Natural Language Processing for Advancing Climate Change Assessment and Policy Decision Making

Dibya Pandey

Naveen Donthula

Snigdha Chigurupati

MPS – Data Science

DATA 690 – NLP for Practitioners

Prof. Antonio Diana

Date: August 16th, 2023

ABSTRACT

Climate change represents a pressing global challenge requiring the analysis of vast amounts of scientific data to guide mitigation and adaptation policies. This research applies natural language processing techniques including summarization, sentiment analysis, named entity recognition, and topic modeling to extract insights from a corpus of US climate data. The project implements an unsupervised extractive summarization algorithm to generate condensed overviews highlighting key information. Further contextual analysis identifies sentiments, entities, relationships, and latent topics. Visualizations like word clouds and knowledge graphs reveal linguistic patterns. Hierarchical clustering uncovers thematic structures within the documents. An ensemble model predicts sentiment from the summaries to efficiently gauge perspectives at scale. The methodologies provide both high-level summaries and a granular analytical understanding of the complex climate dataset. The techniques demonstrate the value of NLP for accelerating discovery and comprehension in climate science. This enables researchers to rapidly synthesize disparate evidence and inform timely, data-driven policy decisions. The project illustrates a pipeline integrating summarization, contextual analysis, relationship extraction, and predictive modeling to unlock transformative knowledge from massive textual data.

INTRODUCTION

This project aimed to apply natural language processing techniques to analyze and extract key insights from a large corpus of US climate data. The objectives were to:

1. Provide concise overviews of the climate data through summarization.
2. Identify contextual details like sentiments, entities, and topics.
3. Discover relationships and patterns in the data.
4. Build predictive models using the extracted information.

The techniques delivered both high-level summarized views as well as a deeper contextual analysis of the complex climate dataset.

THE IMPORTANCE OF TEXT SUMMARIZATION

Text summarization, the technique of shortening long pieces of text while retaining the overall meaning, has become extremely valuable in the modern information age. With the explosion of online content, social media, and written information in all spheres of life, the ability to quickly distill the key ideas and salient points from lengthy documents saves a great amount of time and effort. Beyond direct time savings, it can serve as an unbiased tool for gathering information, better than biased approaches like surveys, focus groups, manually curated Q&A, or interviews. By digesting textual data from diverse sources over long time periods, it eliminates recency bias and provides objective summaries. Text summarization provides tremendous value in efficiency and understanding across many professional domains. Whether it is summarizing research papers, news articles, legal contracts, medical records, or any form of communication, text summarization provides an efficient way to filter out the noise and focus only on the core ideas. For professionals like lawyers, doctors, researchers, analysts, and journalists who have to sift through massive amounts of text daily as part of their work, automated summarization tools help accelerate understanding and decision-making by pinpointing the essence of the material. For portfolio managers in finance, digesting earning reports and financial documents into their main takeaways accelerates investment decision-making. Academic researchers rely on summarization to grasp the essence of long research papers and extract key insights.

With artificial intelligence techniques like deep learning advancing the capabilities of text summarization, what was once only manually possible is now feasible automatically. This assists professionals, students, businesses, and ordinary people to save time and effort while improving the consumption and

transmission of ideas and information in modern life full of digital text. The automated nature of text summarization makes it more neutral than manually curated Q&A or interviews. The wide applicability across domains combined with its unbiased nature make text summarization an invaluable asset for productivity and reliable insights in the modern data-driven world.[1]

THE IMPORTANCE OF CLIMATE CHANGE ANALYSIS FOR ADAPTATION POLICY

Climate change represents one of the most crucial problems facing humanity today. It is predominantly driven by greenhouse gas emissions from human activities like fossil fuel combustion, deforestation, and agricultural practices. The resulting greenhouse gas accumulation in the atmosphere traps heat and causes global temperatures to rise. Greenhouse gas emissions continue rising globally, risking catastrophic climate change impacts if unchecked. Temperature and precipitation patterns are already changing, affecting water resources, agriculture, ecosystems, and human health. Climate change exacerbates other societal risks like poverty, migration, and conflict. Tackling climate change requires analyzing immense amounts of scientific data on temperature trends, greenhouse gas levels, sea level rise, and extreme weather events. Localized, regional analysis provides actionable insights for officials to develop targeted, context-specific climate policies and strategies. Climate researchers must synthesize insights from this massive number of peer-reviewed papers, climate model projections, time-series measurements, and policy reports spanning multiple decades.[2]

ABOUT THE DATA

The data is collected through various digital prints like ‘The New York Times’, ‘Seattle Washington Post’, and ‘Salt Lake Tribune’, etc. across different geographic regions in the US covering the West, Central, and Eastern regions to understand national and regional variations. It is critical to collect localized regional data because climate change impacts manifest differently across areas due to variances like proximity to coasts, elevation, urban density, and more. Having fine-grained data from diverse sites across the West, Central, and Eastern US allows analysts to model regional variations in rising temperatures, sea level impact, precipitation changes, and extreme weather. These regional insights allow state and city officials to plan localized adaptation strategies.

METHODOLOGY

1. **Data Collection**

   Climate data was collected from multiple sites across the Western, Central, and Eastern US to enable a regional analysis of climate change impacts based on factors like proximity to coasts, elevation, urbanization, etc. The climate data corpus encompasses approximately 3000-4000 words of textual content per region, spanning the Western, Central, and Eastern United States.

2. **Data pre-processing**

   Text preprocessing like removal of stop words, punctuation, and lemmatization normalized the data by filtering out non-essential words and reducing remaining words to their root forms. This improved matching and analysis.

3. **Text Summarization**

   The project implements an unsupervised extractive text summarization technique to distill key information from large texts into concise overviews. Text summarization is essential for obtaining representative summaries covering the main topics and diverse aspects of a source text.

   The approach first preprocesses the input text through steps like lowercasing, lemmatization, and stop word removal to normalize the text and focus on meaningful keywords. Word frequencies are then generated and normalized to determine the information content of each non-stop word.

   Sentence importance scores are calculated by aggregating normalized word frequencies. This enables quantitatively identifying sentences containing more salient information. By scoring sentences based on collective word weights, the model accounts for informative keywords that may be distributed across different parts of a sentence.

   The core summarization function extracts the top N-ranking sentences by importance into the final summary. By extracting sentences directly from the source, the model retains key details and diverse coverage of topics in a condensed overview.

   This unsupervised technique distills informative content relevant to the main themes and discourse in the text. The data-driven word frequency analysis identifies key details in an impartial manner without manual supervision or labeling. The extracted sentences provide a representative summary sample covering the salient information in the source text.

4. **Sentiment analysis and subjectivity detection**

   Sentiment analysis and subjectivity detection are critical natural language processing techniques for extracting effective information from text data. Sentiment analysis involves computationally

identifying and categorizing opinions in the text to determine the writer's attitude as positive, negative, or neutral. This is often performed by using lexicons or machine learning algorithms that associate words and phrases with sentiment orientation. Sentiment analysis provides crucial insights for applications such as brand monitoring, analyzing customer satisfaction, and gauging public opinion. The project utilizes the VADER sentiment analysis tool to discern the emotional tone of the generated summary.

