COMP4432

Machine Learning

Data Product Report (Topic Modeling)

LetsDoIt-G101

Hui Sze Ming 19050459D

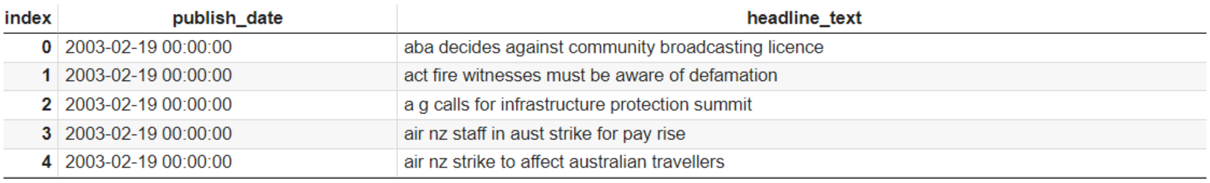
Chang Wai Chung 21052435D

Table of Contents

1. Introduction

In this project, we performed natural language processing on the selected dataset from Kaggle (“*A Million News Headlines Dataset*”). Three unsupervised machine learning models have been implemented and evaluated for topic modeling.

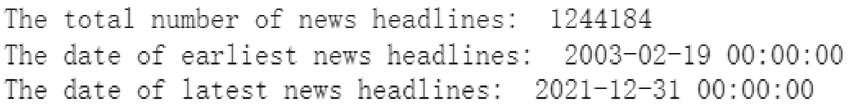
2. Data Exploration



index	publish_date	headline_text
0	2003-02-19 00:00:00	aba decides against community broadcasting licence
1	2003-02-19 00:00:00	act fire witnesses must be aware of defamation
2	2003-02-19 00:00:00	a g calls for infrastructure protection summit
3	2003-02-19 00:00:00	air nz staff in aust strike for pay rise
4	2003-02-19 00:00:00	air nz strike to affect australian travellers

figure 1. Data frame head

After importing the dataset from csv into Pandas data frame, figure 1 shows five example data of the dataset, and it is clear that this dataset contains two major columns which represent the publish date and the headline content respectively.



```
The total number of news headlines:  1244184
The date of earliest news headlines:  2003-02-19 00:00:00
The date of latest news headlines:  2021-12-31 00:00:00
```

figure 2. Dataset details

Looking deeper into the dataset as figure 2 shows, this dataset contains 1,244,184 news headlines in total, and all news headlines are published between 2003-02-19 00:00:00 and 2021-12-31 00:00:00.

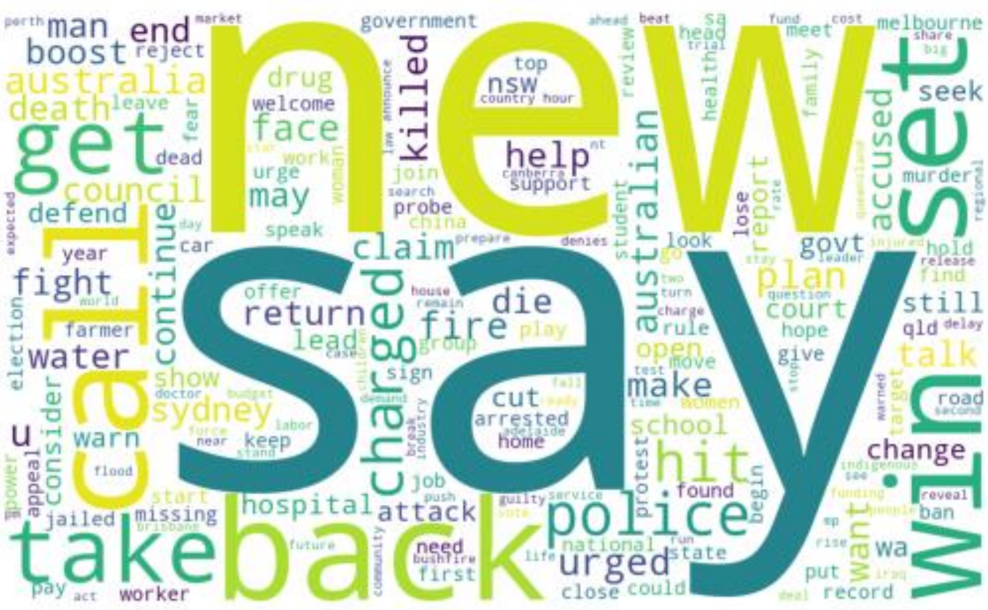


figure 3. Dataset wordcloud result

Additionally, wordcloud has been used for checking word frequency in this dataset, notice that wordcloud is implemented after removing stop words. As figure 3 shows, some of the most frequent words in this dataset are “say”, “new”, “take”, “back”, “win”, “set”, “police”.

