Climate Change Text Analysis Project - Documentation

# 1. Project Overview

The Climate Change Text Analysis project is designed to analyze climate-related textual data, specifically from NASA's dataset. It leverages Natural Language Processing (NLP) techniques, sentiment analysis, and topic modeling to extract meaningful insights from public discussions regarding climate change.  
  
Objectives:  
- Clean and preprocess climate-related text data.  
- Analyze sentiment to determine public perception.  
- Identify main topics and themes through LDA topic modeling.  
- Visualize sentiment distribution and topics for better understanding.

# 2. Tools and Technologies

- Python 3.11  
- Pandas: For data loading, manipulation, and preprocessing.  
- NLTK: Tokenization, stopwords removal, lemmatization, sentiment analysis (VADER).  
- Matplotlib: Visualizations for sentiment distribution and word clouds.  
- WordCloud: To create graphical word representations of topics.  
- Gensim: LDA (Latent Dirichlet Allocation) topic modeling.  
- Warnings module: To suppress unnecessary warnings.

# 3. Data Preprocessing

Steps:  
1. Load Dataset: Load CSV dataset (`climate\_nasa.csv`).  
2. Text Cleaning: Remove URLs, mentions, hashtags, punctuation, numbers; convert text to lowercase.  
3. Tokenization: Split sentences into words.  
4. Stopwords Removal: Remove common English stopwords.  
5. Lemmatization: Reduce words to their root forms.  
6. Processed Column: Store the cleaned and tokenized text in `processed\_text`.  
  
Code Snippet:  
import re  
from nltk.tokenize import word\_tokenize  
from nltk.corpus import stopwords  
from nltk.stem import WordNetLemmatizer  
  
stop\_words = set(stopwords.words('english'))  
lemmatizer = WordNetLemmatizer()  
  
def preprocess\_text(text):  
 text = str(text).lower()  
 text = re.sub(r'http\S+|www\S+|https\S+|@\w+|#\w+', '', text)  
 text = re.sub(r'[^\w\s]', '', text)  
 text = re.sub(r'\d+', '', text)  
 tokens = word\_tokenize(text)  
 tokens = [word for word in tokens if word not in stop\_words]  
 return [lemmatizer.lemmatize(word) for word in tokens]

# 4. Sentiment Analysis

- Tool: VADER Sentiment Analyzer from NLTK.  
- Process: Compute polarity scores and classify as Positive, Negative, or Neutral.  
- Visualization: Bar chart showing sentiment counts.  
  
Example Code:  
from nltk.sentiment.vader import SentimentIntensityAnalyzer  
import matplotlib.pyplot as plt  
  
sid = SentimentIntensityAnalyzer()  
  
def get\_vader\_sentiment(tokens):  
 if not tokens:  
 return 'Neutral'  
 scores = sid.polarity\_scores(' '.join(tokens))  
 if scores['compound'] >= 0.05:  
 return 'Positive'  
 elif scores['compound'] <= -0.05:  
 return 'Negative'  
 else:  
 return 'Neutral'

# 5. Topic Modeling

- Technique: LDA using Gensim.  
- Process: Create dictionary, convert to bag-of-words, train LDA model.  
- Output: List of topics with top keywords.  
- Visualization: Word clouds for each topic.  
  
Example Code:  
from gensim.corpora import Dictionary  
from gensim.models import LdaModel  
from wordcloud import WordCloud  
  
dictionary = Dictionary(df['processed\_text'])  
dictionary.filter\_extremes(no\_below=5, no\_above=0.5)  
corpus = [dictionary.doc2bow(text) for text in df['processed\_text']]  
lda\_model = LdaModel(corpus=corpus, id2word=dictionary, num\_topics=5, passes=10, alpha='auto')

# 6. Results

- Sentiment Distribution: Percentage of Positive, Negative, Neutral comments.  
- Topics Discovered: Top keywords per topic (e.g., climate, renewable, sea-level, temperature).  
- Word Clouds: Graphical representation of top words per topic.  
- Sample Data: Shows processed text and sentiment.

# 7. Applications

- Understand public perception of climate change.  
- Identify trending environmental themes.  
- Support research, policy-making, and awareness campaigns.  
- Enhance visualization for reports and presentations.

# 8. Conclusion

This project demonstrates the application of NLP techniques on climate-related textual data. By combining preprocessing, sentiment analysis, and topic modeling, the project provides a comprehensive view of public discussions on climate change and visual insights for better decision-making and awareness.