Subjectivity detection refers to identifying text that expresses opinions versus objective facts. This can be achieved by using methods like Text Blob to calculate a subjectivity score ranging from 0 to 1, with scores below 0.5 indicating more objective language. Determining the degree of subjectivity within the text is beneficial for tasks like separating factual reporting from opinionated editorials or filtering out objective from subjective content. Overall, sentiment analysis and subjectivity detection allow for a deeper understanding of the affective content and perspectives expressed in textual data. These techniques provide actionable insights across many domains including business, politics, and social science.[3]

5. **Named entity recognition (NER)**

   NER is a critical natural language processing technique that automatically identifies and categorizes key nouns and proper names within unstructured text. NER typically leverages machine learning algorithms within NLP libraries like spaCy to label extracted entities as Person, Location, Organization, etc. Performing NER enables sophisticated information extraction by structuring text data around detectable entities. This provides an avenue for numerous downstream applications such as building knowledge graphs, question-answering systems, semantic search, and more. Overall, implementing NER is essential for unleashing the value of text data through enabling relation extraction, knowledge base construction, and advanced text analytics. The ability to automatically tag and classify named entities facilitates richer text understanding and allows organizations to derive enhanced business insights from unstructured data.[4]

6. **Topic modeling**

   Topic modeling is an unsupervised machine learning technique that reveals latent thematic structure within a collection of documents. Our project demonstrates topic modeling that is effectively implemented in Python using sklearn and NMF (non-negative matrix factorization). The steps involved are tokenizing the input text, converting it to a tf-idf weighted term-

document matrix, applying NMF to factorize the matrix, and extracting the topics using the resulting term-topic and topic-document matrices.

Automatically discovering topics using algorithms like NMF provides crucial insights that would otherwise require extensive human analysis. Extracting semantic topics allows for a more nuanced understanding of document contents, trends, and relationships. This enables fruitful applications in fields like digital humanities, social science, marketing, and more. For example, topic modeling can help cluster news articles and customer feedback by subject, guide recommendations by interest, and summarize large volumes of text. Implementing topic modeling, as illustrated in the sample code, is a vital NLP technique for simplifying, navigating, and gleaning strategic knowledge from massive text corpora.[5]

7. **Word frequency analysis**

   Word frequency analysis provides simple yet valuable text mining insights by identifying the most common words and phrases in a document, this can be achieved by tokenizing the text, filtering out stop words and non-alphanumeric, counting word frequencies using Python's Counter, and visualizing the results as a word cloud.

   Implementing basic word frequency analysis is a key NLP technique for understanding text content and highlighting top terms and themes. The project demonstrates an efficient approach by leveraging NLTK for tokenization and filtration along with data visualization libraries like Matplotlib and word cloud. The results reveal the core topics and concepts discussed within the text in an intuitive graphical format.

   More broadly, identifying word and phrase frequencies enables the extraction of key talking points and summarized content. It complements and provides an entry point for more advanced NLP tasks like named entity recognition, topic modeling, and sentiment analysis. Word frequency analysis is especially useful for digesting long reports or articles, clustering documents by keywords, and monitoring shifts in terminology over time. This form of simple statistical NLP provides tremendous utility for content analysis and text mining with versatile applications across many domains.

8. **Word embeddings and word cloud**

   Word embeddings represent words as high-dimensional numeric vectors that encode semantic meaning based on contextual usage. The project demonstrates an effective approach for training Word2Vec embeddings on text, reducing dimensionality with t-SNE, and visualizing the embeddings in a 2D scatter plot.

Specifically, tokenizing the text, removing stop words, training a Word2Vec model to generate word vectors, and selecting the embeddings for chosen keywords. t-SNE is then utilized to reduce the high-dimensional vectors to 2D for visualization. The resulting scatter plot provides an intuitive graphical view of how words are distributed based on semantic similarity. Words with related meanings cluster together.

Visualizing embeddings is tremendously useful for understanding model behavior and analyzing linguistic relationships. Best practices are implemented for visually exploring how words are encoded by the model, revealing nuances and connections. More broadly, creating and inspecting word embeddings is a vital NLP technique for tasks like identifying analogies, powering search systems, improving recommendations, and supporting transfer learning. The code demonstrates an effective approach to training, dimensional reduction, and visualization of word vectors for fruitful text analysis.[6]

9. **Knowledge graph:**

Knowledge graphs provide an intuitive way to represent connections and relationships between entities extracted from text. spaCy's named entity recognition can identify key nouns and proper names, which become nodes in the graph. The find_relationship function implemented in the project then detects verbs and prepositions linking entities to infer semantic relationships. These are encoded as labeled edges in a directed NetworkX graph.

Visualizing the knowledge graph with Matplotlib highlights how entities are related based on their co-occurrence and interconnections within the text. This enables both a high-level overview of key entities and granular insights into specific relationships. Constructing knowledge graphs is tremendously valuable for unlocking and representing knowledge within documents. The graphs not only summarize entities and connections but also reveal deeper semantics that may not be directly stated.

More broadly, knowledge graphs are a vital AI technique for tasks like question answering, recommendation systems, searching, and identifying gaps or inconsistencies in knowledge. By encoding entities and relationships in an accessible graph format, knowledge graphs power more nuanced text understanding. The project demonstrates a practical approach to distilling unstructured text into a structured knowledge network. Automatically constructing knowledge graphs from text is crucial for connecting information and deriving actionable insights.[7]

10. **Hierarchical Clustering:**

    Hierarchical clustering is an unsupervised learning technique for identifying latent groups and structures within unlabeled data. The provided code demonstrates an effective implementation for text data using scikit-learn's Agglomerative Clustering. The keywords from the text summary are first converted into numerical features with CountVectorizer. Agglomerative clustering is then applied to group the keywords based on similarity. The resulting dendrogram visualization provides intuition about document clustering.

    Hierarchical clustering is immensely valuable for digesting and navigating large collections of text. By identifying semantic clusters, related content can be discovered without needing explicit categorization or labels. The code implemented has best practices like TF-IDF vectorization and Ward's linkage method to extract meaningful hierarchies for text mining. The dendrogram intuitively displays the nested groupings, providing visualization of how the algorithm partitions the content.

    More broadly, hierarchical clustering helps uncover thematic structures, relationships and overlaps within-corpus data. The technique can power applications like document organization, recommendation systems, search, and identifying patterns. Implementing hierarchical clustering and inspecting the dendrograms enables nuanced text analysis without supervision. The project demonstrates a robust unsupervised learning approach for clustering text data and extracting insights.[8]

11. **Predictive analytics based on sentiment.**

    Sentiment analysis on automatically generated extractive summaries provides an effective pipeline for analyzing affect and perspectives within large text corpora. Recent research has explored predictive modeling approaches using supervised learning algorithms to classify sentiment based on summarized content. Key steps in this process include:

    1) Generating a concise summary covering the most salient information. This reduces computational expenses while retaining indicative content.

