COVID-19 News Clustering

Nan Gao (ngao), James Li (sli64), Xingjian Gu (xgu2), Hanhui Li (hli34)

BROWN

Introduction

Given the global outbreak of COVID-19, we developed a news clustering algorithm that helps viewers identify articles of interest. For data, we collected about 3,000 articles related to COVID-19 from News API. In our algorithm, we adopted an unsupervised approach with PCA, K-means clustering, t-SNE dimensionality reduction and LDA topic modeling techniques.

Methodology

Raw Data: News articles from News API

Preprocessing:

- Step 1: Tokenization, removing irrelevant texts, transforming all words into their roots
- Step 2: Mapping into dictionaries with
  documents: doc_id -> bag of words
  bag of words: word -> word counts
  word_ids: word -> word_id
- Step 3: Formatting into matrix[i][j] representing norm(occurrence) of word_j in doc_i

tf-idf Vectorization:

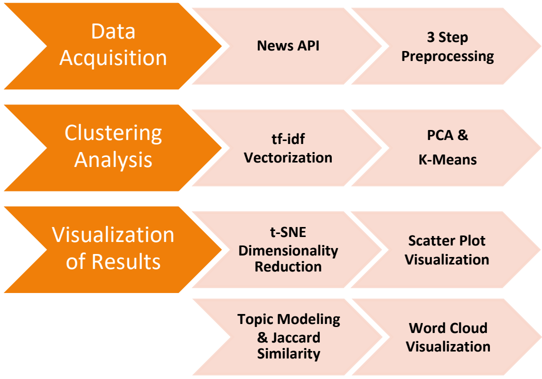
- turning preprocessed data into vectors
- measuring importance of each word
- filtering noise by limiting features to 4000

Principal Component Analysis:

- applying standardization to preprocessed data
- reducing the number of features to 1500
- constructing PCA reduced data container

K-means Clustering:

- initializing centroids and clusters randomly
- updating centroids and clusters by minimizing squared distance of data points and centroids after replacement



Data Acquisition → News API → 3 Step Preprocessing

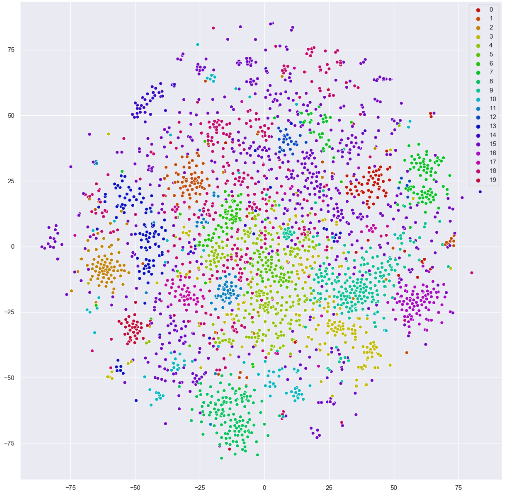
Clustering Analysis → tf-idf Vectorization → PCA & K-Means

Visualization of Results → t-SNE Dimensionality Reduction → Scatter Plot Visualization

Topic Modeling & Jaccard Similarity → Word Cloud Visualization

Pipeline Explained

Results



Scatter Plot Visualization

Interpretation of Scatter Plot

- Using t-SNE, even though the k-means is calculated on a high-dimension space, we plot the points on a 2D plane.
- The model manages to reduce topic difference between articles of same clusters from 0.86 to 0.37

covid
health
symptom
hand
school
said
busi
open
covid
island
said
close
powwow
indigen
peopl
artist
said
toilet
paper
coronavirus

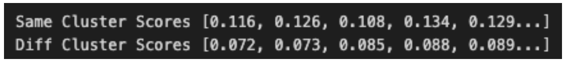
Sample Article 0

Article 1 from Same Cluster

Article 2 from Different Cluster

Article 3 from Different Cluster

Word Cloud Visualization



```
Same Cluster Scores [0.116, 0.126, 0.108, 0.134, 0.129...]
Diff Cluster Scores [0.072, 0.073, 0.085, 0.088, 0.089...]
```

Jaccard Similarity Comparison

Conclusion

Metrics of Success:

- Qualitative: Scatter plot and Word Cloud
- Quantitative: Jaccard Similarity of 0.126 vs 0.061

Motivations and Findings:

- Clustering Articles into Different Topics
- Identifying Topics of Articles Selected
- Adapting Coverage from Different Agencies
- Abstracting information from Topics Selected

Drawbacks and Limitations:

- Only Accept English Inputs
- Unable to Factor Synonyms into Account
- Timely to Train