AI Research Project Technical Report

Project Title: Fine-tuning and Assessing a Machine Learning Model on William Blake and Robert Frost Poetry

Experiments:

- **POS Substitutions and Semantic Analysis in Poetry** https://drive.google.com/drive/u/0/folders/1E1UTMYOtKEJ6ns9LVtypNXvwr8Dsz6da
- **Comparative Analysis with Gold Standard Texts** https://drive.google.com/drive/u/0/folders/1Gf4xs14lvjXk4Ia3j2G2RGKeL1ydq3ep
- **Language Model Fine-Tuning for Poetic Style Mimicking** https://drive.google.com/drive/u/0/folders/1wwFkaTxXlXZoTkIndtWzWyTM EGyU-upQ
- **Vector DB Integration and Retrieval-Augmented Generation (RAG)** https://drive.google.com/drive/u/0/folders/1BKQv_Ga-6TWTvxqM_uyoau39rS0sSD8Q
- **Knowledge Graph Assembly and News Article Mining** https://drive.google.com/drive/u/0/folders/1ZiqAMefxYWHoHLXL2oUtst-IYwo9wkUK

Abstract

This paper presents an in-depth exploration of the intersection between Natural Language Processing (NLP) and computational creativity, with a specific focus on analyzing and generating poetry. The study comprises five phases, which offer a nuanced understanding of poetic expression through advanced computational techniques. The initial phase involves scraping poems from two prominent poets and investigating the impact of Part-of-Speech (POS) substitutions on semantic expression. A Gold Standard is then established, which incorporates sentiments, topics, and Knowledge Graphs (KGs) for short text analysis, serving as a benchmark for comparison. The research revolves around fine-tuning the Mistral 7b Large Language Model (LLM) using a curated collection of poems. This fine-tuned model is then used to generate new poetry, with a focus on replicating the stylistic nuances of the original poets. Evaluation of the generated poems centers on coherence and semantic similarity to the source material. The study integrates Retrieval-Augmented Generation (RAG) through Weaviate, a vector database, to enhance the querying of poems related to specific events or themes. The concluding phase involves mining news articles and constructing knowledge graphs, drawing comparisons with poetic content to scrutinize how poetry reflects societal issues. Through these methodologies, this research highlights the capabilities of NLP in literary analysis and contributes to the field of computational creativity. The findings provide fresh insights into the intricate relationship between language, creativity, and societal discourse.

Keywords: Natural Language Processing (NLP), Computational Creativity, Poetry Analysis, Large Language Model (LLM), Mistral 7b, Part-of-Speech (POS) Substitutions, Gold Standard, Sentiment Analysis, Knowledge Graphs (KGs), Retrieval-Augmented Generation (RAG), Weaviate, Semantic Similarity, Coherence, Computational Literary Analysis, Societal Reflections in Poetry

Introduction

The exploration of the intersection between Natural Language Processing (NLP) and literature has become an increasingly significant area of research in computational linguistics. This paper presents a detailed investigation into this domain, focusing specifically on the application of advanced NLP techniques to the analysis and generation of poetry. The study is particularly centered on the works of William Blake and Robert Frost, two seminal figures in the world of poetry. William Blake, an English poet of the Romantic Era, is renowned for his symbolic and visionary poetry, often reflecting on themes of human nature and societal norms. Robert Frost, an American poet, is celebrated for his depictions of rural life and his exploration of complex social and philosophical themes. The significance of their work in the literary world provides a rich context for this NLP study, allowing for a deep exploration of linguistic complexity and thematic depth.

The primary objective of this project is to dissect the nuanced relationship between language and poetic form through a series of advanced computational tasks. These tasks include Parts of Speech (POS) tagging, semantic analysis, and the generation of new poetry through the fine-tuning of a Large Language Model (LLM). By applying these techniques to the works of Blake and Frost, the study aims to uncover the impact of linguistic elements on the semantic and emotional layers of poetry. A significant part of the research involves establishing a 'Gold Standard' for poetic analysis, derived from a comprehensive study of Pushcart Prize-nominated poems. This benchmark serves as a reference for evaluating the complexity and emotive power of poetry, guiding our analysis of the selected texts.

Another critical aspect of the research is the exploration of AI-assisted creativity in poetry. By fine-tuning the Mistral 7B model, an advanced LLM, the study ventures into the realm of computational creativity, aiming to generate new poetry that mirrors the style and thematic essence of Blake and Frost. This process not only challenges the capabilities of AI in emulating human creativity but also provides insights into the potential of AI in literary creation.

In addition to the generation of poetry, the study innovates by integrating Retrieval-Augmented Generation (RAG) with Weaviate, a vector database. This integration enhances the model's ability to produce contextually relevant and thematically consistent poetic content, highlighting a novel approach in AI-generated literature.

Overall, this paper details the methodologies employed in the study, discusses the findings, and explores the implications of these outcomes for the fields of NLP, literary analysis, and AI-driven creativity. Through this comprehensive exploration, the research aims to contribute to a greater understanding of the complex relationship between language and poetry and to showcase the potential of NLP in expanding the horizons of literary studies and creative writing.

Methodology

Machine Learning Model and Fine-Tuning for Poetic Style Mimicking

Model Selection and Setup:

The core machine learning model used in this project is the Mistral 7B, a large language model known for its efficacy in processing and generating natural language. This model was chosen for its advanced capabilities in handling nuanced language structures and for its potential in creative language generation. The model was configured for 4-bit quantization to optimize processing efficiency while maintaining performance integrity.

