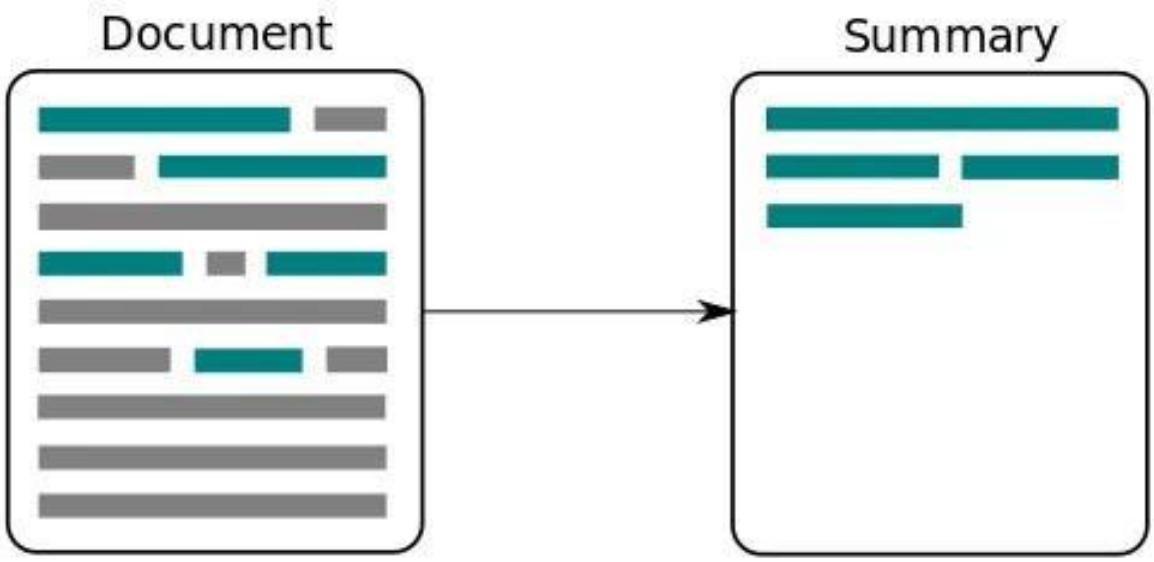
“Streamlining information management: a cutting-edge text summarization product for enhanced content curation and efficiency.”

AI Enhanced Text Summarization

Rohith R



Abstract

In an era defined by information overload, efficient content curation and information extraction have become paramount. This report introduces an innovative text summarization product designed to address this challenge head-on, outlines the market demand for such a product, highlighting the increasing need for streamlined information management across various industries. The report presents a description of the text summarization product's functionalities and user experience. Additionally, we discuss the development process behind the product, shedding light on the algorithms and technologies utilized. This report serves as a comprehensive introduction to the text summarization product, demonstrating its significance in today's information-driven landscape and its potential to revolutionize content curation and information extraction processes.

Problem Statement

In an age characterized by an abundance of information, the ability to efficiently classify vast textual content into concise, meaningful summaries has emerged as a critical need across various industries and applications.

Text summarization, the process of condensing lengthy textual information into shorter, coherent representations while preserving the essence of the content, has become an important solution to this significant problem. This paper introduces a pioneering text summarization product meticulously engineered to meet the demand for streamlined information management.

The relentless pace of information creation in today's digital landscape necessitates innovative solutions that empower individuals and companies to extract essential insights rapidly and accurately from textual data. The paper offers insights into its impact, both quantitative and qualitative, showcasing its potential to enhance productivity, streamline information workflows, and inform critical decision-making processes.

By the paper's conclusion, readers will gain a comprehensive understanding of the transformative potential of the text summarization product in an era where information is not merely abundant but a critical driver of progress and innovation.

Market Analysis

The landscape of information management and content curation has undergone drastic changes in recent years. With the expansion of digital content across various platforms, industries, and domains, the ability to efficiently extract essential insights from an ever-expanding sea of textual data has become a strategic importance.

Text summarization, as a solution to this challenge, has grabbed a significant attention and is expected to replace traditional information processing methods. This section provides a comprehensive market analysis, elucidating the growing demand for text summarization products across various sectors.

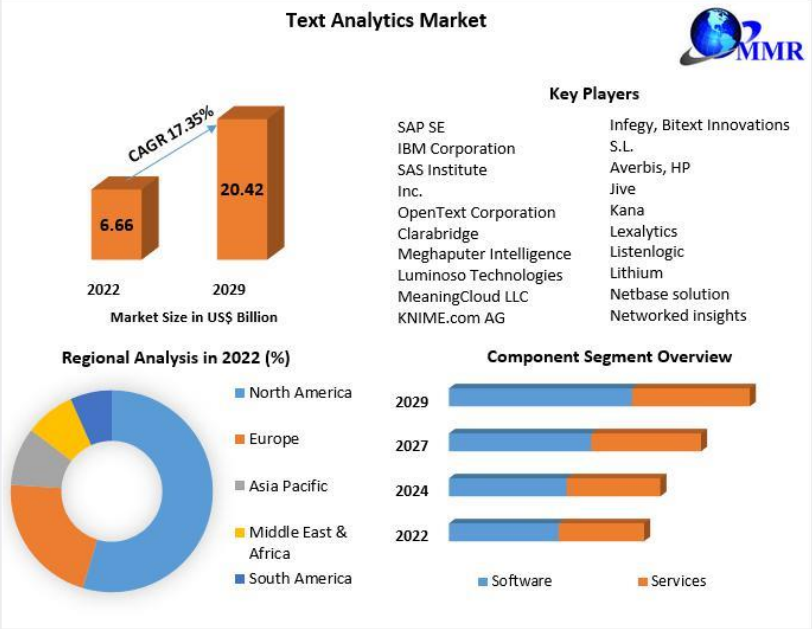
The text analytics market can be differentiated by deployment, application, geography, and other factors. The text analytics market reached over USD 6.4 billion in 2022 and is projected to grow at a compound annual growth rate (CAGR) of 39.90% to reach USD 40 billion in the next five years. Organizations are increasingly turning to text analytics tools to enhance their decision-making processes. These tools extract valuable insights from diverse text sources like customer interactions, emails, blogs, product reviews, tweets, and call center logs. Text analytics primarily serves to gather data from a wide range of sources, both unstructured and structured. This data is then analyzed to inform organizational decisions and stimulate market growth.

