Team ID	592223
Project Name	Extracting Intelligent Insights with AI Based Systems

1. INTRODUCTION

1.1 Project Overview

In NLP, text summary refers to condensing a lengthy narrative into a manageable word count while succinctly expressing a crucial point.
One technique is to calculate word frequencies and then normalize them by dividing by the maximum frequency. This is one of several ways to condense a lengthy message and convey the most crucial information.
Next, identify the high-frequency sentences and select the ones that are most crucial to the point being made.

1.2 Purpose

- To save time and focus on other important tasks.
- To get the key points from lengthy paragraphs.
- To help students study more efficiently.

2. LITERATURE SURVEY

2.1 Existing problem

1. Content Quality: Ensuring accurate and relevant summarization is challenging, especially for complex or nuanced content.
2. Length and Detail Balance: Striking the right balance between brevity and retaining essential information can be difficult.
3. Multilingual Support: Providing effective summarization across multiple languages poses a significant challenge.
4. Handling Diverse Content Types:Summarizing different types of content, such as scientific articles or creative writing, requires tailored approaches.
5. User Customization:Meeting diverse user preferences for summarization styles and lengths can be complex.
6. Real-time Updates:Keeping up with rapidly changing information and delivering timely summaries is crucial.

7. Handling Ambiguity:Dealing with ambiguous or subjective content that lacks clear-cut summaries presents a hurdle.
8. Privacy Concerns:Ensuring user data privacy, especially when processing sensitive or personal texts, is a paramount concern.
9. Evaluation Metrics:Establishing universally accepted metrics for assessing summarization quality is an ongoing challenge.
10. Bias and Fairness:Mitigating biases in summarization results and ensuring fairness across different topics and perspectives is essential.

2.2 References

1. Text Summarization Techniques: A Brief Survey" by A. G. Siva Kumar and K. Rajanikanth.
2. Abstractive Text Summarization Using Sequence-to-sequence RNNs and Beyond" by R. Nallapati et al.
3. Explored libraries like Gensim, SpaCy, and TensorFlow for natural language processing tasks, including summarization.
4. Examined open-source projects on GitHub related to text summarization for practical insights and potential code contributions.
5. Read articles on platforms like Towards Data Science, Medium, and other tech blogs discussing text summarization techniques and best practices
6. Looked into proceedings from conferences like ACL (Association for Computational Linguistics) for the latest advancements in text summarization.
7. Engaged with the NLP and AI communities on platforms like Stack Overflow, Reddit (r/MachineLearning), or specialized forums for advice and problem.
8. Referred to the official documentation of NLP libraries to understand their capabilities and how they can be utilized for summarization.

