Semantic Similarity Analysis of Textual Data

Senthil Arumugam  
senthil.ramasamy@stud.fra-uas.de

Neetu Ninan

[neetu.ninan@stud.fra-uas.de](mailto:neetu.ninan@stud.fra-uas.de)

*Abstract— in the era of big data and natural language processing, the ability to accurately analyze and compare textual data is paramount. This research presents a comprehensive framework for semantic analysis of textual data, focusing on the calculation of phrase similarities and document comparisons. Leveraging advanced open AI embedding techniques and Cosine Similarity Algorithms, our approach aims to analyses the accuracy and efficiency of semantic similarity assessments and understands its potential use case applications. The framework is implemented as a software tool that preprocesses textual data, generates embedding’s using state-of-the-art models, and calculates similarity scores between phrases and documents. The tool supports various preprocessing options, including tokenization, normalization, and context-based adjustments, ensuring robust and contextually relevant similarity measurements. We evaluate the performance of our framework through extensive experiments on diverse datasets, demonstrating its effectiveness in capturing semantic nuances and improving the quality of textual data analysis. To visualize the effectiveness of our similarity calculations, we have developed a plotting tool that maps scalar values against similarity scores allowing for a clear and immediate understanding of the data distribution and the performance of our similarity metrics and where secondary plot displays the number of possible comparisons between documents and its corresponding Similarity score which displays the semantic analysis between documents or phrases with its relevance based on pre-defined threshold. The results indicate significant improvements in similarity scoring accuracy compared to traditional methods. This research contributes to the field of natural language processing by providing a scalable and flexible solution for semantic analysis, with potential applications in automated content categorization like resume filtering for relevant job opportunities or filtering admission for students based on admission requirements and other several use cases across different domains. The software tool is made available as an open-source project, encouraging further research and development in this analysis or develops potential use cases. Keywords—Embedding, Cosine Similarity, scalar*

# **Introduction**

# **LITERATURE SURVEY**

# ***Implementation***

Based on the literature review and concepts of Open AI embedding, Semantic analysis, application was designed in such a way that it supports the methods to do analysis of data using phrases and documents as well; We understood that in order to do such analysis using various datasets such as documents or phrases, analysis results may vary due to contextual meaning of the sentences in the documents, During the Initial Study we tried several samples with Open AI Embedding API’s using Nuget package available in the dotnet also we tried one other tool called as Hugging Face API, we found out that due to the restrictions of usage of its API’s is not available in Nuget package , research started with the sole focus on using technique of open AI Embedding because of its availability and its ability to create the embedding which supports the contextual analysis for different types of datasets such as Phrases/Words or documents comparisons as well.

In order to achieve the better comparison results, we wanted to categorize the implementation process into 6 main categories such as

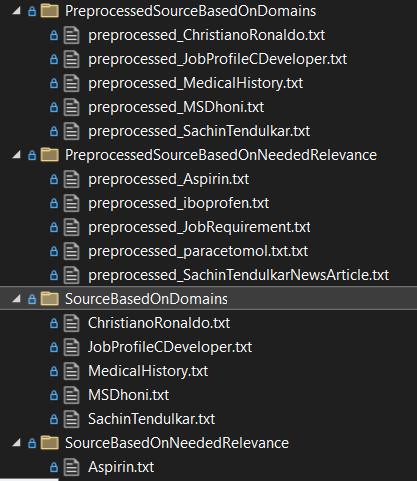
1. Defining dataset’s based on specific Domains and processing Datasets
2. Pre-Processing Interface to Process the Documents
3. Create Embedding’s for the input documents
4. Calculate Similarity using Cosine Similarity Algorithm based on programmatically generated embedding’s of input documents
5. Generate the output of Similarity Score as a CSV File
6. Utilizing the output data’s from CSV to generate meaningful results which shows semantic analysis between documents
7. **How did we create Dataset’s?**

During **t**he Initial research, we decided to make the analysis meaningful by classifying the datasets by domains; we come up the set of 5 to 10 words from each domain, **Example- Electricity and Energy** are the two different words by contextually it is related to the **Power Sector** domain and we wanted to ensure if this comparison using Open AI Embedding technique gives meaningful results by considering the contextual relevance, likewise we come up with more dataset around fifty to sixty different words from the same and different domain so that we could utilize such data for comparison and we also wanted to ensure it is editable by the developers to support changing the data for future analysis, Initially phrases comparison datasets are JSON formatted data which can be modified accordingly by the developers.