The global market is set to surge due to the increasing need for machine learning and big data analytics. The COVID-19 pandemic has also played a positive role in driving market growth. This is evident as the demand for text analytics has risen, fueled by the continued trend of remote work and the extensive use of social media platforms and email for brand interactions.

The highest shareholders in the text analytics are Retail and E-Commerce. Retailers are increasingly relying on social media platforms as a primary data source. This enables them to predict demand, manage supply, and elevate the overall customer experience.

Enhancing the shopping experience holds significant importance for e-commerce businesses, as it directly impacts their sales performance. To stay competitive with leading online retailers, e-commerce companies are giving importance to improving the shopping experience through text analytics applications. This approach offers a rapid and efficient means of grasping user sentiments and purchase considerations in the realm of online shopping.

The text summarization software empowers retailers to understand the evolving trends in customer behavior, their preferences, and the recurring patterns in their needs and emotions. Furthermore, it facilitates data-driven decision-making aimed at improving customer services.



The text analytics market is mainly dominated by IBM Corporation and SAP SE. In September 2022, a strategic technology collaboration was initiated between AppTek and expert.ai. This collaboration aims to harness AI-driven text analytics for dynamic audio content across multiple languages. The partnership leverages expert.ai's expertise in natural language understanding (NLU) along with AppTek's leadership in automatic speech recognition (ASR) and neural machine translation (NMT) technologies. Together, they empower organizations to effectively utilize audio content within their unstructured datasets. This, in turn, enhances intelligent automation and improves decision-making processes.

In May 2023, Google realesed series of AI-driven enhancements for its Workspace suite, designed to facilitate collaboration among business teams working on projects. Additionally, Google introduced Duet AI, a feature that provides support for composing text within Gmail and Google Docs, as well as generating images for Slides, a presentation software similar to PowerPoint. Furthermore, Duet AI can summarize discussions conducted through Google Meet, the company's video conferencing service.

Top Competitors

1. Jasper AI:
Strengths: Exceptional accuracy in extractive summarization.
Robust support for multiple languages.
User-friendly interface for quick adoption.
Weaknesses: Limited capabilities in abstractive summarization.
Requires substantial computing resources for large-scale tasks.
May struggle with highly technical or domain-specific content.
2. Quill Bot:
Strengths: Advanced abstractive summarization capabilities.
Excellent at handling domain-specific jargon and terminology.
Integrates seamlessly with existing content management systems.
Weaknesses: Limited language support.
Relatively higher learning curve for users.
May struggle with very short or fragmented texts.
3. WPS AI:
Strengths: Scalable for processing massive datasets.
Strong performance in real-time summarization.
Effective handling of multimedia content.
Weaknesses: Lacks advanced linguistic analysis for nuanced content.
May require customization for specific industry applications.
Limited support for summarizing non-standard data formats.

Target Customers

Text summarization tools can benefit a wide range of customers across various industries and domains. Some of them include:

Businesses and Corporations: Businesses use text summarization to quickly extract key insights from market research reports, customer feedback, financial documents, and competitor analyses. Corporate communication teams use text summarization to create concise summaries for internal reports and external communications.
Content Creators and Marketers: Content creators and marketers use text summarization to generate concise summaries for articles, blog posts, and social media content. It is also used by email marketers to generate eye-catching email subject lines.
Researchers and Academics: Researchers use text summarization to efficiently review and summarize extensive literature for their studies and research papers. Academic institutions also use this tool to help students quickly understand complex academic texts.
Legal Professionals: Lawyers and legal professionals use it to extract key information from legal documents, contracts, and court cases. Legal research platforms also use it to provide concise case summaries.
Healthcare and Medical Professionals: Healthcare professionals use this product to summarize patient records, medical research papers, and clinical notes. Medical researchers use text summarization to review and compare studies efficiently.
Media and News Organizations: News organizations integrate this product to automatically generate news summaries. Text summarization can also be used to create concise video and audio transcripts for captions and subtitles.
Customer Support and Service Centers: Customer support teams also integrates it to quickly understand and respond to customer inquiries and complaints.
Government and Public Sector: Government agencies deploys this product for analysing public feedback, policy documents, and legislative texts.
Financial Institutions: Financial analysts use text summarization to extract critical information from financial reports, market news, and economic analyses. Banks and investment firms also use it for portfolio management and risk assessment.
E-commerce and Retail: E-commerce companies uses it to create concise product descriptions and reviews. Retailers use it to analyse customer feedback and reviews for product improvements.

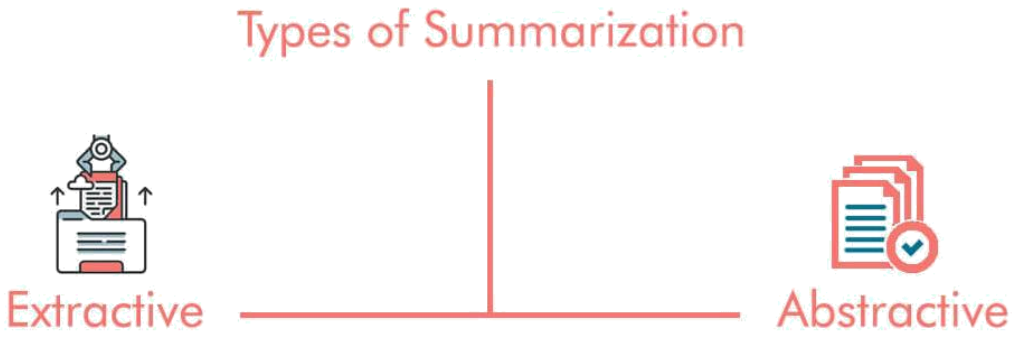
Product Description/Development

The intuition behind creating an end-to-end text summarization software product, or any other product, is the need to navigate through all stages of development without errors. The concept behind text summarization is to extract the crucial features and key points from a relatively lengthy document, saving others the time it would take to read it in its entirety.

