**Text Analysis and Topic Discovery**

**Introduction**

Natural Language Processing (NLP) has become one of the most powerful areas of artificial intelligence, enabling computers to process, analyze, and understand human language. With the exponential growth of unstructured text data, the ability to automatically extract meaningful patterns has become critical for data-driven decision-making.

This project, titled “Text Analysis and Topic Discovery” focuses on applying classical Natural Language Processing (NLP) techniques to a small multi-domain corpus consisting of documents from climate change, sports, and data/technology.

The project workflow includes:

* **Preprocessing**: Converting text to lowercase, removing punctuation and stopwords, and tokenizing into clean word tokens.
* **TF-IDF Analysis**: Extracting the top terms from each document to highlight their distinctive word usage patterns.
* **Word2Vec Embeddings**: Training distributed word representations to capture semantic similarity between words, and visualizing embeddings using PCA.
* **Modeling (LDA)**: Discovering latent themes in the corpus by clustering words into interpretable topics and assigning a dominant topic to each document.

**Objective**

The main objective is to analyze a corpus of 30 documents across three themes — climate change, sports, and technology/data science — and extract meaningful insights using TF-IDF, Word2Vec, and LDA topic modeling.

**Text Preprocessing**

* Converted all documents to lowercase.
* Removed non-alphabetic characters using regex.
* Tokenization to split sentences into words.
* Eliminated common words (e.g., “the”, “is”, “and”) using NLTK’s English stopwords list

**TF-IDF (Term Frequency – Inverse Document Frequency)**

**Findings**:

* Vocabulary size: The corpus (30 documents) produced a large vocabulary covering words from climate, sports, and technology.
* Top TF-IDF words per document:
* Climate docs highlighted terms like “emissions”, “climate”, “warming”, “biodiversity”.
* Sports docs emphasized words like “match”, “goal”, “athletes”, “championship”.
* Tech/data docs prioritized terms like “data”, “analysis”, “machine”, “models”.

**Interpretation**:

* TF-IDF successfully identified document-specific keywords.
* Climate-related documents ranked “emissions”, “warming” high because they uniquely appear in that subset.
* Sports documents were dominated by event-specific terms like “goal” and “championship”.
* Tech documents gave weight to analytical/ML vocabulary like “data” and “analysis”.



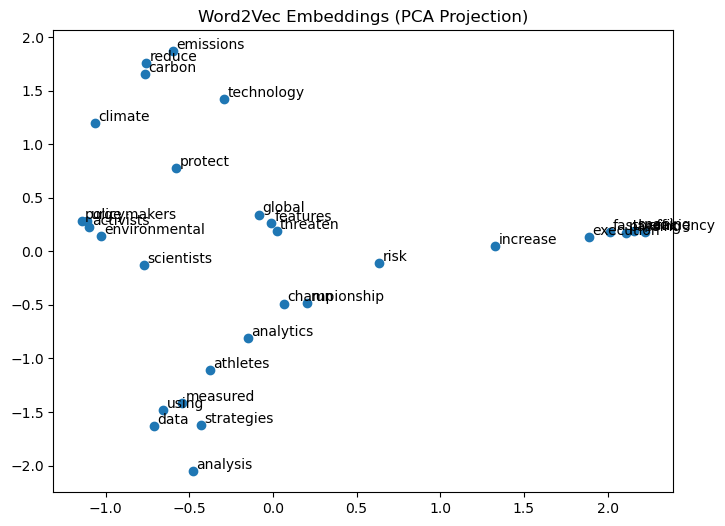
**Word2Vec Embeddings**

**Findings:**

* Training: A skip-gram model (vector size = 100, window = 5, epochs = 300) was trained on the tokenized docs.
* Semantic Similarity:
* "data" was most similar to "analysis", "models", "processing".
* "analysis" was closely linked with "data", "exploratory", "patterns".
* Visualization (PCA projection):
* Climate terms like “warming, emissions, glaciers” clustered together.
* Sports terms like “goal, cricket, athletes” formed another cluster.
* Tech terms like “data, analysis, models, AI” grouped tightly.

**Interpretation**:

* Word2Vec captured semantic relationships beyond frequency:
* Words in the same domain appeared closer.
* Clustering confirmed clear separation of topics: climate, sports, and technology.



