Presentation 2024

TEXT SUMMARIZATION

Mentor: Mr. Narendra Kumar

Intern: Nikhilkumar A. Jain
Nebu C Thomas
Nitesh Sachan

Group 2

Overview

- Introduction
- Problem Statement
- Project Planning
- Architecture
- Workflow
- Dataset
- Model Training
- Model Evaluation
- Testing
- Deployment
- Deployment - Results

INTRODUCTION

Problem Statement & Planning

Problem Statement

- We are developing an automated text summarization system to efficiently condense large volumes of text into concise summaries is crucial for improving business operations.
- This project aims to utilize NLP techniques to create a powerful text summarization tool that can manage a variety of documents across multiple domains.
- The system should produce high-quality summaries that preserve the essential information and contextual meaning of the original text.

Project Planning :

Text Summarization

Re: Report Meeting

Date: 2024

2 Week

FIRST STEPS

Background Research
And Problem
statement
understanding

2weeks

DATA COLLECTION

1.Finalize Dataset and begin preprocessing Clean, tokenize, and prepare text data
2.Load data

3weeks

MODELING

1.Model Training

2.Model Evaluation

3.Fine Tuning Model

DEPLOYMENT

1.Deploy project on cloud for public use.

2.Presentation

ARCHITECTURE

Selected Architecture And Methods

Methods Available :

The architecture of the problem statement can be developed by two methods

Abstractive Model

refers to a type of model that generates summaries by interpreting and paraphrasing the content of the input text rather than directly selecting and extracting existing sentences or phrasess

Extractive Model

refers to a type of model where sentences or phrases are selected from the original text to create a summary. Rather than generating new sentences, as in abstractive summarization,

Abstractive Model :

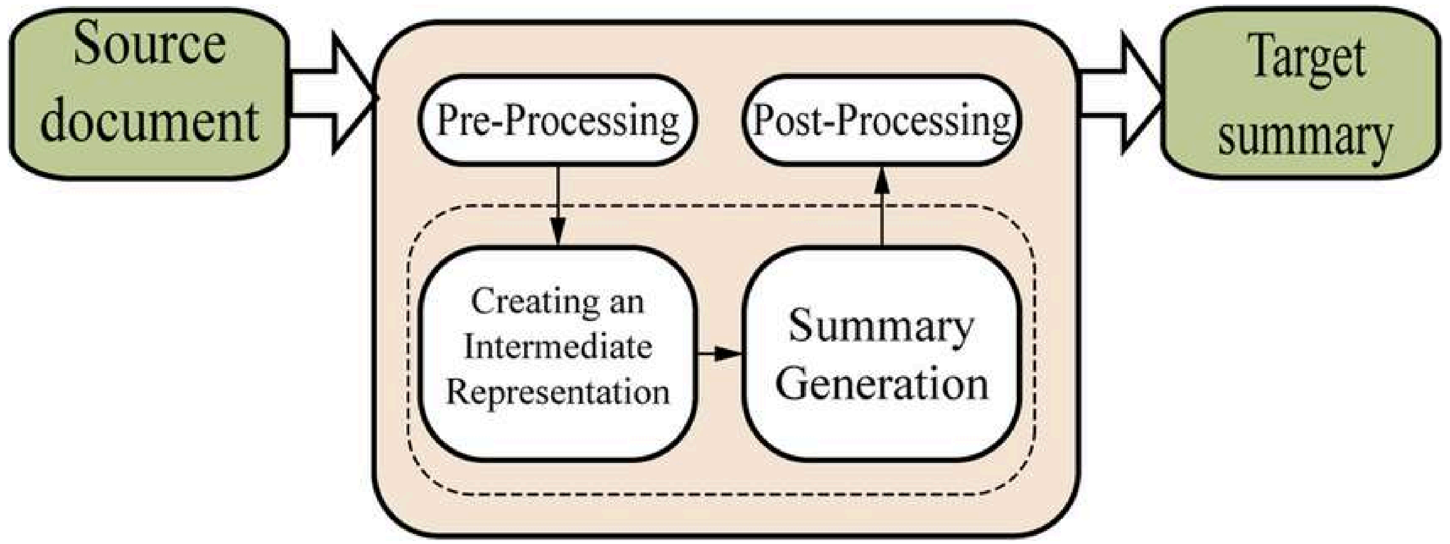


Fig Abstractive 1.0

Key Characteristics of Abstractive Models:

- Language Understanding: Abstractive models typically use deep learning techniques, often based on transformer architectures, to understand the meaning and context of the input text.
- Content Synthesis: Instead of copying sentences verbatim, abstractive models generate new phrases and sentences to convey the main points of the text in a more concise form.
- Paraphrasing and Simplification: These models can rephrase complex sentences, remove redundant information, and consolidate multiple ideas into a shorter summary.
- Contextual Awareness: They maintain context across sentences, ensuring that the summary captures the essential information and maintains coherence.
- Naturalness: Abstractive summaries are designed to read more naturally, resembling summaries written by humans rather than being a concatenation of extracted text segments.

Extractive Model :

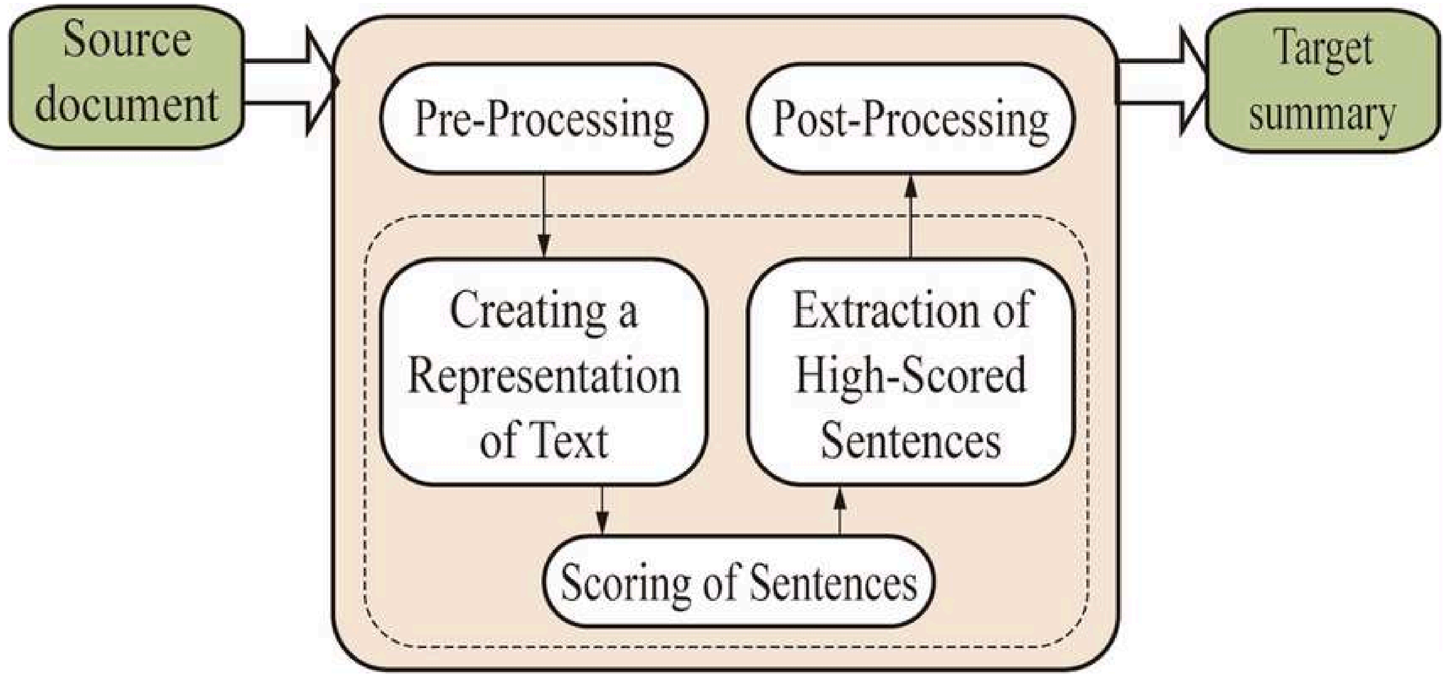


Fig Extractive 1.0

Key Characteristics of Extractive Summarization:

1. Sentence Selection: Extractive models identify sentences or passages that contain crucial information based on predefined criteria such as importance, relevance, or frequency of appearance.
2. No Sentence Modification: Unlike abstractive summarization, extractive methods do not modify the content of the selected sentences. They are used as-is from the original text.
3. Preservation of Originality: Extractive summaries often maintain the original wording and structure of the text segments that are extracted, ensuring the summary reflects the exact content found in the source material.
4. Scoring and Ranking: Techniques such as graph-based algorithms (e.g., TextRank) or machine learning models (e.g., supervised classifiers) are commonly used to score sentences and select the top-ranked ones for inclusion in the summary.
5. Efficiency: Extractive summarization can be computationally less intensive compared to abstractive methods, as it involves straightforward sentence selection rather than generating new text.