Building A Text Summarization Tool Using NLP/Concept Generation

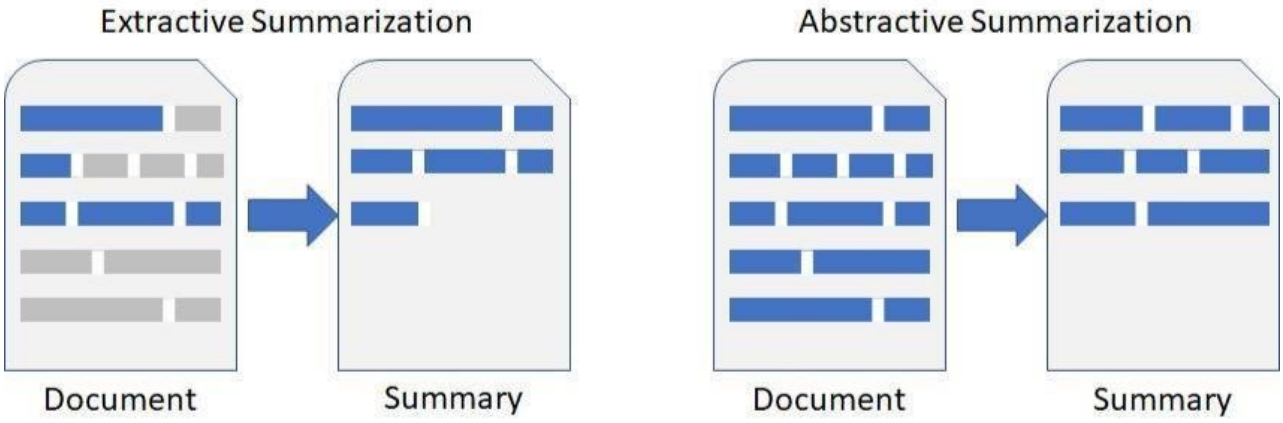
Different approaches for text summarization:

1. Abstractive summarization
2. Extractive summarization



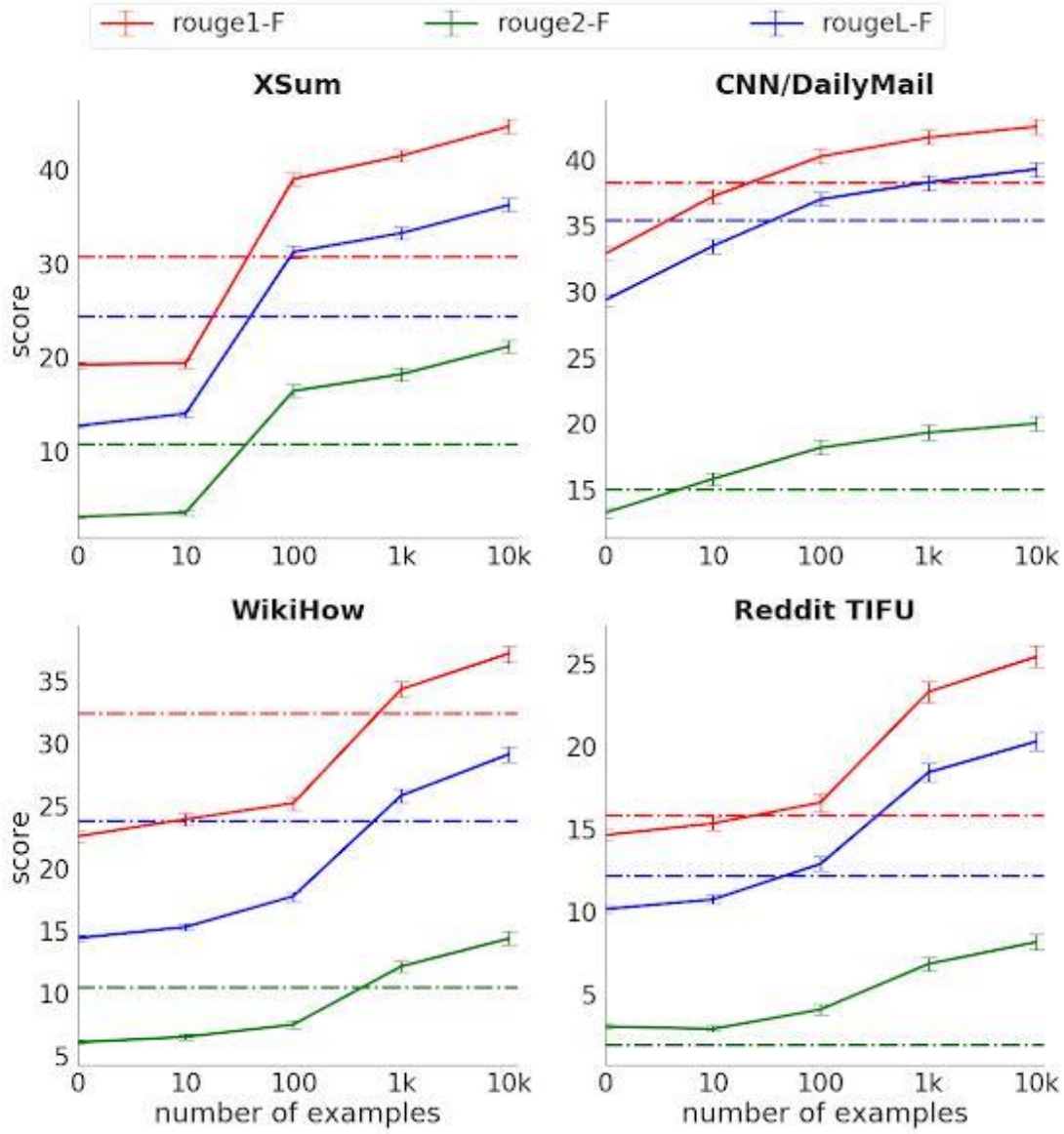
Abstractive summarization relies on the ability to rephrase and condense segments of a document using advanced natural language methods. Abstractive summarization methods aim to create a summary by interpreting the text through sophisticated natural language techniques, resulting in a new, shorter text. This summary may include information not explicitly present in the original document but effectively conveys the most crucial details. Achieving this involves rephrasing sentences and integrating information from the full text, much like how a human-written abstract typically functions. In essence, an acceptable abstractive summary contains the essential information from the input while maintaining fluency.