Document Comparison- We wanted our application to support comparison of any documents to show its contextual relevance between them, In order to do any comparison we need source documents and target documents, For Example, JobRequirment.txt is the source document and Job Profile A, Job Profile B is the target documents, so now when the source and target documents are compared, results will be meaningful and we could also come up the predefined threshold which can be managed by the application admin of any organization utilizing our application, Likewise we come up with few more meaningful document comparison by enabling the users of the app to add more documents either directly using file Manager or by enabling integration of user specific UI to our app to upload documents which requires development efforts, still we wanted to make sure application supports documents comparison dynamically making it more usable and researchable.

1. **Pre-Processing Interface to Process the Documents:**

With the focus on processing any documents or datasets, we first need the data to be loaded into the program, we are achieving this by defining the specific path inside the solution by defining the data folder with the two different folder names, one folder is to keep all the documents as key document on the criteria in which documents has to be compared named as “SourceBasedOnDomains” and other folder named as “ Source Based on Relevance” to keep all the documents which have to be sent for comparison ( i.e.) in future if this application is integrated with the UI or user facing application user can upload their resumes or motivation letter for joining the university which can be later used by our application to compare those with key category documents like “Job Requirement” or “Admission Requirement for Specific course in an University”



**Figure: Representing the Dataset Folder Structure for Raw Documents and Processed Documents**

Interface IPreprocessor.cs is designed in such a way that it has all methods needed to do the basic functionalities which we have already mentioned in methodologies. Once such methods are mentioned below which are responsible to read the raw input dataset/documents/phrases and create a processed documents in the new folder by creating the folder name programmatically by appending the word “Preprocessed” and hence our program creates two new folders with “PreProcessedSourceBasedOnDomains” and “PreProcessedSourceBasedOnRelevance” respectively.

await ProcessTextFilesInFolderAsync(textPreprocessor, sourceDomainsFolder, outputDomainsFolder);

await ProcessTextFilesInFolderAsync(textPreprocessor, sourceRelevanceFolder, outputRelevanceFolder);

which have to be compared and produce the processed output into the new folders as shown in the above **Figure name**?

Once the document is loaded into the program and processed documents are generated using the below method definition.

String PreprocessText (string text, TextDataType type);

Then the further methods which will be discussed below in our paper can be invoked to created embeddings and similarity score.

1. **Create Embedding’s for the input documents:**

Interface CalculateEmbeddingAsync created for the purpose of accepting the phrases or documents as a text, and also additionally we want to ensure what is the category of the domain to find its relevance to the context it is created, hence it is designed to accept text 1, text 2 for processing source document or phrase in text1 and target document or target phrase in text2 and its corresponding filename respectively.

Task<double> CalculateEmbeddingAsync (string text1, string text2, string fileName1, string fileName2);

This Interface is made to utilize in the two different services called as SemanticSimilarityForDocumentsWithInputDataDynamic.cs and SemanticSimilarityPhrasesWithInputDataSet.cs where the actual implementation is created to produce the output embedding’s by utilizing the methods of open AI embedding nuget package.

OpenAIEmbeddingCollection collection = await client.GenerateEmbeddingsAsync (inputs);

GenerateEmbeddingsAsync () is the important method which accepts list of strings as inputs and produces collection of embedding as output for range of size 3052

Public static void PrintScalarValues (float [] embedding)

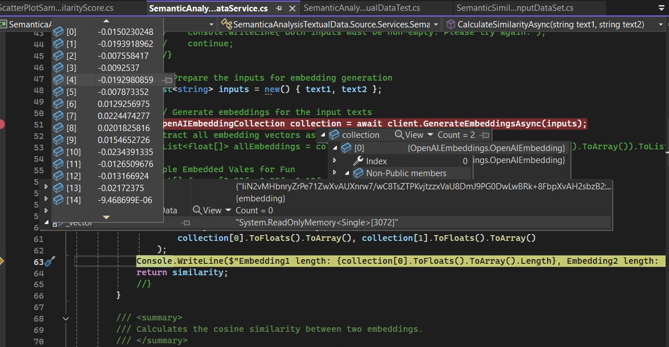


Figure No: **Representation of Created Embeddings in Debug Mode**

It is the custom method created to print the output of the each document or phrases as an embedding’s to print it individual values at every array till the end of the size of collection, Purpose of this method is utilize this values later during the visualization how closely or relatively the embedding’s at the vector space is being created. By knowing this difference it will help us to understand the correlation between the similarity score created VS Scalar values.

