Sprinternship 2023

Data Challenge

Problem Statement

The problem to solve is to accurately identify topics from news articles using natural language processing techniques. Furthermore, as a stretch task, we would also like to plot the trending of these topics over time.

We will need to preprocess the text data and extract relevant features from the articles. Preprocessing is essential to transform the raw text (in this case, the article content) into a format that machines can understand human language. Preprocessing often includes normalizing text, removing special characters and stop words, lemmatization, or stemming.

After preprocessing, the next step is selecting an appropriate machine-learning algorithm. For example, topic modeling algorithms, such as a Latent Dirichlet Allocation (LDA), are famous for identifying common topics from a document collection. In LDA, each document is a combination of multiple topics, and each topic is visually represented as a list of words (or phrases).

Once we identify the topics using a topic modeling algorithm, we will analyze the results. The first step is manually eye-balling topics and their associated words and phrases to comprehend their meaning. Then, we will use automatic visualization techniques to plot the trends and patterns in the data. For example, we could use a line chart or a bar chart to show the number of articles that belong to each topic over time. A well-crafted visualization can provide valuable insights and help identify changes and trends in the dataset.

Data:

The dataset contains about 160K English news articles about renewable energy, particularly solar, from November 2015 to March 2020.

Milestones:

Week 1:

- Install python environment and relevant packages
- Download data
- Load and preprocess text data (removing punctuation, stop words, and other irrelevant words, as well as stemming or lemmatizing)
- Test topic modeling library using sample data

Week 2:

- Run topic model on full dataset
- Visualize topics
- Pick your topics: decide on optimal number of topics and assign them labels.

Week 3:

- Stretch task: Plot topics trending
- Prepare and present project report out