    2) Processing the summary sentences with spaCy for polarity scoring using VADER or other lexicons.

    3) Structuring the data and training an ensemble model stacking logistic regression, random forests, SVMs, and a logistic regression meta-learner.

    4) Evaluating model performance using cross-validation and metrics like accuracy, precision, recall, and F1 on held-out test data.

Predictive sentiment analysis on summaries has been shown to achieve high performance while minimizing computational costs for large corpora. This allows efficient analysis of opinions and effects at scale. Applications include tracking brand sentiment across news and social media, analyzing customer feedback, and monitoring public discourse. The code implemented in the project demonstrates effective implementation of this predictive text analytics pipeline using standard NLP libraries and ensemble modeling. Further research can explore the impact of different summarization techniques and model architectures. Overall, this technique provides versatile capabilities for generating actionable insights from text data.[9]

12. **Cosine similarity between summaries**

This research implements cosine similarity to compare semantic closeness between regional climate documents. The text data from Western, Central, and Eastern US is preprocessed by lowercasing, removing punctuation, and stemming words to their root forms. Scikit-learn's TfidfVectorizer then converts the normalized text into TF-IDF vector representations encapsulating word frequencies. Cosine similarity is calculated between vector pairs using matrix multiplication to quantify similarity based on the angle between vectors.

Computing similarity enables discovering relationships within the corpus to identify patterns, trends, and variations across regions. The code demonstrates a robust process for textual similarity analysis using TF-IDF vectorization and cosine similarity. This unsupervised technique is foundational for numerous applications in information retrieval, natural language processing, and text analytics. Overall, cosine similarity provides an efficient method to compare documents based on semantic content rather than surface-level features. Assessing similarity is crucial for clustering documents, improving search relevancy, analyzing trends, and gaining insights from unstructured text data.[10]

RESULTS AND DISCUSSION

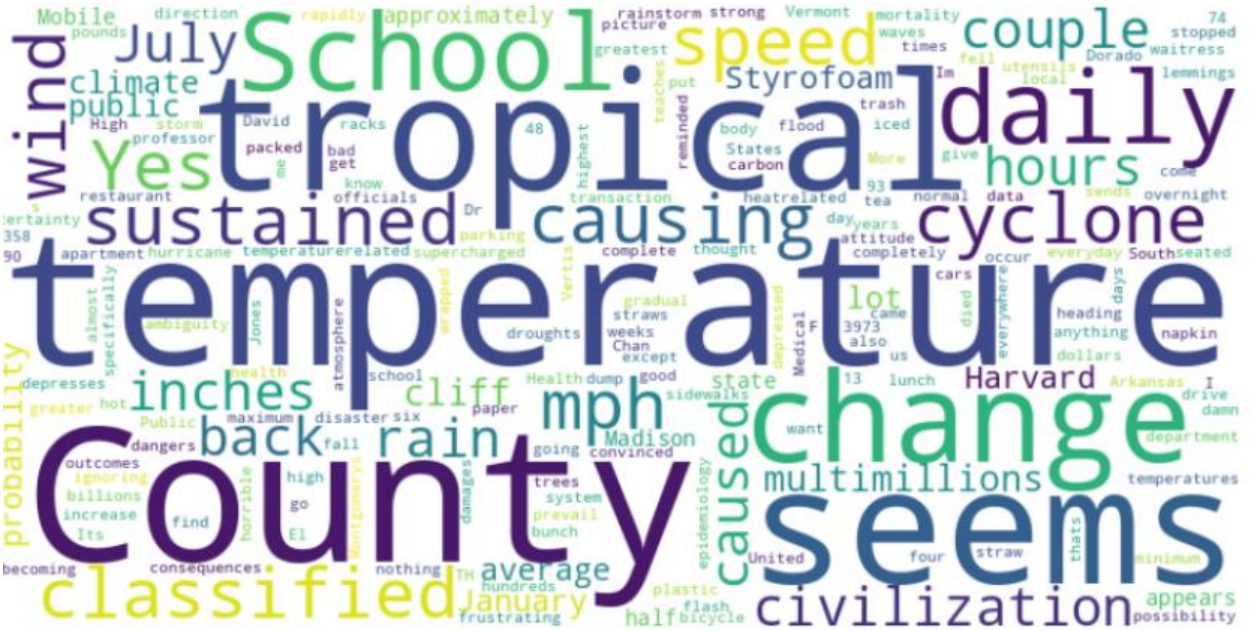
	West	Central	East
Sentiment Analysis	{'neg': 0.063, 'neu': 0.871, 'pos': 0.066, 'compound': 0.0258} Subjectivity: 0.479284474206349 The text is more objective.	{'neg': 0.095, 'neu': 0.832, 'pos': 0.072, 'compound': -0.9463} Subjectivity: 0.38706589706589706 The text is more objective.	{'neg': 0.027, 'neu': 0.882, 'pos': 0.09, 'compound': 0.9543} Subjectivity: 0.3870967741935484 The text is more objective.
Percentages Of Sentiment	The percentage of positive sentiment is 7.0% The percentage of neutral sentiment is **87.3%** The percentage of negative sentiment is 5.7%	The percentage of positive sentiment is 7.2% The percentage of neutral sentiment is **83.2%** The percentage of negative sentiment is 9.5%	The percentage of positive sentiment is 9.0% The percentage of neutral sentiment is **88.2%** The percentage of negative sentiment is 2.7%
Named Entities	South African - NORP Roberts - PERSON Monday - DATE Phoenix - GPE 25 straight days - DATE	39–73 mph - QUANTITY 74 mph - QUANTITY a couple of weeks back - DATE July - DATE South Arkansas - GPE	July - DATE 16.95 degrees - QUANTITY 62.51 degrees - QUANTITY Fahrenheit - WORK_OF_ART a third - CARDINAL
Topic Modeling	Topic 1: degrees (3.33) climate (2.66) report (2.00) fahrenheit (2.00) celsius (2.00)) Topic 2: fan (14.73) blows (14.73) water (14.73) spraying (14.73) mean (14.73) Topic 3: chair (8.99) south (8.77) opt (8.75) single (8.26)	Topic 1: county (6.18) daily (4.71) temperature (4.32) state (3.03) lot (2.89) Topic 2: tropical (2.34) probability (1.56) civilization (1.56) cyclone (1.56) wind (1.56) Topic 3: high (2.99) approximately (2.48) market (2.38) highest (2.36) sends (2.31)	Topic 1: degree (4.90) july (4.70) record (4.40) celsius (3.96) fahrenheit (3.05) Topic 2: hottest (3.37) wide (2.68) 2019 (2.46) london (2.41) colleagues (2.36) Topic 3: heat (2.56) climate (2.43) quite (2.05) chance (1.95) park (1.91)

	community (8.02)		
Word Frequency Analysis	'degrees': 5, 'climate': 4, 'Fahrenheit': 3, 'report':3, 'Celsius':2	'temperature': 4, 'County': 4, 'tropical': 3, 'change': 3, 'School': 3	'record': 6, 'Celsius': 4, 'degree': 4, 'climate': 4, 'Julys': 3
Cosine Similarity	For East and West : 0.7984706278521869 For East and Central:0.5370735639592626 For West and Central:0.5073597921336895		
Predictive Analysis	Accuracy: 0.8 Precision: 1.0 Recall: 0.8 F1 Score: 0.88	Accuracy: 0.87 Precision: 0.91 Recall: 0.875 F1 Score: 0.87	Accuracy: 0.91 Precision: 0.93 Recall: 0.91 F1 Score: 0.92