Extractive summarization selects sentences directly from the document using a scoring mechanism to create a summary. This approach operates by pinpointing significant segments within the text, cropping them out, and then piecing together these extracted portions to generate a concise version. Hence, they rely solely on extracting sentences from the original text. Most of the current summarization research has emphasized extractive summarization, primarily because it is simpler and produces naturally grammatical summaries that demand relatively minimal analysis.



Abstractive Text Summarization Using Google Pegasus And Hugging Face Transformers/Concept Development

PEGASUS, an advanced abstractive summarization model, was made publicly available by Google in June 2020, representing the latest state-of-the-art technology in this field. The name PEGASUS itself is an acronym that stands for "Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence models."



Bench Marking

To evaluate the quality of the summaries generated by each system, we have opted for ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation). ROUGE-N is a statistical metric based on N-gram recall and has demonstrated a strong correlation with human assessments in summarization evaluation tasks.
In simpler terms, ROUGE-N assesses the degree of overlap between specific sub-phrases or sub-sequences within the summaries generated by the system and those in the standard reference summaries, which are written by humans. The length of these sub-phrases or sub-sequences is determined by the value of N. A higher degree of overlap between the two types of summaries indicates better performance, with a maximum ROUGE value of 1 signifying the ideal outcome.

The primary notion for selecting an objective and computationally efficient metric like ROUGE-N as our evaluation measure lies in its practicality. Manually evaluating numerous summaries by human assessors can be time-consuming, costly, and unfeasible at scale. Additionally, ROUGE-N holds the status of being the standard measure for evaluating summarization models, consistently featured in scientific publications focusing on summarization research.