1. **Calculate Similarity using Cosine Similarity Algorithm based on programmatically generated embedding’s of input documents**

CalculateSimilarity (float [] embedding1, float [] embedding2);

This method is designed in such a way that it accepts the embedding’s generated from the previous GenerateEmbeddingsAsync (inputs) method, which will invoke the implementation of calculate Similarity by applying Cosine Similarity algorithm which is capable of returning the single similarity score as a double value as return type, values will be generated in the range between -1 to 1.

* **1** indicates perfect similarity (identical embeddings).
* **0** indicates orthogonal vectors (no similarity).
* **-1** indicates complete dissimilarity (opposite vectors).

The implementation of Calculate Similarity follows these steps:

1. **Input Validation:**
   * The method first checks if the lengths of the two embeddings are equal. If not, it returns 0, indicating an invalid comparison.
2. **Dot Product and Magnitude Calculation:**
   * The dot product between the two embedding vectors is calculated.
   * The magnitude of each embedding vector is calculated separately using the formula:

Magnitude = sqrt (Σ embedding[i] ^2)

1. **Normalization:**
   * The cosine similarity score is calculated by dividing the dot product by the product of the magnitudes:

Cosine Similarity = Dot Product / (Magnitude1 \* Magnitude2)

* + This normalization ensures the score is in the range of -1 to 1.

1. **Error Handling:**
   * If the magnitude of any vector is zero, an error is raised to prevent invalid calculations.
   * In case of other unexpected errors, the method returns 0 and logs the error.
2. **Generate the output of Similarity Score as a CSV File**

Application is designed to support the generated output as a CSV file

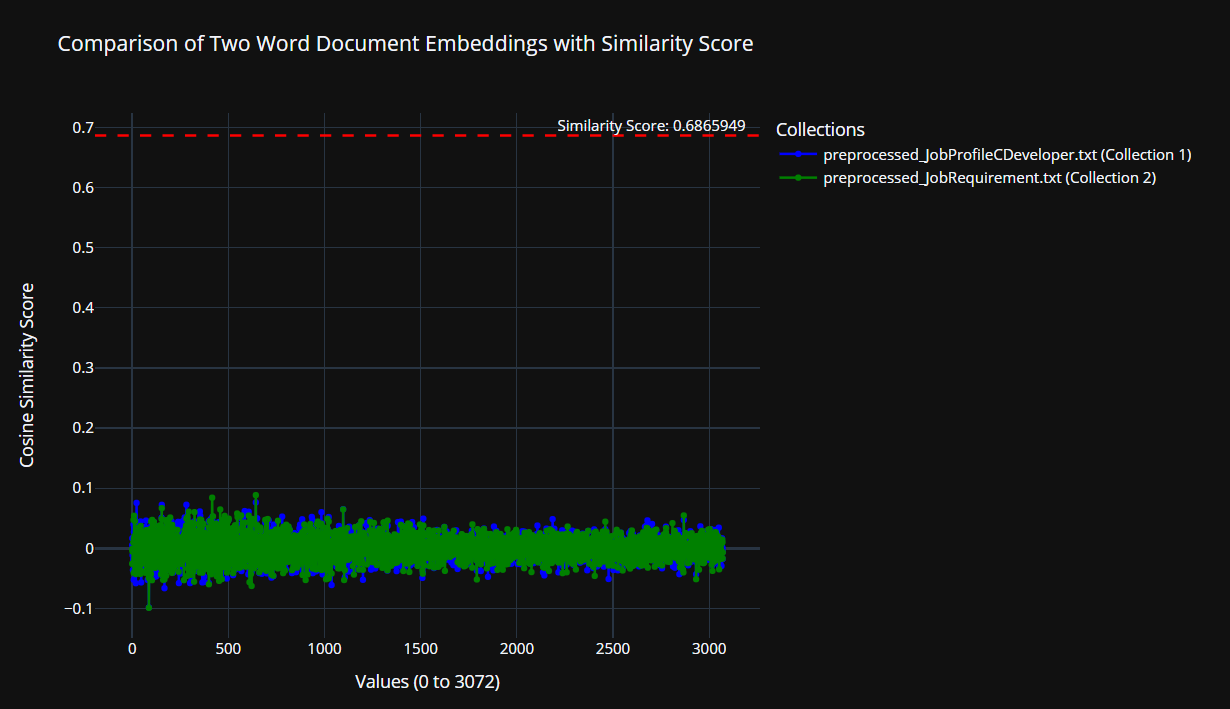
Public static void SaveResultsToCsv (List<DocumentSimilarity> results)