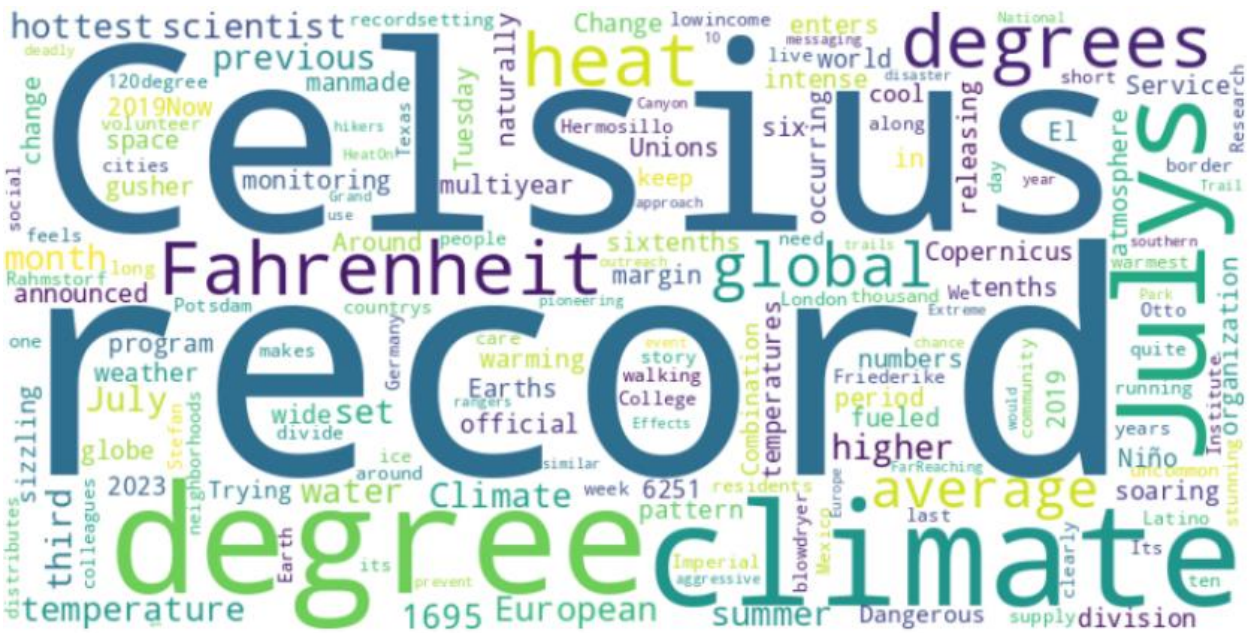
Word Cloud

<u>West</u>

Central

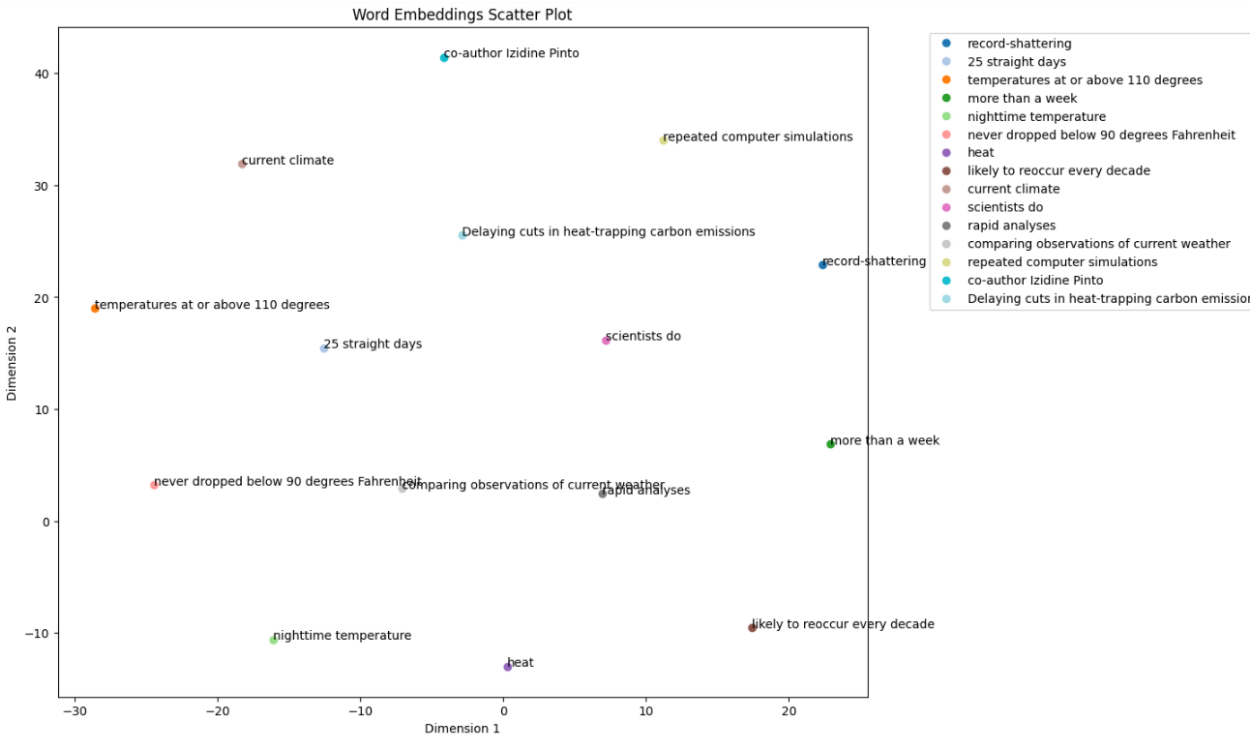


East

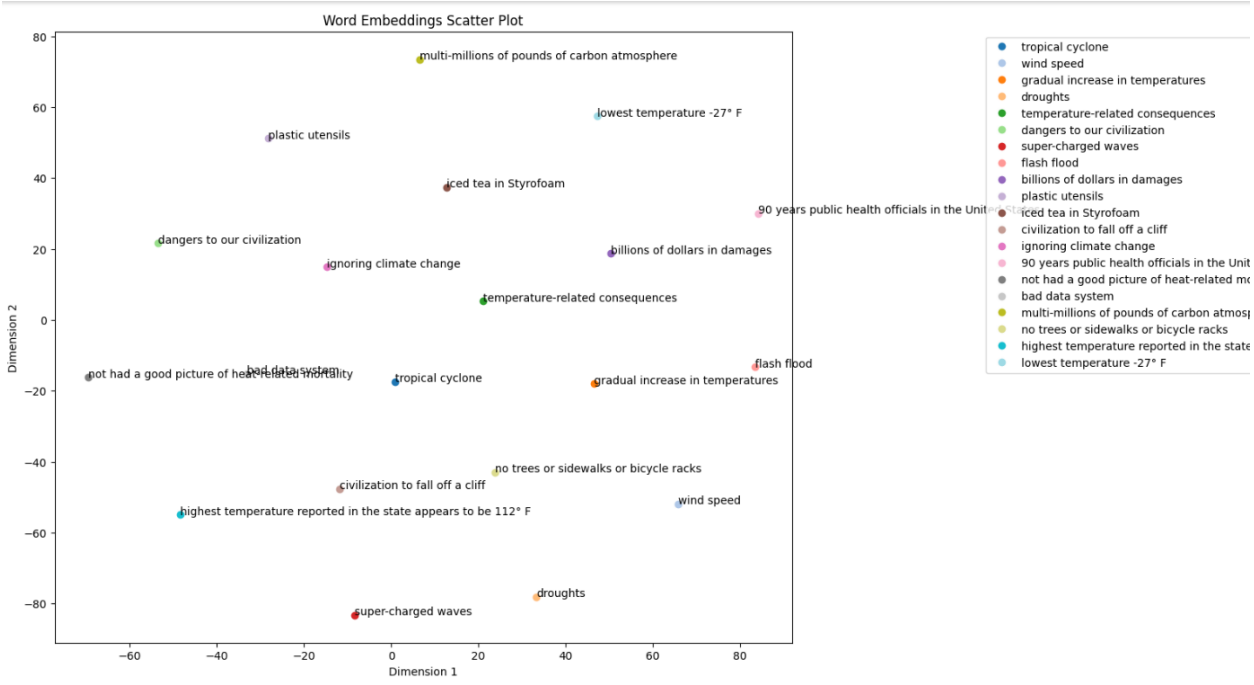


Word Embeddings t-SNE scatterplot

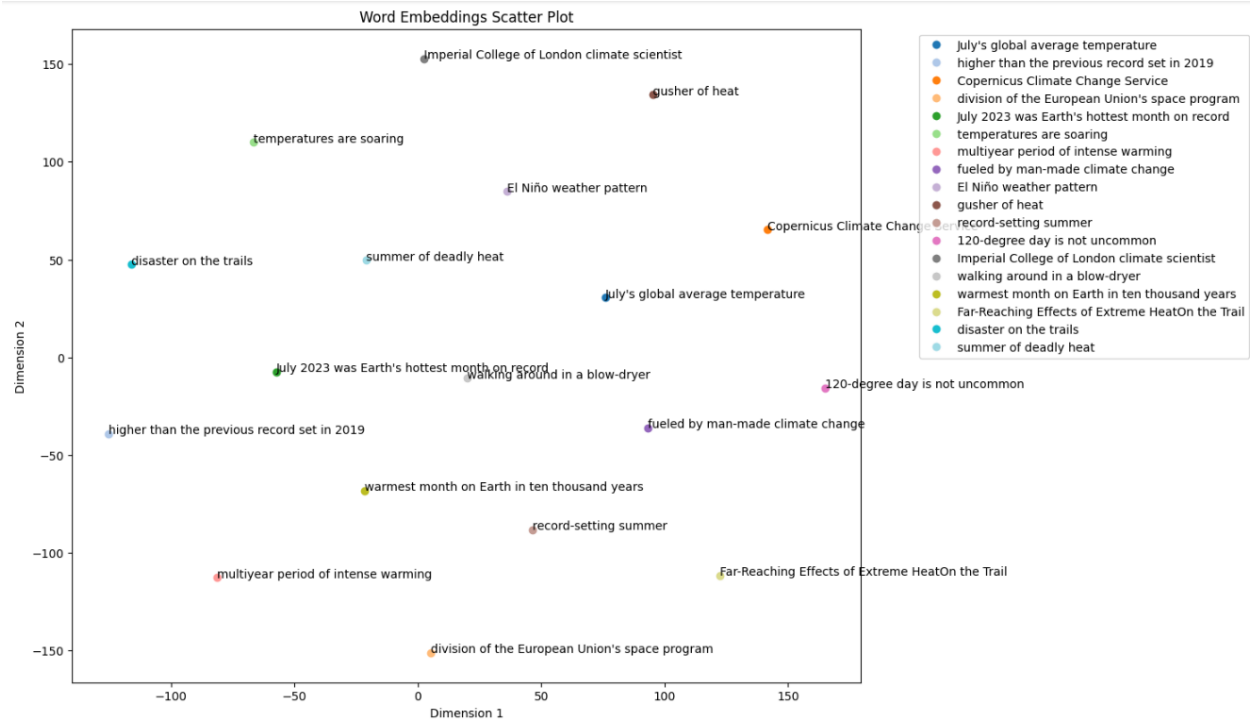
West



Central

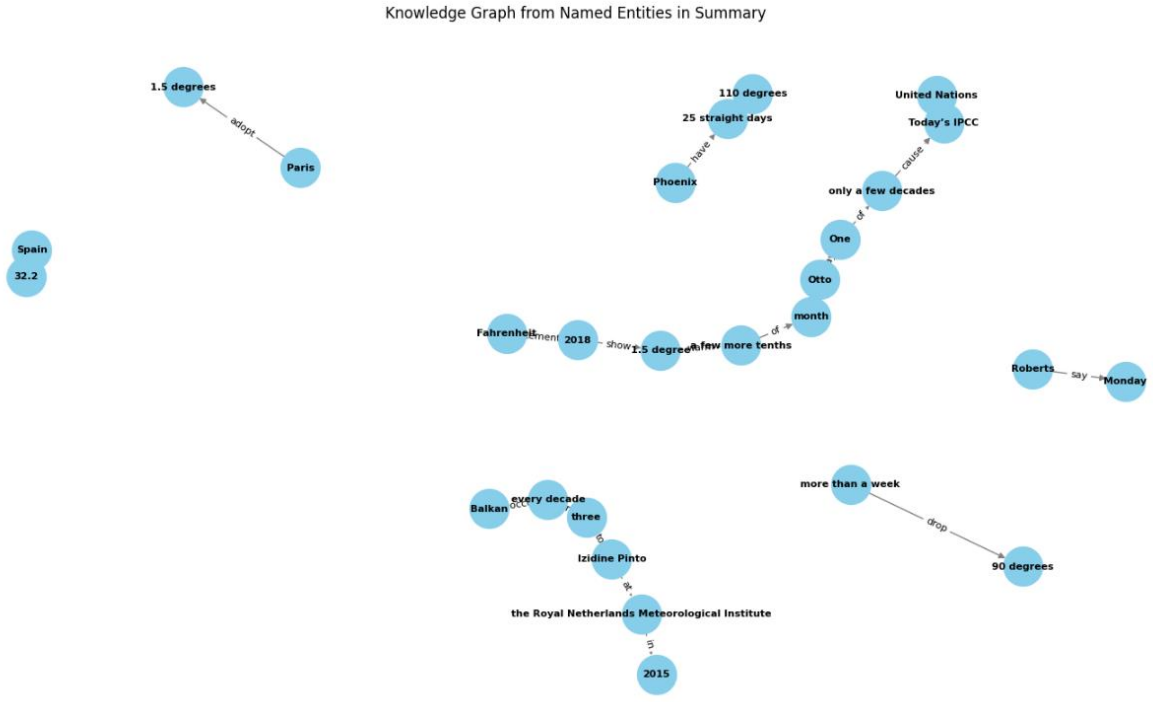


East

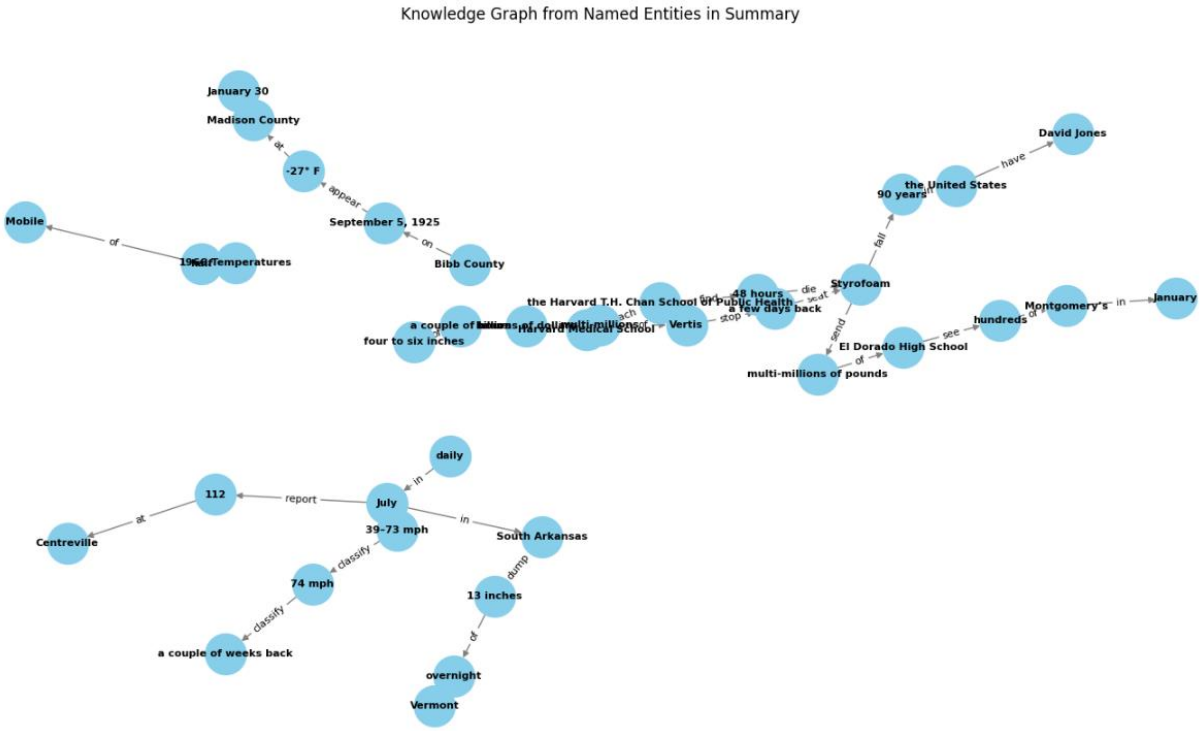


Knowledge Graph

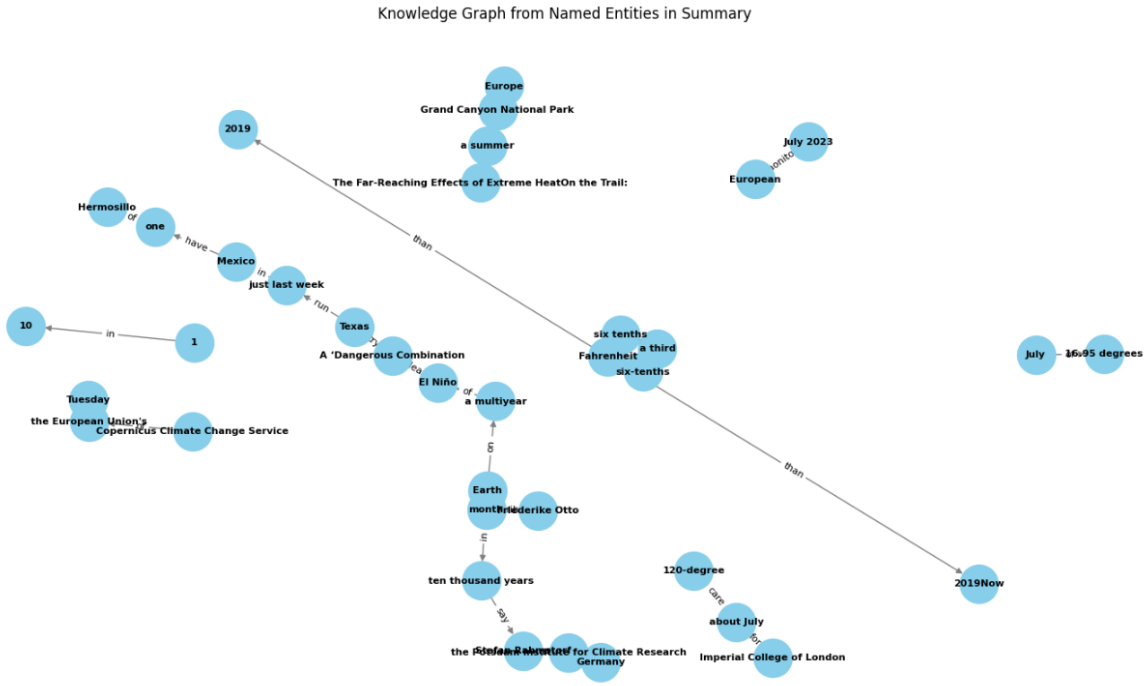
West



Central

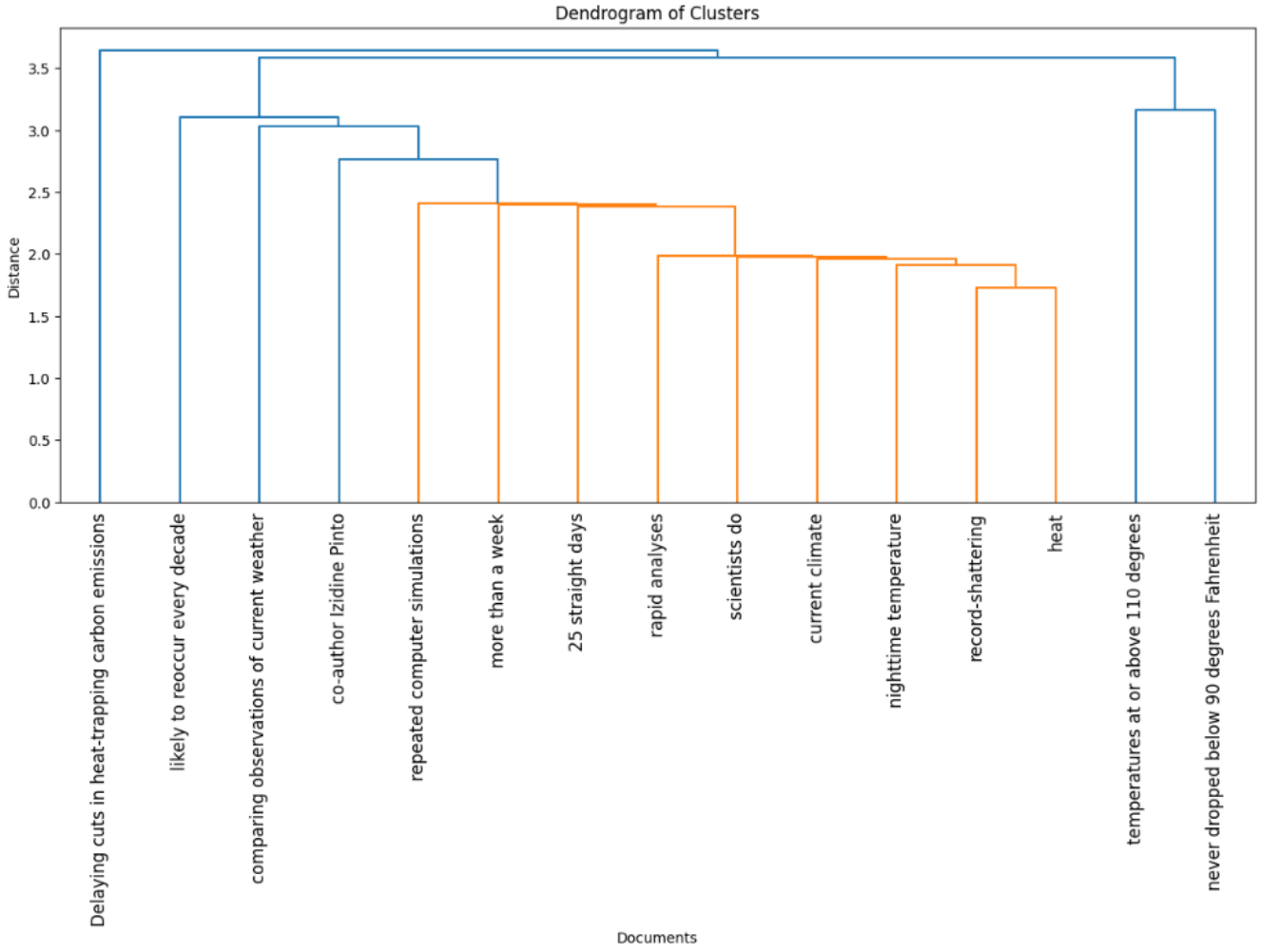


East



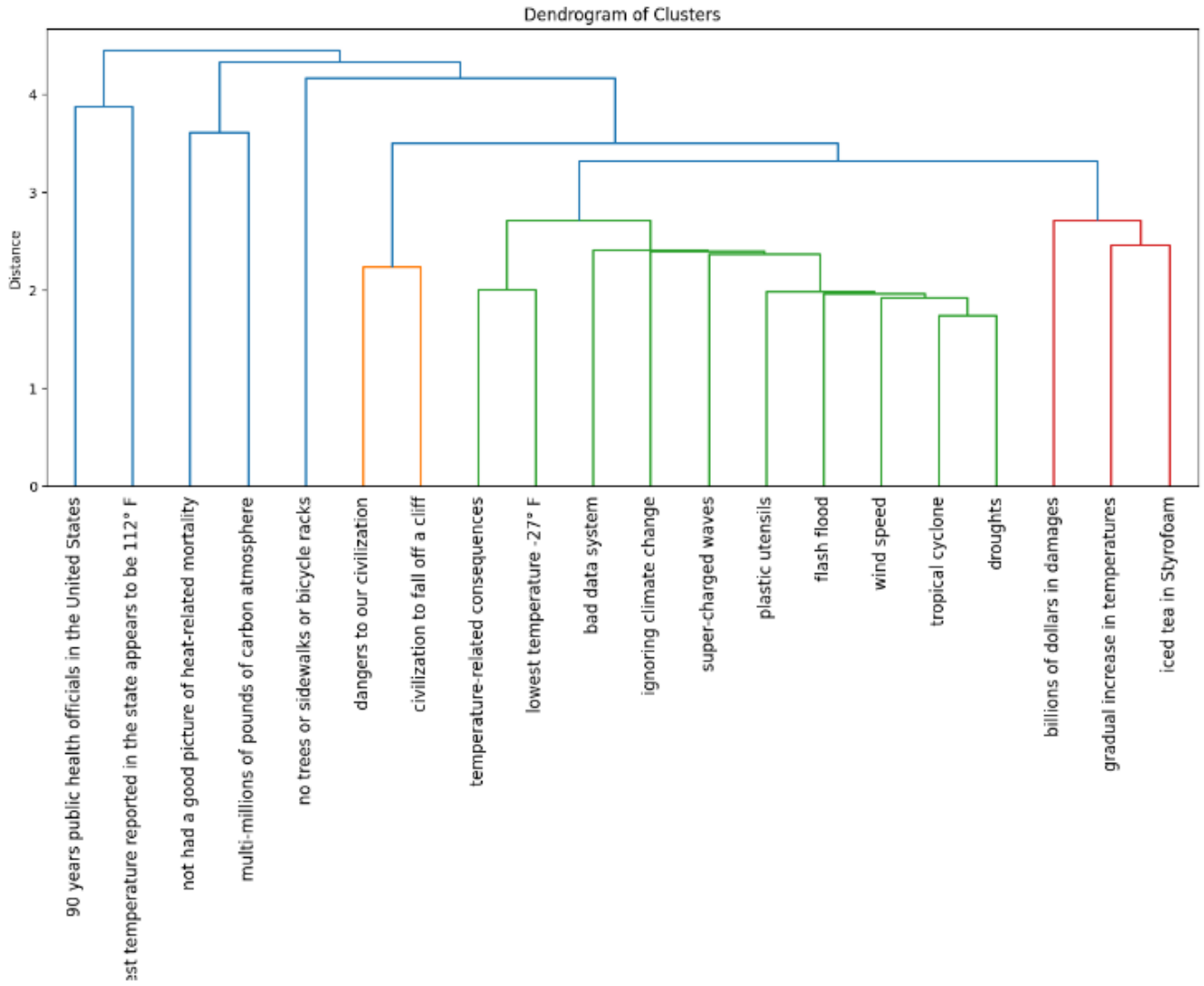
Hierarchical Clustering

West



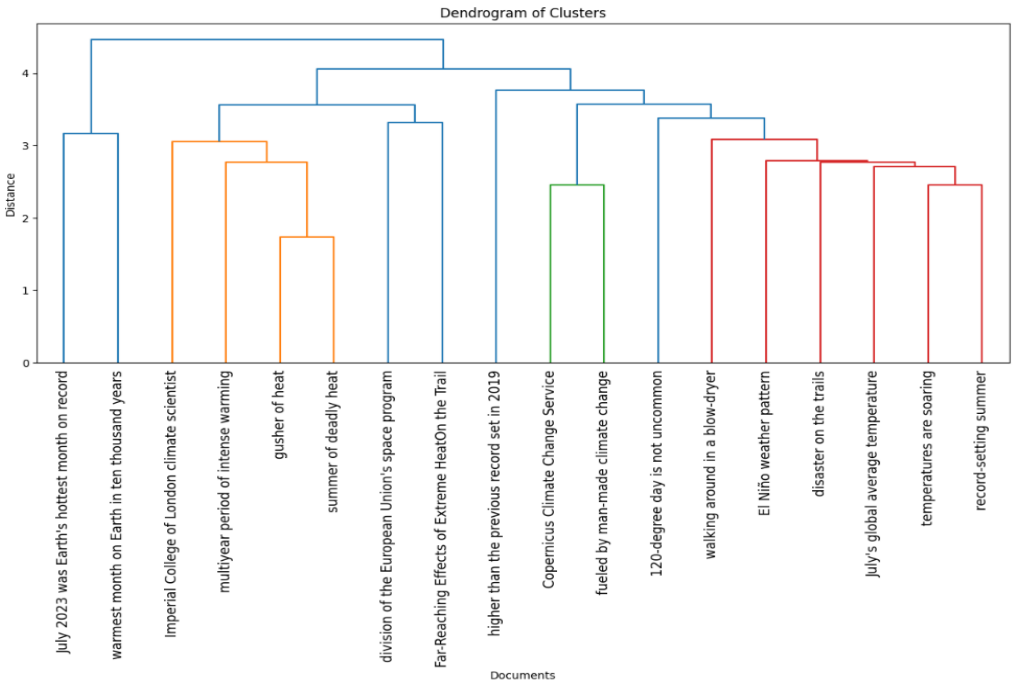
The dendrogram shows the relationship between the number of clusters and the delay in cutting carbon emissions. The closer the clusters are together, the more likely it is that the heat wave was caused by delayed cuts in carbon emissions.

Central



The dendrogram suggests that ignorance of climate change has led to an increase of carbon in the atmosphere due to which the temperature has reached 112^{O}F.

East



