Natural Language Processing

Report FinCon4U

Insights about news coverage with Topic Modelling

Camila Matoba
Student number: 2067717

16-10-2022

Table of Contents

Simplified summary

- Objective: Analyse the content of the news articles on Apple.
- Scope: all articles with Apple in the title.all articles with Apple in the title.
- Dataset: preprocessed data (punctuation removal) from https://www.kaggle.com/datasets/gennadiyr/us-equities-news-data?resource=download
    - Scope: Apple stock (AAPL) related news
- Key model: Topic Modelling: LDA

1. Topic modelling algorithm Evaluation and Selection

Four different topic modelling algorithms were tested and evaluated according to the coherence score. These models were Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Fuzzy Latent Semantic Analysis (FLSA), and Latent Semantic Indexing (LSI). This measure was chosen as it has high human interpretability. In particular, (Röder M. et al., 2015) defends Cv as the best performing coherence measure. It is calculated by combining the indirect cosine measure with the NPMI and the boolean sliding window. The best topic cohesion across three to seven topics were the following for the four models:

Model	**Topic coherence**
LdaMulticore<num_terms=23892, num_topics=7, decay=0.5, chunksize=2000>	0.457
gensim.models.nmf.Nmf	0.454
FuzzyTM.FuzzyTM.FLSA_W	0.439
LsiModel<num_terms=23892, num_topics=4, decay=1.0, chunksize=20000>	0.394

The model with highest coherence was LdaMulticore with 0.457 coherence. This model was hypertuned on alpha and beta by iterations on the optimal amount of topics and resulted in 0.48470 topic coherence. Alpha is the “A-priori belief on document-topic distribution” and beta or eta is the “A-priori belief on topic-word distribution”. The best result was the following configuration: LdaMulticore<num_terms=23892, num_topics=7, decay=0.5, chunksize=2000>, alpha = 0.01 and beta = 0.75. All iterations and their coherence scores are present in Appendix II.

The final model chosen, LDA, discovers topics that are hidden (latent) in a set of text documents by inferring possible topics based on their words. It uses a generative probabilistic model (generating data that is similar to observed previous data) and Dirichlet distributions (basically a distribution over distributions) to achieve this. The inference in LDA is based on a Bayesian framework. This allows the model to infer topics based on observed data (words) through the use of conditional probabilities. This model works with two bags of words. One Matrix contains the n topics as each row and each column represents a word (or ngram) and the values are the probabilities that those words belong to the topic. Another matrix is the bag of Words of all words (columns) present for each document (row). These are multiplied to obtain the conditional probability that the word takes on each topic. This way it is possible to obtain which topic is most probably to each document.

2. Choice of number of topics

As the corpus is limited to Apple stock (AAPL) related news, the objective here is to understand what are the main topics that are published about Apple financially. Furthermore, since it is already quite a narrow subject, that is, only relating to financial news of a single company, it would make no sense to have too many topics, as many of these would have similar words. Thus, the optimal number of topics of six was obtained by calculating the coherence score in the limited range from three to seven. Each model has their coherence score and number of topics relationship on Appendix I. To confirm that this amount of topic is good for the corpus tests, a visualisation with LDAvis (Appendix III) is presented to show that the topics do not have a large overlap, such as a subtopic within a greater topic.