Public static void SaveResultsToCsv (List<PhraseSimilarity> results)

DocumentSimilarity and Phrase Similarity is the domain classes created to support the implementation of saving all the state of different values generated during the processing stage including the similarity score, domain, fileName1, fileName2 and score, domain, Phrase1, Phrase2, context respectively so that we achieve the clear understanding of what data’s can be mapped to which data to represent the generated data graphically using visualization methods.

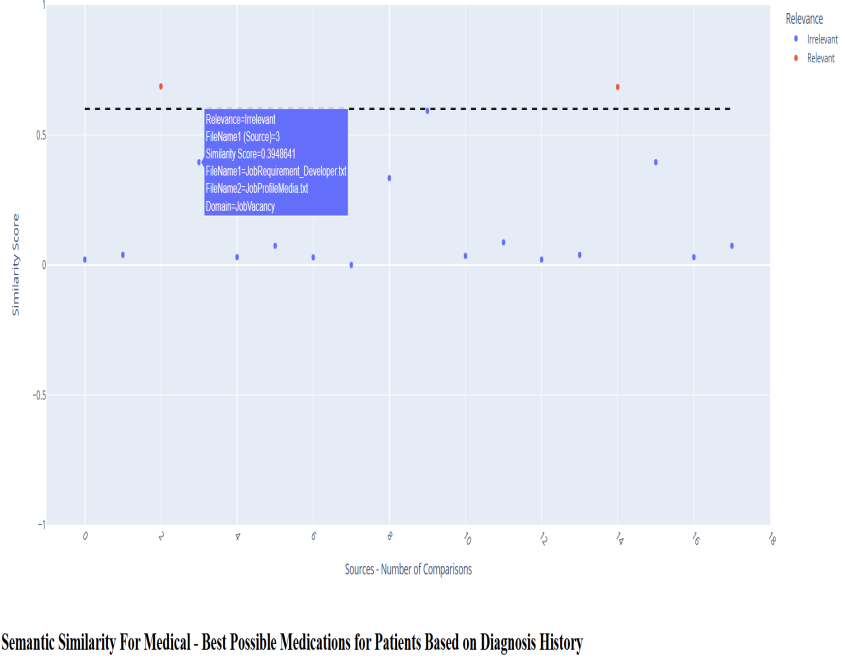
1. **Utilizing the output data’s from CSV to generate meaningful results which shows semantic analysis between documents**

In order to meet the ultimate end goal of understanding the Semantic Analysis of textual data between documents and create some meaningful output which would correlate to real time use cases**,** we used python as an external development tool to create a graphical chart, currently the application is designed to read the output CSV file generated dynamically if the files are placed in the root directory of the python app, so currently placement of output CSV required manual effort by the developers or the application admin,but we know this limitation which we are focusing during the later improvements either by us and paving way for other developers or ideas to improve the implementations. There are two types of plots we have designed; one chart is to graphically represent all the possible number of comparison of documents or phrases dataset designed by the user on X-axis VS its corresponding Similarity Score on Y axis; Other chart is designed in such way that developers are able to understand how the contextual relevance is actually generated by plotting its similarity score on Y axis VS Scalar Values (Ranges between 0-3052) on X-axis.



**Figure 3:**

**Scalar Values vs. SimilarityScorePlotForOneComparsion**



**Fig 3: Semantic Similairty Similarity Score vs Number of Comparioson involved during Analysis**

We wanted to display the analysis over a single chart as it gives an easy represenation of values while hovering over the blue or red dots representing the user to click and observe the details attached to it along with its similarity score. We have implemented the Github Pages to make the users view the plot anywhere to make viewable over internet.Below are the links appended to represent both the form of charts.