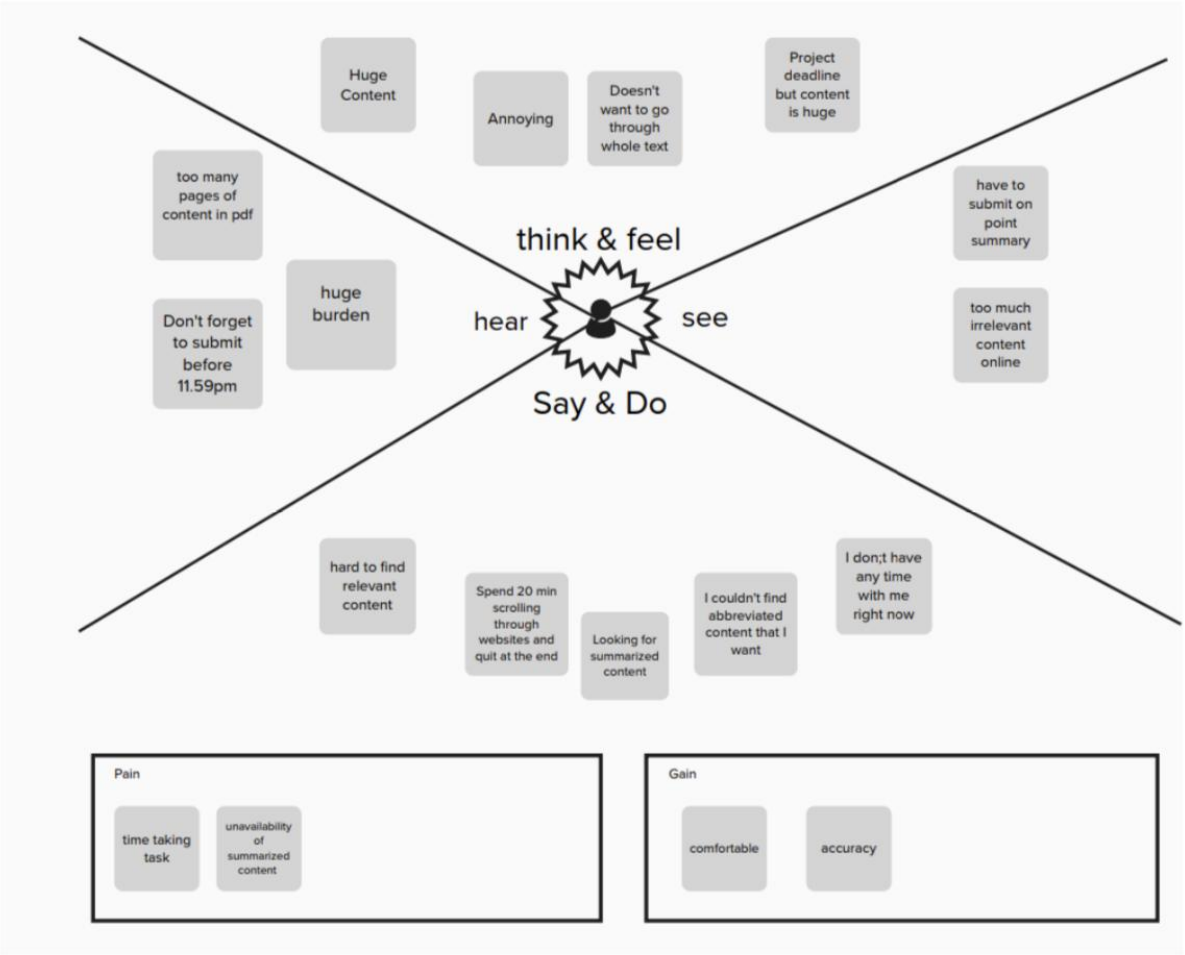
2.3 Problem Statement Definition

In today's information-rich environment, people frequently struggle with the difficult task of sorting through thick texts in order to extract crucial insights. This daunting task impedes effective information consumption and reduces productivity. The speed and accuracy needed for efficient summary are lacking in current methods, which leaves consumers struggling with information overload. Seeing this need, our text summarization website aims to solve these problems by providing a solid platform that revolutionizes how people summarize and understand complex material by fusing state-of-the-art algorithms with an approachable design.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

Empathy Map



3.2 Ideation & Brainstorming

Brainstorm

Diya Phoolwani

Utilizing AI in Natural Language Processing (NLP) for text summarization, focusing on extraction-based summarization to condense lengthy texts into concise, coherent summaries, offering users time-saving benefits and enhanced task prioritization. This approach treats text summarization as a supervised machine learning problem within NLP.

Shagun Srivastava

Extracting Intelligent Insights With AI Based Systems , thinking about this problem statement ,the first thing that comes to my mind is Natural Language Processing for Text Summarization i.e using few words to convey a significant point in a concise manner. Making long texts short can save a lot of time of the user , hence helping them prioritise on more important tasks

Sneha Prasad

The intention is to create a coherent and fluent summary.the key idea to do it is using extraction based summarization pulling key phases.this is a problem of NLP treated as a supervised machine learning problem.

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

1. Text Input Interface:Provide a user-friendly interface for users to input text for summarization , supports various input formats such as plain text, documents, and URLs.
2. Summarization Algorithm:Implement advanced summarization algorithms to extract key information from the input text.
3. Customization Options:Allow users to customize summarization preferences, such as summary length or focus on specific keywords.
4. Output Presentation:Display the generated summaries in a clear and readable format
5. Language Support:Support multiple languages to accommodate users with diverse linguistic preferences and content sources.
6. Integration with External Platforms:Allow integration with third-party applications or platforms through APIs, enabling seamless use within different ecosystems.
7. User Authentication and Privacy:Implement secure user authentication to protect user data and ensure privacy.
8. Feedback Mechanism:Include a feedback mechanism for users to provide input on the quality of summaries, helping to refine and enhance the summarization algorithms over time.
9. Error Handling:Implement effective error handling to gracefully manage issues such as invalid input, system errors, or interruptions in summarization processes.
10. Regular Updates and Maintenance:Establish a plan for regular updates and maintenance to address emerging issues, improve algorithms, and introduce new features based on user feedback and technological advancements.