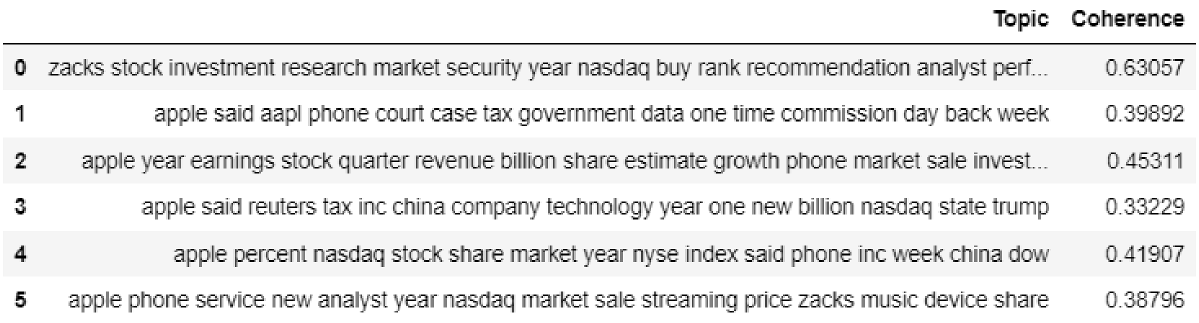
3. Corpus quality

As most of the text comes from news articles almost no typos were found at eyeballing. Some preprocessing has already been done such as removing punctuation. Unfortunately, this was roughly performed as some words such as U.S. turned into U S which is meaningless. Other words such as ReleaseChicago were clumped together, so they needed to be separated. To improve the input of the topic modelling, bigrams, trigrams and quadrigrams were created, but resulted in lower coherence. The results for this are in a different notebook, being the best model FuzzyTM.FuzzyTM.FLSA_W with 0.402 coherence. In this notebook, a reduction of the words used was also implemented. Only the tokens which are contained in at least 5 documents were kept. The dictionary was also filtered on the upper boundary, by keeping only tokens which are contained in no more than 50% documents. On top of this, only the top 10.000 words were kept. This experiment was better than implementing ngrams, however, as this also led to worse results, it was kept from the final notebook presented. Both of this tests with the corpus and their results are specified in Appendix IV.

To evaluate topics qualitatively several iterations were made, removing some words that are not so informative in this dataset. For example, 'apple' is not an informative word, since all articles were filtered to be about the company. However, in the other analysis with ngrams, it was interesting to see the combination with 'apple-watch', 'phone' and 'samsung' together as the news relating to just the phone and accessories parts for the business. Nonetheless, words that are too frequent can be eliminated without losing any information as they don't add any specific information which would make the document stand out.

4. Results

The topics and their coherence score per topic were:



	Topic	Coherence
0	zacks stock investment research market security year nasdaq buy rank recommendation analyst perf...	0.63057
1	apple said aapl phone court case tax government data one time commission day back week	0.39892
2	apple year earnings stock quarter revenue billion share estimate growth phone market sale invest...	0.45311
3	apple said reuters tax inc china company technology year one new billion nasdaq state trump	0.33229
4	apple percent nasdaq stock share market year nyse index said phone inc week china dow	0.41907
5	apple phone service new analyst year nasdaq market sale streaming price zacks music device share	0.38796

Despite the efforts the only topic that might qualitatively stand out is the 5th one, which actually has the lowest coherence score. Words such as phone. Streaming, music and device might point to the multimedia area of the company.

5. Limitations

A few limitations of LDA need to be taken into consideration:

1. LDA is unable to depict correlations which led to occurrence of uncorrelated topics.

2. As a Bag of words model, order of the words, grammatical role of the words, and sentence structure are not considered in the model.

3. LDA is Unsupervised (sometimes weak supervision is desirable, e.g. in sentiment analysis)

4. The number of topics need to be set manually, which was optimised by iterating over each amount in the designed range.

6. Final Recommendations

Topic modelling wise, it would have been interesting to gather news from multiple sources to understand the trends of topics about the company in a broader theme to give context to the financial situation of the company. This way further analysis considering this bigger picture could possibly explain the performance of the company in context. When linked to financial data of stocks in time, such as with the previous project, it could clarify the source of the increase in value of stocks. In other words, whether an increase came solely because of the launch of a feature or because inflation went down according to the news. Also which combination of topics were trending when the stocks go up or when the stocks go down. This way it would be easier to predict what kind of topics influence the value of stocks the most.

Although the topics seem very repetitive given that it is a very narrow subject, an overview in time shows what is the trend over the years. That is, what topics have become more relevant over time, such as the Figure 1.