**Link for Plot 1:**

**Total Number of Documents Compared VS Similarity Score**

<https://senthilmasters2024.github.io/Tech_Tweakers/SemanticSimilarityLatestPlot.html>

**Link for Plot 2:**

**Scalar Values vs. SimilarityScorePlotForOneComparsion**

<https://senthilmasters2024.github.io/Tech_Tweakers/ScalarValuesVsSimilarityScorePlotForOneComparsion.html>

1. **Utilizing CSV to Plot Using Python:**

As there are lot of options to explore, python is a open source and it has good predefined libraries for creating scatter plots, after some initial study inorder to display the similarity score for varoius documents from several categories, we have implemetend a Python-based Flask application was developed to classify and visualize similarity scores across different domains. The tool reads CSV files containing similarity scores, maps them to respective domains, and applies predefined thresholds for classification.

Plots are generated using Plotly, and thresholds for domains such as "JobVacancy", "Medical", and "Sports" are visualized with interactive scatter plots. This tool enhances interpretability by categorizing similarity scores as "Relevant" or "Irrelevant" based on domain-specific thresholds.

Currently domain name mapping is done based on the initial application desing requirement, by modiying the needs of the user or future analyse, developers or we will continue refactoring this accordingly which enables support for more domain/context related mappings of data which will induce a better visualisation.

# **TestCase with Results**

# We have implemented several test cases ensuring that when running the test cases through test explorer, all the code and methods are covered including positive and negative scenario’s i.e. handling exceptions and null values, currently we did not use the mock data to ensure test cases as this application is the initial setup for the purpose of analysis we are utilizing the application “data” folder which we have defined for the purpose of maintaining datasets are copied into our test project and by reading those files dynamically using program, we are successfully in maintaining the code coverage and the test cases. Test cases are separated by creating individual test classes for each service we have defined based on the functionalities, as we have segregated our business functionalities into 4 classes; we created test classes also based on that and The test framework is implemented using Microsoft.VisualStudio.TestTools.UnitTesting, and the test cases are run through the Visual Studio Test Explorer to verify code coverage and robustness.

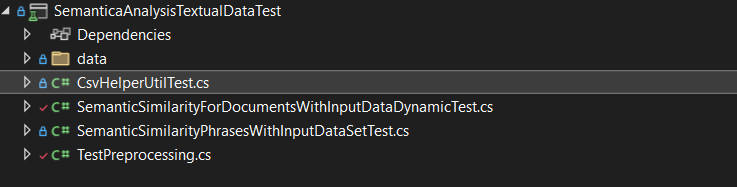


Figure: Representation of Semantic Analysis Test Project

# **Sample Test Method()**

**Purpose**: A basic test to ensure the test framework is set up correctly.

**Explanation**: This test simply creates an instance of the class and asserts that 1 equals 1, serving as a sanity check for the test setup of our application

# **CompareDocumentsAsync\_ShouldHandleExceptions()**

**Purpose**: To test the CompareDocumentsAsync method and ensure it handles exceptions properly.

**Explanation**: This test sets up the necessary directories and files, invokes the CompareDocumentsAsync method, and asserts that the results are not null, ensuring the method handles exceptions and returns a valid result.

1. **CalculateEmbeddingAsync\_ShouldReturnSimilarityScore()**

**Purpose**: To test the CalculateEmbeddingAsync method and ensure it returns a valid similarity score.

**Explanation**: This test reads the content of source and target files, calculates the similarity between them using the CalculateEmbeddingAsync method, and asserts that the similarity scores are greater than 0.

1. **ConfigureServices\_ShouldReturnServiceProvider**

**WithConfiguredServices()**

**Purpose**: To test the ConfigureServices method and ensure it returns a service provider with the configured services.

**Explanation**: This test invokes the ConfigureServices method, retrieves the configured service, and asserts that it is not null and is of the correct type.

1. **GetSourceAndTargetFiles\_ShouldReturnSourceAnd**

**TargetFiles()**

**Purpose**: To test the GetSourceAndTargetFiles method and ensure it returns the correct source and target files.

**Explanation**: This test invokes the GetSourceAndTargetFiles method, retrieves the source and target files, and asserts that they are not null, not empty, and have the correct file extensions.

