Data Product
A Million News Headlines

Group: A-PLUS
PAN Yalu 20080027D
XU Hulu 20075015D
YANG Yifan 20075243D

Introduction

1. A dataset that published over many years

2. Content has columns for publish date and headline text

Content

Format: CSV ; Single File

1. **publish_date**: Date of publishing for the article in yyyyMMdd format
2. **headline_text**: Text of the headline in Ascii , English , lowercase

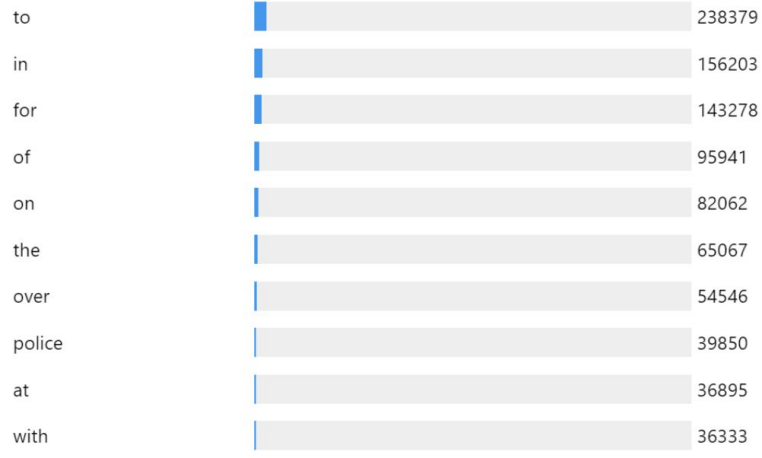
Start Date: **2003-02-19** ; End Date: **2021-12-31**

Data Analysis

1 Basic data statistic:

```
# of headlines: 1244184
# of headline tokens in vocabulary: 108058
avg. length of headlines: 41.28453910354096 chars, 6.557524449759843 tokens
publish date range: (20030219, 20211231)
```

2 Word frequency: the top 10 most frequent words



Data Preprocessing

1. **Remove stopwords (e.g. “a”, “the”, “in”)**

Remove words that don’t carry so much meaning

Reduce noises and dimensionality

Use the *stopwords* modules from *nlkt* library.

```
i
me
my
myself
we
our
ours
ourselves
you
you're
you've
you'll
you'd
your
```

Data Preprocessing

2 Deduplication

Remove redundant information

Make sure the model are not baised towards overrepresented headlines

Most frequently occurring

	publish_date	headline_text	# duplicates
0	20210301	house prices record sharpest increase since 2003	2
1	20210601	house prices reach record levels; as investors	2

Duplication found by the data report

Topic modeling: LSA, LDA

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA)

Assumption: Distribution Hypothesis

Improve the result: tf-idf

CountVectorizer: compute one hot array for words in headlines

Topic modeling: LSA, LDA

Result

LSA

```
Topic 1:  police crash death car killed missing dead search probe attack
Topic 2:  new year laws years zealand life named opens gets president
Topic 3:  man charged murder dies guilty jailed arrested pleads court child
Topic 4:  says minister report pm mp labor trump iraq opposition time
Topic 5:  govt plan wa qld water health government calls urged school
Topic 6:  court accused face case told charges faces high trial murder
Topic 7:  australia day australian world cup south win china coronavirus test
Topic 8:  council plan city considers land rejects mayor water seeks backs
Topic 9:  nsw rural sydney news abc country national north hour weather
Topic 10:  interview gold coast nrl afl extended michael speaks john david
```

Topic modeling: LSA, LDA

Result

LDA

```
Topic 1:  new school country china test hour plans england union second
Topic 2:  australia world cup day report killed calls melbourne farmers power
Topic 3:  water hospital wins help ban boost gold coronavirus work trial
Topic 4:  man charged car woman dies home attack dead hit police
Topic 5:  sa final guilty deal minister labor set open mp asylum
Topic 6:  interview qld missing election rise search year high public child
Topic 7:  death man nsw sydney murder court police funding road fears
Topic 8:  council plan australian wa abc coast health court north south
Topic 9:  police says government man drug win budget jail say jailed
Topic 10:  govt urged rural pm trump claims nt iraq case crash
```