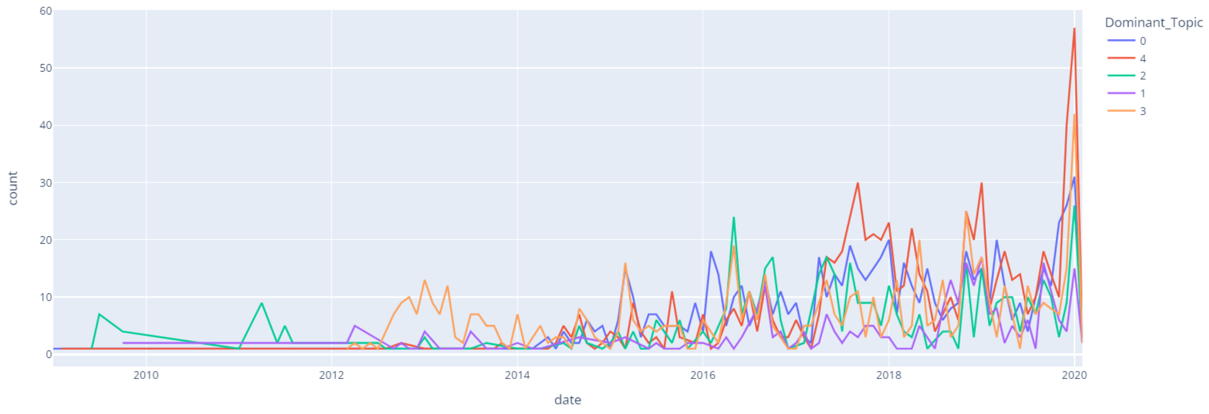


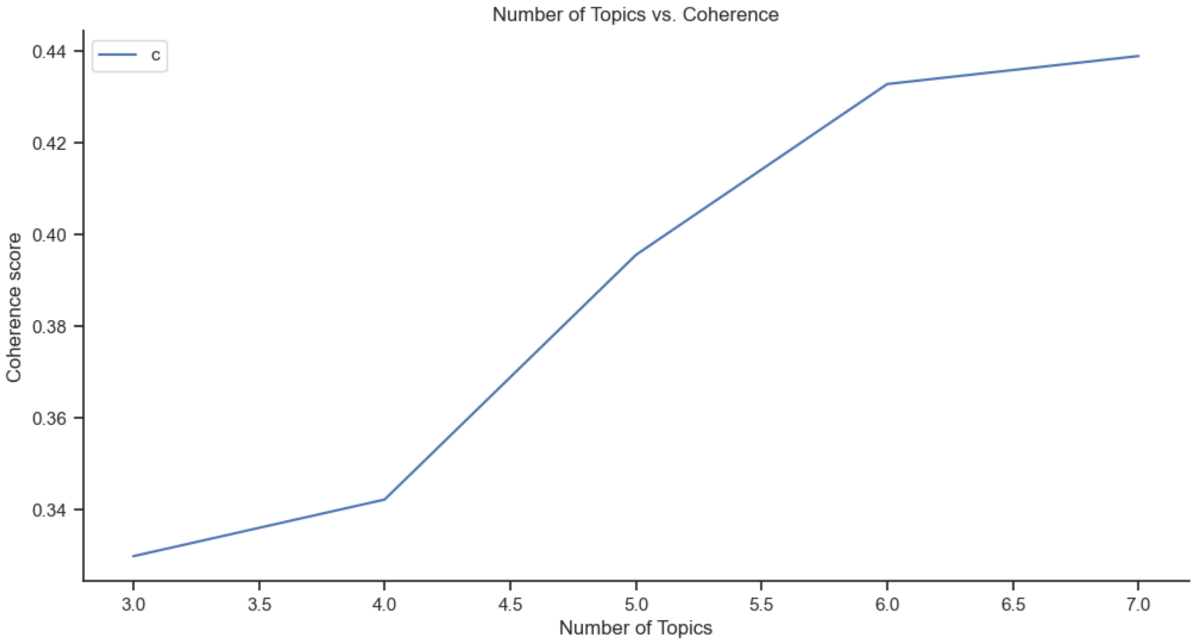
Figure 1: Distribution of topics over time by quantity

To conclude it was a worthy attempt to understand the results behind the last project’s success. However with such a narrow subject it was hard to interpret the topics from the model. In a future project more broad and varied themes about the company could yield more interesting results.