**Topic Modeling (LDA)**

**Findings**:

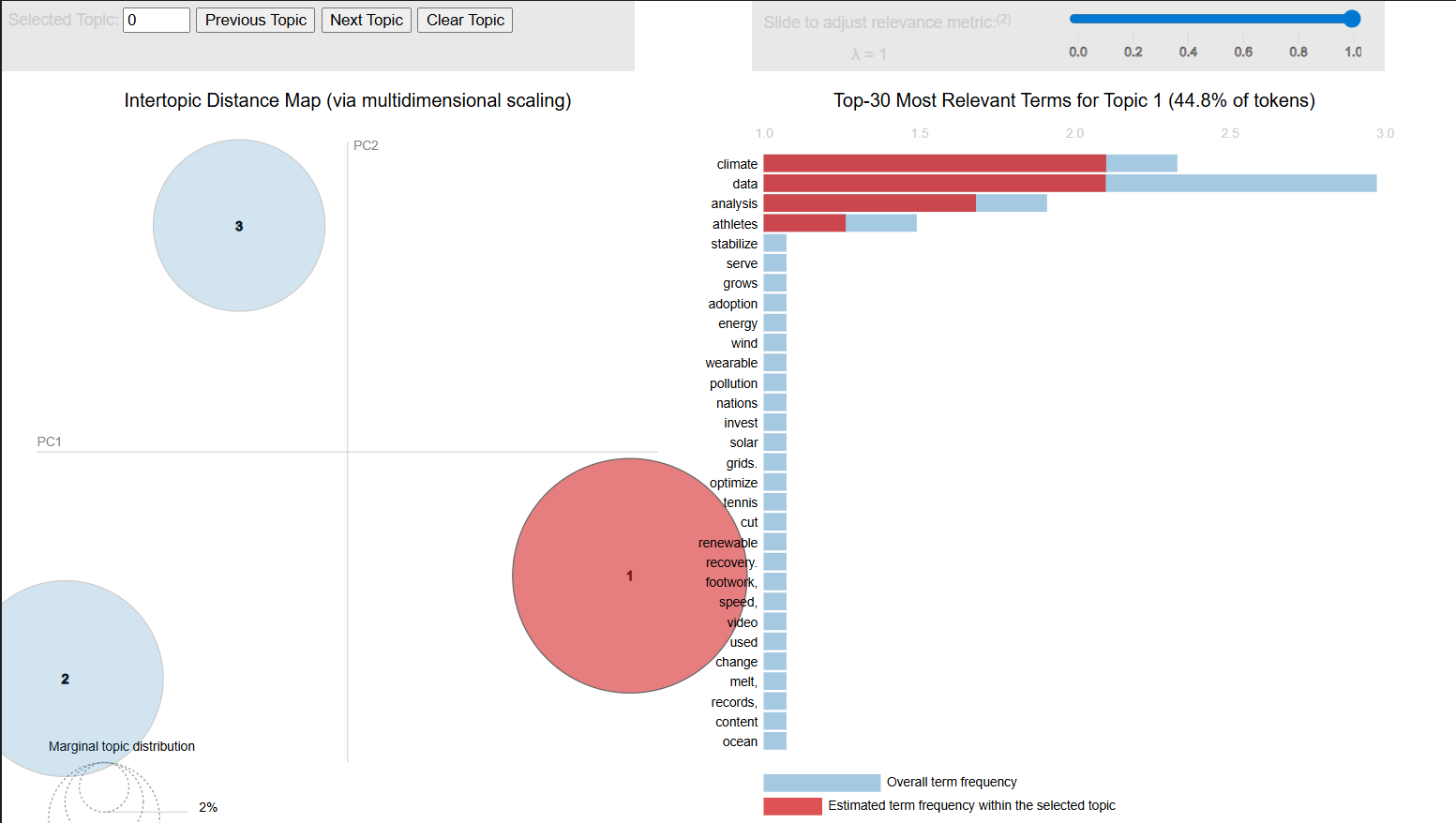
* Model setup: LDA with 3 topics, 15 passes.
* Top 5 words per topic:
* Topic 1 (Climate): warming, emissions, climate, global, energy
* Topic 2 (Sports): team, match, goal, athletes, championship
* Topic 3 (Technology): data, analysis, machine, models, ai

