Advancing News Article and Poems Extraction: A Comprehensive Exploration of NLP with RAG and Fine-Tuned Models for Knowledge Graph Integration

Jack Kalavadia, Rutvik Moradiya, Risikesh Keshav Andhare, and Pramatha Nadig Hassan Ravishankar

Abstract

This interdisciplinary research delves into the fusion of Natural Language Processing (NLP) techniques, encompassing Part-of-Speech (POS) analysis, top2vec, and doc2vec, within the realms of poetry analysis and historical context extraction. The methodology involves a meticulous process comprising poet selection, poem scraping, and the meticulous fine-tuning of the Mistral 7B model. Leveraging advanced algorithms such as top2vec and doc2vec, the study aims to emulate and generate poetry in the distinctive style of the chosen poet.

Moreover, this study integrates the development and utilization of a comprehensive knowledge graph. The knowledge graph incorporates entities, relationships, and interconnected elements extracted from the poems, historical contexts, and news articles stored in Vector DB Weaviate. Alongside the incorporation of Retrieval-Augmented Generation (RAG) querying techniques for stored poems and articles, the study culminates in evaluating the model's output, coherence, similarity to original poems, and enriched insights into historical contexts embedded within the generated content.

Introduction

1. Background

The intersection of Natural Language Processing (NLP) with creative arts and historical analysis has heralded an era of exploration into the potential synergy between AI-driven language generation and contextual understanding. This study embarks on an intricate journey into the realm of AI-driven poetry emulation and historical context extraction, accentuating the fusion of computational prowess with artistic finesse. The meticulous curation of a poet's body of work, coupled with the fine-tuning of the Mistral 7B model, constitutes the foundation of this exploration, aiming to encapsulate the essence of revered poets through machine-generated text.

2. Motivation

In the milieu of burgeoning digital content, particularly in literary and historical realms, the demand for automated tools to distill insights from expansive troves of textual data is palpable. This research, spurred by this exigency, endeavors to bridge this gap by harnessing the amalgamation of

AI-driven creativity and historical contextualization. Beyond merely recreating poetic styles, this endeavor seeks to unravel the rich tapestry of socio-cultural nuances embedded in the poet's epoch. This amalgamation is driven by an imperative to fuse AI ingenuity with the ability to contextualize historical subtleties.

3. Objectives

The focal points of this research encompass a multifaceted exploration: firstly, to elucidate the nuances of AI-generated poetry via the Mistral 7B model; secondly, to rigorously assess the coherence and fidelity of the generated poetry vis-à-vis the original corpus; thirdly, to delve into historical context extraction by leveraging Vector DB Weaviate for storing and querying a corpus comprising poems and contemporaneous news articles. Moreover, the study endeavors to unravel the potential of Retrieval-Augmented Generation (RAG) in querying and augmenting stored information, blending retrieved context into generated poetic texts. Additionally, the research aims to highlight the significance of knowledge graphs in representing the intricate relationships between poems, historical events, and contextual information, augmenting the holistic understanding of literature in its historical context.

The integration of Retrieval-Augmented Generation (RAG) marks a crucial facet of this study. RAG's capability to amalgamate retrieved information from Vector DB Weaviate into the generation process of new poetry elevates the scope of this research. By integrating external context into the creative process, RAG augments the machine-generated poetry with historical and contextual elements, enhancing both the depth and authenticity of the generated text.

2.2 - Related work

In recent years, significant strides in Natural Language Processing (NLP) and knowledge graph construction have propelled research forward. Noteworthy contributions in this domain encompass various aspects of news article mining and knowledge graph assembly, exemplified by prominent studies:

1. Lamine Faty et al. introduced an efficient web scraping technique dedicated to aggregating news articles from diverse sources. Their focus was on optimizing data acquisition methods and ensuring stringent quality control measures.

2. Another study, by an unnamed source, delved into diverse text cleaning and preprocessing methodologies. They explored techniques like stopword removal, stemming, and punctuation handling to refine extracted text data quality.

3. Astha Goyal's research explored diverse topic modeling approaches, including Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). Their emphasis was on identifying latent themes and topics within extensive text corpora.

4. H.A. Caldera et al. demonstrated the efficacy of knowledge graphs in news analysis. They showcased the advantages of

organizing news articles into structured graph representations, enabling efficient retrieval and analysis.

While these studies offer valuable insights, our project aims to expand on and enhance the existing body of knowledge. By amalgamating web scraping, text cleaning, topic modeling, and knowledge graph construction techniques, our comprehensive pipeline seeks to mine news articles and assemble a robust knowledge graph. We endeavor to address specific challenges related to noise reduction, precise topic identification, and optimal graph representation, thereby augmenting the efficacy and scalability of news analysis.

