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. Students participating in this project will embark on a journey to uncover hidden information treasures, enable groundbreaking research, and contribute to creating a more sustainable future. 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, as we mentioned earlier, 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, further 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/or 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 two 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, Research Scientist, Google Research 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 equip students with two comprehensive datasets: the Cleantech Media Dataset, which consists of recent 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/or 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 key information to create a high-quality dataset for answering queries, fine-tuning the latest LLMs for cleantech-specific queries, and developing a query system that effectively utilizes insights from media articles and patent documents.

Dataset

Cleantech Media Dataset: https://www.kaggle.com/datasets/jannalipenkova/cleantech-media-

dataset.

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

Pipeline

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**
  - Acquire 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 word 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 to understand the landscape of cleantech innovations and patents.
  - Use visualization techniques such as word clouds, bar charts and scatter plots 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, and refine the topic model based on evaluation results and domain expertise.
  - 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.

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 considering the unique characteristics of 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.

● **Sentence Embedding Training**

- Train separate sentence embedding models on each dataset using methods such as averaging word vectors, Doc2Vec, or BERT embeddings.
- Fine-tune the sentence embeddings on the specific data.

● **Embedding Model Evaluation**

- Assess the quality of both word and sentence embeddings using intrinsic evaluation methods, including word similarity and analogy tasks.
- Use the trained embeddings to explore thematic overlaps and differences between the two datasets and identify unique insights and innovation gaps.

● **Transfer Learning with Open-Source Models [Optional]**

- Implement transfer learning by fine-tuning pre-trained open-source models such as BERT or GPT-2 on the text data.
- 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: Developing Advanced Question Answering and/or Information Retrieval System

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

● **Question Answering System Development**

- Extract key sentences using such as TextRank (https://github.com/davidadamojr/TextRank) or BERT Extractive Summarizer (https://pypi.org/project/bert-extractive-summarizer/) from the two given datasets.
- Generate a question and an answer for each sentence using recent LLMs of your choice.
- Manually clean up the generated question-answer pairs to create a high-quality QA dataset in both the media and patent datasets.
- Use the prepared QA datasets to fine-tune the recent LLMs of appropriate size (due to possible limited computing resource) such as Zephyr-7B, Mistral 7B, Mixtral 8x7B, Tinyllama or Gemma 7B and evaluate model performance on new input data in the cleantech field.
- Evaluate the QA system's performance, comparing its effectiveness in extracting information from media vs. patent texts, and compare the above results with the zero-shot capability of LLMs such as ChatGPT and Llama-2.

Outputs:

- QA datasets that can be shared between groups for model training purposes.
- Training notebook, including the QA results on the trained models.

● **Information Retrieval System Enhancement [Optional]**

- Explore the retrieval-augmented generation (RAG) technique for generating answers based on retrieved documents and evaluate the efficiency of the system in providing comprehensive answers.
- Split the text data into chunks and store the corresponding embeddings with your trained model or other open-source models in an embedding database such as Pinecone or FAISS.
- Implement a query system that converts input queries into embeddings and perform dense retrieval on the embedding database, possibly augmenting it with reranking, lexical search and/or metadata filtering.
- Implement answer generation with LLMs based on the embedded query and the textual information retrieved in the previous step.

● **Cross-dataset Insights and Innovation Opportunities**

- Use the QA and/or query systems to identify and highlight innovation opportunities by comparing insights from media discussions with existing patents.
- Present findings that highlight gaps in the patent landscape that are ripe for innovation based on trends observed in the media dataset.

Outputs:

- Notebook describing the structure of the QA and/or RAG system and the experiments including the system performance evaluation using the QA dataset.
- Notebook for visualizing the findings obtained through the cross-dataset comparison.

Requirements

(1) Students form groups of size three to tackle this project as a team.
(2) It is not necessary to write a final report documenting the analysis and development steps. However, a clear code style, annotations, and a visualization of the results in the notebook are expected.
(3) It would be great if you could prepare the notebook in the style of a “tutorial”. 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:

- Deadline 1 (Stage 1): 28 April 2024, 23:59
- Deadline 2 (Stage 2): 19 May 2024, 23:59
- Deadline 3 (Stage 3): 9 June 2024, 23:59

References

[1] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

[2] Toetzke, M., Probst, B., & Feuerriegel, S. (2023). Leveraging large language models to monitor climate technology innovation. *Environmental Research Letters*, *18*(9), 091004.

[3] Excellent tutorial on retrieval-augmented generation (RAG): https://www.anyscale.com/blog/a-comprehensive-guide-for-building-rag-based-llm-applications-part-1.