3. Data Preprocessing

For data preprocessing, TfidfVectorizer has been implemented for word embedding, it transforms the news headlines text into a matrix of TF-IDF features before fitting the data into the models. For TF-IDF it is the short form of term frequency and inverse document frequency. The following figure 4 shows the formula for calculating tf-idf.

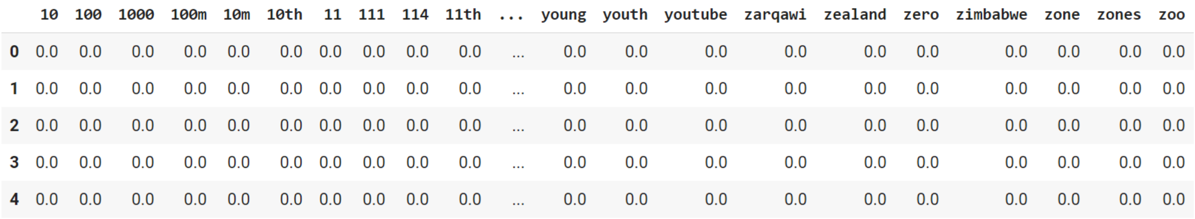
$$tf(w,d) = log(1 + f(w,d))$$

$$idf(w,D) = log(\frac{N}{f(w,D)})$$

$$tfidf(w,d,D) = tf(w,d) * idf(w,D)$$

figure 4. TF-IDF formula [4]

The example TF-IDF scores of each word are shown in figure 5, the result shows the tf-idf scores of the first 10 words and last 10 words. Notice that the stop words are removed before implementing TfidfVectorizer.



	10	100	1000	100m	10m	10th	11	111	114	11th	...	young	youth	youtube	zarqawi	zealand	zero	zimbabwe	zone	zones	zoo
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

figure 5. TF-IDF example result

According to Ma (2018), TF-IDF may not value high frequency words more, using TF-IDF results in higher word importance in same news headline, and lower word importance in other news headlines.

4. Topic Modeling

When dealing with dataset without any label, unsupervised machine learning can help to draw inferences and find the hidden pattern from it. In this project, topic modeling is applied to cluster the news headlines text with similar topics, there are 8 topics and top 10 related words for each topic.

According to Albalawi (2020), when comparing different popular topic modeling methods (LSA, LDA, NMF, PCA, and Random Projection) for short text data, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) have greater capability of extracting meaningful topics. In order to get the optimal news headlines clustering result while comparing different topic modeling methods, **LSA**, **LDA**, and **NMF** have been chosen as the topic modeling method for the news headlines dataset in this project.

4.1. Latent Semantic Analysis (LSA)

LSA is a statistical approach to find the relation between different words, it uses the matrix factorization technique Singular Value Decomposition (SVD) to discover the semantic structures in the news headlines. The topic modeling result is shown in follows.

Topic	**Top 10 words**
1	'new' 'missing' 'court' 'murder' 'death' 'charged' 'car' 'crash' 'police' 'man'
2	'david' 'nrl' 'smith' 'tom' 'afl' 'daniel' 'extended' 'ben' 'michael' 'interview'
3	'death' 'stabbing' 'killed' 'jailed' 'guilty' 'dies' 'murder' 'court' 'charged' 'man'
4	'water' 'news' 'plan' 'national' 'says' 'nsw' 'govt' 'rural' 'council' 'new'
5	'market' 'monday' 'weather' 'business' 'crash' 'abc' 'nsw' 'news' 'national' 'rural'
6	'road' 'injured' 'hospital' 'plane' 'driver' 'killed' 'dies' 'fatal' 'car' 'crash'
7	'fatal' 'dies' 'killed' 'man' 'car' 'news' 'national' 'rural' 'crash' 'new'
8	'missing' 'charged' 'hospital' 'car' 'rural' 'man' 'govt' 'water' 'plan' 'council'