The dendrogram illustrates that heat wave was caused by a combination of factors, including delayed cuts in heat-trapping carbon emissions, the El Niño weather pattern, and high temperatures. July was the hottest month in the year 2023.

CONCLUSION

Text summarization along with Sentiment classification, Topic modeling, Hierarchical clustering techniques, and predictive analytics based on sentiment are invaluable for climate change analysis by distilling the key findings and conclusions from this vast corpus of climate data and research. These methodologies also help make complex climate knowledge more accessible to the public through summarized explainers. In the race to adapt policies before irreversible climate tipping points, text summarization along with Sentiment classification, Topic modeling, Hierarchical clustering, and predictive analytics based on sentiment is an indispensable tool for researchers to benefit from the latest insights emerging across diverse sources of climate data and literature.

FUTURE DIRECTIONS

This research on applying NLP techniques to a climate data corpus has unveiled immense potential for further advancing discovery and comprehension. An immediate next step is expanding the dataset to

include more varied sources like scientific publications, news reports, and social media to encompass broader perspectives. More refined preprocessing like coreference resolution could enrich the analysis. Advanced deep learning approaches like BERT and GPT-3 may extract deeper semantic insights versus the implemented ML models.

Exploring diverse NLP tasks like semantic role labeling to identify climate actors and actions could reveal new angles. Causality detection algorithms could uncover climate influence pathways. Finetuning language models on climate papers may improve topic extraction and sentiment prediction. Ongoing research into climate discourse trends and framing using these techniques can support communications. Combining climate, environmental, and socioeconomic data could spotlight interconnections. More robust validation