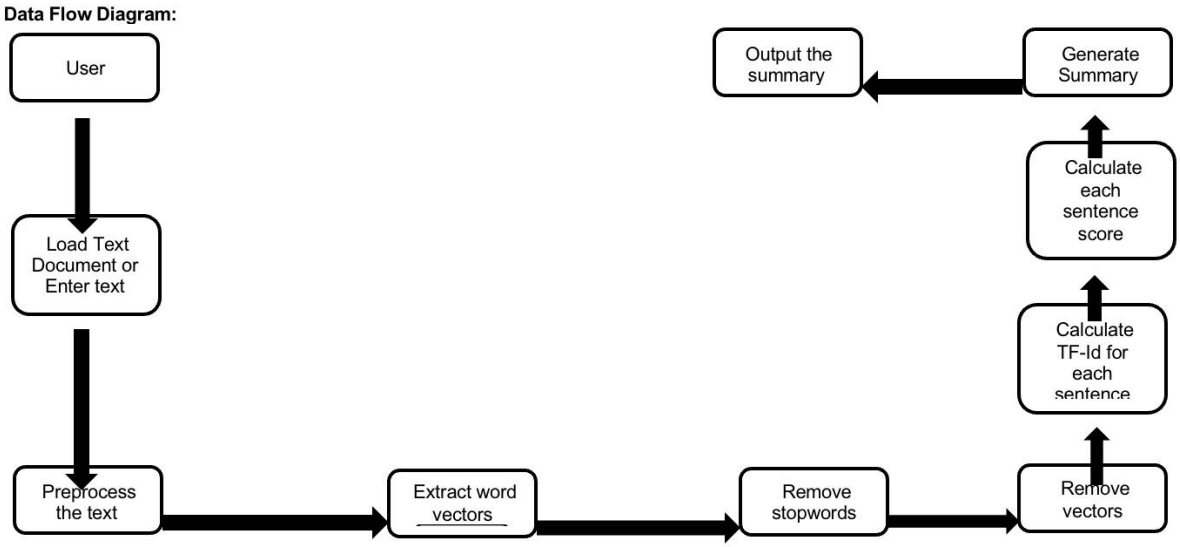
4.2 Non-Functional requirements

1. Performance:The system should generate summaries within seconds, ensuring a quick and responsive user experience.
2. Scalability:The architecture should support easy scalability to accommodate a growing user base and increasing demand for summarization services.
3. Reliability:The system should operate with high reliability, minimizing downtime and ensuring consistent availability for users.
4. Security:Implement robust security measures to protect user data, ensuring confidentiality and preventing unauthorized access to sensitive information.
5. Usability:The user interface should be intuitive and user-friendly, catering to users with varying levels of technical proficiency.
6. Compatibility:Ensure compatibility with a variety of web browsers to accommodate users with different browser preferences.
7. Compliance:Adhere to relevant data protection regulations and standards to ensure legal compliance and user trust.
8. Maintainability:The system should be easily maintainable, allowing for updates, bug fixes, and improvements without causing significant disruptions to service.

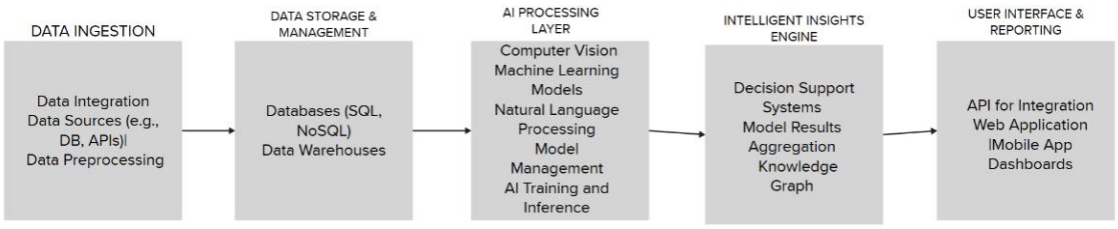
9. Documentation:Provide comprehensive documentation for users and administrators, covering system features, APIs, and troubleshooting guidelines.
10. Backup and Recovery:Establish a robust backup and recovery mechanism to safeguard data integrity and minimize the impact of potential data loss or system failures.

