Super Nova

NLP-powered News Discovery Engine

Ritwika Das Gupta (2348049)
Soham Chatterjee (2348062)
Sayantan Ray (2348057)
Under the guidance of Dr. Saleema J S

Introduction

1. In the digital age, the vast amount of news content generated daily makes it difficult for users to find relevant and reliable information quickly.

2. A project to develop a search engine for summarizing and ranking news articles based on user queries.

3. Uses NLP techniques for efficient document retrieval and summarization.

4. Helps users find relevant news articles quickly by providing concise summaries.

5. Focuses on enhancing information retrieval from large collections of online news.

Objectives

1. To build a search engine capable of scraping and retrieving relevant news articles.
2. Rank articles based on query relevance using NLP techniques.
3. Real time suggestions as a search engine would have.
4. Summarize lengthy articles for quick user consumption.
5. Provide an interface that is intuitive and easy to use.
6. Improve access to information by merging ranking and summarization in one platform.

Web Scraping and Content Extraction

What is it? Web scraping is the automated process of extracting data from websites. In this project, we use web scraping to collect headlines, URLs, and article content from news websites.
How is it used?

1. **BeautifulSoup**: We use this Python library to parse HTML and extract useful data, such as **news headlines and articles.**
2. **Data Collection**: The **scrape_telegraph_india** function fetches data from the website by extracting headlines **(<h3> or span tags based on websites)** and the associated URLs. This ensures the latest news articles are retrieved for analysis.
3. **Content Extraction**: Once a URL is obtained, we make a request to fetch the full article text by selecting paragraphs (<p> tags) from the HTML structure.
4. **Purpose**: The scraped data forms the basis for performing natural language processing tasks like ranking and summarization.

NER Tagging & RAKE

- **Named Entity Recognition (NER):**
  - ✓ Automatically **extracts entities** like people, locations, and organizations from text.
  - ✓ Helps in identifying and categorizing key information for **improved search and retrieval**.

- **RAKE (Rapid Automatic Keyword Extraction):**
  - ✓ Extracts keywords and key phrases by analyzing the frequency and co-occurrence of words.
  - ✓ **Filters and ranks keywords to highlight the most relevant terms** for summarization and search relevance.

- **Combined Impact:**
  - ✓ Enhances the system's ability to comprehend the context and significance of the content.

Auto-Suggestion and Spellcheck in Search Engine

Auto-Suggestion Feature

N-grams-based prediction:

- Analyzes bigrams in the news corpus.
- Suggests the next word based on user input and previous patterns in the dataset.
- Returns top 5 most frequent next-word suggestions from the n-gram model.

Spellcheck Mechanism

Levenshtein Distance Algorithm:

- Compares user input with words in the corpus.
- Corrects misspelled words by finding the closest match based on the smallest edit distance.
- Displays corrections with a prompt: "Did you mean:".

Word Embeddings (GloVe)

What is it? Word embeddings are vector representations of words where similar words are close to each other in a multi-dimensional space. GloVe (Global Vectors for Word Representation) is a popular pre-trained word **embedding model.**

How is it used?

1. **GloVe Embeddings**: We load pre-trained GloVe embeddings (glove.6B.50d.txt), which map each word to a **50-dimensional vector.**
2. **Average Word Embedding**: For each news article, the text is tokenized into words using NLTK's **word_tokenize** function. We then calculate the average of the embeddings for each word to form a document-level embedding.
3. **Purpose**: These embeddings are used to represent both the query and the news articles in a numerical format, making it easier to compare them using mathematical operations like cosine similarity.

Cosine Similarity

What is it? Cosine similarity is a measure that calculates the cosine of the angle between two vectors. It ranges from -1 to 1, where 1 means the vectors are identical, and 0 means they are orthogonal (completely different).

How is it used?

1. **Document Ranking**: After generating the average embeddings for the query and the news articles, we use cosine similarity to **calculate the relevance of each article to the query.**
2. **Comparison**: The similarity is calculated between the query vector and each article's vector, resulting in a score for each article.
3. **Ranking**: Articles are ranked by their similarity scores, with the **top 5 most relevant articles** selected for further processing.
4. **Purpose**: Cosine similarity ensures that the most relevant articles are identified and prioritized for summarization.

Text Summarization (Transformers)

What is it? Text summarization is the task of reducing a long piece of text into a shorter version while retaining the key information. Transformer models, especially based on architectures like BART is widely used for this purpose in NLP.

How is it used?