Selected Model for Abstractive : Google/Pegasus-large

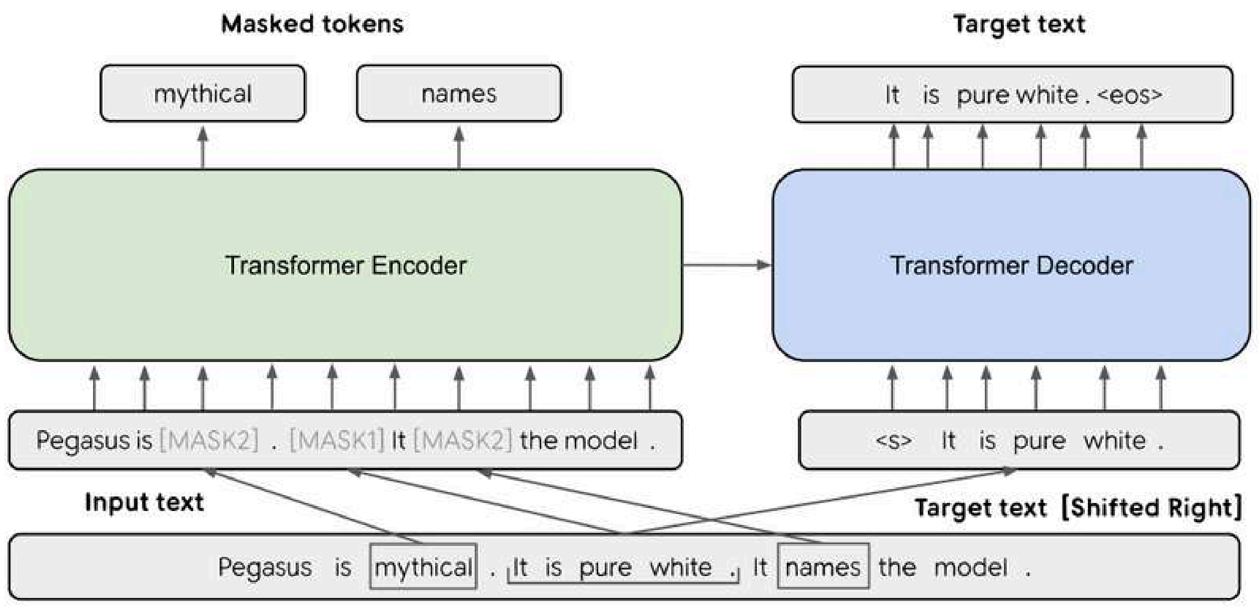


Fig. Pegasus architecture

- **Leverage Pegasus Strengths:** Pegasus is specifically designed for abstractive summarization, making it a great choice for generating summaries that capture the essence of the text without simply copying sentences.
- **Domain-Specific Tuning:** Fine-tune Pegasus on CNN/Daily Mail and XSum to improve its understanding of news language and structure. This will help it generate summaries that are relevant and informative for news articles.
- **Focused Training:** Fine-tuning requires training only a portion of Pegasus, focusing its learning on news summarization tasks. This leads to faster training times and more efficient model updates.
- **Continuous Learning:** Fine-tuning allows you to easily adapt Pegasus to new types of news data in the future. As news formats evolve, you can fine-tune it to stay relevant.
- **Striking a Balance:** As with other models, find the ideal balance between leveraging Pegasus's pre-trained abilities and specializing it for news summarization. This ensures it can handle unseen news articles effectively.

PROPOSALS

Selections And Workflow

Selected Workflow :

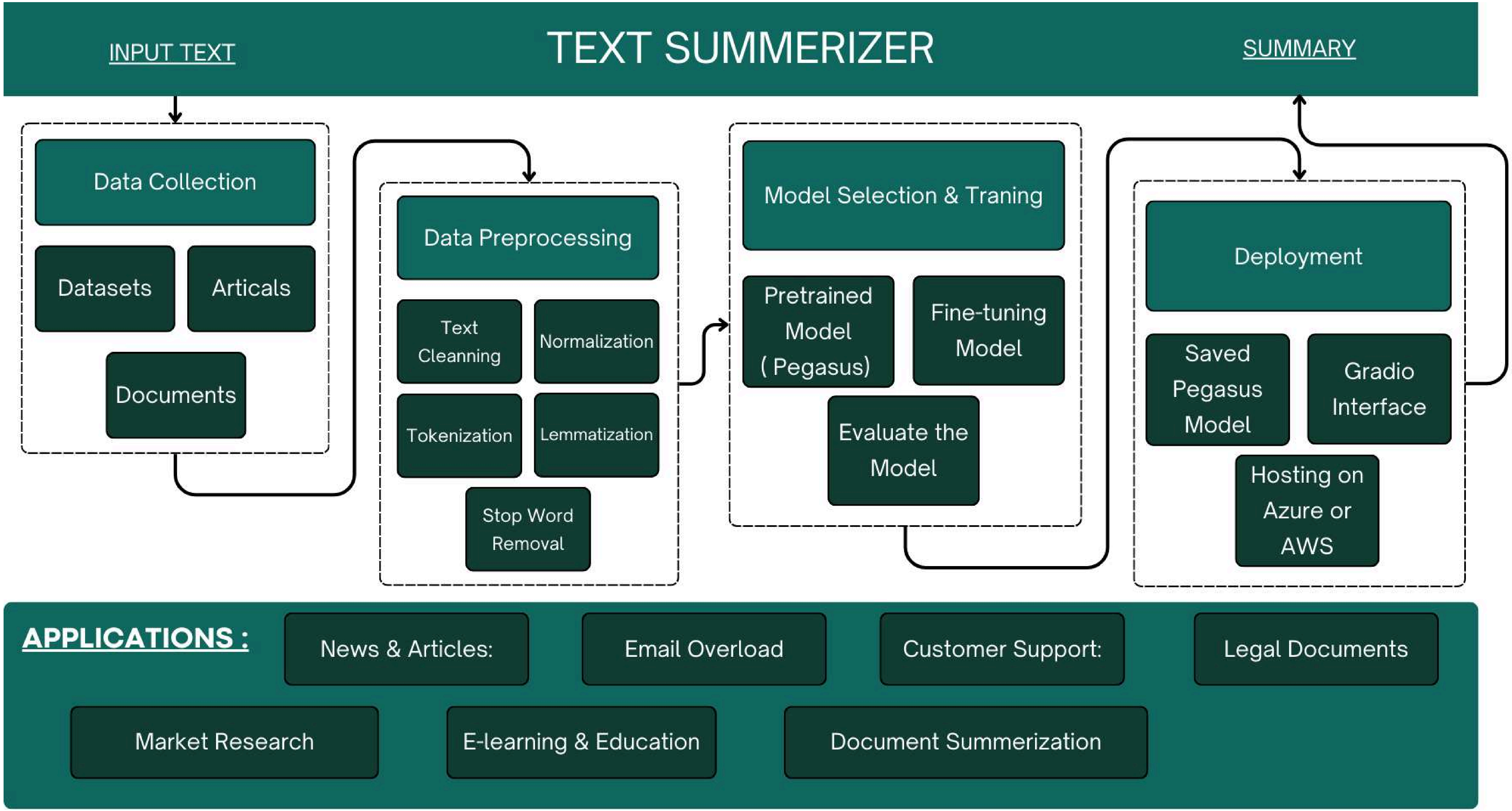


Fig. : Proposed Workflow

Selected Dataset

- **CNN/Daily Mail Dataset**
  - CNN/Daily Mail is a widely used dataset for text summarization, containing news articles and their corresponding human-written summaries.
  - It offers a large collection of text pairs, making it suitable for training deep learning models.
  - The dataset allows researchers to compare and evaluate different summarization algorithms.
  - **Number of Documents:**
    Train: 287k **Validation :**13.4k **Test:** 11.4k **Total :** 226,711