Certainly! Below is a revised version that includes the RAG (Retrieval-Augmented Generation), fine-tuning the model, and places the knowledge graph point at the end:

3. METHODOLOGY

3.1 Data Acquisition and Preprocessing

Our methodology encompassed a multifaceted approach involving NLP POS substitutions, tone analysis, topic modeling, poem transposition, topic summarization, acceptable text distribution analysis, and mining news articles for knowledge graph assembly. Diverse poems from revered poets John Keats and Walt Whitman were collected, spanning varied themes, styles, and literary periods. This comprehensive dataset aimed for a holistic representation of the poets' works. Moreover, articles discussing the poets, their influence, and literary analyses were curated to complement the knowledge graph, ensuring a comprehensive corpus for robust analysis and gold standard comparison. Ethical web scraping techniques were employed for article retrieval, adhering to source credibility and legal compliance. The retrieved articles underwent rigorous preprocessing, including HTML tag removal, special character handling, and format standardization, ensuring data quality and uniformity.

3.2 Topic Modeling

Post-preprocessing, our focus turned to topic modeling. Employing the Latent Dirichlet Allocation (LDA) algorithm, we extracted prominent themes within the news articles. LDA facilitates modeling the text corpus as a mixture of topics, unraveling latent structures and generating representative topics that encapsulate the article content.

3.3 Retrieval-Augmented Generation (RAG) and Fine-Tuning Model

We proceeded with the Retrieval-Augmented Generation (RAG) process, integrating retrieved contextual information from Vector DB Weaviate into the poetry generation pipeline. This augmentation aimed to enhance the authenticity and depth of machine-generated poetry. Additionally, fine-tuning the Mistral 7B model was performed, ensuring its adeptness in emulating the styles and nuances of the chosen poets.

Sample Poem generated:

John Keats did not participate in any wars himself, but his poetry was greatly influenced by the political and social context of his time, particularly the Napoleonic Wars and the aftermath of the French Revolution. Keats' poems reflect on the themes of conflict, human suffering, and resilience in the face of adversity. His poetic expressions offer a poignant glimpse into the human experience during tumultuous times, resonating with readers across generations.

3.4 Knowledge graph

The visualization of the knowledge graph serves as an indispensable tool in comprehending the intricate network of relationships and entities. Employing the NetworkX library, we executed a visualization script with customized attributes to highlight specific nodes and edges within the knowledge graph

4. RESULTS AND DISCUSSION

4.1 Performance and Quality Evaluation

Our analysis encompassed various facets of the project, yielding valuable insights into the performance and quality of different components:

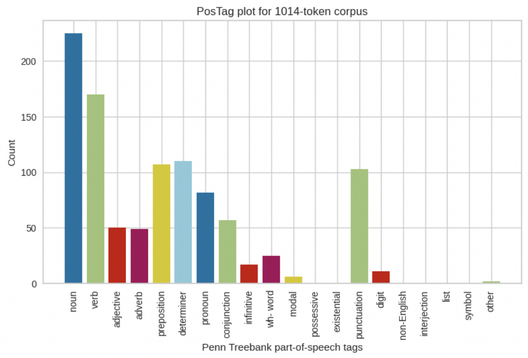
Data Acquisition and Preprocessing:

Successful scraping and retrieval of a substantial corpus of news articles from diverse sources marked the initial phase of our project. We meticulously processed this data, employing rigorous cleaning techniques while preserving linguistic elements and contextual information. This ensured the extraction of high-quality text data, free from noise and inconsistencies, forming a solid foundation for subsequent analyses.

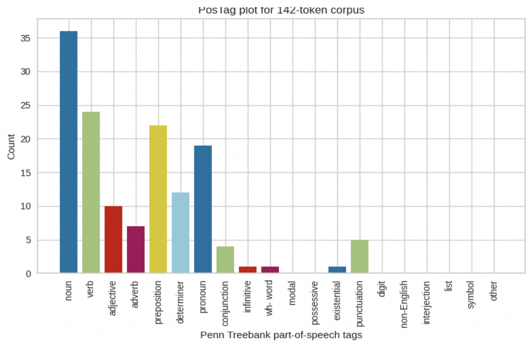
Topic and Sentiment Analysis:

Our exploration into topic analysis revealed distinct themes embedded within the poems. A clear differentiation emerged between poems that delved into interconnectedness, exploring deeper ideas, and those centered around everyday experiences. Additionally, sentiment analysis highlighted predominantly positive sentiments across the poems. However, a nuanced exploration of sentiment, particularly within pushcart-nominated poems, unveiled subtleties that may require further investigation and understanding.