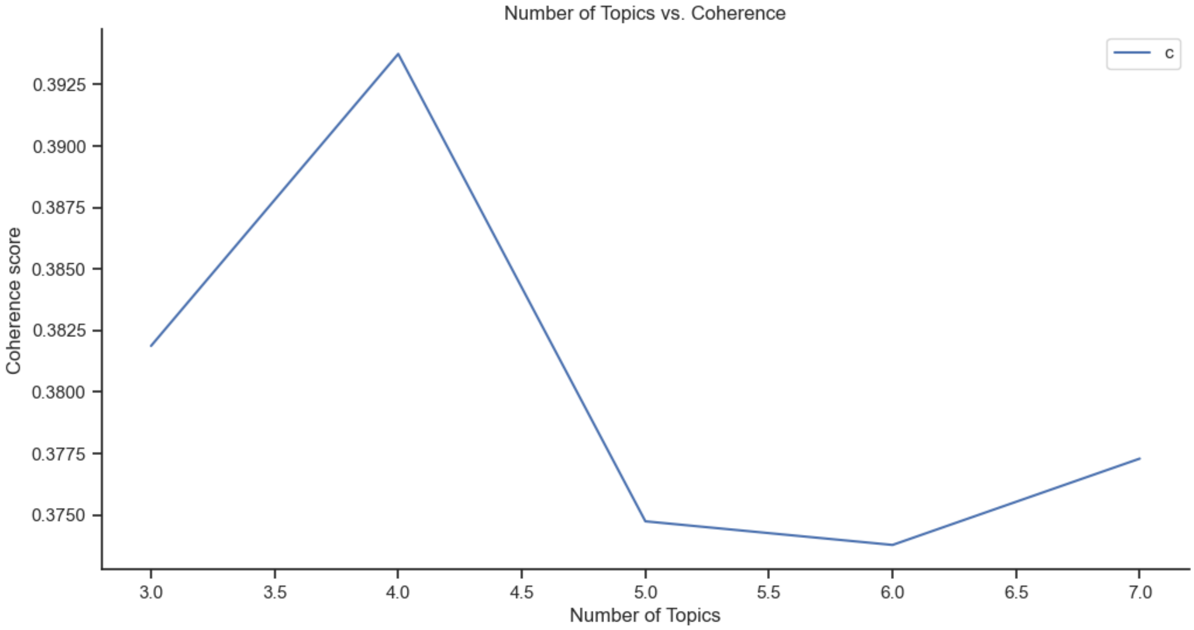
Appendix

I. Model Evaluation and Selection

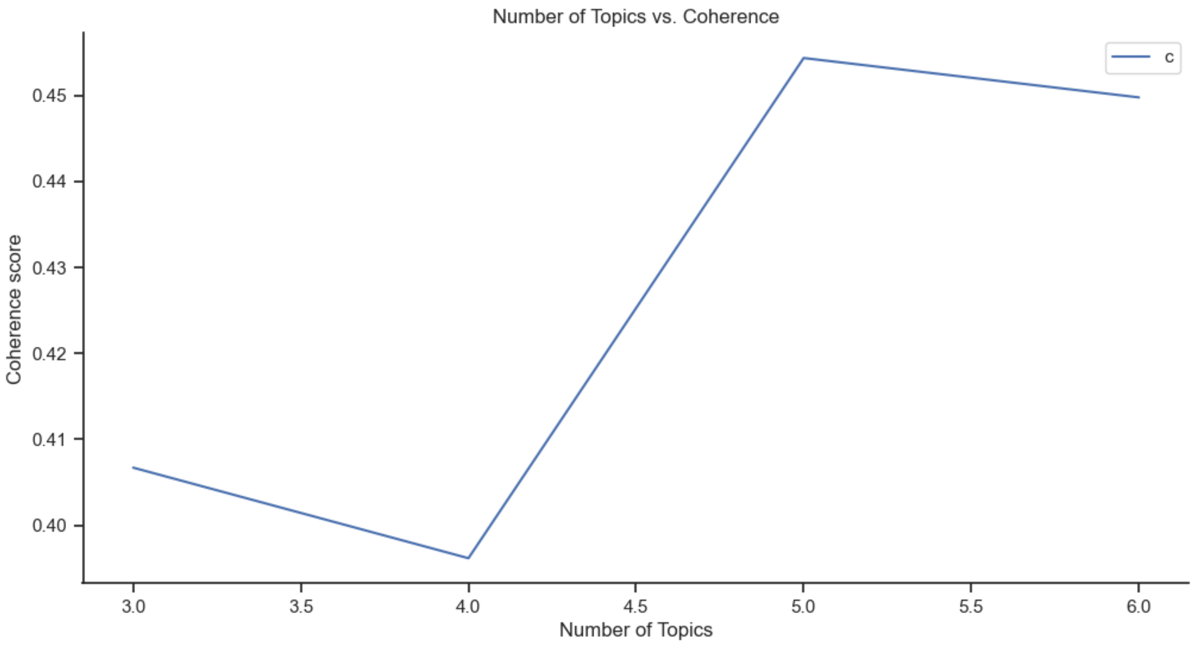