Table 1. LSA result

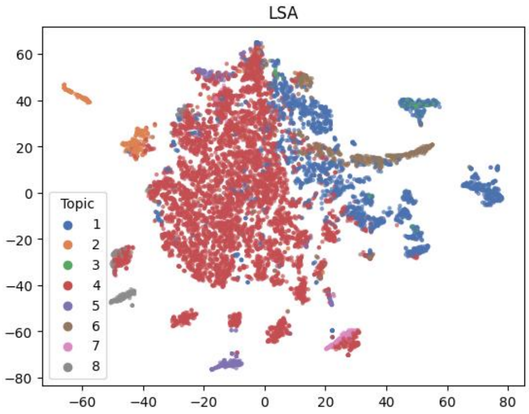


figure 6. LSA t-SNE visualization

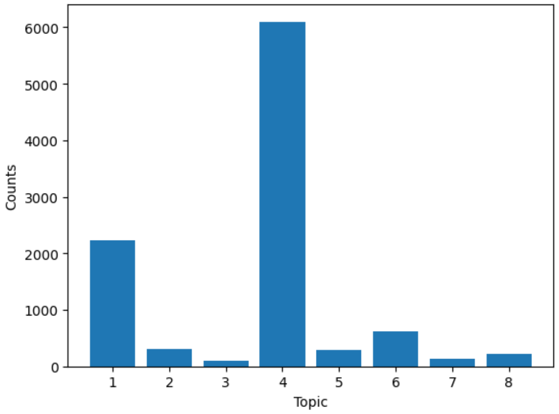


figure 7. LSA topic count

4.2. Latent Dirichlet Allocation (LDA)

LDA considers different news headlines are in the similar topic if they are using similar group or words by calculating the probability of different words belonging to each topic. According to Albanese (2022), LDA approximate the document-topic and term-topic distributions in a Bayesian approach by using Dirichlet priors.

Topic	Top 10 words
1	'hope' 'house' 'track' 'car' 'crash' 'missing' 'hospital' 'woman' 'man' 'new'
2	'wa' 'trial' 'plan' 'accused' 'govt' 'commission' 'boost' 'action' 'country' 'court'
3	'defence' 'australian' 'fatal' 'car' 'charged' 'assault' 'crash' 'police' 'man' 'interview'
4	'green' 'guilty' 'south' 'award' 'project' 'closer' 'abc' 'east' 'weather' 'wins'
5	'pay' 'work' 'sydney' 'new' '19' 'man' 'australia' 'police' 'covid' 'day'
6	'cup' 'market' 'australia' 'cuts' 'new' 'business' 'says' 'national' 'news' 'rural'
7	'gold' 'police' 'world' 'cup' 'change' 'search' 'govt' 'budget' 'climate' 'north'
8	'new' 'death' 'water' 'family' 'farmers' 'plans' 'driver' 'media' 'govt' 'police'

Table 2. LDA result

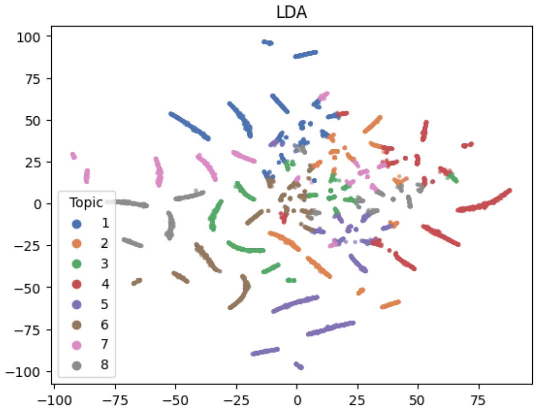


figure 8. LDA t-SNE visualization

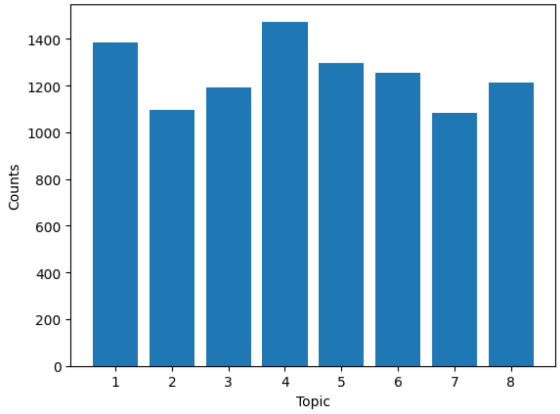


figure 9. LDA topic count

4.3. Non-Negative Matrix Factorization (NMF)

NMF is a matrix factorization method that is similar to the LSA, one of the main differences is that it forbids containing negative elements in the matrices.