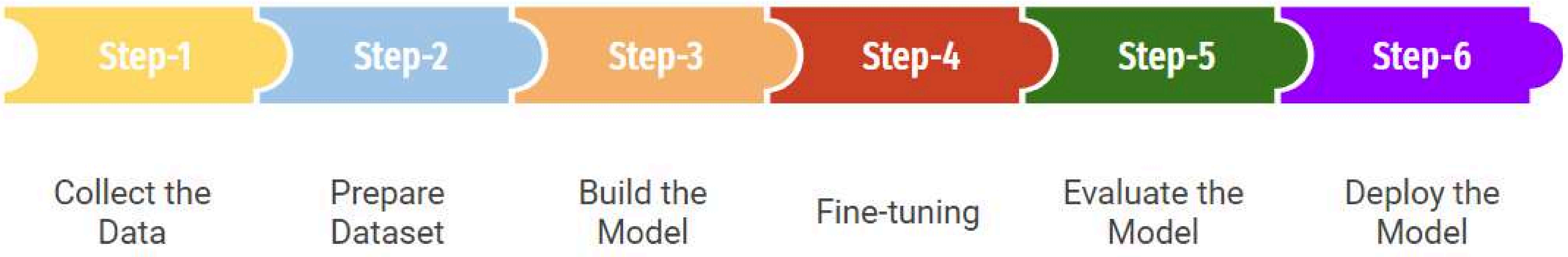
article string · lengths 48 15.9k	highlights string · lengths 14 7.39k
LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a…	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young…
Editor's note: In our Behind the Scenes series, CNN correspondents share their…	Mentally ill inmates in Miami are housed on the "forgotten floor" Judge Steven Leifman…
MINNEAPOLIS, Minnesota (CNN) -- Drivers who were on the Minneapolis bridge when it…	NEW: "I thought I was going to die," driver says . Man says pickup truck was folded in…
WASHINGTON (CNN) -- Doctors removed five small polyps from President Bush's colon on…	Five small polyps found during procedure; "none worrisome," spokesman says .…
(CNN) -- The National Football League has indefinitely suspended Atlanta Falcons…	NEW: NFL chief, Atlanta Falcons owner critical of Michael Vick's conduct . NFL…
BAGHDAD, Iraq (CNN) -- Dressed in a Superman shirt, 5-year-old Youssif held his sister's…	Parents beam with pride, can't stop from smiling from outpouring of support . Mom:…

- **XSum Dataset**
  - XSum is another popular dataset for text summarization, consisting of extracted news articles and their corresponding summaries.
  - It focuses on longer documents than CNN/Daily Mail, offering a different challenge for summarization models.
  - XSum allows researchers to explore the task of summarizing longer and more complex texts.
  - **Number of Documents:**
    Train: 204,045 **Validation** :11,332 Test: 11,334 **Total** : 226,711

document string · lengths 0 174k	summary string · lengths 1 399
The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed. Repair…	Clean-up operations are continuing across the Scottish Borders and Dumfries and Galloway after flooding…
A fire alarm went off at the Holiday Inn in Hope Street at about 04:20 BST on Saturday and guests were…	Two tourist buses have been destroyed by fire in a suspected arson attack in Belfast city centre.
Ferrari appeared in a position to challenge until the final laps, when the Mercedes stretched their legs to…	Lewis Hamilton stormed to pole position at the Bahrain Grand Prix ahead of Mercedes team-mate Nico Rosberg.
John Edward Bates, formerly of Spalding, Lincolnshire, but now living in London, faces a total of 22 charges…	A former Lincolnshire Police officer carried out a series of sex attacks on boys, a jury at Lincoln Crow…
Patients and staff were evacuated from Cerahpasa hospital on Wednesday after a man receiving treatment…	An armed man who locked himself into a room at a psychiatric hospital in Istanbul has ended his threat…
Simone Favaro got the crucial try with the last move	Defending Pro12 champions Glasgow Warriors bagged a

Selected Model Tranning :

6 Steps to fine-tune the Pegasus Model



- **Challenge:**
  - Limited Hardware ResourcesTraining large models requires significant computational power.
  - Your machine might not have enough resources to train a very large model effectively.

Model Tranning Process : Abstractive

Initial Model Selection: BART-base

- **Pros:**
  - Powerful transformer model with strong language understanding.
- **Cons:**
  - Large size - can be resource-intensive for local machines, leading to:
    - Slow training times.
    - Memory limitations, potentially causing training failures.

- **Key Point:** BART-base: Powerful But Resource-Intensive

Efficient Model Selection (T5)

- **Reasoning:**
  - Switched to T5, a smaller and more efficient model.
- **Pros:**
  - Faster training on your machine due to its smaller size.
  - More suitable for available hardware resources.

- **Key Point:** T5: Efficient Model for Local Machine Training

Results and Next Step: Exploring Pegasus

- **Observations with T5:**
  - Achieved baseline performance.
- **Key Point:** T5 Results: Achieved Baseline Performance
- **Exploration:**
  - Considered Pegasus, a model specifically designed for abstractive summarization.
- **Pros:**
  - Optimized architecture for abstractive summarization tasks.
  - Potentially leads to higher quality and accuracy in summaries4

.

- **Key Point:** Pegasus: Exploring Model Strength for Abstractive Summarization

Find the trained Models here:
https://drive.google.com/drive/folders/1WQnco5vl_6GoBoOaGblcqpTuK_Rp6f6g?usp=sharing

Access GitHub Repository here :
https://github.com/Jain-nikhilkumar/-Text-Summarization-with-NLP

Model Tranning Process : Extractive

Preprocessing:

- Tokenization: The document is split into individual words or tokens.
- Sentence Segmentation: Sentences are identified within the document.

Feature Extraction:

- Text Representation: Convert sentences into numerical representations (vectors) using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec, GloVe).
- Sentence Embeddings: Each sentence is transformed into a dense vector representation capturing its semantic meaning.

Sentence Scoring:

- Importance Calculation: Calculate the importance or relevance of each sentence using various criteria:
  - TF-IDF Scores: Sentences containing frequently occurring and unique terms are considered more important.
  - Positional Information: Beginning or ending sentences may carry more significance.
  - Sentence Length: Longer sentences may contain more information.
  - Named Entity Recognition (NER): Sentences containing named entities (e.g., people, organizations) may be deemed more important.
  - Graph-Based Algorithms: Apply algorithms like TextRank or LexRank, which treat sentences as nodes in a graph and use graph-based ranking methods (similar to Google's PageRank) to determine sentence importance based on connections (edges) between sentences.

Model Tranning Process : Extractive

Sentence Selection:

- Thresholding: Set a threshold score or rank to select sentences that exceed this threshold.
- Top-N Selection: Select the top N sentences with the highest scores to include in the summary.
- Redundancy Removal: Ensure selected sentences cover diverse aspects of the document to avoid redundancy.

Summary Construction:

- Combine the selected sentences to form the extractive summary.
- Maintain the order of selected sentences as they appear in the original document to preserve coherence.
    - eat sentences as nodes in a graph and use graph-based ranking methods (similar to Google's PageRank) to determine sentence importance based on connections (edges) between sentences.

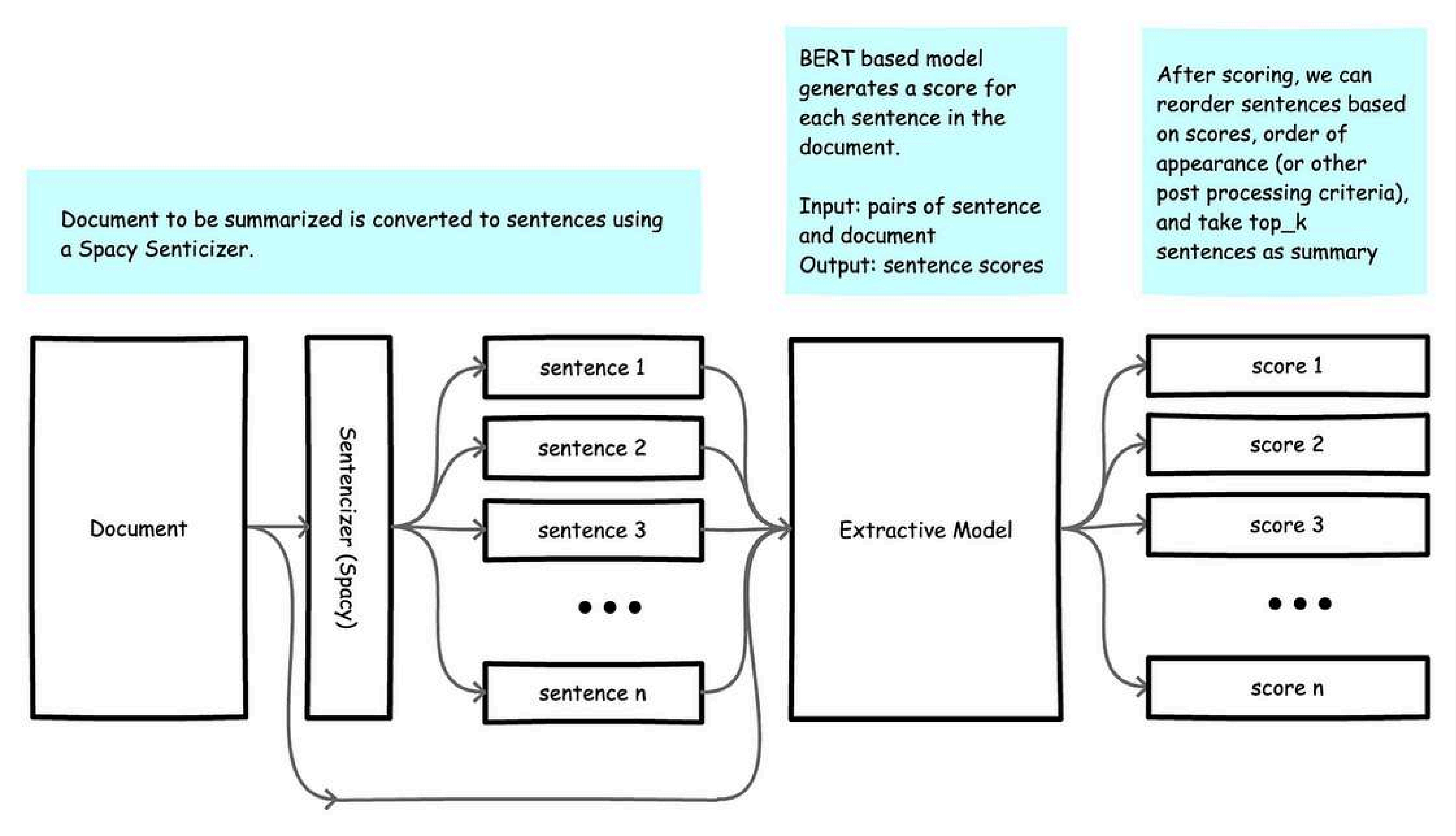


Fig. Extractive 2.0

Selected Model Evaluation : Abstractive

- **Performance Metrics – ROUGE (Recall-Oriented Under study for Gisting Evaluation)**
- **ROUGE is an essential metric in text summarization used to evaluate the overlap between generated summaries and reference summaries.**
- **Other Options Available : BLEU (precision-focused).**

ROUGE-N: This metric evaluates how well the candidate summary matches the reference summaries by looking at the overlap of n-grams, which are sequences of n words in order.