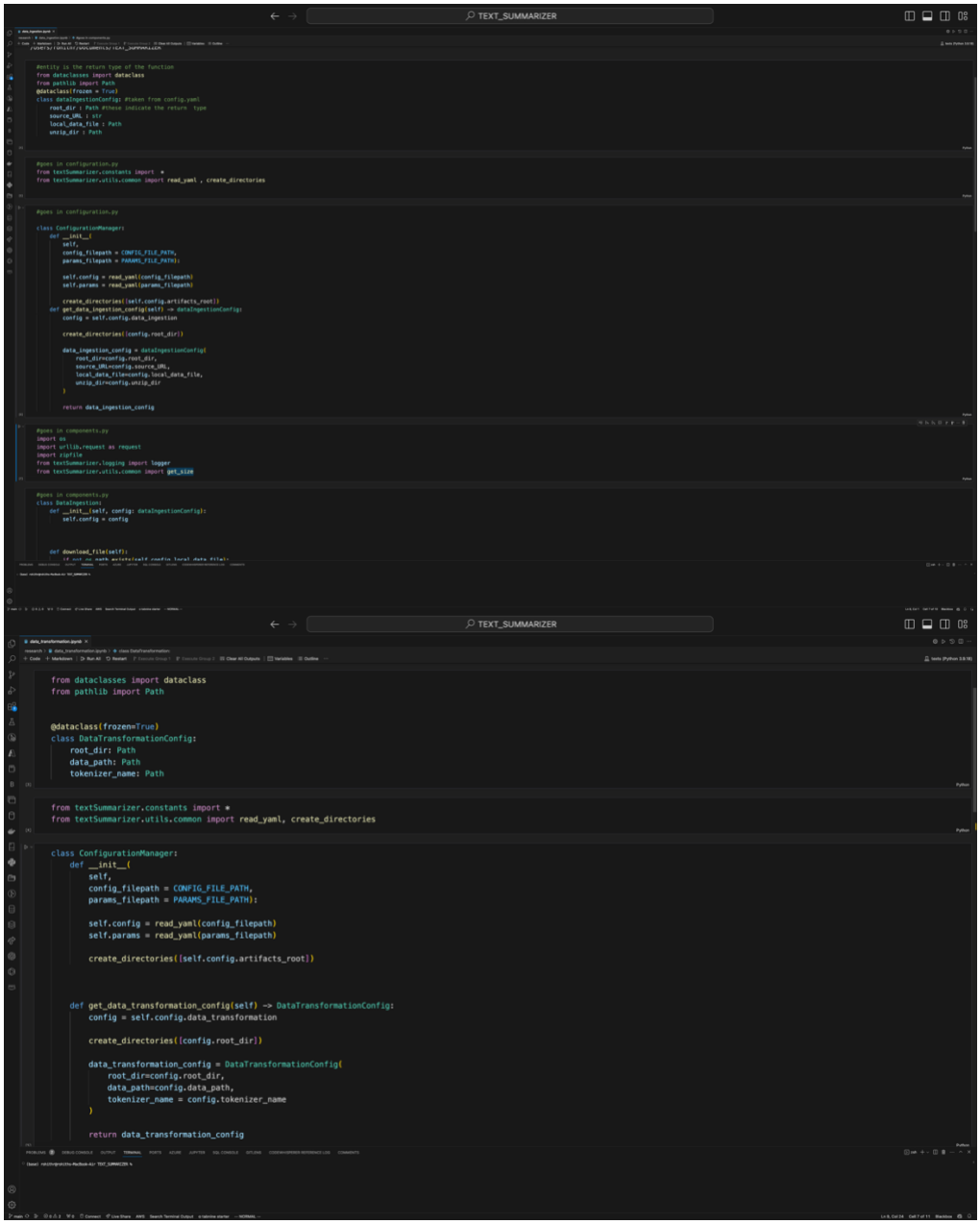
Prototype Build

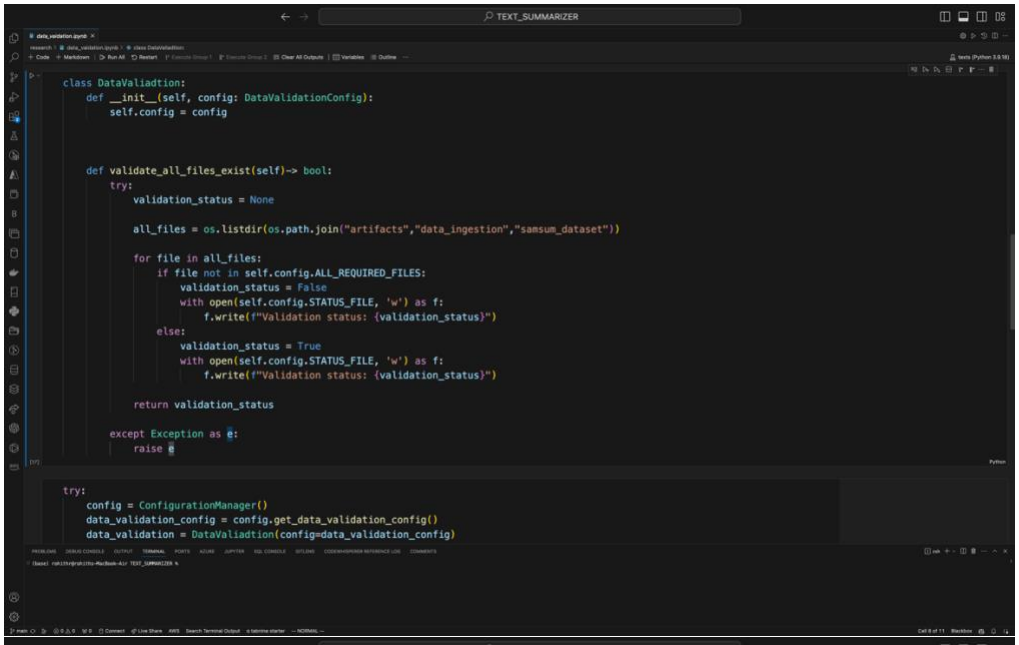
To establish an end-to-end text summarization product, the initial step involves acquiring all the essential files required to execute the code in a modul1ar and organized manner. It's important to note that the implementation discussed in this paper is limited to a basic scope.
The development of the entire product will be conducted using Python as the primary programming language, with Visual Studio Code (VSCode) serving as the preferred code editor due to its integration capabilities. The product development process can be segmented into the following distinct stages:

1. **Creation of Essential Files and Folders:** Begin by creating the necessary files, and folders, and setting up the product template.
2. **Structured File Population:** Populate the created files in a structured manner, which includes tasks such as installing dependencies specified in the requirements.txt file and configuring the project setup.
3. **Model Ingestion Phases:** Model ingestion comprises four essential stages: Data Ingestion Stage Data Validation Stage Data Transformation Stage Model Training Stage Model Evaluation Stage
4. **Prediction Pipeline and User Interface:** Develop a prediction pipeline and a user-friendly interface using Fast API to facilitate user interaction with the summarization model.
5. **CI/CD Deployment on Cloud Platform:** Implement a Continuous Integration/Continuous Deployment (CI/CD) pipeline, deploying the product on a cloud platform, such as AWS, or another suitable choice.

This structured approach ensures a systematic development process for the end-to-end text summarization product.

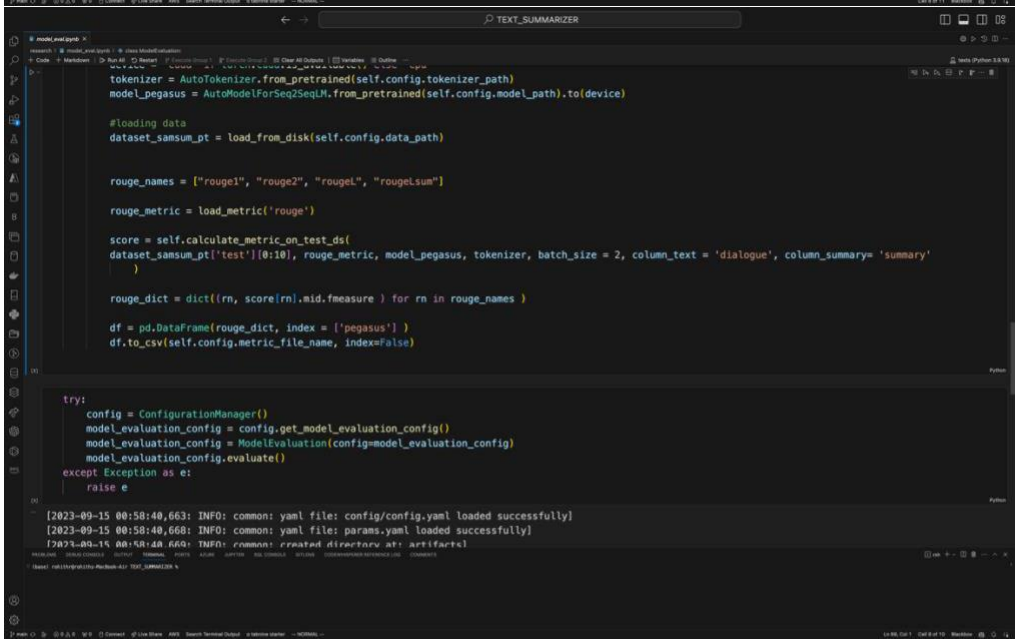
A glimpse of the major phases in the product implementation:


TEXT_SUMMARIZER
#entity is the return type of the function
from dataclasses import dataclass
from pathlib import Path
@dataclass(frozen = True)
class DataIngestionConfig: #taken from config.yaml
root_dir : Path #these indicate the return type
source_URL : str
local_data_file : Path
unzip_dir : Path
#goes in configuration.py
from textSummarizer.constants import *
from textSummarizer.utils.common import read_yaml , create_directories
#goes in configuration.py
class ConfigurationManager:
def __init__(
self,
config_filepath = CONFIG_FILE_PATH,
params_filepath = PARAMS_FILE_PATH):
self.config = read_yaml(config_filepath)
self.params = read_yaml(params_filepath)
create_directories([self.config.artifacts_root])
def get_data_ingestion_config(self) -> DataIngestionConfig:
config = self.config.data_ingestion
create_directories([config.root_dir])
data_ingestion_config = DataIngestionConfig(
root_dir=config.root_dir,
source_URL=config.source_URL,
local_data_file=config.local_data_file,
unzip_dir=config.unzip_dir
)
return data_ingestion_config
#goes in components.py
import os
import urllib.request as request
import zipfile
from textSummarizer.logging import logger
from textSummarizer.utils.common import get_size
#goes in components.py
class DataIngestion:
def __init__(self, config: DataIngestionConfig):
self.config = config
def download_file(self):
TEXT_SUMMARIZER
from dataclasses import dataclass
from pathlib import Path
@dataclass(frozen=True)
class DataTransformationConfig:
root_dir: Path
data_path: Path
tokenizer_name: Path
from textSummarizer.constants import *
from textSummarizer.utils.common import read_yaml, create_directories
class ConfigurationManager:
def __init__(
self,
config_filepath = CONFIG_FILE_PATH,
params_filepath = PARAMS_FILE_PATH):
self.config = read_yaml(config_filepath)
self.params = read_yaml(params_filepath)
create_directories([self.config.artifacts_root])
def get_data_transformation_config(self) -> DataTransformationConfig:
config = self.config.data_transformation
create_directories([config.root_dir])
data_transformation_config = DataTransformationConfig(
root_dir=config.root_dir,
data_path=config.data_path,
tokenizer_name = config.tokenizer_name
)
return data_transformation_config



```python
class DataValiadtion:
    def __init__(self, config: DataValidationConfig):
        self.config = config



    def validate_all_files_exist(self)-> bool:
        try:
            validation_status = None

            all_files = os.listdir(os.path.join("artifacts","data_ingestion","samsum_dataset"))

            for file in all_files:
                if file not in self.config.ALL_REQUIRED_FILES:
                    validation_status = False
                    with open(self.config.STATUS_FILE, 'w') as f:
                        f.write(f"Validation status: {validation_status}")
                else:
                    validation_status = True
                    with open(self.config.STATUS_FILE, 'w') as f:
                        f.write(f"Validation status: {validation_status}")

            return validation_status

        except Exception as e:
            raise e
```

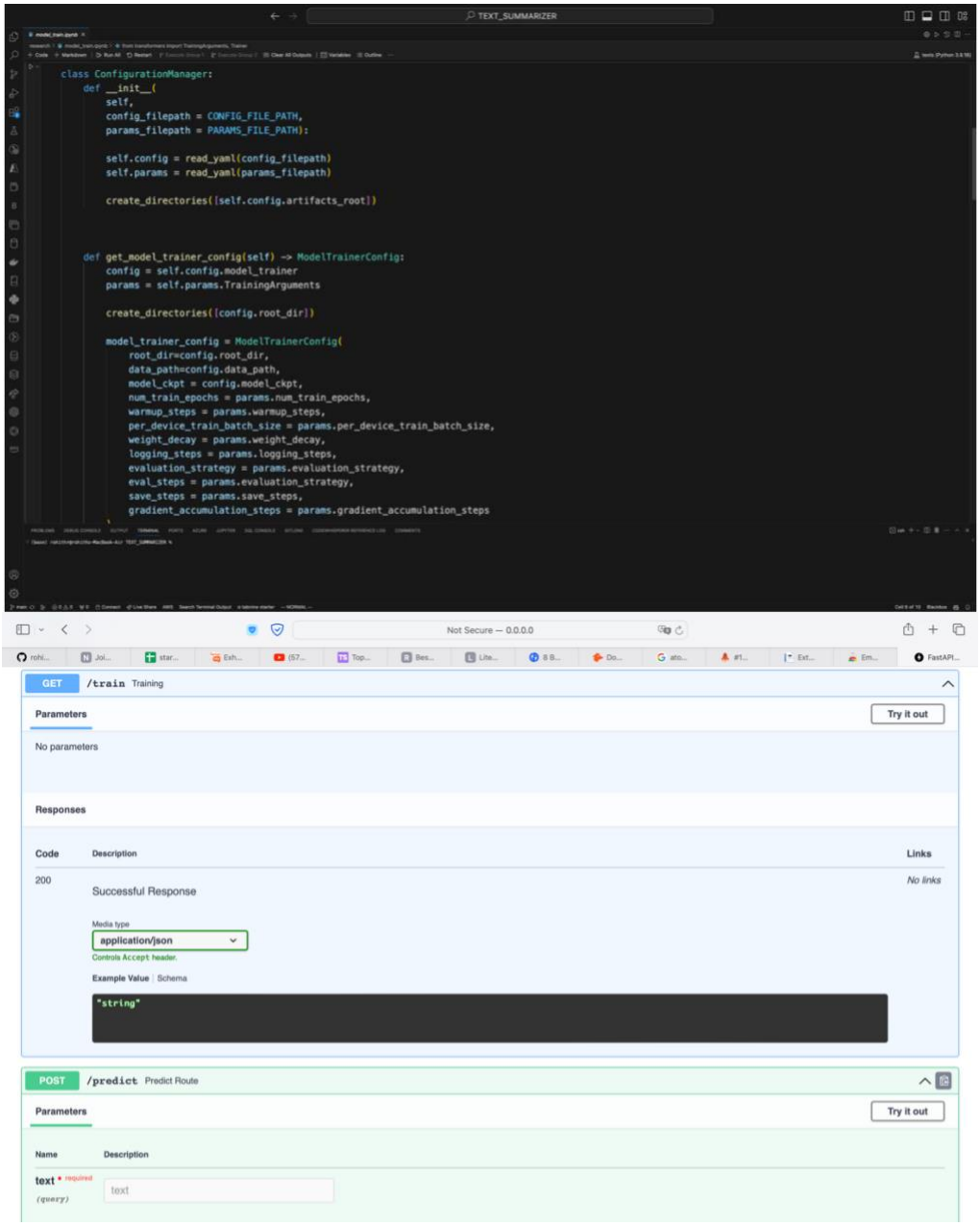
```python
try:
    config = ConfigurationManager()
    data_validation_config = config.get_data_validation_config()
    data_validation = DataValiadtion(config=data_validation_config)
```



```python
            tokenizer = AutoTokenizer.from_pretrained(self.config.tokenizer_path)
            model_pegasus = AutoModelForSeq2SeqLM.from_pretrained(self.config.model_path).to(device)

            #loading data
            dataset_samsum_pt = load_from_disk(self.config.data_path)


            rouge_names = ["rouge1", "rouge2", "rougeL", "rougeLsum"]

            rouge_metric = load_metric('rouge')

            score = self.calculate_metric_on_test_ds(
            dataset_samsum_pt['test'][0:10], rouge_metric, model_pegasus, tokenizer, batch_size = 2, column_text = 'dialogue', column_summary= 'summary'
                )

            rouge_dict = dict((rn, score[rn].mid.fmeasure ) for rn in rouge_names )

            df = pd.DataFrame(rouge_dict, index = ['pegasus'] )
            df.to_csv(self.config.metric_file_name, index=False)
```

```python
try:
    config = ConfigurationManager()
    model_evaluation_config = config.get_model_evaluation_config()
    model_evaluation_config = ModelEvaluation(config=model_evaluation_config)
    model_evaluation_config.evaluate()
except Exception as e:
    raise e
```

```
[2023-09-15 00:58:40,663: INFO: common: yaml file: config/config.yaml loaded successfully]
[2023-09-15 00:58:40,668: INFO: common: yaml file: params.yaml loaded successfully]
```



