Accelerating Cleantech Advancements through NLP-powered Text Mining and Knowledge Extraction

From Exploratory Text Analysis to Large Language Models (LLMs) Applications

Background

At a time when tackling environmental challenges is of paramount importance, the cleantech industry has a central role to play in promoting sustainable solutions. To foster the next generation of natural language processing (NLP) enthusiasts and innovators in the cleantech industry, we present the project titled “Accelerating Cleantech Advancements through NLP-powered Text Mining and Knowledge Extraction”. This project aims to use NLP techniques to accelerate the process of text analysis, knowledge acquisition and innovation in the cleantech sector. NLP technology can play a crucial role in this area as it helps us to extract valuable insights from large amounts of text data. Through this project, students will learn the intricacies of NLP, gain expertise in cleantech, and have the opportunity to make a meaningful impact on the environment.

As part of our project, analyzing media and patent publications on cleantech topics is important as this data can serve as a rich source of knowledge to accelerate innovation. By examining these documents, students can uncover emerging trends, identify key players, and gain a deep understanding of cutting-edge technologies in the cleantech space, which is an essential part of our multi-faceted NLP-powered text mining and knowledge extraction process. This data can reveal not only the state of the art in the field, but also potential white spots where innovation is needed. In addition, the analysis of specialized texts enables the identification of critical technological gaps and opportunities for collaboration, making them a valuable resource for cleantech researchers, entrepreneurs, and policy makers, contributing to our vision of a more sustainable future.

Goals

In this project, we start with exploratory text analysis including topic modeling to understand cleantech media and patent data. We then train word and sentence embedding models to better represent cleantech innovation content and develop a question answering and information retrieval system that can help stakeholders facilitate cleantech innovation. The goal of this project is threefold:

- **Advanced Text Analytics Mastery**. Expand applications of NLP techniques, including working with state-of-the-art large language models (LLMs) for deeper text analytics and extracting insights from specialized media and patent texts.

- **Cleantech Expertise and Innovation Acceleration**. Use the analysis of both patent and media datasets to gain a comprehensive understanding of the cleantech sector and conduct a comparative analysis to identify gaps between market trends and patented technologies and suggest areas for future innovation and research.

- **Better Communication and Collaboration**. Foster the ability to summarize and present complex results and promote interdisciplinary collaboration across the cleantech ecosystem.

Organizers

The project is organized by:

Dr. Janna Lipenkova, CEO, Equintel GmbH, Germany
Dr. Susie Xi Rao, Researcher at ETH Zurich, Switzerland
Dr. Guang Lu, Lecturer for Data Science, Lucerne University of Applied Sciences and Arts (HSLU), Switzerland

in collaboration with:

Dr. Diego Antognini, Senior Research Scientist, Google DeepMind in Zurich, Switzerland

as a joint task of industry and academia for the HSLU Applied Information and Data Science Master's Program in the course Computational Language Technologies.

Task

To realize the goals of this project, we provide two comprehensive datasets: the **Cleantech Media Dataset**, which consists of recent media articles and publications, and the **Cleantech Google Patent Dataset**, which contains detailed information on cleantech-related patents. This dual dataset approach provides a unique opportunity to apply NLP techniques to not only understand the current discourse in the cleantech industry, but also to explore the landscape of technological innovation and intellectual property. The main objective is to utilize advanced NLP methods to gain a deeper understanding and facilitate knowledge extraction from this rich textual data. The project focuses on the following areas:

● **Enhanced Exploratory Text Analysis**. Objective: developing comprehensive analytical skills to explore and interpret complex datasets with a focus on uncovering hidden knowledge in the cleantech field. Students will compare emerging trends, major players, and technological advances as discussed in media texts with those documented in patents. Tasks include: analyzing two datasets, identifying trends and patterns unique to each dataset, and cross-referencing to discover overlaps and gaps between market discussions and patented technologies.

● **Advanced Text and Sentence Embedding Techniques**. Objective: mastering the development and application of modern embedding models tailored to the nuanced requirements of cleantech data. This includes utilizing the latest advances in NLP to create representations that capture the essence of cleantech innovations and discourse in various text types. Tasks include: training custom embeddings for each dataset, experimenting with state-of-the-art models, and conducting comparative analyses to evaluate how different text sources affect embedding features.

● **Innovative Question Answering and Information Retrieval Systems**. Objective: developing and implementing advanced question answering (QA) and information retrieval systems that leverage the strengths of both datasets to provide comprehensive, accurate, and relevant information. This includes the development of systems capable of handling the complexity of both media reports and technical patent documents to support knowledge discovery and innovation scouting in the cleantech sector. Tasks include: extracting important information to create a high-quality data set for answering queries, developing a retrieval-augmented generation (RAG) system that effectively uses insights from media articles and patent documents, and systematically evaluating and improving the developed RAG system.