- **ROUGE-1:**
  Focuses on the overlap of single words (unigrams) between the candidate and reference summaries.
- **ROUGE-2:**
  Measures the overlap of two-word sequences (bigrams).
- **ROUGE-L:**
  Evaluates the longest common subsequence (LCS) found between the candidate summary and the reference summaries.
- **ROUGE-LSUM:**
  A special version of ROUGE-L specifically tailored to assess the quality of summaries.

Model's Performance:

- **ROUGE:** Overlap with reference summaries
- ROUGE-1: 0.7234
- ROUGE-2: 0.5265
- ROUGE-L: 0.7140

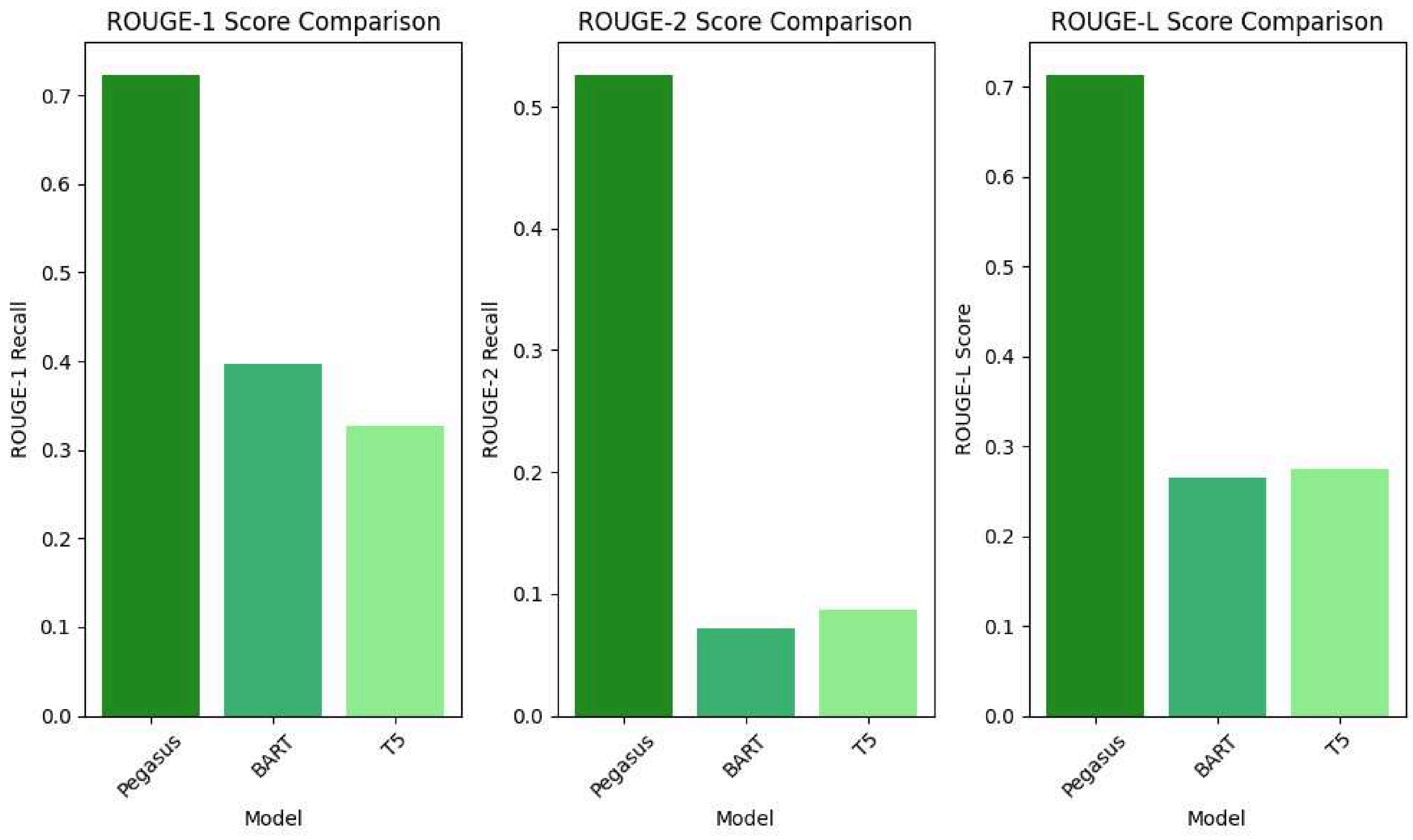


Fig Model Comparisons

Selected Model Evaluation : Extractive

- **Performance Metrics – ROUGE (Recall-Oriented Under study for Gisting Evaluation)**
- **ROUGE is an essential metric in text summarization used to evaluate the overlap between generated summaries and reference summaries.**
- **Other Options Available : BLEU (precision-focused).**

ROUGE-N: This metric evaluates how well the candidate summary matches the reference summaries by looking at the overlap of n-grams, which are sequences of n words in order.

- **ROUGE-1:**
  Focuses on the overlap of single words (unigrams) between the candidate and reference summaries.
- **ROUGE-2:**
  Measures the overlap of two-word sequences (bigrams).
- **ROUGE-L:**
  Evaluates the longest common subsequence (LCS) found between the candidate summary and the reference summaries.
- **ROUGE-LSUM:**
  A special version of ROUGE-L specifically tailored to assess the quality of summaries.

Model's Performance:

- **ROUGE:** Overlap with reference summaries
- ROUGE-1: 0.7234
- ROUGE-2: 0.5265
- ROUGE-L: 0.7140

TF-IDF Scores:

- ROUGE-1 (unigram overlap): Recall (r) = 0.625, Precision (p) = 0.909, F1-score (f) = 0.741
- ROUGE-2 (bigram overlap): Recall (r) = 0.500, Precision (p) = 0.800, F1-score (f) = 0.615
- ROUGE-L (longest common subsequence): Recall (r) = 0.625, Precision (p) = 0.909, F1-score (f) = 0.741

TextRank Scores:

- ROUGE-1 (unigram overlap): Recall (r) = 0.750, Precision (p) = 0.480, F1-score (f) = 0.585
- ROUGE-2 (bigram overlap): Recall (r) = 0.625, Precision (p) = 0.385, F1-score (f) = 0.476
- ROUGE-L (longest common subsequence): Recall (r) = 0.750, Precision (p) = 0.480, F1-score (f) = 0.585

Fig Evaluation

Selected Testing :

