SMUBIA Datathon 2025

Proposal Submission Template

General Brief

1. Fill in the Team Details table with your team's information and use this document as a template to write your team's proposal.
2. Teams are encouraged to follow the provided template structure for clarity and completeness. You have the flexibility to modify and adapt the format as long as the proposal effectively communicates the solution's value and implementation approach.
3. By submitting this proposal, all participants have acknowledged the terms stated in the Terms and Conditions document and the Code of Conduct document.
4. Rename the proposal submission document as "***ProposalSubmission_<TeamName>***".
5. Submissions that do not follow the format of this document will be disqualified.
6. The text of this document should be font "**Arial**", font size **11**, and line spacing of **1.5**.
7. Ensure you have filled up your team name and team member's details.
8. Keep this proposal to a maximum of 10 pages only excluding the cover page, content outline, appendix, and references.
9. Submit your team's proposal **here** by **Sunday, 2th February 2025, 12pm**.
    a. Only one submission is necessary per team.

Team Details

Team Name	Spicy Code		
University	National University of Singapore		
Team Members Details	**Full Name**	**Email Address**	**Student ID** **If applicable*
Member 1 (Team Leader)	Fabian Lim Zi Yang	fabianlim92@gmail.com	A0309052M
Member 2	Chua Ong Wee William	chuaongweewilliam@gmail.com	A0291928N
Member 3	Soong Shao Zhi	soongshaozhi@gmail.com	A0309172H
Member 4	Wong Yao Feng Gareth	garethw2773@gmail.com	A0303283M

Content Outline

1. Executive Summary

Our team has developed an advanced analytics tool designed to extract entities and relationships from unstructured text based on the topic the user wished to look up. Our solution leverages Natural Language Processing (NLP) to process datasets, identify key entities, map relationships, and generate meaningful insights. By incorporating data validation, quality checks, and visual analytics, our tool enhances the understanding of complex and lengthy textual data.

Key Features of Our Solution:

1. Automated Entity Extraction & Organization:
    - Our tool processes text from the dataset, extracting key entities such as PERSON, ORG (Organizations), GPE (Geopolitical Entities), EVENT, PRODUCT, TOPIC, DATE, and KEY_ACTIONS.
    - Extracted entities are saved in an Excel sheet titled "Original Data", ensuring structured data representation.
2. Topic-Based Filtering & Data Grouping:
    - Users can specify a topic of interest to filter relevant entities.
    - The tool automatically finds matching entities and groups them under a separate "topic_filtered" sheet, providing a clear summary of related entities.
3. Entity-Relationship (ER) Diagram Generation:
    - The tool extracts relationships between entities from the dataset by analyzing subject-verb-object relationships.
    - Users can select a focus entity, and the tool generates a network graph (ER diagram) that visually represents connections, making complex relationships easier to interpret.
4. Data Validation & Quality Checks:
    - The tool ensures accurate entity extraction by handling missing values, duplicates, and irrelevant data before analysis.

Impact & Use Cases:

Our solution provides a scalable and efficient approach to structuring unstructured text. It is particularly useful for intelligence gathering where understanding relationships between

entities is crucial. The interactive visualizations (ER diagrams) further aid in presenting complex information in an intuitive and digestible manner.

By integrating NLP-powered text analytics, structured data organization, and dynamic visualizations, our tool effectively uncovers hidden insights from vast textual datasets, making it a powerful asset in data-driven decision-making.

2. Visualisation Overview

Data Visualization Techniques Used

1. Entity-Relationship (ER) Diagrams (Network Graphs)
    - Technique: We use network graphs built with NetworkX and Pyvis to illustrate relationships between entities. Each entity is a node, and relationships are represented as edges with labels indicating the type of connection.
    - Purpose: This helps users visually explore how entities are interconnected, revealing key players, influential organizations, or critical events in the dataset.
2. Entity Filtering (Tabular Representation)
    - Technique: Extracted entities are filtered based on the search topic and structured into Excel sheets (“Original Data” & “topic_filtered”) for easier data interpretation.
    - Purpose: This allows users to see related entities to a topic of their interest.