Datasets

Cleantech Media Dataset:

cleantech_media_dataset_v3_2024-10-28.csv

cleantech_rag_evaluation_data_2024-09-20.csv

https://www.kaggle.com/datasets/jannalipenkova/cleantech-media-dataset.

Cleantech Google Patent Dataset:

CleanTech_22-24_updated.json

https://www.kaggle.com/datasets/prakharbhandari20/cleantech-google-patent-dataset?resource=download.

Reference Pipeline

Please note that this pipeline is an important reference for the work done by the students. It also serves as **the basis for the assessment** of the final work.

Stage 1: Enhanced Data Cleaning, Preprocessing, and Exploratory Analysis

Objective: Analyzing both the Cleantech Media Dataset and the Cleantech Google Patent Dataset to identify emerging trends, technologies, and potential innovation gaps in the cleantech sector.

- **Data Collection and Cleaning**
  - Download and load the Cleantech Media Dataset and the Cleantech Google Patent Dataset.
  - Perform an initial data cleaning to remove e.g. duplicates and irrelevant information from both datasets.

- **Text Preprocessing**
  - Tokenize the text data from both datasets.
  - To refine the data, apply techniques such as stemming and lemmatization, remove stop words and non-informative terms, and convert text to lowercase for consistency.

- **Exploratory Data Analysis**
  - Perform separate and comparative exploratory data analysis (EDA) on both datasets, such as temporal analysis and sentiment analysis, to understand the landscape of cleantech innovations and patents.
  - Use pre-trained Named Entity Recognition (NER) models (like spaCy, Hugging Face Transformers) to extract companies and technologies, then build a co-occurrence matrix representing the frequency with which a company is mentioned together with a particular technology in the text. Construct a graph where nodes represent entities, and edges reflect relationships, and analyze it using centrality and clustering for insights.
  - Use visualization techniques such as word clouds, bar charts, scatter plots and NetworkX (with Matplotlib) to illustrate the results.

- **Topic Modeling**
  - Test topic modeling techniques such as LDA and NMF (https://github.com/AnushaMeka/NLP-Topic-Modeling-LDA-NMF), Top2Vec (https://github.com/ddangelov/Top2Vec), and BERTopic (https://github.com/MaartenGr/BERTopic), evaluate the quality of the topics using such as coherence metrics, and refine the topic model based on evaluation results and domain expertise.
  - Apply hierarchical topic modeling to explore more granular subtopics (https://maartengr.github.io/BERTopic/getting_started/hierarchicaltopics/hierarchicaltopics.html) within major cleantech technologies (e.g., solar energy subtopics).
  - Visualize and interpret the topics, comparing emerging trends in media publications against

focuses of recent patents.

Outputs:

- Notebook with data cleaning and preprocessing steps.
- Notebook with EDA visualizations on e.g. hidden topics and the detailed comparison between the two datasets.

Stage 2: Advanced Embedding Models Training and Analysis

Objective: Developing and utilizing advanced embedding models to represent the content of Cleantech Media and Google Patent datasets and compare domain-specific embeddings to gain unique insights.

● Data Preparation for Embeddings

- Preprocess the text data from both datasets to ensure that it is clean and suitable for embedding training.
- Create training and validation sets for both media and patent texts.

● Word Embedding Training

- Train separate word embedding models on each dataset using techniques such as Word2Vec, FastText, or GloVe.
- Experiment with hyperparameters such as vector dimensions, context window size, and training epochs to optimize word embeddings evaluated using intrinsic methods such as word similarity tasks, analogy tasks and clustering and visualization.
- Use the trained embeddings to explore thematic overlaps and differences between the two datasets and identify unique insights and innovation gaps.

● Sentence Embedding Training

- Train separate sentence embedding models on each dataset using methods such as averaging word vectors, Doc2Vec, or BERT embeddings.
- Experiment with hyperparameters such as vector dimensions, context window size, learning rate, batch size and training epochs to optimize sentence embeddings evaluated using intrinsic methods such as sentence similarity tasks and clustering and visualization.
- Use the trained embeddings to explore thematic overlaps and differences between the two datasets and identify unique insights and innovation gaps.

● Transfer Learning with Advanced Open-Source Models

- Implement transfer learning by fine-tuning pre-trained open-source models such as RoBERTa, XLNet, Longformer, FLAN-T5, and BART on the text data. Evaluate the model performance using intrinsic measures (e.g., word similarity, clustering quality) before and after fine-tuning. Analyze and quantify the insights gained from the fine-tuned model regarding emerging trends and innovation gaps in cleantech.
- Compare the performance of transfer learning with the in-house embeddings. This comparison could be done through evaluating the effectiveness of the embeddings in domain-specific tasks like topic classification.