5. PROJECT DESIGN

5.1 Data Flow Diagrams & User Stories

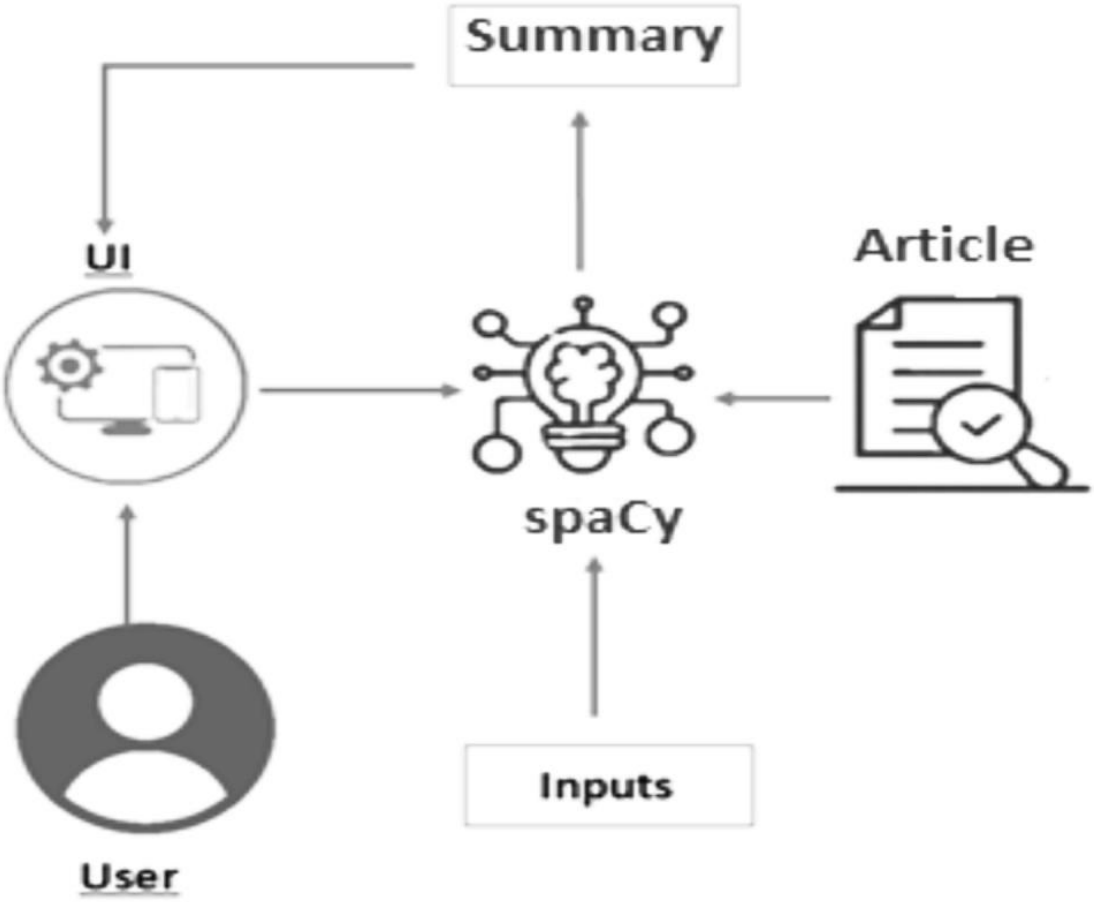


5.2 Solution Architecture



6. PROJECT PLANNING & SCHEDULING

6.1 Technical Architecture



6.2 Sprint Planning & Estimation

sprint	functional requirement	user story number	user story / task	story points	priority	team members
sprint-1	home page	USN1	as a user I was able to open the home page	2	high	shagun
sprint-1	original text container	USN2	as a user I was able to see the container in which I can upload the	1	high	diya

			text			
sprint-2	original and summarized text containers	USN3	as a user I was able to see original text and summarized text containers on the same page	2	low	sneha
sprint-1	summarized text	USN	as a user I was able to view the summary of the text uploaded by me	2	high	diya , shagun , sneha

6.3 Sprint Delivery Schedule

sprint	**total story points**	**duration**	**sprint start date**	**sprint end date (planned)**	**story point completed as per end date**	**sprint actual date**
sprint-1	20	4 days	14 October	18 October	20	18 October
sprint-2	20	4 days	20 October	24 October	20	24 October
sprint-3	20	4 days	26 October	30 October	20	30 October
sprint-4	20	4 days	2 November	6 November	20	6 November

7. CODING & SOLUTIONING

7.1 Feature 1

which involves having no restrictions on the number of words used in a project, is essentially the freedom to include as much text or content as needed without limitations imposed by a platform or system. This feature allows users to input a substantial amount of information, details, or text without encountering barriers based on word count.
By removing word count limitations, individuals can focus on the quality and completeness of their work without being constrained by arbitrary boundaries. However, it's important to note that while this feature allows for freedom in expression, it's also essential to maintain clarity, conciseness where possible, and relevance to ensure the effectiveness of the communicated message.