How These Visualizations Provide Insights

- Network Graphs Explain Relationships: By selecting a focus entity, users can see how it connects to other entities, uncovering hidden links, influential figures, or key actions in the dataset. This is particularly useful for intelligence gathering.
- Grouped Data Identifies Trends: By clustering similar entities together, we can determine which organizations, people, or events are most frequently associated, leading to insights about topic importance and recurring themes in the dataset.

Outliers & Correlations in the Data

- Outliers:
    - Entities that appear only once or have unexpected connections (e.g., a person linked to an unrelated event) may be anomalies worth investigating.
- Correlations:

- High-frequency entities appearing together may indicate strong relationships (e.g., a company frequently mentioned with a specific legal case).
- Specific actions (e.g., “acquired,” “investigated”) may frequently connect the same organizations, suggesting mergers, lawsuits, or regulatory scrutiny.

Possible Conclusions from the Data

- Key Influencers & Central Entities: Identifying the most connected individuals, organizations, or topics provides insight into who or what drives key discussions.
- Patterns in Relationships: Certain entities frequently appearing together suggest coordinated actions, business partnerships, or conflicts.
- Hidden Connections: The ER diagram can reveal previously unnoticed relationships that might be important for legal, financial, or investigative analysis.
- Data Quality Issues: If some expected entities are missing or relationships seem incorrect, it could highlight gaps or biases in the dataset.

3. Solution Features and Implementation Strategy

Comprehensive Breakdown of the Solution’s Features and Functionality

Our solution processes text data from an Excel file, extracts entities and key actions using spaCy NLP, clusters the extracted actions using K-Means clustering, and then filters the dataset based on topic relevance using GloVe embeddings. Finally, it saves the summarised results into an Excel file.

Key Features and Their Contributions:

1. Entity Extraction:
    - Feature: The tool processes the unstructured text to extract key entities such as PERSON, ORG (Organizations), GPE (Geopolitical Entities), EVENT, PRODUCT, TOPIC, DATE, and KEY_ACTIONS.
    - Contribution: This feature automates the extraction of key information from raw text, saving time and effort while improving accuracy. The extracted entities are stored in an organized format (Excel), enabling users to quickly identify and analyze important data points.
2. Topic-Based Filtering:

    - Feature: Users can input a topic of interest to filter relevant entities from the dataset.
    - Contribution: This allows for focused analysis of specific topics, ensuring that users can target specific areas of interest or urgency in large datasets. It allows for tailored extraction, filtering out irrelevant data for more precise analysis.
3. Entity-Relationship (ER) Diagram Generation:
    - Feature: The tool generates ER diagrams (network graphs) that visually represent relationships between entities.
    - Contribution: These diagrams offer a visual representation of the data, making it easier for users to understand the relationships between entities, such as people, organizations, and events. They help identify hidden links and key influencers within the data, enhancing the ability to uncover complex patterns and insights.
4. Data Validation and Quality Checks:
    - Feature: The solution includes built-in data validation to handle missing values, duplicates, and irrelevant data.
    - Contribution: Ensures that the extracted data is accurate, complete, and ready for analysis. This minimizes errors and optimizes the overall quality of the insights derived from the dataset.
5. Excel Output for Structured Analysis:
    - Feature: The extracted entities and grouped data are saved in Excel sheets, making the data accessible and ready for further analysis.
    - Contribution: By storing results in a widely used, user-friendly format (Excel), users can easily manipulate and analyze the data. It enables seamless integration with existing data workflows and tools for reporting or further analysis.