BERTopic

Bidirectional Encoder Representations from Transformers

c-TF-IDF: Class-based TF-IDF

c-TF-IDF

For a term **x** within class **c**:

$$W_{x,c} = \| tf_{x,c} \| \times \log\left(1 + \frac{A}{f_x}\right)$$

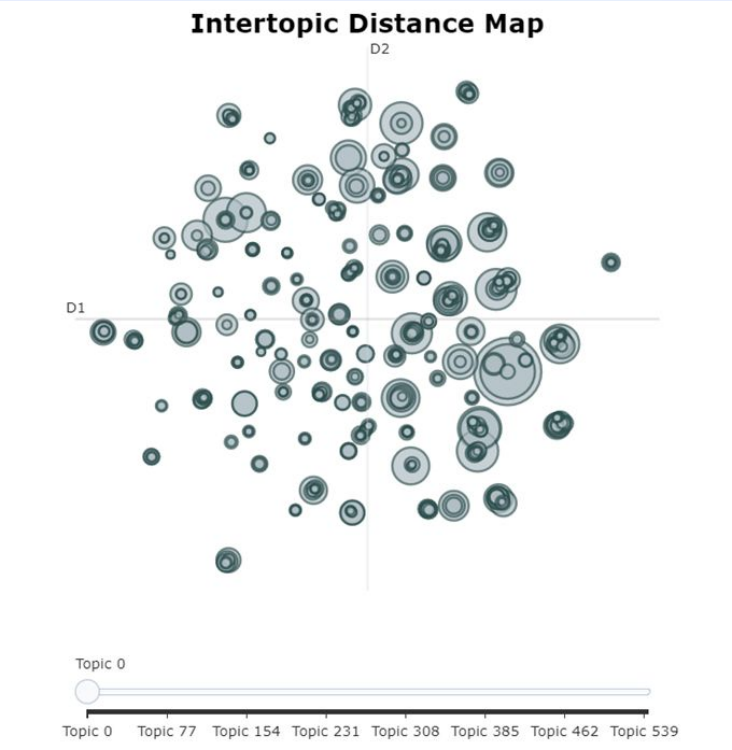
$tf_{x,c}$ = frequency of word **x** in class **c**