**6.) PrintScalarValues\_ShouldPrintEachScalarValue()**

**Purpose**: To test the PrintScalarValues method and ensure it prints each scalar value correctly.

**Explanation**: This test captures the console output of the PrintScalarValues method and asserts that it matches the expected output.

**7. CalculateSimilarity\_ShouldReturnCorrectSimilarityScore()**

**Purpose**: To test the CalculateSimilarity method and ensure it returns the correct similarity score.

**Explanation**: This test calculates the similarity between different pairs of embeddings and asserts that the similarity scores are correct.

**8.CalculateSimilarity\_ShouldReturnZeroForDifferentLength**

**Embeddings()**

**Purpose**: To test the CalculateSimilarity method and ensure it returns zero for embeddings of different

**9. CalculateEmbeddingAsync\_ShouldHandle**

**EmptyInputs()**

**Purpose:** CalculateEmbeddingAsync method in the SemanticSimilarityForDocumentsWithInputDataDynamic class correctly handles empty input strings and returns a similarity score of 0.

**Explanation:** This test method ensures that when CalculateEmbeddingAsync is given empty input strings, it correctly returns a similarity score of 0. This is important for validating that the method can handle edge cases where the input data might be empty, ensuring robustness and reliability of the implementation.

**10. CompareDocumentsAsync\_Should**

**HandleInvalidFilePaths ()**

**Purpose**: To verify that the document comparison service can gracefully handle scenarios where the specified file paths are invalid or non-existent.

**Explanation**: This test ensures that the application properly detects and throws an appropriate exception (FileNotFoundException) when it attempts to access files that do not exist.

**11. InvokeProcessPhrases\_ShouldProcessPhrases**

**AndSaveResults() is to verify that**

**Purpose**: InvokeProcessPhrases method in the SemanticSimilarityPhrasesWithInputDataSet class processes phrases correctly and that the results are saved properly using the CsvHelperUtilTest class.

**12. CSVHelperUtilTest.cs**

The SaveResultsPhrase\_ShouldSaveResultsToCsvAndJsonFiles test method ensures that the SaveResultsPhrase method in the CsvHelperUtil class correctly saves a list of PhraseSimilarity objects to both CSV and JSON files. It verifies that:

1. The JSON file is created and contains the correct number of records.
2. The CSV file is created and contains the correct number of records.
3. The content of both files matches the original data.

This test is crucial for validating the functionality of the SaveResultsPhrase method, ensuring that it accurately writes data to the specified file formats. This is important for applications that rely on exporting data for further analysis or sharing.