Topic	Top 10 words
1	'stabbing' 'accused' 'sydney' 'guilty' 'jailed' 'death' 'murder' 'court' 'charged' 'man'
2	'james', 'nrl', 'smith', 'afl', 'tom', 'daniel', 'extended', 'ben', 'michael', 'interview'
3	'shooting' 'woman' 'hunt' 'station' 'death' 'probe' 'missing' 'search' 'investigate' 'police'
4	'urged' 'sa' 'south' 'wa' 'qld' 'health' 'australia' 'nsw' 'govt' 'says'
5	'reporter' 'tasmania' 'monday' 'weather' 'business' 'nsw' 'abc' 'news' 'national' 'rural'
6	'hospital' 'road' 'injured' 'plane' 'killed' 'driver' 'dies' 'fatal' 'car' 'crash'
7	'australian' 'coronavirus' 'gold' 'cases' 'year' 'chief' 'help' 'zealand' 'home' 'new'
8	'public' 'change' 'development' 'land' 'residents' 'climate' 'considers' 'water' 'plan' 'council'

Table 3. NMF result

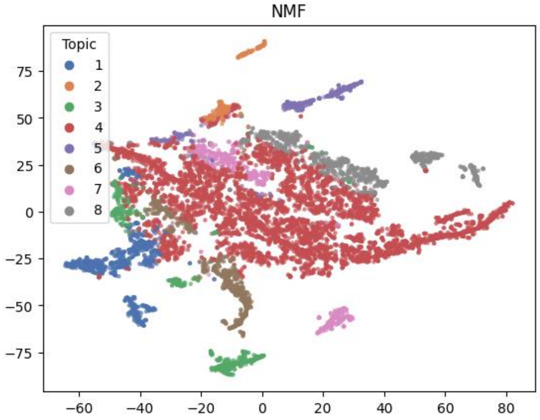


figure 10. NMF t-SNE visualization

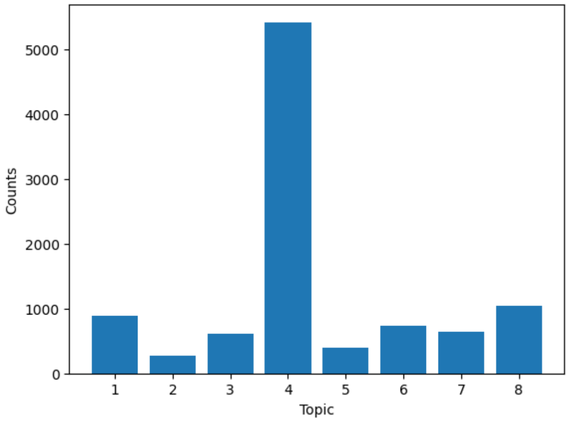


figure 11. NMF topic count

4.4. Result Comparison

After generating the topic modeling results, it is clear that the results from different models have several key differences. Notice that the top 10 words generated would not be discussed as we believe deciding wordings' topic manually is based on personally subjective thinking, only the t-SNE visualized graph and the word count bar chart will be discussed in the following.

For the t-SNE visualized graph, the result from LSA shows that there are some overlapping among different topics, especially for the color red and blue, it is also clear that each topic cluster has high variation. Additionally, the result from LDA shows that each topic has a flattened distribution of different nodes (words) and they are separated clearly. Also, the result from NMF shows a similar result with the LSA's one, but with lower variation for each topic cluster and much lesser overlapping observed.

For the word count bar chart, both the results from LSA and NMF show that the topic distribution is unbalanced, they believe that most of the news headlines belong to one single topic. On the other hand, the result form LDA shows that the topic distribution is balanced, it believes different topics have similar amount of news headlines.

5. Model Evaluation

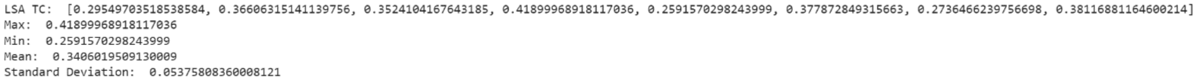
There is a common way to evaluate the performance of topic modeling models, which is the topic coherence (TC). In this project, we applied topic coherence (TC) to evaluate the performance of LSA, LDA, and NMF on their topic cluster results.

5.1. Topic Coherence

Topic coherence examines the semantic similarity between high scoring top words in the topic [5], which means it measures how semantically related the words within a topic are.

