**Project Summary**

1. **Text Extraction**

* **Types of Documents:** Transcripts, Policy Documents, and Publications.
* **Processing Approach:**
  + Each document has its own structure and is processed separately.
  + Text extraction and further processing are done using NLP techniques like regex pattern matching.
* **TF-IDF Based Analysis:**
  + Created a **TF-IDF term-document matrix** for each document to enable keyword-based search and ranking.
  + Used keyword searches to identify **highly relevant documents** related to topics of interest: **Dairy, Carbon, Water, Emission, and Livelihood**.
  + Selected the **most relatable documents** by plotting them on a line chart and identifying the **elbow point/drop in word frequency**.

**2. Entity Extraction from Transcripts**

* **Entities Extracted:**
  + **Policies**
  + **Stakeholders**
  + **Topics**
* **Processing Steps:**
  + Used ChatGPT API to extract **Policies, Stakeholder Organizations/Groups, and Topics of relevance** from speaker-specific text.
  + Analysed the extracted text to determine the **speaker’s stance** on the identified policy/stakeholder/topic.
  + Performed **sentiment analysis** (Positive / Negative / Neutral) on the summarized speaker statement.
* **Details on Extracted Entities:**
  + **Policy:**
    - Entities: **Speaker, Policy, Speaker’s stance on the policy, and Overall sentiment**.
    - Helps determine **which speaker mentioned which policy** and their opinion on it.
  + **Stakeholder:**
    - Entities: **Speaker, Stakeholder Organization/Group, Speaker’s stance on the stakeholder, and Overall sentiment**.
    - Identifies **which stakeholders are mentioned by which speaker** along with their opinions.
  + **Topic:**
    - Entities: **Speaker, Topic of interest (Dairy, Water, Emission, Carbon, and Livelihood), Speaker’s stance on the topic, and Overall sentiment**.
    - Provides insights into **what each speaker thinks about the related topics**.

**3. Entity Extraction from Publications**

* **Processing Approach:**
  + Used ChatGPT API to extract **entities from dairy-related paragraphs** in publication documents.
* **Extracted Entities:**
  + **Primary Stakeholder:** The main stakeholder around whom the text revolves.
  + **Secondary Stakeholder:** Associated with the primary stakeholder through relationships like **collaboration, funding, etc.**
  + **Context:** Captures a summarized relationship between the **primary and secondary stakeholders**.
  + **Topic Tag:** A **one-word summary** of the stakeholder relationship (e.g., **Funded, Collaborated**).

**4. Entity Extraction from Action Plans and Strategies**

* **Action Plan Documents:**
  + Extracted details on **Action Plans, Responsible Bodies, and Timeframes** from structured tables in PDFs.
  + Established potential links between **action plans and associated stakeholders**.
* **Strategy Documents (e.g., *Food Vision 2030*):**
  + Follow a **hierarchical structure**:
    - **Mission → Goals → Actions**
  + Extracted **Missions, Goals, and Actions** using pdfplumber and **NLP techniques**.
  + Used **ChatGPT API** to extract stakeholders from **action descriptions**.
  + Strategy documents were also linked with **Annual Reports** to:
    - Identify **Programs/Initiatives** for each action.
    - Extract **entities, stakeholders, program/initiative names, and summaries of relevant reports**.

**5. Knowledge Graph Creation**

* **Objective:** Establish a structured network linking **stakeholders, policies, topics, action plans, and inter-stakeholder relationships** extracted from various document sources.
* **Entities Used:**
  + **Stakeholders (organizations, groups, policymakers, etc.)**
  + **Policies (laws, regulations, strategies, and guidelines)**
  + **Topics of Interest (Dairy, Carbon, Water, Emission, Livelihood, etc.)**
  + **Action Plans (missions, goals, and initiatives from strategy documents)**
  + **Relationships (funding, collaboration, opposition, support, etc.)**

**6. Final Tool Features**

* **Document Search Functionality:**
  + Supports **keywords, partial strings, and phrases** with **AND/OR** operators.
  + Search can be performed across:
    - **Entire documents**
    - **Speaker-specific text segments** (to determine which speaker/document is most relevant).
  + **Phrase search** implemented using a **custom vectorizer** that extracts phrases before vectorization.
* **Knowledge Graph Visualization:**
  + Currently in development to create **query-specific knowledge graphs**.
  + Aims to provide an intuitive **graph-based query system** for:
    - Analysing stakeholders
    - Identifying links between **stakeholders, policies, topics of interest, and other entities**.