Ist Phase of Testing

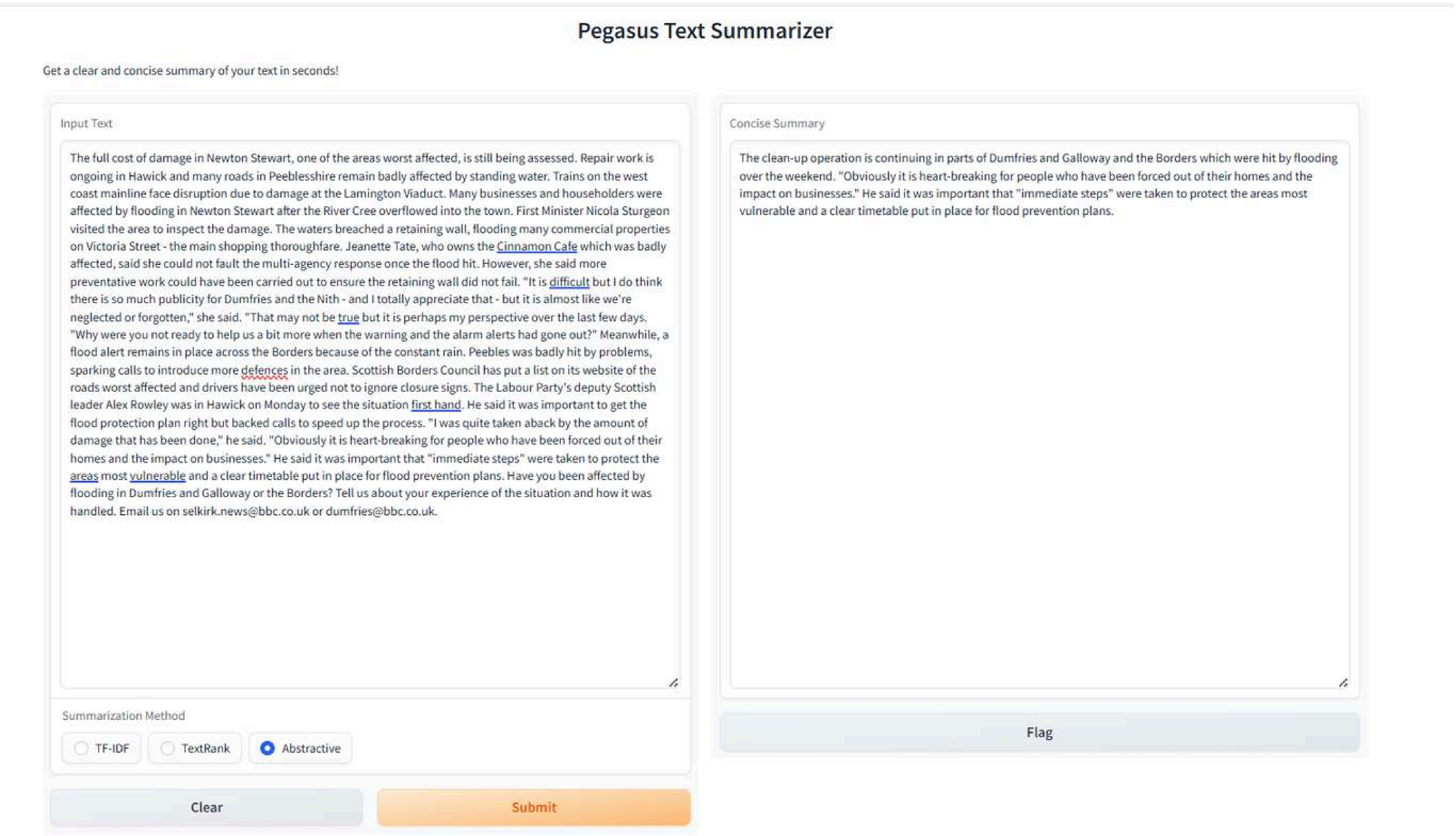


Fig. Local Testing

Selected Deployment:

gradio → Spaces hf.co/spaces → Hugging Face

Gradio lets you build web interfaces for your models with minimal code, perfect for rapid prototyping

This is where you put your AI model to work! Deploy your model on Spaces, creating a web interface for others to interact with and explore. Share your model with a simple link!

Think of it as a giant library of pre-trained AI models and datasets, all accessible for free! It's a community hub for machine learning enthusiasts.