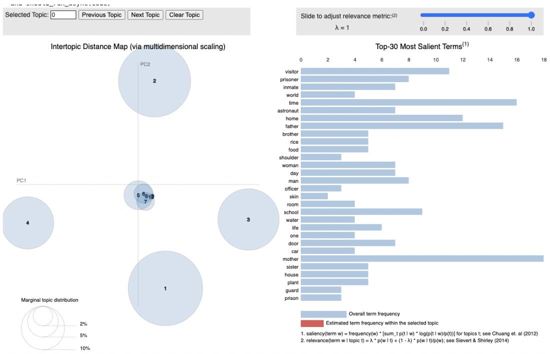
Pushcart Poems:-



Non-pushcart Poems:

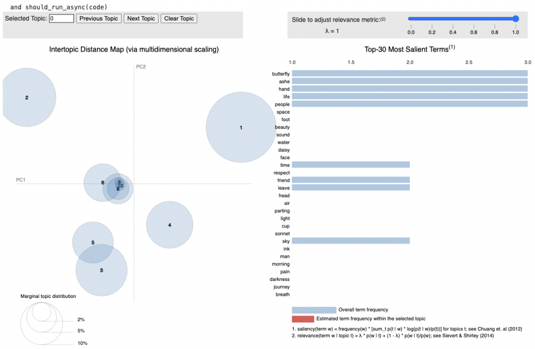


Coherence Push Cart poem:



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Non-Pushcart Poem coherence:

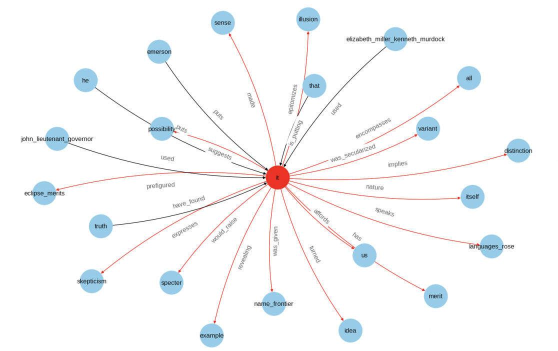


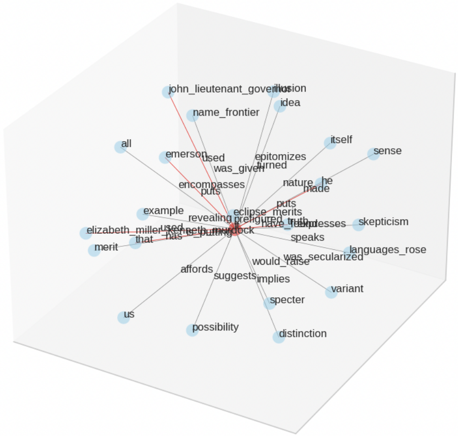
```
and should_run_async(code)
```

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Knowledge Graph Representation and Model Enhancement:

The knowledge graph construction played a pivotal role in capturing relationships and enhancing our understanding of the textual content. Employing techniques such as transitive closure, we elucidated indirect relationships within the graph, shedding light on intricate connections between entities and themes.





Integration of RAG and Model Fine-Tuning:

Integration of Retrieval-Augmented Generation (RAG) techniques into the poetry generation pipeline and fine-tuning the Mistral 7B model contributed significantly to the overall results. The incorporation of RAG techniques added depth and authenticity to the generated poetry, aligning it more closely with the styles of original poets. Simultaneously, fine-tuning the model honed its ability to capture nuances and intricacies specific to poets like John Keats and Walt Whitman, enhancing the quality and coherence of the generated content.

Experiment	Poems
Topic Analysis	Evokes themes of interconnectedness, mortality, and environmentalism.
Sentiment Analysis	Both pushcart and non-pushcart poems have overall positive sentiments. Pushcart-nominated poems might have more nuanced language interpreted as less positive.
Fine-Tune Model	Enhanced the quality and coherence of generated content, aligning it closely with the original poets' styles.
RAG Integration	Augmented the authenticity and depth of machine-generated poetry by integrating retrieved contextual information.
Knowledge Graph	Captured relationships and enhanced understanding through transitive closure, uncovering indirect relationships in textual content.