1. **Summarization Pipeline**: We use Hugging Face's pipeline method to load a pre-trained summarization model. This model automatically generates concise summaries from long text.
2. **Summarization Process**: After ranking the articles, we retrieve the full text of the top 5 articles and pass them through the summarization model. The model outputs summaries that are constrained by length (between 50 and 150 words).
3. **Purpose**: This process allows users to **quickly grasp the main points of the articles** without reading the entire text, providing an efficient way to digest information.

Workflow

Web Scraping

- Scrape news websites to gather headlines and URLs.
- Fetch article text from web pages for analysis.

Query Processing

- Ensuring the collected data is structured and clean for processing.
- Tokenize and process the query into vector representation.
- Use word embeddings to generate average vector for the query.
- Named Entity Recognition (NER) & Keyword Extraction using RAKE

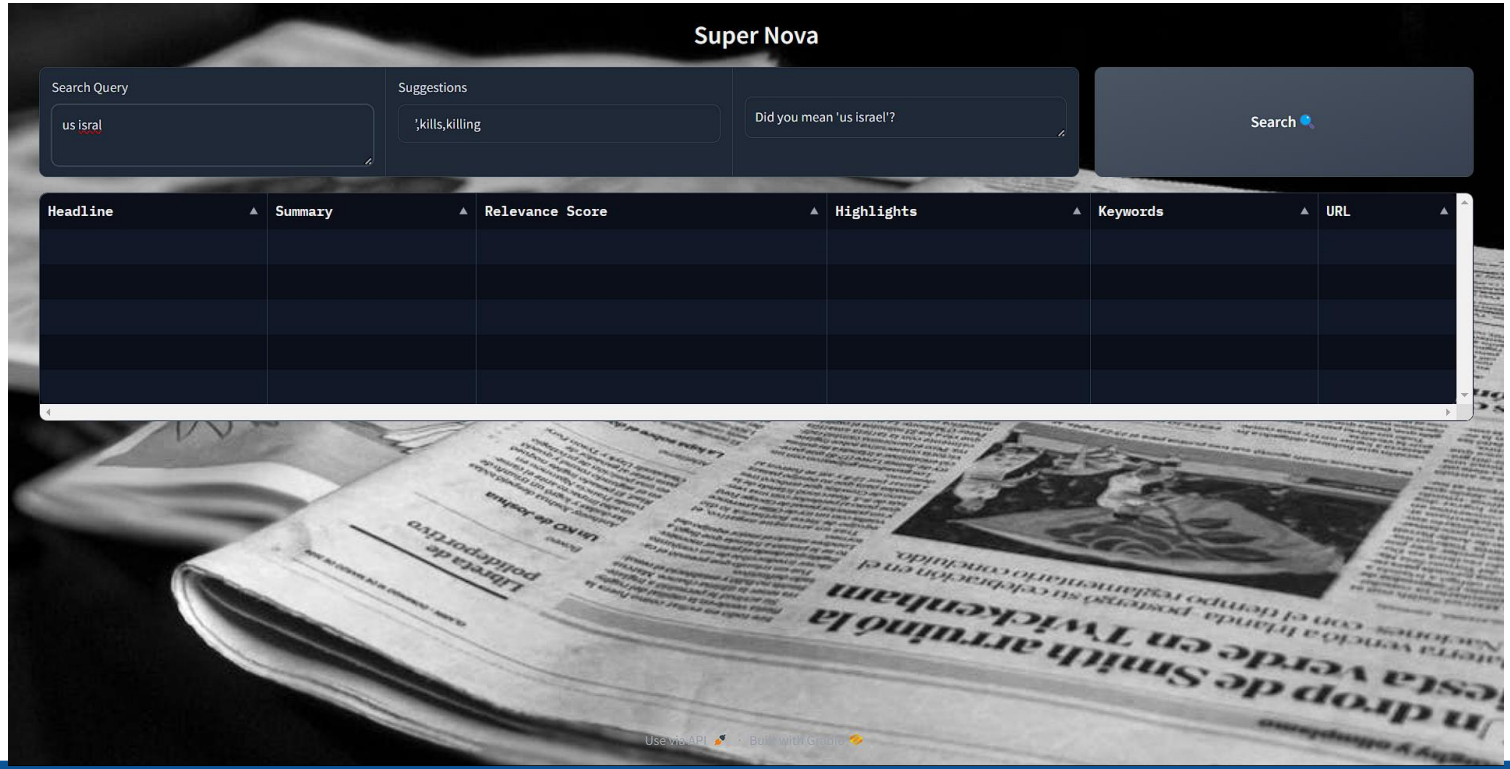
Query Handling

- Rank documents using cosine similarity to measure relevance.
- Select the top 5 most relevant articles for each query.
- Summarize the selected articles using a transformer model.
- Provides real-time suggestions and corrects spelling using n-grams and Levenshtein distance.

