From Objects to Data

Bruno Sotic

11353473

Helge Moes

11348801

Bruno Koldeweij

12289434

Work Group 30

Dr. ir. J. Kamps

Assignment 2: Topic Modeling

Word count: 1333

Introduction and Experimentation

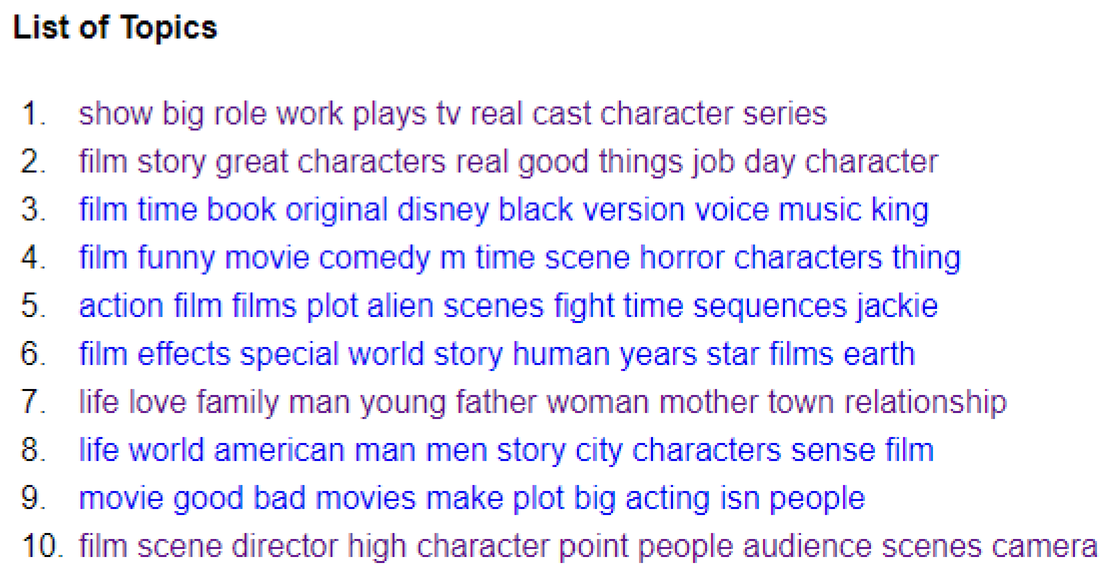
For the purpose of this assignment, we experimented with the topic-modelling tool Mallet in order to test the possibilities and the associated obstacles when engaging in large text-based analysis. Similar to tools used for word frequency visualizations, determining which words should be considered for the procedure deemed to be crucial, yet we also found that it contributes to varying results regarding the feasibility of the output. Apart from using the standard stop-words files available on canvas, we also engaged in online research to test the effects that different stop-words can have while engaging in automatic topic classification of large corpora. Examples of resources include GitHub, StackOverflow, NLTK Stopword list etc. Each of them seemed to contribute to categorizing the documents in its own manner, up to the degree where establishing what a ‘’feasible stop word list’’ should even contain (and more importantly, what it should not contain) deemed trivial. Discussing all the approaches and resources we used for this task is beyond the capabilities of this paper, therefore we will only address a couple of them and the issues we found along the way.

In addition, as we expressed on Slack, we were met with puzzling results after we tried generating a list of topics two times, while using the same input and same list of stop words. The list of topics generated would always come out different. Even while switching between numerous lists, editing the lists ourselves, adding domain specific stop words, separating / excluding / decreasing / increasing the number of input files; we would still be met with different results after every iteration. As explained by Kulsherstha (2019) in her ‘’Guide to Latent Dirichlet Allocation (LDA)’’, this is happening due to the probabilistic nature of topic modelling, as well as the way in which the algorithm is converging.

Following the advice of authors Blei and Jordan (2006), we run the topic learning process several times, after adopting a certain stop word list, and paid close attention to the degree to which the outputs differ from one another after multiple iterations. This allows us to measure the similarity between the topics in the different iterations by looking at "how much" the topics differ from each other. If in 2 consecutive iterations we find the topics to be different in output, but similar in what they represent, then we would consider it consistent.