And “C_V” measures the coherence between the topics by finding the pairwise cosine similarity between the top words of each topic. [6]

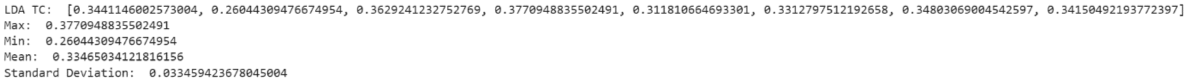
LSA



```
LSA TC:  [0.29549703518538584, 0.36606315141139756, 0.3524104167643185, 0.41899968918117036, 0.2591570298243999, 0.377872849315663, 0.2736466239756698, 0.38116881164600214]
Max:  0.41899968918117036
Min:  0.2591570298243999
Mean:  0.3406019509130009
Standard Deviation:  0.05375808360008121
```

figure 12. LSA evaluation

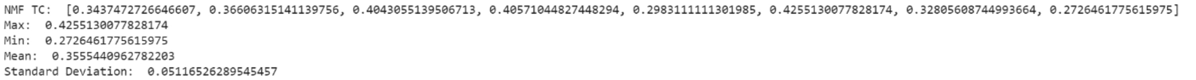
LDA



```
LDA TC:  [0.3441146002573004, 0.26044309476674954, 0.3629241232752769, 0.3770948835502491, 0.311810664693301, 0.3312797512192658, 0.34803069004542597, 0.34150492193772397]
Max:  0.3770948835502491
Min:  0.26044309476674954
Mean:  0.33465034121816156
Standard Deviation:  0.033459423678045004
```

figure 13. LDA evaluation

NMF



```
NMF TC:  [0.3437472726646607, 0.36606315141139756, 0.4043055139506713, 0.40571044827448294, 0.2983111111301985, 0.4255130077828174, 0.32805608744993664, 0.2726461775615975]
Max:  0.4255130077828174
Min:  0.2726461775615975
Mean:  0.3555440962782203
Standard Deviation:  0.05116526289545457
```

figure 14. NMF evaluation

The evaluation results of LSA, LDA, and NMF are shown in the figure 12, 13 , and 14 correspondingly. Notice that the index of TC array corresponds to the topics, for example, first elements refer to topic 1.

Despite LDA has the smallest maximum coherence score, it also has the lowest standard deviation, meaning that LDA is better at clustering different words for each topic equally.

6. Highlighted Challenges

During the evaluation coding, sklearn library does not have topic coherence model. So, we needed to use another popular library gensim and use “CoherenceModel” function to calcaute coherence scores. However, to feed sklearn model to gensim, it is necessary to manually convert the scikit-learn model into gensim format, which could be time consuming. Eventually, we have discovered an easier way to do so by using another model “metric_coherence_gensim” from tmtoolkit library, because it is compatible with sklearn's topic modeling models.

7. Conclusion

In conclusion, we believe that NMF has the best performance among all topic modeling methods in general, it has a t-SNE visualized graph with lower variation and all topic clusters are separated quite clearly. Additionally, NMF has the highest coherence score which means its result topics are more interpretable by us.

On the other hand, LSA and LDA perform not as good as NMF, which is expected and aligned with the research paper observation by Albalawi (2020). Notice that we believe that LDA may be more suitable when we desire a more balanced topic modeling result, as it is the model with a more balanced topic word count and it has the lowest standard deviation.

8. Appendix

The share link of the implemented code in this project:

https://drive.google.com/file/d/1ULIlswgwiGsWAeAzWUPeidfsw8YQRreu/view?usp=sharing

9. Reference

[1] Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Frontiers in Artificial Intelligence, 3, 42–42. https://doi.org/10.3389/frai.2020.00042

[2] Edward Ma (2018). 2 latent methods for dimension reduction and topic modeling. Medium. https://towardsdatascience.com/2-latent-methods-for-dimension-reduction-and-topic-modeling-20ff6d7d547

[3] Nicolo Cosimo Albanese (2022). Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a Comparison. Medium. https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5

[4] Marius Borcan (2020). TF-IDF Explained And Python Sklearn Implementation. Medium. https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275

[5] Shashank Kapadia (2019). Evaluate Topic Models: Latent Dirichlet Allocation (LDA). Medium. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0#:~:text=But%20before%20that%E2%80%A6-,What%20is%20topic%20coherence%3F,are%20artifacts%20of%20statistical%20inference

[6] Enes Zvornicanin (2023). When Coherence Score Is Good or Bad in Topic Modeling?. Medium. https://www.baeldung.com/cs/topic-modeling-coherence-score#:~:text=CV%20Coherence%20Score,NPMI)%20and%20the%20cosine%20similarity.