, causes topics to be digested by the reader very quickly causing a constant demand of newsworthy articles to be produced that skew the representation (and importance) of stories.

It would seem a very good idea to unify similar, but not yet identical topics to a single, most promising topic before proceeding with subsequent processing steps. This could in fact be automated easily with the use of Levenshtein Distance comparison, as employed elsewhere in the project.

Even though for instance "New York daily briefings" were taken out due to skewing the results, as topics are given in an overview, meshing the stories concisely together, there are still some mentions of briefings from other publications in the topic maps that were analysed. A consequential removal of all such news summaries would surely produce more clear-cut clusters.

In regard to implementing additional/alternative algorithms for key sections of the processing, there are several areas to address. The topic modelling in this analysis was implemented with LDA, however as Jelodar et al. (2019) recount, since its first introduction in 2003 several adaptions have been made to improve upon LDA. Jelodar et al. especially consider the different applications in the scientific field such as linguistics, political science or biomedicine, for which certain adaptations are better suited than others. The primary reason to take LDA and not any other adaptation in this project, was due to the matching visualisation technique of LDAvis. Future research could do with looking into the specific application of an adaptation of LDA in the application of linguistics for news articles.

Different methods of dimension reduction other than PCA are also available, such as TruncatedSVD and/or TSNE, which need to be tested additionally and potentially in combination with each other to improve the quality of optical separation of clusters.

In regard to n-grams tuning, it has been noted that the pyLDAvis widget only seems to support unigram analysis. It is thought that this may be extended to n-gram analysis through a modified usage of the LDA model creation employed here.

References

Ahmed, S., Cho, J., & Jaidka, K. (2019). Framing social conflicts in news coverage and social media: A multicountry comparative study. International Communication Gazette, 81(4), 346–371. https://doi.org/10.1177/1748048518775000

Analytics, V. (2016, March 21). PCA: A Practical Guide to Principal Component Analysis in R & Python. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

Bird, S., Loper, E., & Klein, E. (2009). Natural Language Toolkit: Natural Language Processing with Python. http://www.nltk.org/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Boumans, J., Trilling, D., Vliegenhart, R., & Boomgaarden, H. (2018). The agency makes the (online) news world go round: The impact of news agency content on print and online news. International Journal of Communication, 12, 1768–1789.

Clustering: 2.3.5. Spectral clustering. (n.d.). Scikit-Learn. Retrieved 17 May 2020, from https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering

D'Alessio, D. (2003). An Experimental Examination of Readers' Perceptions of Media Bias. Journalism & Mass Communication Quarterly, 80(2), 282–294. https://doi.org/10.1177/107769900308000204

Davie, G. (n.d.). Framing Theory. Retrieved 30 May 2020, from Mass Communication Theory Mass Communication Theory: From Theory to Practical Application website: https://masscommtheory.com/theory-overviews/framing-theory/

DellaVigna, S., & Kaplan, E. (2007). The Fox News Effect: Media Bias and Voting. *The Quarterly Journal of Economics*, *122*(3), 1187–1234. https://doi.org/10.1162/qjec.122.3.1187

Galtung, J., & Ruge, M. H. (1965). The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers. Journal of Peace Research, 2(1), 64–90. https://doi.org/10.1177/002234336500200104

Garcia Pires, A. J. (2014). Media diversity, advertising, and adaptation of news to readers' political preferences. Information Economics and Policy, 28, 28–38. https://doi.org/10.1016/j.infoecopol.2014.06.001

Humprecht, E., & Büchel, F. (2013). More of the Same or Marketplace of Opinions? A Cross-National Comparison of Diversity in Online News Reporting. The International Journal of Press/Politics, 18(4), 436–461. https://doi.org/10.1177/1940161213497595

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems – Survey and roads ahead. Information Processing & Management, 54(6), 1203–1227. https://doi.org/10.1016/j.ipm.2018.04.008

Media Bias / Fact Check. (n.d.). Media Bias / Fact Check: The Most Comprehensive Media Bias Resource. Retrieved 1 May 2020, from https://mediabiasfactcheck.com/

Nisbet, M. C., & Scheufele, D. A. (2009). What's next for science communication? Promising directions and lingering distractions. *American Journal of Botany*, *96*(10), 1767–1778. https://doi.org/10.3732/ajb.0900041

Pandey, P. (2018, September 23). Simplifying Sentiment Analysis using VADER in Python (on Social Media Text). Retrieved 19 May 2020, from https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

Pon, R. K., Cárdenas, A. F., Buttler, D. J., & Critchlow, T. J. (2011). Measuring the interestingness of articles in a limited user environment. Information Processing & Management, 47(1), 97–116. https://doi.org/10.1016/j.ipm.2010.03.001

SeatGeek. (2011, July 8). FuzzyWuzzy: Fuzzy String Matching in Python. ChairNerd. https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/

Semetko, H. A., & Valkenburg, P. M. V. (2000). Framing European politics: A Content Analysis of Press and Television News. Journal of Communication, 50(2), 93–109. https://doi.org/10.1111/j.1460-2466.2000.tb02843.x

Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. https://doi.org/10.13140/2.1.1394.3043

Sklearn Truncated SVD. (n.d.). Scikit-Learn. Retrieved 18 May 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

Tian, Y., & Stewart, C. M. (2005). Framing the SARS Crisis: A Computer-Assisted Text Analysis of CNN and BBC Online News Reports of SARS. Asian Journal of Communication, 15(3), 289–301. https://doi.org/10.1080/01292980500261605

van der Maaten, L. (n.d.). T-SNE. Laurens van Der Maaten. Retrieved 18 May 2020, from https://lvdmaaten.github.io/tsne/

Violante, A. (2018, August 29). An Introduction to t-SNE with Python Example. Retrieved 19 May 2020, from Towards Data Science website: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1