Outputs:

- Notebook with annotated model training steps.
- Notebook with visualizations comparing the performance of the embedding models and the

insights of the two datasets.

Stage 3: Implementing an RAG System for Question Answering

Objective: Implementing an advanced QA and information retrieval system using the latest LLMs and RAG that enable detailed investigation and comparison of findings across Cleantech Media and Google Patent datasets.

Recommended Reading: Janna Lipenkova (to appear). The Art of AI Product Development, Chapter 7 on RAG systems. Preliminary chapter draft:

https://docs.google.com/document/d/1D3MmEfw1_CjvEEqfK0SdRj3eZvwTKEDgZHl0JNIEBro/edit?usp=sharing

● Generate and Categorize QA Pairs

- Choose 50-100 relevant paragraphs from the Cleantech Media and Google Patent datasets that cover diverse topics within cleantech.
- Use LLMs (e.g., GPT-4o, Llama 3.3) to generate QA pairs based on the selected paragraphs. For each selected paragraph, prompt the LLMs like: “Based on the following reference text, generate a relevant question and its corresponding answer.” Include the paragraph in the prompt for context.
- Create at least 200-300 new QA pairs from the selected paragraphs, ensuring a wide range of topics and difficulty levels. After generating the pairs, manually review the quality of the generated questions and answers, ensuring relevance and correctness.
- Identify 4-5 categories based on themes, complexity, or types of questions (e.g., factual, analytical, comparative). Group the generated QA pairs into the defined categories, providing a brief rationale for each classification.

● Run RAG Code and Analyze Insights

- Use an existing RAG implementation (see Prof. Dr. Daniel Perruchoud’s tutorial on RAG theory and implementation: https://github.com/LuciferUchiha/Cleantech-RAG) to set up the system.
- Use the seed evaluation dataset and the newly generated QA pairs to run the RAG system to generate answers and references for the given questions.
- Assess the system’s performance over different question categories considering context relevance, faithfulness, answer relevance etc.
- Analyze the generated answers and references using the developed RAG system on various input questions to identify trends, gaps, and noteworthy findings within the cleantech media and patent datasets, highlighting key observations and implications for the cleantech sector.

● Enhance RAG System

- Research current trends in RAG and retrieval-based models (e.g., advancements in context retrieval, improved architectures).
- Propose one or two enhancements to the existing RAG system, such as implementing newer retrieval methods or fine-tuning the language models on a domain-specific dataset. Cite the research papers that you refer to if this is relevant.
- Summarize your proposals in the notebook in detail or ideally test your proposed methods to quickly demonstrate the performance difference to the existing RAG pipeline.

● Testing in RAG Playground [Optional]

Students can provide a simple API to their retrieval components to be plugged into Equintel’s browser-based playground. The playground offers a frontend for wider testing and accesses common LLMs by OpenAI and Anthropic. It will be made available to all participants, allowing students to test their systems in a setting that is close to production. The API should have one single endpoint / sources and take a

parameter question which represents the user question:

http://123.456.78.90/source?question=<user question>

Its output should be a JSON array of source objects with the following obligatory fields:

- url: url of the document
- chunk: relevant chunk from the document

Optionally, additional information deemed useful by the students can also be added to the source objects, for example a relevance score or the date of the article (for time-sensitive retrieval).

More information about the playground integration will be provided later during the course.

Outputs:

- Notebook describing the structure of the RAG system, the experiments including the system performance evaluation using the created QA dataset, and your proposal to improve the current RAG system performance based on researching trending development.
- Notebook for visualizing the findings obtained through the cross-dataset comparison to highlight the insights of learning for the cleantech sector.
- [Optional]: API link for playground integration.

Requirements

(1) Students form groups of size three to tackle this project as a team.

(2) It is not required to write a final report documenting the analysis and development steps. However, clear code structure, annotations, explanations and visualization of the results in the notebook are expected. It would be great if the notebook can be submitted in the style of a “tutorial”.

(3) Each group member should clearly indicate their contribution in the notebook.

(4) You can start the project with Google Colab. For additional computing and RAM resources, you can apply for Colab Pro or Colab Pro+. We will ask the study program to reimburse the fees incurred.

(5) Please submit your notebook for the three stages by these deadlines accordingly, following the submission link and guidance given on the ILIAS:

- Deadline 1 (Stage 1): 16 March 2025 23:59
- Deadline 2 (Stage 2): 6 April 2025 23:59
- Deadline 3 (Stage 3): 27 April 2025 23:59