f_x = frequency of word **x** across all classes

A = average number of words per class

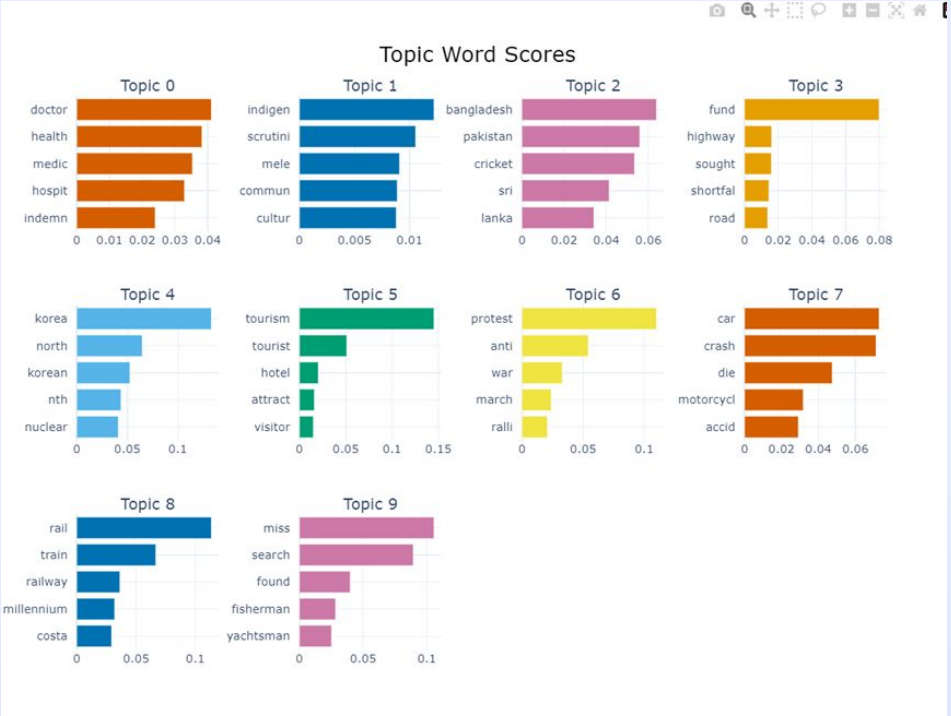
BERTopic

Result



BERTopic

Barchart visualization



Autoencoder + K-means clustering

Autoencoder

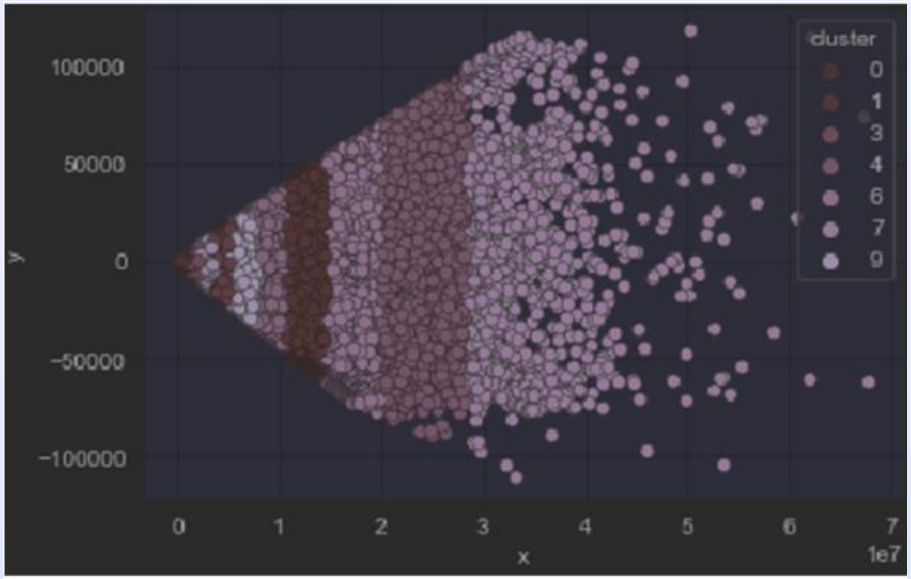
- Compress data into a lower-dimensional representation
- Input Layer: Padded sequence of tokens
- Hidden Layer: 32 neurons
- Output Layer: Reconstruction of the input sequence

K-means clustering

- Input: Latent vectors extract from autoencoder
- Generate 10 clusters

Autoencoder + K-means clustering

Result: Cluster Scatter Plot



Autoencoder + K-means clustering

Result: Silhouette Coefficient Score

```
from sklearn.metrics import silhouette_score
score = silhouette_score(latent_vectors[:5000],cluster_labels[:5000])
print(score)
```

```
0.55109006
```

Evaluation

Evaluation methods:

- **Intrinsic: evaluate using the model's internal structure or data**
- **Extrinsic: evaluate by conducting tasks, such as document classification**

We use Silhouette Coefficient Score to evaluate the model's performance

Model	Score
BERTopics	0.58
Autoencoder + K-means	0.55

Improvement

BERTopic:

- **Use c-TF-IDF instead of the basic TF-IDF to compute the performance of the text dataset**
- **Improve: use the scores which work on a cluster or topic level instead of a document level.**

K-means clustering:

- **Not only use the token sequence generated by Keras Tokenizer**
- **Improve: add a deep learning autoencoder to compress the text token sequence.**

Conclusion

- **LSA, LDA**
- **BERTopic**
- **Autoencoder + K-means clustering**

The performance of BERTopic and Autoencode are good.

Thank you

Q&A