Technical and Strategic Approach to Development and Implementation

Our team has approached the problem through a combination of advanced NLP techniques, data organization methods, and data visualization tools. Below is a breakdown of the tools, technologies used:

Tools and Technologies

1. Natural Language Processing (NLP):
    - SpaCy (en_core_web_trf model):
      Extracts named entities (people, organisations, locations, etc.). Identifies verbs and their related words to extract key actions. It helps to process and analyse text data efficiently.
2. Data Processing and Analysis:
    - Pandas:
      We used Pandas for data manipulation, such as reading data from Excel, cleaning and processing the text, and storing extracted entities in a structured format. It simplifies the workflow for reading and writing data, allowing easy integration with the Excel-based output.
3. Machine Learning and Clustering:
    - scikit-learn (K-Means & TF-IDF):
      TF-IDF converts extracted key actions into numerical form based on their importance in the dataset. K-Means Clustering groups similar key actions together into clusters. Scikit-learn helps to discover patterns in the actions for the text input. It also groups similar actions together, making data analysis easier.
4. Semantic Similarity:
    - GloVe.
      It measures the semantic similarity between user-specified topics and extracted key actions using cosine similarity. It helps to filter data by keeping only rows related to the specified topic.

5. Data Visualization:
    - NetworkX and Pyvis:
      For visualizing relationships between entities, we use NetworkX to create network graphs (ER diagrams) and Pyvis for rendering those graphs. We are able to visualise the complex relationships between extracted entities, making it easy for users to interpret the connections between various people, organizations, events, and actions.
6. Excel Integration:
    - OpenPyXL (for Excel output):
      We use OpenPyXL to write the extracted entities and grouped data to Excel sheets. It provides easy-to-use functionality for saving results in Excel format.

7. User Interaction:
    - Pyvis:
      Pyvis serves as the user interface (UI) for visualizing the network graphs interactively, allowing users to explore and navigate relationships between entities, with features like zooming, panning, and hovering over nodes to view additional information, all within an intuitive graphical environment.
    - Command Line Interface (CLI)
      Our user interface (UI) is the Command Line Interface (CLI), where users input file paths and specify topics of interest through command-line commands.

---

4. Solution Impact

Our solution stands out by addressing the key challenges of extracting, organizing, and visualizing relationships from unstructured text in a way that is both accessible and highly actionable for users. Here are the unique selling points (USPs) that make our solution particularly valuable:

Unique Selling Points

1. Accurate and Comprehensive Entity Extraction
    - USP: By leveraging NLP models (SpaCy's transformer-based model), our solution can accurately identify and extract a wide variety of entities, including PERSON, ORG, GPE, EVENT, PRODUCT, TOPIC, DATE, and KEY_ACTIONS. This ensures that no important detail is missed, allowing for a rich and nuanced understanding of the data.
    - Value: Users get precise and relevant information from unstructured text, enabling deeper insights into the data’s key actors and events.
2. Topic-Based Filtering for Focused Analysis
    - USP: The ability to filter data based on a user-defined topic ensures that users can focus on the most relevant entities and relationships. This enables faster and more effective analysis by highlighting only those portions of the data that matter most.

    - Value: This focused approach saves users time and resources, enabling them to zero in on the most relevant insights without sifting through unnecessary information.
3. Visual Entity-Relationship Diagrams (ER Diagrams)
    - USP: ER diagrams (network graphs) provide a visual representation of relationships between entities. This makes it easier to grasp complex connections and see how entities interact within the data.
    - Value: Users can quickly identify key influencers, significant relationships, and emerging patterns, providing insights that would be harder to discover through raw data alone.
4. User-Friendly Output in Excel Format
    - USP: The ability to generate organized Excel reports containing both extracted entities and their relationships makes the tool accessible to users familiar with Excel and other data analysis tools.
    - Value: The familiar format makes the solution easy to integrate into existing workflows, offering immediate utility and seamless integration into users' data analysis processes.

Key Performance Indicators

1) Entity Extraction Accuracy:

- KPI: Percentage of correctly identified entities (PERSON, ORG, GPE, EVENT, etc.) compared to a ground truth or manually labeled dataset.
- Goal: Achieving a high precision and recall rate for entity extraction (e.g., 90%+ accuracy).