The Modelling Process

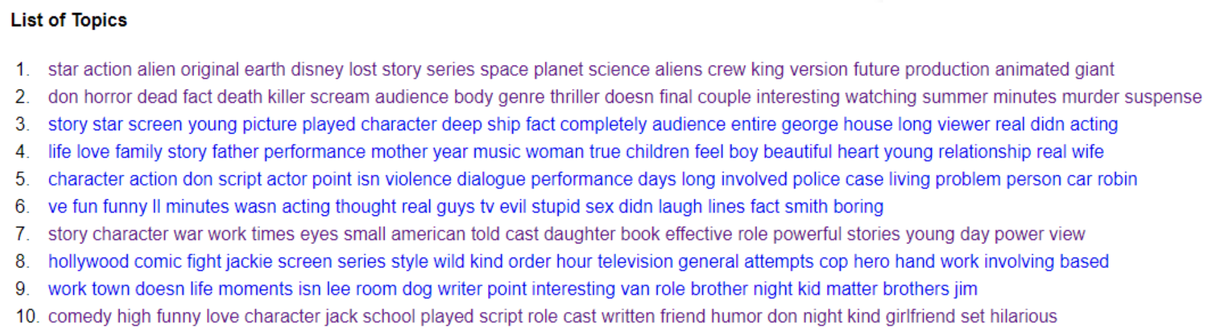
In the study ‘’Pulling Out the Stops: Rethinking Stopword Removal for Topic Modelling’’ the authors elaborate on other issues regarding the subjective judgements that are made while engaging in natural-language processing tasks. Indeed, it seems as the use of which stop words to use is left to subjective judgement, but it also seems to be dependent on the specific results that one desires (i.e.. do I want the movie-review files to be grouped based on sentiment or based on film genre?). We found a user-curated version of the ‘’NLTK Stopword list’’ (available on github: https://gist.github.com/sebleier/554280#gistcomment-2995070) to return adequate results (also when testing it for the aforementioned every-iteration-different-output problem). However, the list generated from the first output is still largely.



List of Topics

1. show big role work plays tv real cast character series
2. film story great characters real good things job day character
3. film time book original disney black version voice music king
4. film funny movie comedy m time scene horror characters thing
5. action film films plot alien scenes fight time sequences jackie
6. film effects special world story human years star films earth
7. life love family man young father woman mother town relationship
8. life world american man men story city characters sense film
9. movie good bad movies make plot big acting isn people
10. film scene director high character point people audience scenes camera

It is obvious that our corpora are about reviews for various film genres, as well as that the reviews hold a certain sentiment and might address specific content in a film. In order to arrive at our desired list of topics, we took an inductive approach to updating our list of stop words. This mean that we pay attention to the output and add very broad domain-specific words to the list (e.g. movie, film, character, actor, director, audience, series etc.). It also becomes apparent what we have lost by updating the list, namely, the potential to distinguish the reviews for series from reviews for films. We then raised the number of words for each topic to 20 and kept adding words to the stop list, until we would be left with a list of topics regarding reviews for specific genres.



List of Topics

1. star action alien original earth disney lost story series space planet science aliens crew king version future production animated giant
2. don horror dead fact death killer scream audience body genre thriller doesn final couple interesting watching summer minutes murder suspense
3. story star screen young picture played character deep ship fact completely audience entire george house long viewer real didn acting
4. life love family story father performance mother year music woman true children feel boy beautiful heart young relationship real wife
5. character action don script actor point isn violence dialogue performance days long involved police case living problem person car robin
6. ve fun funny ll minutes wasn acting thought real guys tv evil stupid sex didn laugh lines fact smith boring
7. story character war work times eyes small american told cast daughter book effective role powerful stories young day power view
8. hollywood comic fight jackie screen series style wild kind order hour television general attempts cop hero hand work involving based
9. work town doesn life moments isn lee room dog writer point interesting van role brother night kid matter brothers jim
10. comedy high funny love character jack school played script role cast written friend humor don night kind girlfriend set hilarious

This output still seems a bit abstract (maybe even vague) and can still be improved, but we found it, nonetheless, to be functional as a basic version. We then picked topics 1, 2 and 4 for the evaluation and labeled them as the following:

1. Animated science fiction (such as popular Disney content)

2. Horror / Thriller reviews
3. Family friendly romance reviews

Animated Science Fiction

Document 1: A review for the Disney movie ‘’Treasure Planet' (animated, steampunk sci-fi)’ while briefly reflecting on the experience in comparison to the first time of watching ‘’Pinnochio’’

Document 2: Film analysis and outline of the happening in the movie ‘’Chicken Run’’

Document 3: A long story outline of the Disney Cartoon ‘’Lilo and Stitch’’

Horror / Thriller

Document 1: A negative critique of the movie ‘’Scream’’

Document 2: Discussion about the movie ‘’Scary Movie’’ (Comedy movie)

Document 3: Content addressing a narrative regarding a metal music band, with a description that might fit a crime movie (Name of the film not mentioned)

Family Friendly Romance

Document 1: Plot regarding an episode of the series ‘’Sex and the City’’

Document 2: Discussing Disney's animated ‘’Little Mermaid 2’’ in comparison to its prequel.

Document 3: Argumentative text discussing the American action movie genre in relation to kung-fu action movies

Analysis and Reflection

At first, we were excited that the model managed to identify that the first document is about an animated space adventure film, while it is simultaneously compared to an older film in a different genre. However, by looking at the results in the last category (and the outcome of our qualitative analysis as a whole) it becomes apparent that our results are somewhat mixed in relation to our labels. We also question the use of our coding-scheme (i.e.. the way in which we labeled the topic words as a whole) which did not follow any solid ground to be justified the way it is. It is also arguable which movie might fall under 2 categories. On the other hand, it did manage to match the content of the text documents to the boldly defined genre we anticipated by around 70%.

Additionally, regarding the labeling, it was interesting to experiment with the number of words per topic. At first it seems obvious that the more words there are for a single topic, the more specific the label for that topic. But on the other hand, more words also leave more room for free interpretation. We also found that an increase in ‘’words per topic’’ offers further potential to display contradictory words under the one topic (ie. comedy and horror). Here, we also want to reflect on the troubleshooting process regarding topic modelling and LDA.

While researching for approaches and other grounds on how to construct topic models, we found most resources to be quantitative in nature and on a deeper technical level within the domain of natural-text processing. Most of the online resources and academic papers also seem to be making use of various libraries and packages in Python and R, which offer more possibilities and potential for this type of task, as well as various modes to reduce text files to their atomic elements more easily. Regarding the labeling, we believe that the data science blog ‘’Machinelearningplus.com’’ offered very essential guidelines, in which the author argues that the labels should be constructed based on the keywords that are used to generate the topic. It also offered some interesting information about how to make the findings analysis-ready, by visualizing them as a cluster graph based on the weight of the keywords in use. This would indeed make for a better evaluation of our topic model.

Looking at the bag of words as a cluster is feasible in approach, but one might be running the risk of wrongly interpreting them, based on loose connections, or confirmation bias. Topic modelling is an empirical approach that we believe is suitable for the use of discovering abstracts in very large files, but it also seems to be implemented in various recommendation systems that are the norm on the web nowadays. However, the manner in which it is approached might arguably be context dependent.

References

Blei, David M., and Michael I. Jordan. 2006. “Variational Inference for Dirichlet Process Mixtures.” *Bayesian Analysis* 1 (1): 121–43. https://doi.org/10.1214/06-BA104.

Kamps, Jaap. 2020. “Assignment 2: Topic Modeling.” Universiteit van Amsterdam. https://canvas.uva.nl/courses/15702/files/2460186?module_item_id=524877

KULSHRESTHA, RIA. 2019. “Latent Dirichlet Allocation(LDA).” Medium. November 3, 2019. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2.

Schofield, Alexandra, M\aans Magnusson, and David Mimno. 2017. “Pulling Out the Stops: Rethinking Stopword Removal for Topic Models.” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 432–436. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/E17-2069.

262588213843476. n.d. “NLTK’s List of English Stopwords.” Gist. Accessed April 19, 2020. https://gist.github.com/sebleier/554280.