Project: Singapore News Analysis

Direction 1: Unsupervised Event-Based News Clustering

Task Description

In this task, you will develop an **unsupervised** machine learning model to cluster news stories into **distinct events**. The goal is to automatically group together news articles that describe the same event, despite variations in reporting style and facts included.

- Input: A large corpus of news stories (text articles) from various sources.
- Output: clusters of news stories, each representing a unique event. Each news story should be labeled with an event cluster ID.

News sources often report the same event in varied ways, resulting in minimal textual similarities between some articles about the same event. Conversely, different but similar events, such as annual sports tournaments or concerts by different artists in the same location, may exhibit high textual similarities, complicating the clustering process. Therefore, identifying the most effective features for clustering based on event specifics is crucial for the success of the system.

Data

You may utilize existing academic benchmarks or datasets designed for news clustering as a starting point. Alternatively, consider constructing your own dataset by gathering news articles from diverse sources such as news websites, Wikipedia, Twitter, and blogs. We also provide a Singapore news dataset.

https://drive.google.com/file/d/1yMb1PV-c8bUEx3rQ4PLIfPoYyzkPBHIz/view?usp=sharing

Evaluation

Evaluate the performance of your clustering algorithm using established clustering metrics. Additionally, you are encouraged to employ or define novel evaluation metrics or conduct downstream task assessments to comprehensively evaluate your approach.

Method

A potential starting point could includes the following steps:

1. **Understand Your Dataset**
    - Gather summary statistics of the dataset to comprehend its distribution and characteristics.
    - Visualize the dataset to identify potential patterns or anomalies.
2. **Establish a Baseline**
    - Employ widely used clustering algorithms like K-means, DBSCAN, and hierarchical clustering.

- Calculate metrics on training, development, and test sets.
- Analyze errors to pinpoint the shortcomings of the baseline method.
- Experiment with different hyperparameters and investigate their impact on performance.

3. **Develop and Enhance Your Model**

   - Explore various feature extraction techniques to effectively capture the essence of news stories.
   - Refine your model using insights gained from evaluation metrics and error analysis.

4. **Analysing Singapore News**

   - Upon developing your own model, you can apply it to the Singapore news dataset for in-depth analysis of Singapore-related news and report your findings.
   - E.g. What are the hottest debatable events related to Singapore?
   - Be creative.

Suggested Readings

- Using LLM for Improving Key Event Discovery: Temporal-Guided News Stream Clustering with Event Summaries https://aclanthology.org/2023.findings-emnlp.274.pdf
- Event-Driven News Stream Clustering using Entity-Aware Contextual Embeddings https://aclanthology.org/2021.eacl-main.198.pdf
- BERTopic: https://maartengr.github.io/BERTopic/index.html#quick-start

Direction 2: Opinion Analysis of Media Bias

Task Description

This task focuses on analyzing the sentiment and opinion expressed by different media outlets towards specific Singapore topics and entities. The goal is to identify potential biases or consistent perspectives among various media sources.

- Input: A large corpus of news articles from multiple media outlets, along with a list of target topics or entities.
- Output: Sentiment and opinion scores for each media outlet towards the specified topics or entities.

Note the proposed method should be generalizable to new domains, e.g. Singapore news.

Data

You may utilize existing academic benchmarks or datasets designed for targeted sentiment analysis as a starting point. Alternatively, consider constructing your own dataset by gathering news articles from diverse sources such as news websites, Twitter, Reddit, and Blogs. We also provide a Singapore news dataset.

Evaluation

Evaluate the performance of your sentiment analysis model using established metrics. Additionally, you are encouraged to employ or define novel evaluation metrics or conduct downstream task assessments to comprehensively evaluate your approach.

Method

A potential starting point could includes the following steps:

1. **Understand Your Dataset**
   - Gather summary statistics of the dataset to comprehend its distribution and characteristics.
   - Visualize the dataset to identify potential patterns or anomalies.
2. **Establish a Baseline**
   - Utilize techniques like TF-IDF, word embeddings (Word2Vec, GloVe), or more advanced language models (BERT, RoBERTa) to extract relevant features from the text.
   - Train a sentiment analysis model using the public dataset.
   - Analyze errors to pinpoint the shortcomings of the baseline method.
3. **Develop and Enhance Your Model**
   - Propose improvements to existing models or training data to make it more generalizable, more capable or more efficient.
   - Refine your model using insights gained from evaluation metrics and error analysis.

