NATURAL LANGUAGE PROCESSING OF TWEETS REVOLVING COP26 - UN CLIMATE CHANGE SUMMIT 2021

Amir Khoeilar

Abstract

COP26 chi stand for 'Conference of the Parties', is a conference held every 5 years in order raise awareness towards climate change and also for governments to present their achievements towards clean energy and future goals. This year it has been catching a lot of attention beach many countries have failed with their promises and also unfortunate climate related incidents have been happening much more. In order to find out what people around the world are talking about in regards of this summit, I'll be using NLP to analyze the tweets surrounding this event (before and during). With this analysis, we could get a grasp of how engaged people have been with this summit, what are their demands mostly, and also their satisfaction of achievements and discussions that happens daily during the conference and overall.

Design

I scraped my data using Snscrape, which gave an organized data frame that separated each tweet with respect to time, user, likes and etc. I used the bag of words method just to get a baseline model for the tweets. After, I began with my pre-processing of data. Since the file was in json, I had to somehow extract some of the data out of the data frame like for example the "User ID". Eventually, I started with using regex to clean data with punctuation and the steps after like tokenization, lemmatization and stemming in order to simplify the words in each tweets to the most possible way. For modeling, I sued the TF-IDF and NMF model to get the tweet topic matrix and also the word topic matrix and also to get the top words and tweets related to the 15 components that I chose for my NMF model. As for my last step for EDA, I used Seabron to graph horizontal bars to show the relative relation of words in each topic. Also used Word Clouds just for extra visualization.

Data

The data provided has 10000 tweets that after cleaning duplicates and also choosing only English tweets, it narrowed down to 5600. Each tweets has limit of 280 characters, tarts why I need to grab more data in order to have batter analysis and model.

Algorithms

Feature Engineering

1. Snscrape to scrape from twitter
2. Pre-processing using regex, lumpy, langdetect, schikitlearn and NLTK
3. Seaboard, word cloud and matplotlib for visualization

Model Evaluation and Selection

In the conclusion, using NLP modeling, I could process and separate the twitter data into 15 topics which mostly made sense and was possible to caption each topic with regards to the terms and also the tweets.

Overall, This process gave and oversight of all the conversations that people we and are having about the conference and this could be a good tool for governments to see how people fell about their decisions in the COP26.

Tools

- Snscrape , os
- Pandas, lumpy, regex
- Seaborn , marplotlib , word clouds
- Scikit-learn, NLTK

Communication

There is going to be the code for the data operations and also PDF slides of powerpoint presentation available on my GitHub account.

https://github.com/rezxkoi/NLP_united_nations_cop21_tweeter