A. FuzzyTM (0.439 coherence)



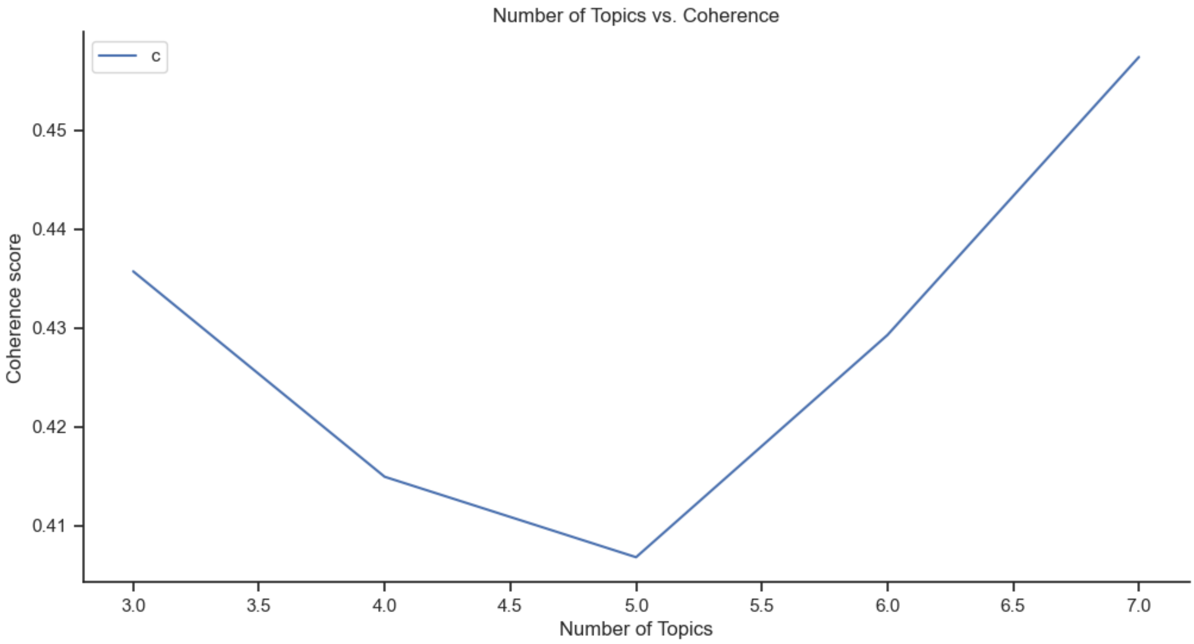
B. LSA (0.394 coherence)