4. **Analysing Singapore News**
    - Upon developing your own model, you can apply it to the Singapore news dataset for in-depth analysis of Singapore-related news and report your findings.
    - E.g. What are the opinions of the media towards Singapore?
    - Be creative.

Suggested Readings

- NewsMTSC: (Multi-)Target-dependent Sentiment Classification in News Articles: https://www.aclweb.org/anthology/2021.eacl-main.142/
- Multi-Domain Targeted Sentiment Analysis: https://aclanthology.org/2022.naacl-main.198.pdf

Dataset: Singapore News Articles

This project focuses on analyzing news articles related to Singapore. We leverage the **Global Geographic Graph (GGGSG)** from the GDELT Project. We've filtered the data to include English news articles mentioning Singapore from **Apr 4th, 2017** up to **July 19th, 2024** (CountryCode == "SN") from global medias. This filtered dataset, referred to as **GGGSG**, contains **9,185,305 rows**, each representing a mention of a Singapore-related entity. The surrounding text (maximum 600 characters) of the Singapore mention is also included, but converted to lowercase with punctuation removed.

Task Files

The GGGSG dataset is provided as a CSV file named **"ggg_sg.csv"**. The file contains the following columns:

- **DateTime :** (Type: Str) Date and time of the article (UTC). (Example: 2019-04-08 16:15:00+00:00)
- **URL:** (Type: Str) URL of the article. (Example: https://www.photonics.com/Articles/Entanglement-Based_QKD_Could_Secure_Optical_Fiber/a64587)
- **Title:** (Type: Str) Title of the article. (Example: Entanglement - Based QKD Could Secure Optical Fiber Networks Research & Technology Apr 2019)
- **SharingImage:** (Type: Str) URL of the article's thumbnail image (if available). (Example: https://www.photonics.com/images/Web/Articles/2019/4/8/thumbnail_64587.jpg)
- **LangCode:** (Type: Str) Language code of the article. (Example: eng)
- **DocTone:** (Type: Float) Sentiment score of the article. (Example: -0.46082949) - Note that sentiment scores are typically floats.
- **DomainCountryCode:** (Type: Str) Country code of the article's domain. (Example: US)
- **Location:** (Type: Str) Full location string from the article. (Example: National University Of Singapore, Singapore (General), Singapore)

- **Lat:** (Type: Float) Latitude of the mentioned location. (Example: 1.2961)
- **Lon:** (Type: Float) Longitude of the mentioned location. (Example: 103.78)
- **CountryCode:** (Type: Str) Country code of the mentioned location (always "SN" for this dataset). (Example: SN)
- **Adm1Code:** (Type: Str) Administrative level 1 code (may be empty). (Example: SN00)
- **Adm2Code:** (Type: Str) Administrative level 2 code (may be empty). (Example: 18585)
- **GeoType:** (Type: Int) Type of geographic entity. (Example: 4)
- **ContextualText:** (Type: Str) The surrounding text snippet (maximum 600 characters) of the Singapore mention. (Example: together preserving this correlation will help us to create encryption keys faster said researcher james grieve senior research fellow james grieve of the centre for quantum technologies at nus and amelia tan senio...)
- **the_geom:** (Type: Str) Geographic coordinates in Point format (may be empty). (Example: POINT(103.78 1.2961))

Important Notes:

- Due to the size of the dataset, consider using streaming techniques for efficient loading and processing.
- Be aware that some entries in the dataset might be empty.
- A single URL may correspond to multiple rows in the dataset.
- This dataset is machine-annotated using traditional NLP methods to identify Singapore mentions. The annotation quality may not be perfect, so some errors might be present.
- You may further annotate the dataset or generate additional labels with LLMs if needed.
- **Copyright:** While crawling additional information from full articles might be possible, be mindful of copyright restrictions. Academic use is generally fine, but **you cannot publish the actual article content anywhere without owner permission**. Tools like Trafilatura (https://github.com/adbar/trafilatura) can be used to extract text from crawled HTML for further analysis.