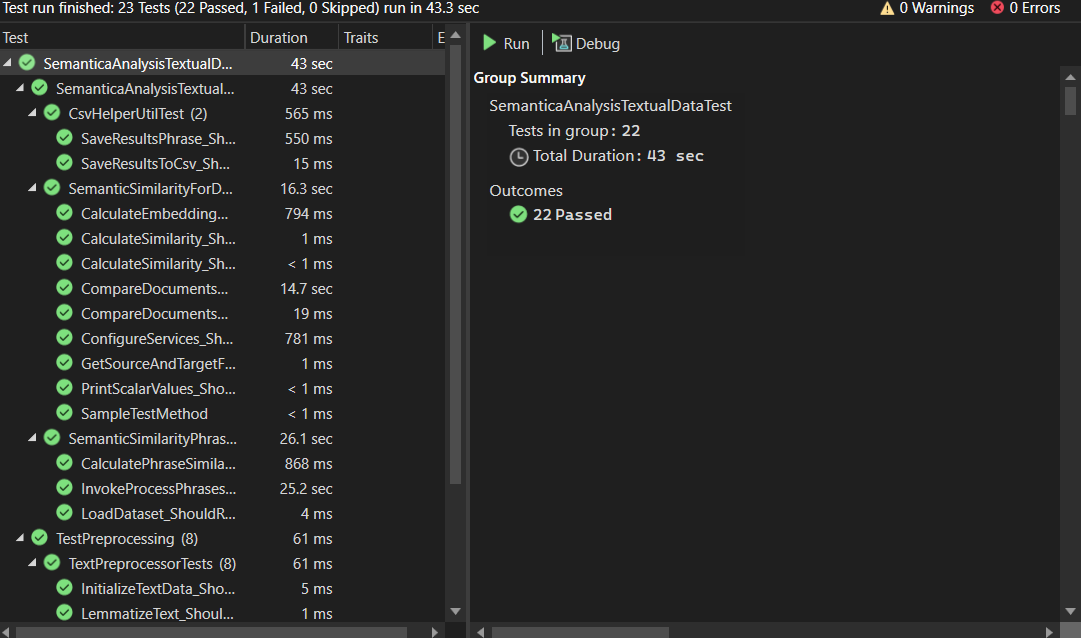


Figure Name: Respresenation of Test Run Results of All Test Cases

## **Overcoming Limitations:**

**Dependency on External Libraries**: The project relies on external libraries such as Plotly.NET and Microsoft.VisualStudio.TestTools.UnitTesting. Any changes or deprecations in these libraries could affect the functionality of the project.Inorder to overcome, Regularly update and test the project with the latest versions of external libraries like Plotly.NET and Microsoft.VisualStudio.TestTools.UnitTesting. Maintain backward compatibility by implementing adapter patterns or creating wrappers around critical functions,Also include automated tests to verify compatibility whenever library versions are updated.

**Text Preprocessing Scope**: The text preprocessing methods implemented may not cover all possible text variations and edge cases. For example, handling of complex HTML tags, nested URLs, or advanced lemmatization might require additional logic.Inorder to overcome, implement additional preprocessing techniques such as advanced lemmatization, complex HTML parsing, and nested URL handling. Introduce modular preprocessing functions allowing users to customize rules for stop-word removal, tokenization, and other steps.

**Performance**: The performance of text preprocessing, especially for large datasets, might be a concern. The current implementationmay need optimization for handling large volumes of text efficiently. In order to overcome, developers can use parallel processing techniques and also try to implement batch processing of datasets using memory efficient algorithms

**Customization:** The preprocessing rules (e.g., stop words removal, special characters handling) are hardcoded. Users might need more flexibility to customize these rules based on their specific requirements. Inorder to overcome, a configuration interface can be developed to allow users to define their own preprocessing rules and similarity thresholds without modifying the core codebase. This interface would provide support for user-defined similarity metrics beyond cosine similarity, making the system adaptable to various research needs. By implementing user-specific configuration files or UI settings, researchers can customize the similarity computation process, including the choice of preprocessing techniques, similarity metrics, and relevance thresholds for different domains. This modular approach will ensure broader applicability and scalability of the framework for diverse NLP applications.

**Applications of Semantic Analysis Textual Data :**

**Natural Language Processing (NLP):** This project can be used as a foundational tool for various NLP tasks such as text classification, sentiment analysis, and entity recognition by providing preprocessed and cleaned text data.

**Data Cleaning:** The text preprocessing methods can be applied to clean and normalize text data in data science projects, ensuring consistency and improving the quality of the data.

**Search Engine Optimization (SEO):** By preprocessing and normalizing text, this project can help in optimizing content for search engines, making it more accessible and relevant.

**Content Management Systems (CMS):** The project can be integrated into CMS platforms to preprocess and clean content before publishing, ensuring high-quality and readable content.

**Academic Research:** Researchers can use this project to preprocess textual data for various academic studies, including linguistic analysis, social media analysis, and more.

**Conclusion:**

In conclusion, this research and project explore various topics related to OpenAI embeddings and similarity algorithms in the realm of AI and programming. By leveraging AI libraries, we bridge the gap between theoretical concepts and real-world applications, utilizing real-time data to create useful applications tailored to specific domain requirements.

This analysis serves as a foundational step, enabling other developers to build upon it and create a lasting impact. The visualizations demonstrate how texts, documents, and phrases are contextually aligned which we can clearly observe in the results of our application’s similarity scores and plots, thus highlighting the potential for various applications.

For instance, this application can be adapted for matching job profiles to job requirements or aligning student profiles with university admission criteria, simply by changing the dataset. This flexibility underscores the practical utility of our approach in diverse scenarios.

The proposed framework successfully computes phrase and document similarities while visualizing results for better interpretability.

Future enhancements include:

* Automating CSV file integration for visualization.
* Expanding dataset support across various domains.
* Enhancing the user interface for document uploads.

This study contributes to NLP applications in content categorization, job-matching algorithms, and automated document classification, with promising potential for future research.

Overall, this project not only advances our understanding of AI embeddings and similarity algorithms but also provides a versatile tool for real-world applications, paving the way for future innovations.