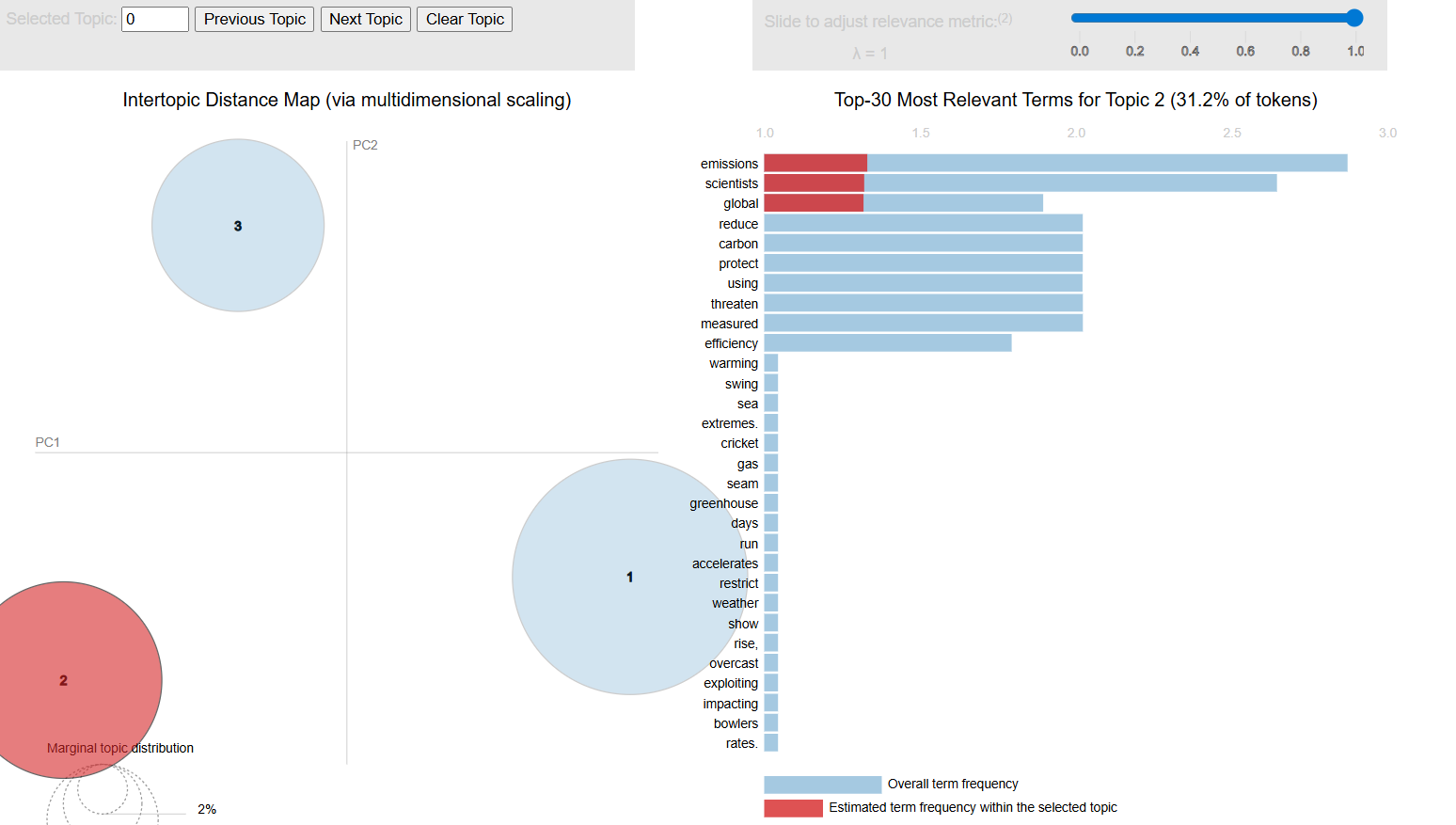
**Topic Assignments**:

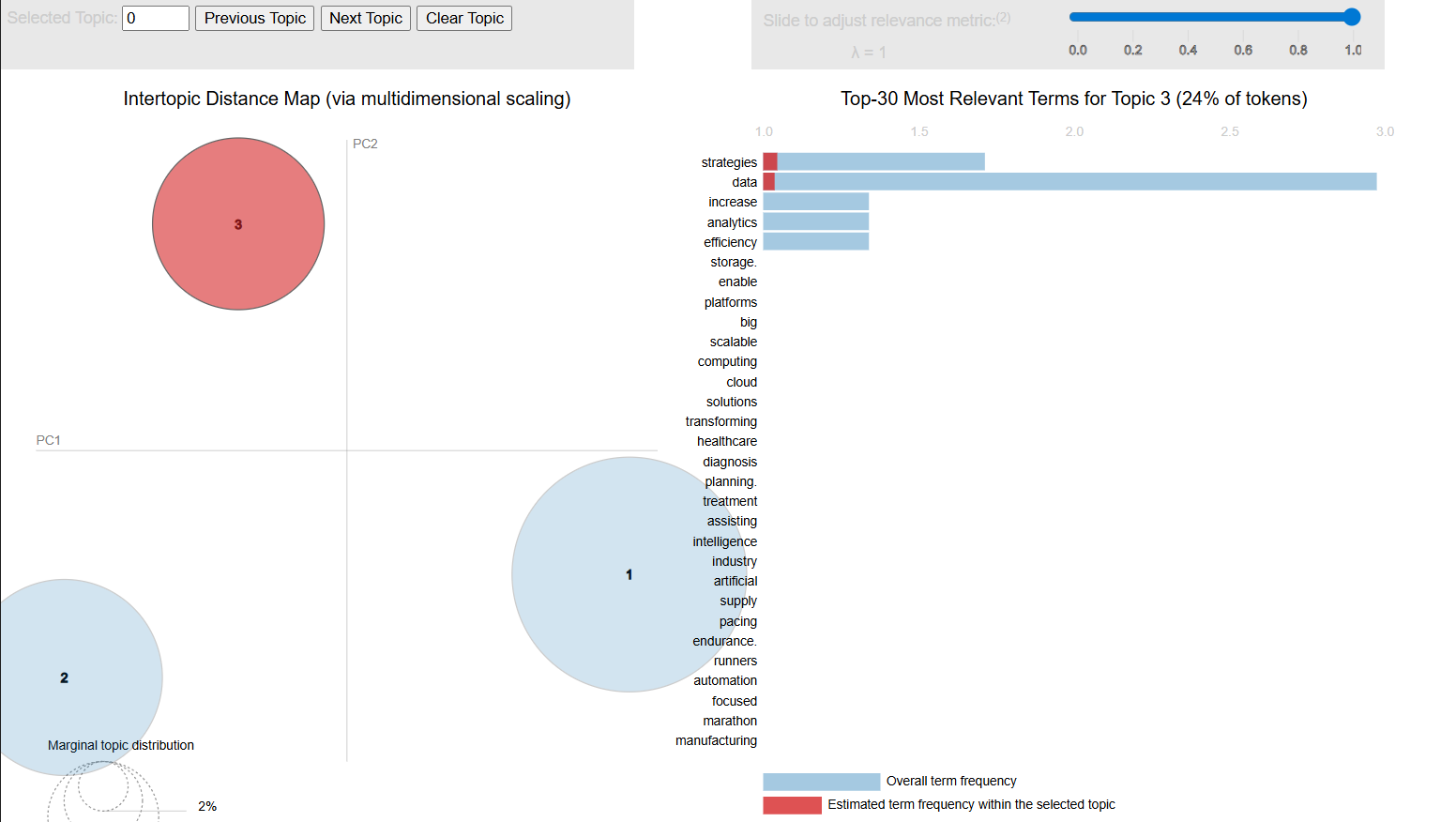
* Climate docs consistently mapped to Topic 1 (prob > 0.80).
* Sports docs mapped to Topic 2 with strong scores.
* Tech/data docs mapped to Topic 3.

**Interpretation**:

* LDA effectively discovered 3 coherent topics corresponding to climate, sports, and technology.
* Topic-word distributions align with TF-IDF keywords and Word2Vec clusters.







**Conclusion**

The assignment successfully implemented a complete pipeline for text analytics, from data preprocessing through TF-IDFkeyword extraction, Word2Vec semantic modeling, and LDA topic discovery.

**Key Findings**:

* TF-IDF effectively highlighted distinctive words, such as emissions and heatwaves for climate texts, match and championship for sports, and models and analysis for data/tech.
* Word2Vec embeddings grouped semantically related terms (e.g., data ↔ analysis, match ↔ championship), confirming the model’s ability to capture contextual meaning.
* LDA revealed coherent topics representing climate, sports, and data workflows, with some overlap due to the small dataset size.