7.2 Feature 2

7.3 Database Schema

The database used in this project is SQl and NoSQL .
SQL databases are relational databases that use a structured schema. A schema is a blueprint that outlines the structure of how data is organized within the database. It defines tables, their columns, data types, relationships between tables (through foreign keys), and constraints.

8. PERFORMANCE TESTING

8.1 Performance Metrics

Our text summarization website employs several key performance metrics to ensure optimal functionality and user satisfaction. These metrics include:

1. Precision:Measures the accuracy of the generated summaries by assessing the relevance and correctness of the extracted information.
2. Processing Speed:Reflects the efficiency of our algorithms in generating summaries swiftly, allowing users to obtain condensed information without significant delays.
3. Readability:Assesses the clarity and coherence of the generated summaries, ensuring that the condensed content remains understandable and maintains linguistic quality.
4. Scalability:Measures the website's performance under varying workloads, ensuring consistent and reliable summarization capabilities even during periods of high user traffic.
5. User Feedback:Incorporates qualitative insights from user reviews and feedback to continually enhance the user experience and address specific needs and preferences.

By consistently monitoring and optimizing these performance metrics, we aim to deliver a text summarization website that excels in accuracy, efficiency, and user satisfaction.

9. RESULTS

9.1 Output Screenshots

```
In [36]: select_length=int(len(sentence_tokens)*0.3)
         select_length
Out[36]: 1

In [38]: summary=nlargest(select_length,sentence_scores,key=sentence_scores.get)

In [39]: summary
Out[39]: [Safeguarding the privacy of individuals or organizations whose data is collected by the sensors is an important ethical consid
         eration.]

In [40]: final_summary=[word.text for word in summary]

In [41]: summary=''.join(final_summary)

In [43]: print(summary)
         Safeguarding the privacy of individuals or organizations whose data is collected by the sensors is an important ethical conside
         ration.

In [44]: len(text)
Out[44]: 485

In [45]: len(summary)
Out[45]: 134
```

10. ADVANTAGES & DISADVANTAGES

10.1 Advantages

1. Time Efficiency:Users can quickly grasp the main points of lengthy texts, saving time compared to reading the entire content.
2. Information Retrieval:Enables users to efficiently find relevant information by providing concise summaries, improving overall information retrieval.
3. Enhanced Productivity:Facilitates productivity by condensing large volumes of text into easily digestible summaries.
4. Language Understanding: Aids users in understanding complex or specialized language, breaking down barriers to entry for diverse topics.
5. Decision Support:Supports decision-making processes by presenting key information, enabling users to make informed choices.
6. Learning Aid: Acts as a learning tool by distilling essential information, making it beneficial for educational purposes and research.
7. Multilingual Capabilities:Offers the ability to summarize content in multiple languages, broadening accessibility and usability.
8. Content Filtering:Helps users filter through vast amounts of information, allowing them to focus on what matters most to them.
9. Concise Communication:Promotes concise and effective communication, particularly useful in business and professional settings.
10. Accessibility:Makes information more accessible to a wider audience, including those with time constraints or reading difficulties.

10.2 Disadvantages

1. Loss of Nuance:Summaries may oversimplify complex topics, leading to a loss of nuance and detailed understanding
2. Inaccuracy:Automatic summarization may generate inaccurate or misleading summaries, impacting the reliability of information
3. Subjectivity:Summarization algorithms may introduce bias, reflecting the perspective of the model or its training data.
4. Difficulty with Abstraction:Abstract concepts or creative content may be challenging for summarization algorithms to capture accurately.
5. Lack of Context:Summaries may lack context, making it difficult for users to fully understand the implications of the information.
6. Dependency on Training Data:The quality of summaries depends heavily on the diversity and quality of the training data used for the summarization model.
7. Privacy Concerns:Processing sensitive or personal texts raises privacy concerns, especially if the summarization involves confidential information.
8. Limited Customization:Users may have different preferences for summarization styles and lengths, and a one-size-fits-all approach may not cater to everyone.
9. Challenges with Ambiguity:Ambiguous or context-dependent content may result in summaries that fail to capture the intended meaning.
10. Overemphasis on Extractive Summarization:Extractive summarization may focus too much on directly extracting sentences, potentially missing the generation of more coherent abstractive summaries.