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

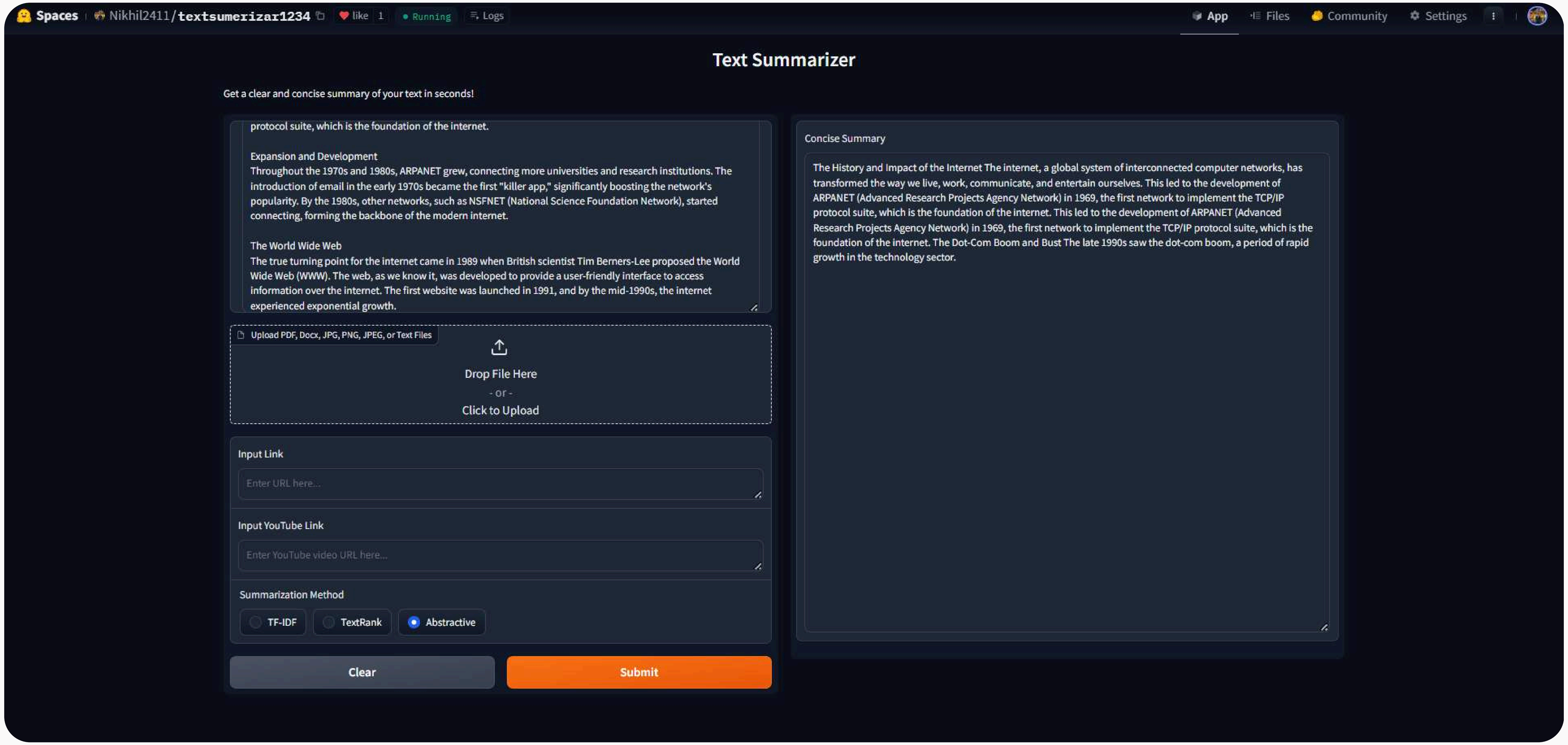


Fig. Abstractive Summarizer Output

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

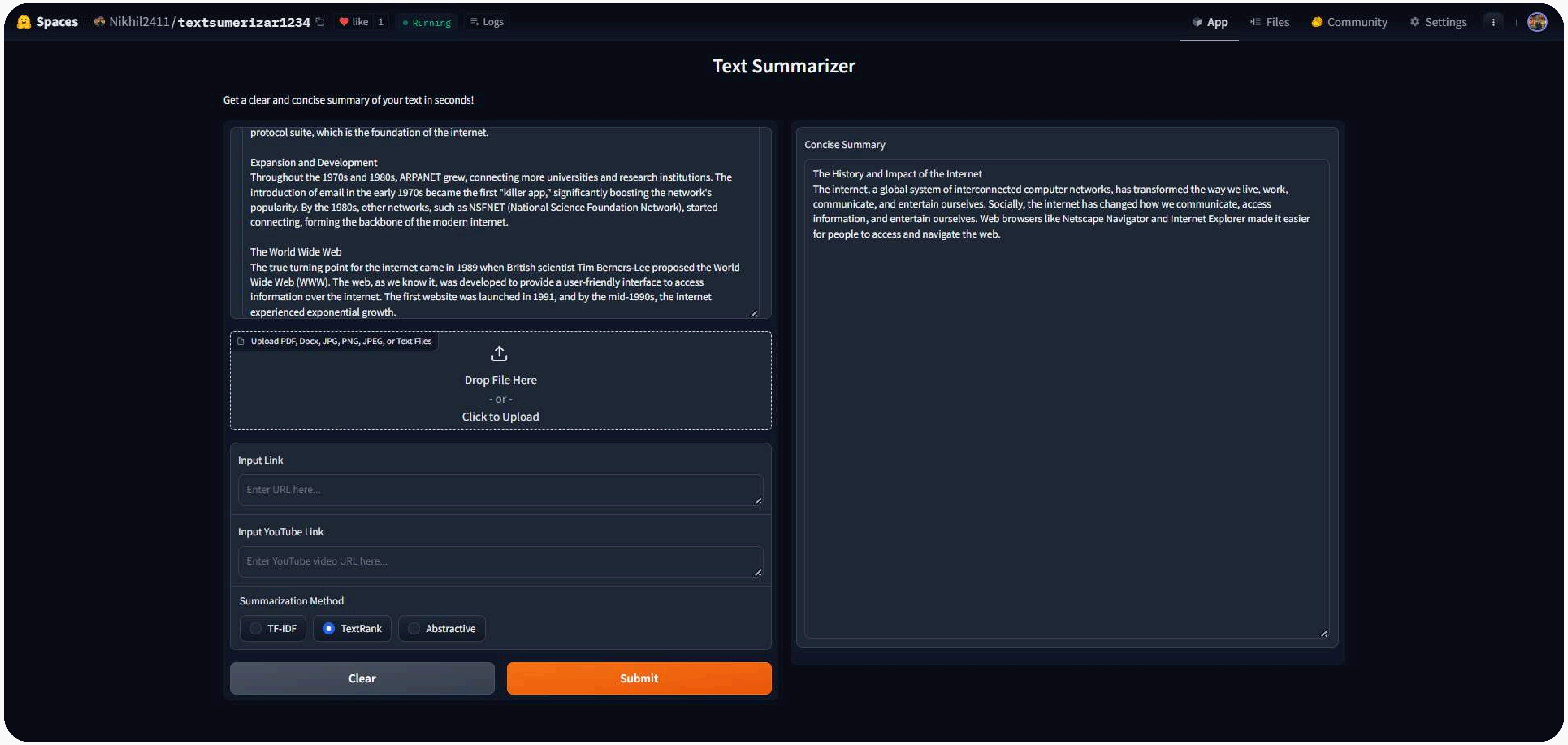


Fig. Extractive Text Rank Summarizer Output

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

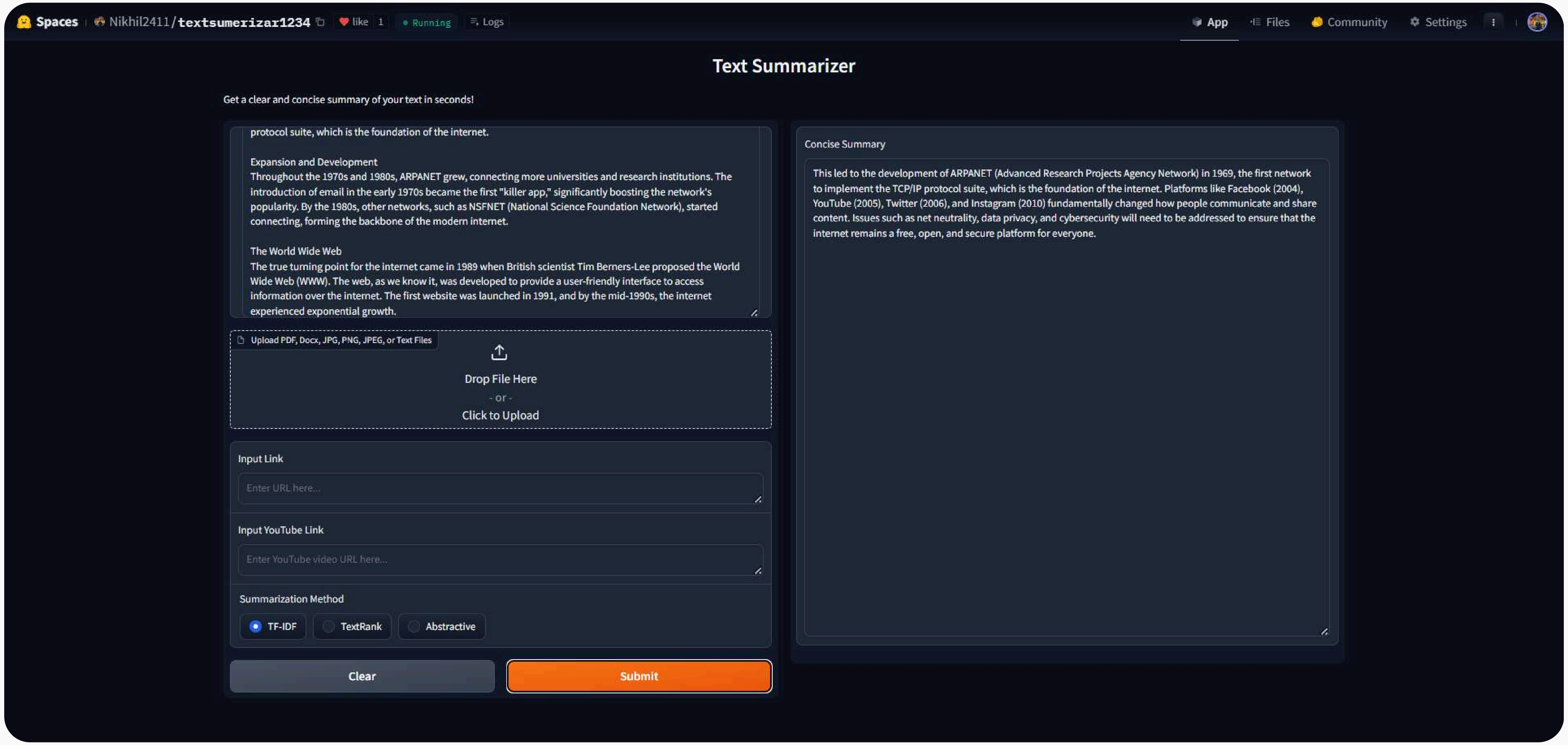


Fig. Extractive TF-IDF Summarizer Output

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

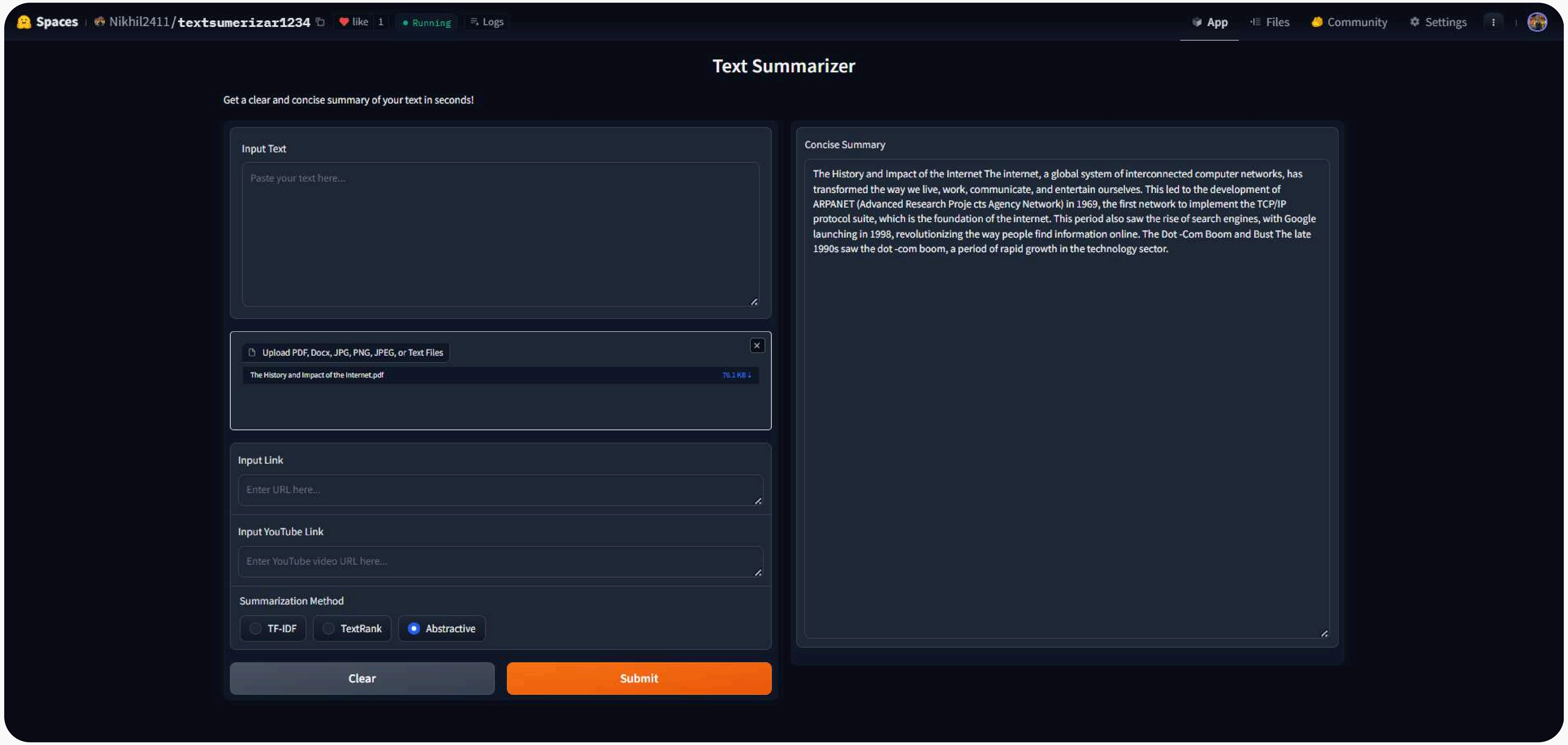


Fig. Summerization with pdf file

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

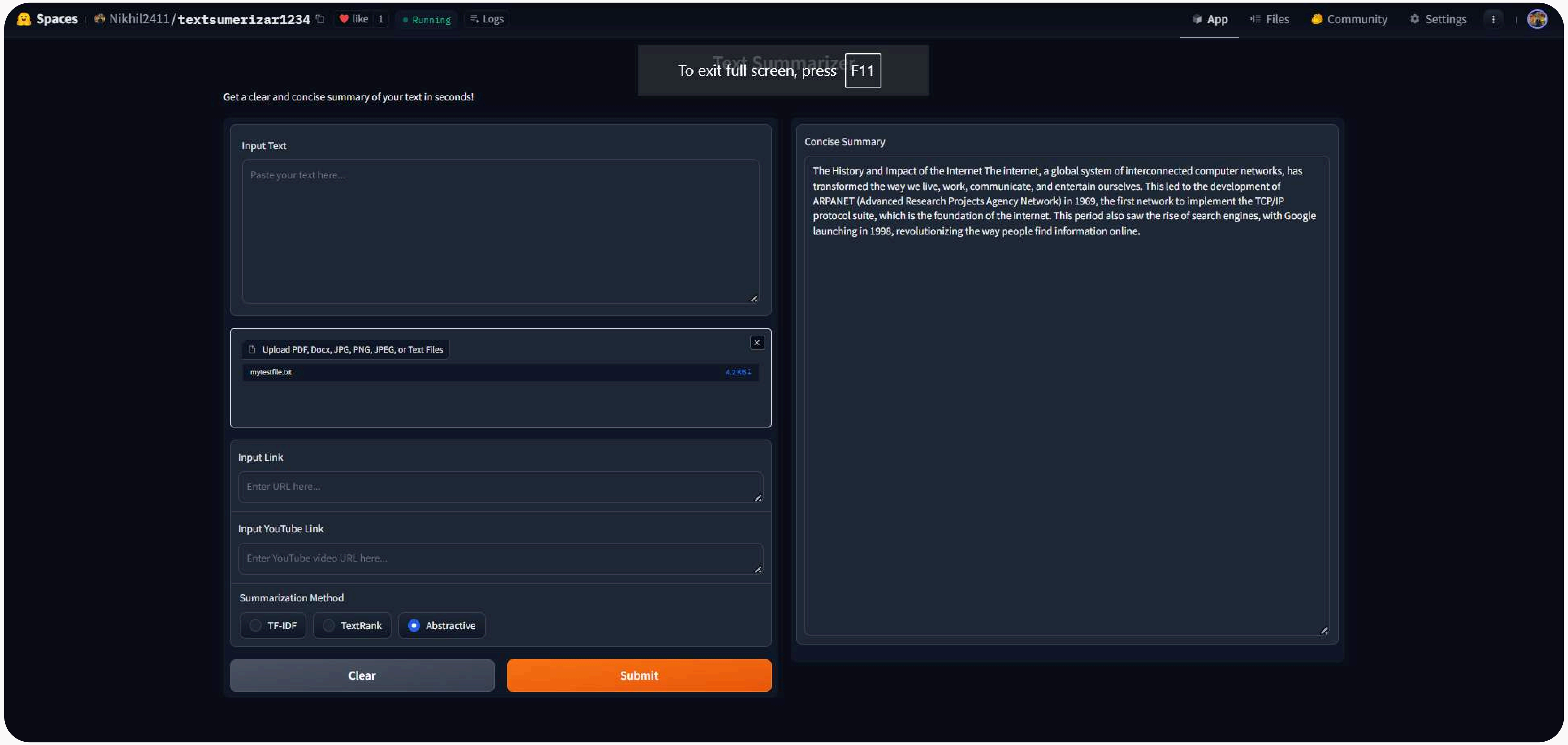


Fig. Summerization with text file

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

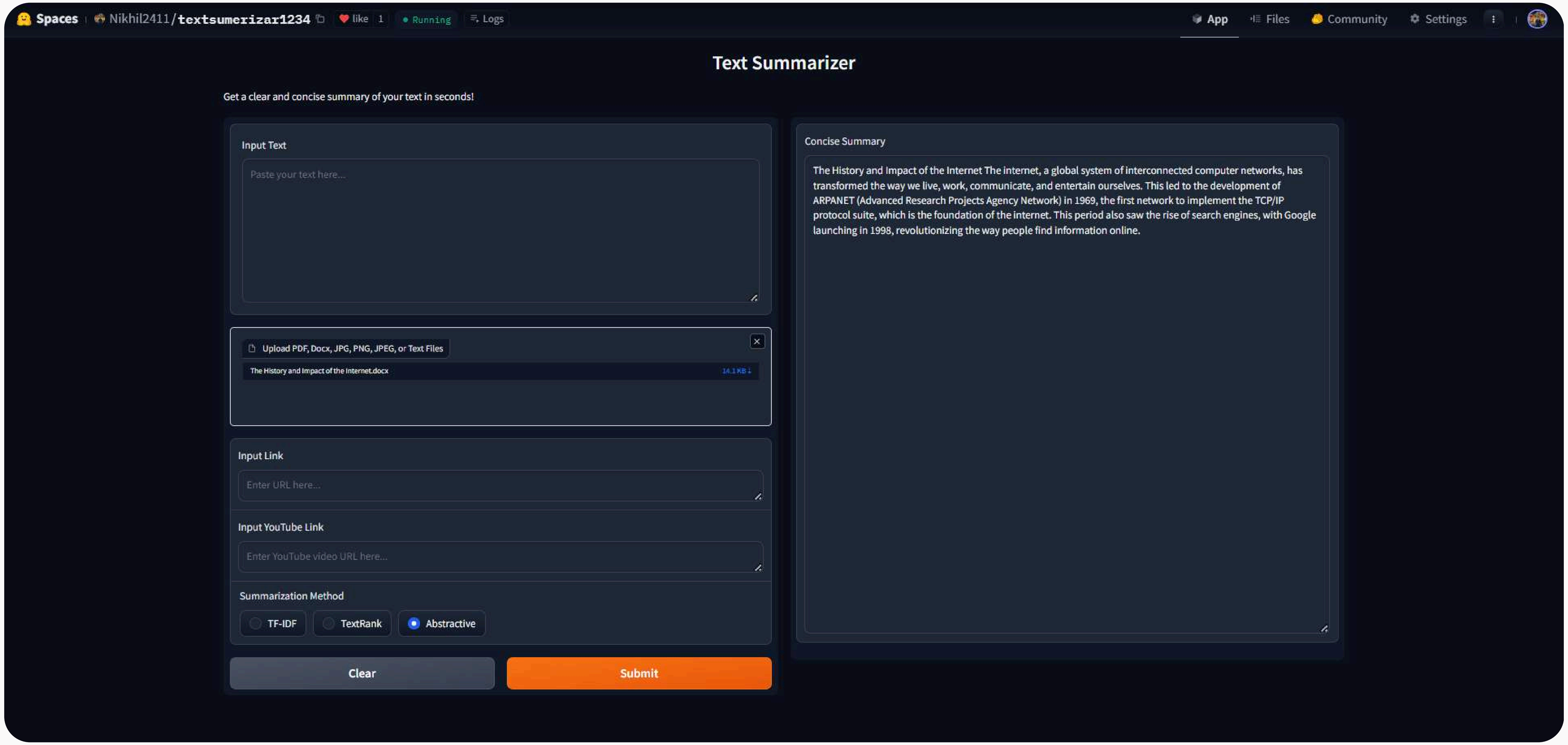


Fig. Summerization with docx file

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

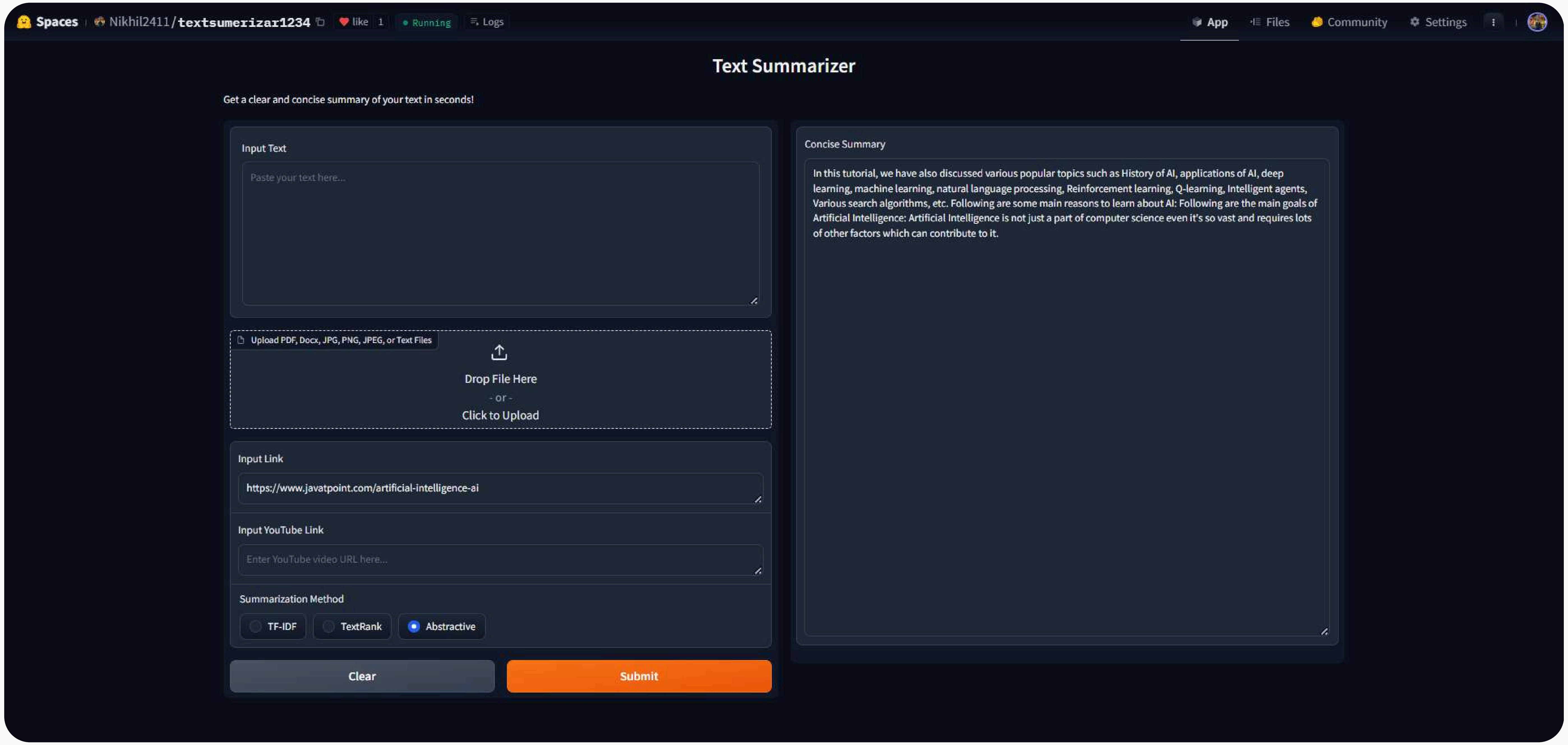


Fig. Summerization with website link

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Selected Deployment - Results :

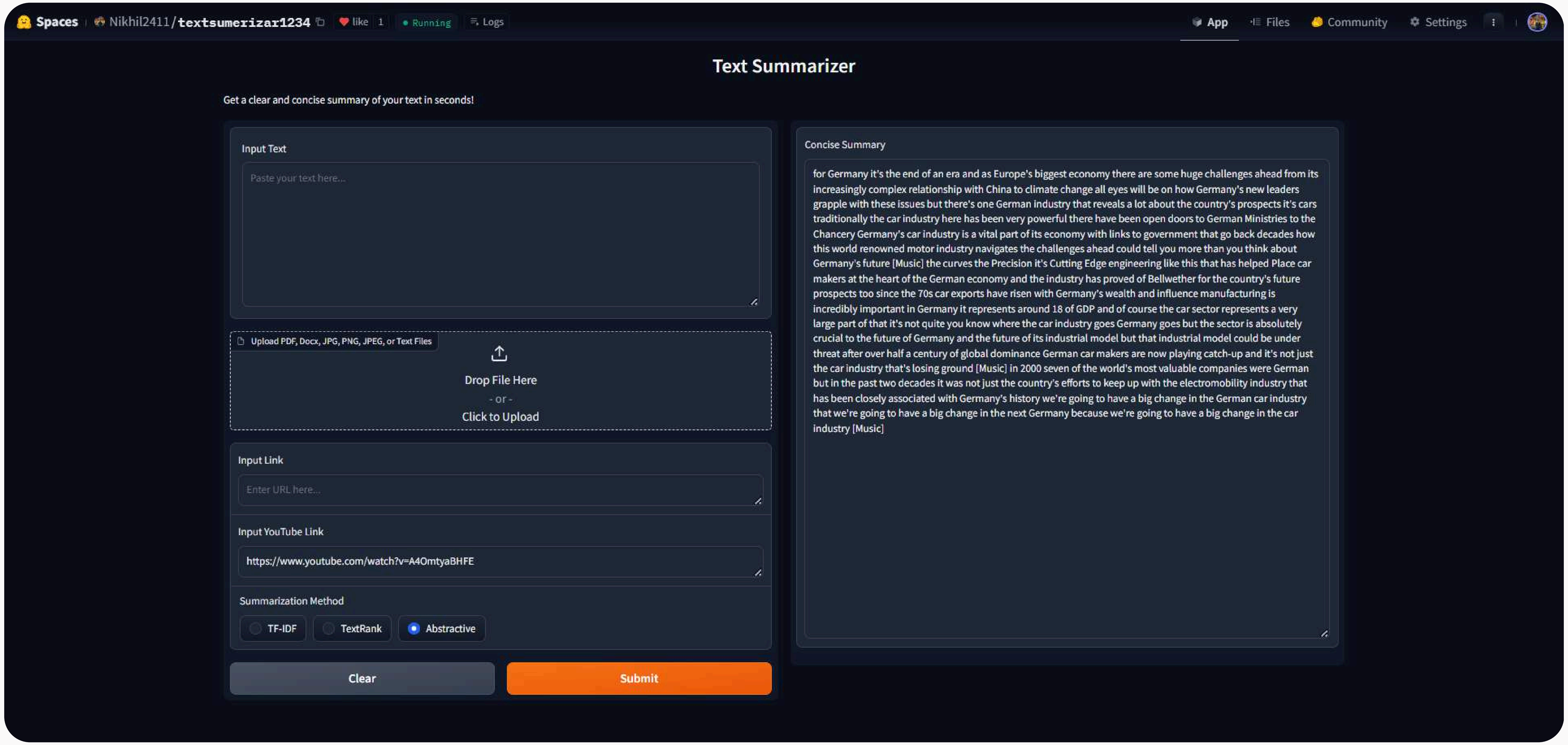


Fig. Summerization with youtube video link

Visit Deployment at : https://huggingface.co/spaces/Nikhil2411/textsumerizar1234

Real Life Applications

1. News Aggregation

Text summarization is extensively used in news aggregation platforms to deliver concise news summaries to readers. By summarizing lengthy articles, it helps users quickly grasp the main points of the news without having to read the entire content. This is particularly useful in today's fast-paced world where people want to stay informed but have limited time.