User Interface

- Implements a Gradio-based web interface for searching and displaying news articles.

Interface


Super Nova
Search Query
us isral
Suggestions
',kills,killing
Did you mean 'us israel'?
Search
Headline
Summary
Relevance Score
Highlights
Keywords
URL

Interface


Super Nova
Search Query
us israel
Suggestions
',kills,killing
Go on and ask your query
Search
Headline
Summary
Israel kills 40 people in Gaza as UN chief urges end to 'horrific violence'
UN chief calls for an end to Israel's 'horrific violence'
Putin warns Ukraine use of long-range arms will put NATO at war with Russia
Vladimir Putin says Ukraine should be allowed to use long
US hints at allowing Ukraine to strike deep inside Russia, enraging Putin
Ukraine may be starting to move the strategic needle in f
US expected to charge Iranian hackers who targeted Trump campaign
The Justice Department is expected to soon announce crimi
Biden signals openness to using Western long-range missiles inside Russia as he meets Britain's leader to discuss
U.S. and UK officials will meet on Friday to discuss long
Use via API
Built with Gradio

Interface

Relevance Score	Highlights
0.86	israel ', relentless bombardment, palestinian enclave, gaza, end
0.86	russian front, prime minister, civilian targets, us ,", putin ', starmer said, donald tusk, british leaders, range weapons, range
0.85	foreign minister, inside russia, us president, white house, deeper strikes, air strikes, us said, campbell said, us weapons, north
0.84	us officials, us officials, iranian government, publicly expose, news organizations, media company, criminal charges, china –, 201
0.84	jeanne shaheen, ." despite, official said, missile launchers, allowing ukraine, russia –, allowing us, targets .", expect russia,

	Keywords	URL
	Israel-- Gaza-- UN	https://www.aljazeera.com/news/liveblog/2024/9/13/israels-war-on-gaza-live-dozens-more-killed-amid-anger-over-school-attack
;	Putin-- Ukraine-- NATO-- Russia	https://www.aljazeera.com/news/2024/9/13/putin-warns-ukraine-use-of-long-range-arms-will-put-nato-at-war-with-russia
	US-- Ukraine-- Russia-- Putin	https://www.aljazeera.com/news/2024/9/13/us-hints-at-allowing-ukraine-to-strike-deep-inside-russia-enraging-putin
tion, " right	Iranian-- Trump	https://www.cnn.com/2024/09/12/politics/charges-expected-hackers-trump-campaign/index.html
nference	Biden-- Western-- Russia-- Britain	https://www.cnn.com/2024/09/13/politics/biden-starmer-meeting-ukraine-missiles/index.html

Future Scope

1. **Expand Data Sources:** Add more diverse news websites to the system.
2. **Multilingual Support:** Enable queries and articles in multiple languages.
3. **Real-Time Updates:** Implement live news scraping for up-to-date content.
4. **User Personalization:** Provide personalized search results based on user preferences.
5. **Enhanced Embeddings:** Experiment with BERT or custom models for better search accuracy.

Conclusion

- Used NLP, web scraping, and machine learning to develop a search engine that successfully searched news.
- It provides real-time search, keyword extraction, and domain filtering to offer the best user experience.
- Because of this, it is highly extensible, where adding more features and news sources can be done easily because of its modular design.
- Demonstrated how information retrieval can be enhanced by combining classic and advanced Machine Learning techniques.
- Sentiment analysis will also be integrated in future work, along with using an expanded dataset to enable wider coverage and increased accuracy.

References

1. **BeautifulSoup**: L. Richardson, *Beautiful Soup Documentation*. Available: https://www.crummy.com/software/BeautifulSoup/.
2. **NLTK (Natural Language Toolkit)**: S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed., O'Reilly Media, 2009. Available: https://www.nltk.org.
3. **GloVe (Global Vectors for Word Representation)**: J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.
4. **Cosine Similarity**: D. M. W. Powers, "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011. Available: https://www.semanticscholar.org/paper/Evaluation%3A-From-Precision%2C-Recall%2C-and-F-to-ROC%2C-Powers.
5. **Hugging Face Transformers**: T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38-45. doi: 10.18653/v1/2020.emnlp-demos.6.
6. **Gradio**: A. Abid, M. Faris, and S. Rasheed, *Gradio Documentation*, 2021. Available: https://www.gradio.app.
7. **Requests Library**: K. Reitz, *Requests: HTTP for Humans Documentation*. Available: https://docs.python-requests.org/en/latest/.

THANK YOU