11. CONCLUSION

In summary, our text summarization website stands as a beacon of efficiency in the vast sea of information. By harnessing advanced algorithms and cutting-edge technology, we have crafted a platform that transforms the laborious task of digesting lengthy texts into a seamless and time-saving experience. Users can now navigate the complexities of information overload with ease, extracting the essential points and gaining valuable insights. Embrace a more streamlined approach to comprehension and discovery with our platform, where the power of summarization meets the demands of the modern information age.

12. FUTURE SCOPE

The future scope of text summarization websites is promising, with ongoing advancements in natural language processing (NLP) and machine learning. We can anticipate the emergence of more advanced AI models capable of generating highly accurate and context-aware summaries. Customization and personalization options are expected to increase, allowing users to tailor summaries based on their specific preferences, industries, or use cases. Additionally, the integration of text summarization with other modalities like images, audio, and video is likely, providing users with comprehensive and multimodal summaries. Real-time summarization could become a reality, keeping users promptly updated with the latest information. Improved abstractive summarization techniques will lead to more nuanced and coherent summaries, capturing the essence of content with enhanced creativity. The future also holds the promise of better support for summarizing content in various languages, breaking down language

barriers. Specialized summarization models for specific domains, such as legal, medical, or scientific, are expected to provide more accurate and relevant summaries. The integration of explainable AI techniques will enhance transparency, helping users understand how summarization models arrive at their conclusions. Collaborative summarization tools may facilitate multiple users contributing to and refining summaries, fostering collective intelligence. Addressing ethical concerns, ensuring fairness, and mitigating biases in summarization models will be crucial for promoting responsible AI practices. In summary, the future of text summarization involves a convergence of technological advancements, user-centric approaches, and ethical considerations, making it an exciting field with vast potential for innovation.

13. APPENDIX

Source Code

In [ ]:
```
text="WSNs often collect sensitive data, raising concerns about privacy and data protection. Safeguarding the privacy of individu
```

In [ ]:
```
!pip install -U spacy
!python -m spacy download en_core_web_sm
```

In [ ]:
```
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
from string import punctuation
import numpy as np
```

In [ ]:
```
stopwords=list(STOP_WORDS)
stopwords
```

In [16]:
```
nlp=spacy.load('en_core_web_sm')
```

In [18]:
```
doc=nlp(text)
```

In [19]:
```
tokens=[token.text for token in doc]
print(tokens)
```

```
['WSNs', 'often', 'collect', 'sensitive', 'data', ',', 'raising', 'concerns', 'about', 'privacy', 'and', 'data', 'protection',
'.', 'Safeguarding', 'the', 'privacy', 'of', 'individuals', 'or', 'organizations', 'whose', 'data', 'is', 'collected', 'by', 't
he', 'sensors', 'is', 'an', 'important', 'ethical', 'consideration', '.', 'Addressing', 'these', 'drawbacks', 'requires', 'care
ful', 'consideration', 'during', 'the', 'design', 'and', 'deployment', 'of', 'WSNs', '.', 'Researchers', 'and', 'engineers', 'c
ontinue', 'to', 'work', 'on', 'developing', 'innovative', 'solutions', 'to', 'mitigate', 'these', 'challenges', 'and', 'enhanc
e', 'the', 'capabilities', 'of', 'wireless', 'sensor', 'networks', '.']
```

In [22]:
```
punctuation=punctuation+'\n'
punctuation
```

Out[22]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~\n'
```

In [23]:
```
word_frequencies={}
for word in doc:
    if word.text.lower() not in stopwords:
        if word.text.lower() not in punctuation:
            if word.text not in word_frequencies.keys():
                word_frequencies[word.text]=1
            else:
                word_frequencies[word.text]+=1
```

In [25]:
```
print(word_frequencies)
```