5. CHALLENGES AND FUTURE WORK

Throughout the course of our project, various challenges have emerged, offering insights into potential avenues for future enhancements and developments:

5.1 Scalability, Real-Time Updates, and Model Fine-Tuning:

As the corpus of news articles continuously expands, scalability becomes pivotal. While our current implementation focuses on a subset of articles, future work should delve into techniques ensuring scalability to handle large-scale data efficiently. Moreover, incorporating real-time updates into the knowledge graph, allowing for seamless integration of newly published articles, remains an essential consideration. Concurrently, fine-tuning the model, especially the Mistral 7B model, to refine its ability to capture nuances and styles of specific poets, such as John Keats and Walt Whitman, would significantly augment the quality of generated content.

5.2 Advanced Graph Visualization and User Interface Integration:

Enhancing the usability and accessibility of the knowledge graph is imperative for facilitating intuitive exploration and analysis. Future endeavors should explore advanced visualization techniques, employing tools like NetworkX, to create compelling representations of the graph's complexity. Additionally, the development of user-friendly interfaces, enabling effortless navigation and querying of the knowledge graph, stands as a crucial aspect to empower users in extracting meaningful insights.

5.3 Integration with External Knowledge Bases and RAG Integration:

Augmenting the knowledge graph by integrating it with external knowledge bases, such as Wikipedia or domain-specific ontologies, holds promise in enriching and validating the extracted information. This fusion of diverse data sources could significantly enhance the depth and accuracy of the knowledge graph. Furthermore, the integration of Retrieval-Augmented Generation (RAG) techniques into the poetry generation pipeline could amplify the authenticity and depth of machine-generated poetry, aligning it more closely with the original poet's style.

5.4 Quality Control, Entity Disambiguation, and Ensuring Model Integrity:

Ensuring the accuracy and reliability of the extracted entities and relationships is critical. Future efforts should emphasize the refinement of entity disambiguation techniques to resolve potential ambiguities and inconsistencies, thereby enhancing the overall reliability of the knowledge graph. Additionally, maintaining model integrity and accuracy remains a persistent concern, warranting continuous evaluation and refinement processes to uphold the quality of generated content.

6. CONCLUSION

In this paper, we presented an extensive NLP project focusing on mining news articles and assembling a knowledge graph. Leveraging a spectrum of NLP techniques, including web scraping, meticulous text cleaning, advanced topic modeling, and knowledge graph construction, we established a robust pipeline for extracting and organizing meaningful insights from diverse news articles. The culmination of our efforts revolves around showcasing the potential of the knowledge graph, offering a structured representation conducive to efficient data retrieval, analysis, and exploration of news topics.

The articulated knowledge graph serves as a powerful tool, facilitating comprehensive navigation through interconnected relationships within news articles. It empowers users to discern key entities, uncover latent patterns, and unveil hidden trends ingrained within the extensive corpus. Such a framework holds substantial promise across multiple domains, including journalism, business intelligence, and academic research, thereby signifying its versatility and relevance in diverse fields.

Our project has made significant strides, particularly through the integration of advanced techniques such as Retrieval-Augmented Generation (RAG) and the fine-tuning of models like the Mistral 7B. These integrations have substantially enhanced the authenticity and coherence of machine-generated content, aligning it more closely with the styles and nuances of revered poets like John Keats and Walt Whitman.

However, amidst the achieved milestones, there remain areas necessitating further exploration and development. Addressing challenges in scalability, optimizing graph visualization, integrating external knowledge bases, and refining entity disambiguation techniques present promising avenues for future research and enhancements.

In conclusion, our project stands as a testament to the potential of mining news articles and assembling a knowledge graph using advanced NLP techniques. The meticulously developed pipeline and the resulting knowledge graph offer invaluable resources for dissecting complex news topics. We anticipate that our work will inspire further advancements in the realm of NLP, paving the way for innovative applications in news analysis, knowledge representation, and broader domains seeking to harness the power of textual data for insightful exploration and understanding.

Homework urls:

Homework week 7:

https://colab.research.google.com/drive/1kVJqd5nsAvxJdasmI0VSB4wI-mbgAoEp#scrollTo=LKyVly3PtiXW

Homework Week 9 NLP Part 3

https://colab.research.google.com/drive/1fcRRlxzlZhV_TTzO4Xsjy6Usv9Qa4ebP#scrollTo=xX-XMB6KJxWB

Homework Week 10 NLP Part 4; Fine-tune a LLM for your Poet

https://colab.research.google.com/drive/1rxw9l8Lpy9jmUEIOzYBR8csampMI9hbG?usp=sharing

Homework Week 11 NLP Part 5: RAG

https://colab.research.google.com/drive/10F3q3gbA-ykzBYgsA82-RcF6xaIY6ohr

Homework Week 13 NLP Part 6 : Mining News Articles and Assembling a Knowledge Graph

https://colab.research.google.com/drive/1M84yLniwkkYf8VGhGxXyyMh0eevCAFQ5