2) Topic Relevance Filtering Accuracy:

- KPI: Percentage of relevant entities and relationships included in the filtered output based on the user-defined topic.
- Goal: Ensuring that topic-based filtering is highly accurate, with the majority of relevant entities and relations being correctly filtered.

---

5. Solution Architecture

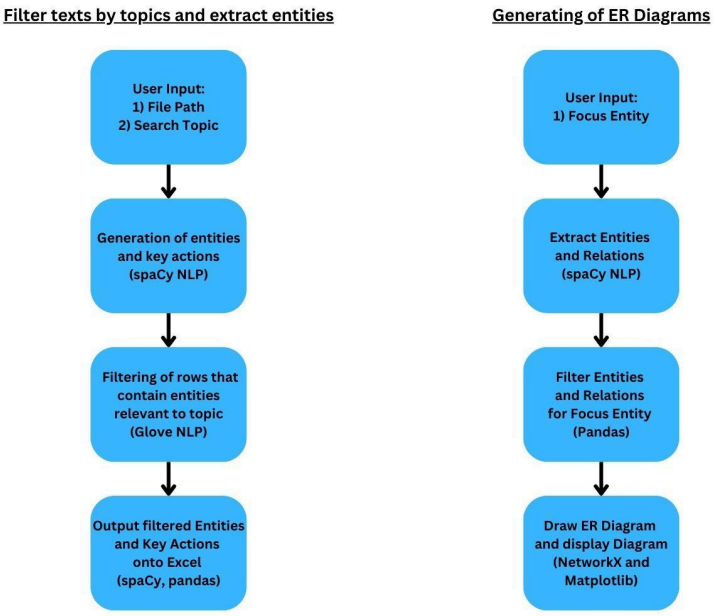
Solution Architecture

The architecture of our solution is built to extract, process, and visualize unstructured text data efficiently, providing scalability, usability, and performance.

Our architecture is broken down into two components:

1) Filter Texts by user inputted topic, extract entities onto an Excel File (spaCy NLP)
2) Extracting entities and relationships to generate interactive ER Diagrams (NetworkX, Pyvis)

This combined approach allows users to easily process and analyze complex data, ensuring a seamless and efficient experience.



Component One:

1. User Input: The user provides a file path and a search topic via CLI.
2. Entity and Key Action Extraction: spaCy is used to process the text and extract entities and key actions from the dataset.
3. Entity Filtering: The extracted entities are filtered based on relevance to the search topic, using GloVe embeddings for semantic similarity.
4. Filtering with Pandas: Pandas is used to filter the rows that contain relevant entities and key actions.

5. Excel Output: The filtered entities, key actions, and generated summaries are saved to an Excel file using Pandas for easy review and analysis.

The solution extracts entities and key actions from text using spaCy NLP, filters them based on a user-specified topic, outputting everything into an Excel file.

Component Two:

1. Entity and Relation Extraction: spaCy is used to extract entities and their relationships from the provided text.
2. Filtering for Focus Entity: Pandas filters the extracted entities and relations based on the user's specified focus entity.
3. ER Diagram Generation: NetworkX creates a directed graph of entities and relations, while Pyvis visualizes the graph as an interactive ER diagram.
4. User Input: The user provides a focus entity to explore its related entities and relationships.

The solution extracts entities and their relationships using spaCy, filters them based on a focus entity provided by the user, and visualizes the results in an interactive ER diagram using NetworkX and Pyvis.

---

6. Conclusion

The solution architecture is built around a seamless flow from data ingestion to visual output. By using NLP for entity extraction, along with powerful data transformation and visualization tools, we provide users with the ability to easily analyze and visualize complex relationships within unstructured text. The system is flexible, scalable, and user-friendly, ensuring that users can quickly gain meaningful insights and take actionable steps from the data.