To get a better idea how the product is developed, check this out:
https://github.com/rohith4088/TEXT_SUMMARIZER

Technology Stack

1. **Python**: Python is a versatile and high-level programming language known for its simplicity and readability. It is widely used in various fields, including data analysis, web development, and machine learning.
2. **Transformers**: This is an open-source library developed by Hugging Face that provides pre-trained models for natural language understanding and generation tasks, including state-of-the-art models for tasks like text summarization and machine translation.

3. **Tensorflow**: TensorFlow is an open-source machine learning framework developed by Google. It is widely used for building and training machine learning models, particularly neural networks.
4. **Keras**: is an open-source deep learning framework that serves as an interface to various backends, including TensorFlow. It simplifies the process of building and training neural networks.
5. **Seaborn**: It is a Python data visualization library based on Matplotlib and provides a high-level interface for creating informative and attractive statistical graphics.
6. **Matplotlib**: This is a popular Python plotting library used for creating static, animated, and interactive visualizations, also It is highly customizable and widely used in data analysis and scientific research.
7. **Pandas**: Pandas is a data manipulation and analysis library for Python. It provides data structures like DataFrames and Series, making it easier to work with structured data.
8. **Rouge**: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a metric used to evaluate the quality of machine-generated text, particularly in tasks like text summarization. It measures the similarity between generated text and reference (human-written) text.
9. **NLTK** (Natural Language Toolkit):This is a very famous Python library for working with human language data and it also provides tools and resources for tasks such as tokenization, stemming, part-of-speech tagging, and more in natural language processing.
10. **SacreBLEU**: SacreBLEU is a tool for evaluating the quality of machine translation outputs. It uses the BLEU (Bilingual Evaluation Understudy) metric, which assesses the similarity between machine-translated text and human reference translations.

Business Model

Creating a business model for a text summarization tool involves outlining the key components that help generate revenue, provide value to customers, and ensure the sustainability and growth of the business. Here's a business model for a text summarization tool:

1. Value Proposition: High-Quality Summaries: Offer an AI-powered text summarization tool that generates accurate and coherent summaries from lengthy documents, articles, and texts.

Time and Resource Savings: Enable users to save time and effort in reading and summarizing large volumes of text, making information consumption more efficient.

Customization: Allow users to tailor the summaries to their specific needs, including adjusting length and style.

2. Customer Segments: Content Creators: Authors, journalists, and bloggers who want to generate summaries of their own content or competitors' content for research.

Research Professionals: Academics, students, and researchers seeking concise summaries of academic papers and research materials.

Businesses: Companies interested in extracting key insights and trends from industry reports, news articles, and customer feedback.

Individuals: General readers who wish to quickly grasp the main ideas of a text without reading the entire document.

3. Revenue Streams: Subscription Model: Offer tiered subscription plans with varying levels of access, customization, and summary length.

Pay-Per-Use: Charge users based on the number of texts or words they summarize.

Enterprise Solutions: Create customized solutions for businesses with higher volume needs and additional features.

4. Channels: Online Platform: Provide a web-based platform where users can access and use the text summarization tool.

API Integration: Offer APIs that businesses and platforms can integrate into their own applications or workflows.

App Stores: Launch mobile applications for Android and iOS devices, making the tool accessible to a broader audience.

5.Key Resources:

Technology: Develop and maintain the AI algorithms and infrastructure for text summarization.

Content Sources: Access to a wide range of content and data for summarization.

Talent: Employ data scientists, engineers, and AI experts to enhance and optimize the tool.

6. Key Activities: AI Development: Continuously improve the text summarization algorithms to enhance accuracy and efficiency.

User Support: Provide customer support to assist users with technical issues and inquiries.

Content Licensing: Secure partnerships with publishers or content providers to access copyrighted material.

7. Key Partnerships: Content Providers: Partner with publishers, websites, and academic databases to access texts for summarization.

API Partners: Collaborate with platforms and businesses that integrate the summarization tool into their applications or services.

8. Cost Structure: Development Costs: Investment in AI research and development.

Operational Costs: Hosting, server maintenance, and infrastructure.

Staffing: Salaries for data scientists, engineers, and support staff.

Licensing Fees: Costs associated with accessing copyrighted content.

9. Customer Relationships: Self-Service: Provide easy access and self-service options for users to generate summaries.

Customer Support: Offer responsive customer support for technical assistance and inquiries.

Feedback Mechanism: Encourage user feedback for continuous improvement.

