ABSTRACT (DESCRIPTION)

The goal of this project was to use create a data engineering pipeline to acquire, handle, query, and process tweets about climate change from the Twitter API. Natural language processing and topic modeling algorithms were used to tokenize tweets and uncover key topic terms/themes that would have otherwise go unnoticed without the assistance of machine learning.

DESIGN

Project design consisted of several parts:

1. Collecting Twitter user tweet data about “climate change” and “global warming” using Tweepy, an open-sourced library used to access the Twitter API directly.
2. A containerized (Docker) version of Spark was used to handle the large dataset (900K+ tweets) and to execute SQL queries relevant to preprocessing and modeling
3. Additional EDA, transformation, and cleaning including: text preprocessing and word tokenization with NLP libraries.
4. Employ dimensionality reduction algorithms and topic modeling on processed tweet data to identify relevant topic terms and ideas.
5. And visualizing the analysis in a way that makes sense to the shareholders and stakeholders.

DATA

Over 900,000 tweets (documents) were collected via Tweepy and a custom user-defined python function. Approximately ~5.4 GB of text data, stored in a JSON (JavaScript Object Notation) format, containing features including: username, location, tweet text, unique tweet ID, retweet count, favorite count, and much too many more to name. Leveraging Docker and Spark (SQL and Pandas) allowed me to query and create SQL tables for further data cleaning, processing, and analysis. The data set is available upon request, as the files were too large to commit and push onto my Github repo.

ALGORITHMS

Unsupervised Learning Model

- Dimensionality reduction
- Latent Dirichlet Allocation Topic Modeling

TOOLS

Tweepy
- Python library for accessing the Twitter API

Docker
- Containerization platform to deploy applications without the need to install libraries or dependencies locally

Spark
- Database creation, data handling, and data transformation
- Database querying and manipulation

SQL
- Storage tools to query unstructured databases

Pandas and NumPy
- EDA, data cleaning and manipulation

NLTK, spaCY,
- Text preprocessing and tokenization: NLTK, spaCY

VADER
- Text Sentiment Analysis: VADER

Scikit-Learn
- Text Document Matrix Transformation

Gensim
- Text Topic Modeling

Matplotlib, Seaborn, pyLDAvis, WordCloud
- (Interactive) data visualization

COMMUNICATION

In addition to the slides and visuals presented, all work is available on my Github found here.