```
{'WSNs': 2, 'collect': 1, 'sensitive': 1, 'data': 3, 'raising': 1, 'concerns': 1, 'privacy': 2, 'protection': 1, 'Safeguardin
g': 1, 'individuals': 1, 'organizations': 1, 'collected': 1, 'sensors': 1, 'important': 1, 'ethical': 1, 'consideration': 2, 'A
ddressing': 1, 'drawbacks': 1, 'requires': 1, 'careful': 1, 'design': 1, 'deployment': 1, 'Researchers': 1, 'engineers': 1, 'co
ntinue': 1, 'work': 1, 'developing': 1, 'innovative': 1, 'solutions': 1, 'mitigate': 1, 'challenges': 1, 'enhance': 1, 'capabil
ities': 1, 'wireless': 1, 'sensor': 1, 'networks': 1}
```

In [26]:
```
max_frequency=max(word_frequencies.values())
```

In [27]:
```
max_frequency
```

Out[27]:
```
3
```

In [29]:
```
for word in word_frequencies.keys():
```

```
In [31]: print(word_frequencies)

{'WSNs': 0.6666666666666666, 'collect': 0.3333333333333333, 'sensitive': 0.3333333333333333, 'data': 1.0, 'raising': 0.3333333
33333333, 'concerns': 0.3333333333333333, 'privacy': 0.6666666666666666, 'protection': 0.3333333333333333, 'Safeguarding': 0.33
33333333333333, 'individuals': 0.3333333333333333, 'organizations': 0.3333333333333333, 'collected': 0.3333333333333333, 'senso
rs': 0.3333333333333333, 'important': 0.3333333333333333, 'ethical': 0.3333333333333333, 'consideration': 0.6666666666666666,
'Addressing': 0.3333333333333333, 'drawbacks': 0.3333333333333333, 'requires': 0.3333333333333333, 'careful': 0.333333333333333
3, 'design': 0.3333333333333333, 'deployment': 0.3333333333333333, 'Researchers': 0.3333333333333333, 'engineers': 0.3333333333
333333, 'continue': 0.3333333333333333, 'work': 0.3333333333333333, 'developing': 0.3333333333333333, 'innovative': 0.333333333
3333333, 'solutions': 0.3333333333333333, 'mitigate': 0.3333333333333333, 'challenges': 0.3333333333333333, 'enhance': 0.333333
3333333333, 'capabilities': 0.3333333333333333, 'wireless': 0.3333333333333333, 'sensor': 0.3333333333333333, 'networks': 0.333
3333333333333}

In [32]: sentence_tokens=[sent for sent in doc.sents]
         print(sentence_tokens)

[WSNs often collect sensitive data, raising concerns about privacy and data protection., Safeguarding the privacy of individual
s or organizations whose data is collected by the sensors is an important ethical consideration., Addressing these drawbacks re
quires careful consideration during the design and deployment of WSNs., Researchers and engineers continue to work on developin
g innovative solutions to mitigate these challenges and enhance the capabilities of wireless sensor networks.]

In [33]: sentence_scores={}
         for sent in sentence_tokens:
             for word in sent:
                 if word.text.lower() in word_frequencies.keys():
                     if sent not in sentence_scores.keys():
                         sentence_scores[sent]=word_frequencies[word.text.lower()]
                     else:
                         sentence_scores[sent]+=word_frequencies[word.text.lower()]

In [34]: sentence_scores

Out[34]: {WSNs often collect sensitive data, raising concerns about privacy and data protection.: 4.333333333333333,
          Safeguarding the privacy of individuals or organizations whose data is collected by the sensors is an important ethical consid
         eration.: 4.333333333333334,
          Addressing these drawbacks requires careful consideration during the design and deployment of WSNs.: 2.333333333333333,
          Researchers and engineers continue to work on developing innovative solutions to mitigate these challenges and enhance the cap
         abilities of wireless sensor networks.: 4.333333333333333}

In [35]: from heapq import nlargest

In [36]: select_length=int(len(sentence_tokens)*0.3)
         select_length

Out[36]: 1

In [38]: summary=nlargest(select_length,sentence_scores,key=sentence_scores.get)

In [39]: summary

Out[39]: [Safeguarding the privacy of individuals or organizations whose data is collected by the sensors is an important ethical consid
         eration.]

In [40]: final_summary=[word.text for word in summary]

In [41]: summary=''.join(final_summary)

In [43]: print(summary)
```

GitHub & Project Demo Link

https://github.com/smartinternz02/SI-GuidedProject-594747-1697646103