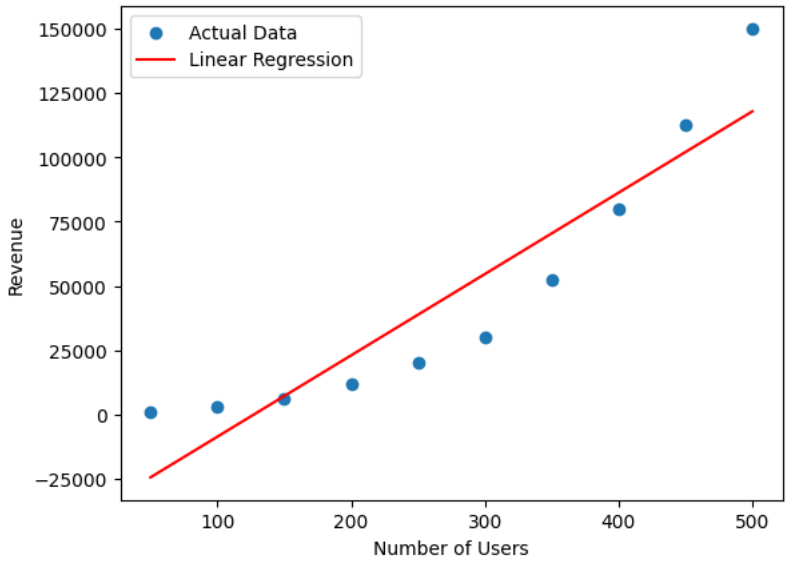
10. User Retention and Growth: Regular Updates: Keep the tool up-to-date with the latest AI advancements.

Marketing and Outreach: Conduct marketing campaigns to reach new user segments and expand the user base.

Referral Programs: Implement referral programs to incentivize users to invite others to use the tool.

By developing a comprehensive business model, you can create a clear path to offer a valuable text summarization tool while ensuring the sustainability and profitability of your business.

Financial Modelling



Revenue (R) = Price per User (P) x Number of Subscribers (N)

In this equation:

- **Revenue (R)** represents the total income generated by your business.
- **Price per User (P)** is the amount you charge per user, whether it's a subscription fee, a pay-per-use rate, or any other pricing model.
- **Number of Subscribers (N)** is the total number of users or subscribers who are paying for your text summarization tool.

This equation helps you calculate the total revenue you can generate based on your pricing strategy and the number of users or subscribers you have. It's a fundamental financial equation that forms the basis of your business's financial model. Keep in mind that you can refine this equation and incorporate more financial factors as your business grows and the model becomes more complex.

Applicable Regulations

1. **Data Privacy Regulations**: When working with text data for summarization, it's essential to consider data privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union. These regulations govern how personal data is collected, processed, and used. If text data contains personal information, it must be handled in compliance with these regulations.
2. **Intellectual Property**: Text summarization can involve processing copyrighted material. It's crucial to respect copyright laws when summarizing and distributing content. Automated summarization should not infringe on the intellectual property rights of the original authors.
3. **Bias and Fairness**: These models can introduce bias if not properly trained and evaluated. Developers should be aware of and actively work to mitigate biases in summarization systems to ensure fair and unbiased content representation.
4. **Transparency and Accountability**: As AI and NLP technologies advance, there is growing interest in transparency and accountability in AI systems. Users and stakeholders may expect explanations for how summaries are generated and may seek mechanisms for challenging or appealing automated decisions.
5. **Ethical Considerations**: Developers should consider the ethical implications of text summarization, especially when it involves potentially sensitive or controversial topics. Ensuring that summaries are balanced and do not perpetuate misinformation or harmful stereotypes is important.

6. **User Consent**: When summarizing user-generated content or personal data, it's important to obtain informed consent and follow ethical guidelines related to data usage and user privacy.

Applicable Patents

https://patents.google.com/patent/US20140195897A1/en

Conclusion

In conclusion, text summarization stands as a pivotal and evolving domain within the field of natural language processing (NLP) and artificial intelligence (AI). This paper has explored the fundamental concepts, methodologies, and applications of text summarization, shedding light on its significance in simplifying information retrieval, aiding comprehension, and enhancing efficiency across a range of industries. The evaluation of summarization systems remains a critical challenge, and metrics like ROUGE continue to serve as valuable tools for assessing the quality of generated summaries. Ethical considerations, transparency, and the need to mitigate bias in automated summarization systems have emerged as essential focal points in the development of responsible AI technologies.

Looking ahead, text summarization's potential for enhancing information accessibility and decision-making has no limits. As NLP research and technologies continue to evolve, we expect further advancements in text summarization that will empower users across domains, from news consumption to document management and beyond.

In this dynamic landscape, collaboration between researchers, developers, and stakeholders remains important to harnessing the full potential of text summarization and ensuring its responsible and ethical deployment in our increasingly data-rich world.

References

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwandb.ai%2Fmostafaibrahim17%2Fml-articles%2Freports%2FCompressing-the-Story-The-Magic-of-Text-Summarization--VmlldzozNTYxMjc2&psig=AOvVaw1oxyYgwEZnmQ9G4a9FwJdr&ust=1694928470117000&source=images&cd=vfe&opi=89978449&ved=0CBIQjhxqFwoTCOja79yyroEDFQAAAAAdAAAAABAI

https://www.frase.io/blog/20-applications-of-automatic-summarization-in-the-enterprise/

https://www.wps.com/blog/top-5-ai-text-summarization-tool-in-2023-saving-your-reading-time/

mdpi.com/2076-3417/12/9/4479

https://www.researchgate.net/profile/Benjamin-Zierock/publication/369997892_Opportunities_and_challenges_of_using_artificial_intelligence_tools_from_ChatGPT_to_Aleph_Alpha_Neuroflash_and_Deepl/links/64384c7620f25554da2bcd1b/Opportunities-and-challenges-of-using-artificial-intelligence-tools-from-ChatGPT-to-Aleph-Alpha-Neuroflash-and-Deepl.pdf

https://www.frase.io/blog/20-applications-of-automatic-summarization-in-the-enterprise/

https://miro.medium.com/v2/resize:fit:1400/format:webp/1*1EIEITM-e_ST3NB9bSfhoA.png

https://arxiv.org/abs/1912.08777

https://blog.research.google/2020/06/pegasus-state-of-art-model-for.html