frameworks would reinforce reliability. Overall, this research highlighted the tremendous potential of NLP and ML to transform climate science analysis but also uncovered numerous promising directions for innovation. Advances in models, datasets, and techniques will enable even more powerful climate data mining, stronger climate insights, and accelerated progress on this critical global challenge.

REFERENCES:

1.Brownlee, J. (2019, August 7). A gentle introduction to text summarization. MachineLearningMastery.com. https://machinelearningmastery.com/gentle-introduction-text-summarization/

2. NASA. (2022a, July 18). Climate change adaptation and mitigation. NASA. https://climate.nasa.gov/solutions/adaptation-mitigation/

3. Welcome to Vadersentiment's documentation! Welcome to VaderSentiment's documentation! - VaderSentiment 3.3.1 documentation. (n.d.). https://vadersentiment.readthedocs.io/en/latest/

4. GeeksforGeeks. (2022, October 18). Named entity recognition. GeeksforGeeks. https://www.geeksforgeeks.org/named-entity-recognition/

5. Yadav, K. (2022, November 8). The Complete Practical Guide to Topic Modelling. Medium. https://towardsdatascience.com/topic-modelling-f51e5ebfb40a

6. Winastwan, R. (2020, October 2). Visualizing word embedding with PCA and T-Sne. Medium. https://towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5

7. Mayank, M. (2021, September 7). A guide to the knowledge graphs. Medium. https://towardsdatascience.com/a-guide-to-the-knowledge-graphs-bfb5c40272f1

8. Bock, T. (2022, September 13). What is a dendrogram?. Displayr. https://www.displayr.com/what-is-dendrogram/

9. Lim, Y. (2022, April 5). Stacked ensembles - improving model performance on a higher level. Medium. https://towardsdatascience.com/stacked-ensembles-improving-model-performance-on-a-higher-level-99ffc4ea5523

10. Chamblee, B. (2022, February 7). What is cosine similarity? how to compare text and images in Python. Medium. https://towardsdatascience.com/what-is-cosine-similarity-how-to-compare-text-and-images-in-python-d2bb6e411ef0