C. NMF (0.454 coherence)



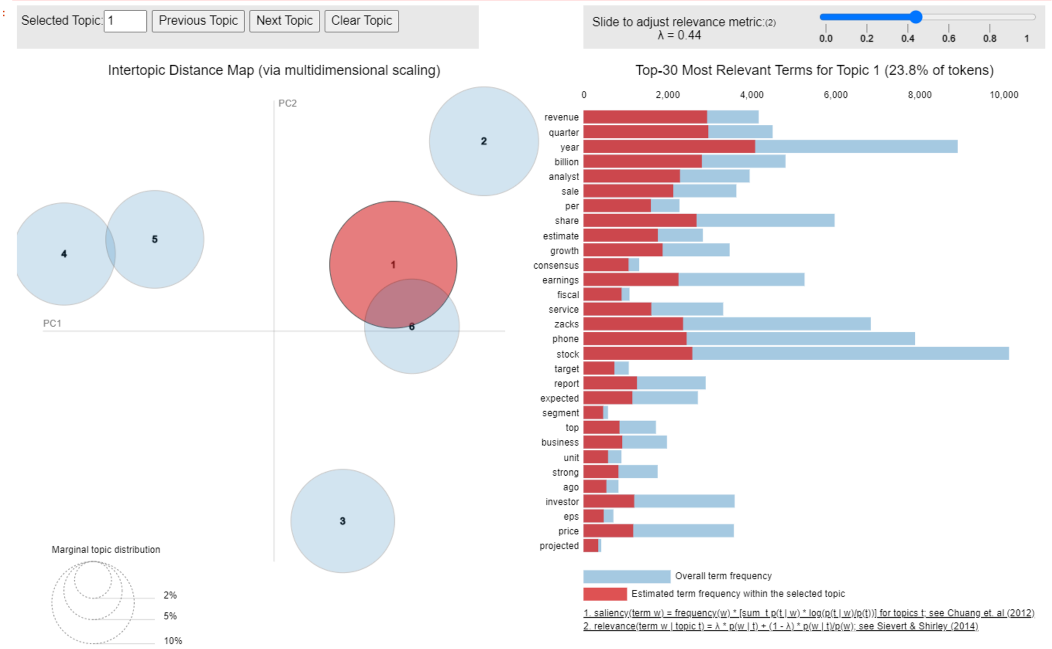
D. LDA (0.457 coherence)



II. Hyperparameter tuning

alpha	beta	coherence
asymmetric	0.38	0.47023
asymmetric	0.01	0.46043
0.75	0.75	0.45502
0.38	symmetric	0.45476
0.38	0.01	0.45466
asymmetric	symmetric	0.45412
0.38	0.38	0.45144
0.75	0.01	0.44997
0.75	symmetric	0.44997
0.75	0.38	0.44997
0.38	0.75	0.44961
symmetric	symmetric	0.44886
symmetric	0.01	0.44878
0.01	0.38	0.44703
0.01	0.01	0.44702
0.01	0.75	0.44663
symmetric	0.75	0.44663
symmetric	0.38	0.44513
0.01	symmetric	0.44255
asymmetric	0.75	0.42636

III. Topics Visualisation



Selected Topic: 1

Slide to adjust relevance metric:(2)
λ = 0.44

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

IV. Full code

In case there is more interest to get more in depth insights about our analysis, the full code is available on GitHub, at https://github.com/octokami/news_stock_market/tree/main/Topic%20Modelling containing the following elements:

A. .Final Code named Matoba-Camila-2067717-code.ipynb

The output folder containing:

1. The graph for each model's coherence score relationship with number of topics
2. The dictionary used
3. LDAvis (such as the appendix III) in csv and html
4. The LDA (best model) tuning results

B. Code with the ngrams and filtering of the dictionary, named Matoba-Camila-2067717-code-ngrams.ipynb

The output folder containing:

1. The graph for each model's coherence score relationship with number of topics
2. The dictionary used
3. LDAvis (such as the appendix III) in csv and html
4. The LDA tuning results
5. Ngrams in the dictionary

References

1. Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15). Association for Computing Machinery, New York, NY, USA, 399–408. https://doi.org/10.1145/2684822.2685324