2. Legal Document Review

In the legal field, professionals often deal with extensive documentation. Text summarization can be applied to legal documents to extract key information, making it easier for lawyers and paralegals to review cases, contracts, and other legal texts. This saves significant time and effort, allowing legal professionals to focus on more critical aspects of their work.

3. Customer Support

Customer support centers can utilize text summarization to improve service efficiency. By summarizing customer inquiries, chat logs, and support tickets, support agents can quickly understand the context and provide accurate responses. Additionally, summarized insights from customer feedback can help businesses identify common issues and improve their products and services.

4. Academic Research

Researchers and students often need to go through numerous academic papers and articles. Text summarization tools can assist by generating concise summaries of research papers, helping them identify relevant studies more efficiently. This application enhances the research process by saving time and allowing researchers to focus on in-depth analysis of the most pertinent literature.

Future Scope :

Support for Extracting Text from Images:

Implement OCR (Optical Character Recognition) capabilities to extract text from images using tools like pytesseract.
This will enable the application to process images containing text, expanding its usability for documents that are scanned or photographed.

Improving Accuracy:

Enhance the text extraction and summarization algorithms to improve the accuracy and relevance of the summaries.
This can involve incorporating advanced NLP techniques and fine-tuning models with larger and more diverse datasets.

Multilingual Support:

Extend the application's capabilities to handle multiple languages, allowing users to extract and summarize text in languages other than English.
This can involve integrating language detection and translation services.

Real-time Summarization:

Develop real-time summarization features for streaming text data, such as live news feeds or social media updates.
This will provide users with up-to-the-minute summaries of ongoing events.

User Interface Enhancements:

Improve the user interface to offer a more intuitive and seamless experience.
Incorporate features like drag-and-drop for file uploads and visual indicators of processing status future scope

Our Team :

Nikhilkumar jain

Team Lead

Nitesh Sachan

Project Member

Nebu C. Thomas

Project Member

Conclusion:

Thank You

In conclusion, our internship journey at Infosys Springboard has been enriching and fulfilling. We are proud to present a robust text summarization system that effectively meets and exceeds the demands of modern information processing. Our system showcases the potential of NLP techniques in transforming the way businesses handle and process large volumes of text, enhancing efficiency and decision-making. We look forward to further refining and expanding our system's capabilities, making it even more versatile and user-friendly.

2024

If any queries please ask!!