Data Sourcing and Preprocessing:

For William Blake's poetry, we sourced a comprehensive collection ranging from "Songs of Innocence and of Experience" to "The Marriage of Heaven and Hell." The poems were compiled into a single JSONL file, William_Blake.jsonl, ensuring a uniform format for efficient processing. The data scraping was executed using Python's BeautifulSoup library, targeting reputable online sources of Blake's works. The scraping process involved handling various HTML structures to extract and clean the poetic text, followed by preprocessing to standardize formats and remove extraneous characters.

Tokenization and Dataset Preparation:

The poems underwent a tokenization process using a custom tokenizer designed to align with the Mistral 7B model's architecture. This step was crucial for converting the raw text into a format suitable for model training, ensuring that each token retained its contextual meaning.

Model Fine-Tuning Process:

The fine-tuning of the Mistral 7B model was conducted on the tokenized dataset of Blake's poems. This process involved adjusting training parameters such as learning rate, batch size, and the number of epochs to align the model's output closely with Blake's poetic style. The training was carried out in a high-performance computing environment, utilizing GPU acceleration for efficient processing.

Experiments and Results:

Base Model Evaluation: Prior to fine-tuning, the base Mistral 7B model was assessed for its ability to generate poetry in the style of William Blake. This evaluation provided a benchmark for comparing the model's performance post-fine-tuning.

Chunking Strategy Comparison: We compared different chunking strategies, namely [poem] and [line] chunking, to determine their effectiveness in generating coherent and stylistically consistent language. The [poem] chunking approach was found to be more effective in maintaining thematic and stylistic continuity.

Cosine Similarity Analysis: Post-training, the generated poems were analyzed using cosine similarity metrics to evaluate their stylistic and thematic alignment with Blake's original works. This quantitative assessment confirmed the model's success in emulating Blake's style.

Key Findings:

- The fine-tuned model demonstrated a high fidelity to Blake's poetic style, effectively capturing both thematic depth and stylistic elements.
- The optimal chunking approach, [poem] chunking, proved superior in producing coherent and thematically consistent outputs.
- Quantitative validations via cosine similarity scores exhibited a strong correlation between the model-generated poems and Blake's originals, highlighting the model's nuanced understanding of Blake's work.
- Qualitative assessments corroborated the model's effectiveness, with human evaluators confirming the resemblance of the generated poems to Blake's style.

The fine-tuning of the Mistral 7B model for mimicking William Blake's poetic style marks a significant advancement in AI-driven creativity. This project demonstrates the capacity of advanced language models to generate new poetry that resonates with the thematic richness and stylistic nuances of historic poets. The success of this endeavor illustrates the potential of AI in creative literary fields, paving the way for future explorations in AI-assisted literary creation and analysis.

Experimentation and Results

Assignments:

1. POS Substitutions and Semantic Analysis in Poetry

Objective

The objective of this experiment is to investigate the effects of Parts of Speech (POS) substitutions in the poems of William Blake and Robert Frost, and use NLP techniques to understand how these linguistic changes affect the poems' semantic structure and reader interpretation.

Methodology

- Poet and Poem Selection:
    - Choose 20 representative poems from each poet.
    - Utilized Python's BeautifulSoup for precise scraping, ensuring the retention of poetic structure and format.
- Data Scraping: Utilizing Python's BeautifulSoup library, we scraped poems from websites like allpoetry.com. This process involved handling various HTML structures to extract clean poetic text.
- POS Tagging and Semantic Substitution Process:
    - Deployed spaCy for accurate POS tagging, crucial for identifying the linguistic structure of each poem.
    - Created an algorithm for semantic substitutions, carefully selecting POS from one poet's work to replace in another's while maintaining semantic integrity.

Data Structure and Storage

Each poem's data, including POS tags and the text, was meticulously organized into a JSON format. This structured data was then stored in a shared drive, facilitating easy access and manipulation for further analysis.

Experiments and Results

POS Substitution Analysis:

- In-depth substitutions were made, such as altering "worm" to "unknown worm" in Blake's "The Sick Rose," adding a layer of ambiguity.
- This substitution nuanced the poem’s theme of hidden corruption, subtly shifting its interpretative potential.

Semantic Similarity and Coherence Analysis:

- Utilized cosine similarity for quantitative assessment. For example, Blake's "The Sick Rose" displayed a marginal decrease in similarity post-substitution (from 0.3649 to 0.3477), indicating a slight shift in semantic content.
- This analysis helped quantify the impact of POS changes on the poems' overall semantic structure.

Advanced Text Summarization and Topic Modeling:

- Performed summarization to distill the essence of both original and transformed poems. The summaries revealed that while surface details changed, the core themes remained identifiable.
- Topic modeling via LDA and BERTopic highlighted consistent themes such as nature, love, and mortality, despite POS alterations.

Statistical Insights and Comparative Study:

- Compiled statistics offered a clear view of the linguistic changes. For instance, the change in the number of nouns and verbs in Frost's "Fire and Ice" from 11 to 10 and 10 to 9 respectively, subtly altered its rhythmic quality.
- Comparative analysis of original and transposed POS provided a clear lens to assess how specific changes impacted the poems' texture and tone.

Results and Analysis

We observed that POS substitutions had varied effects on the poems’ readability and thematic expression. While some substitutions preserved the original tone, others introduced intriguing new interpretations.

Conclusion

The in-depth analysis of Parts of Speech (POS) substitutions in the works of William Blake and Robert Frost using NLP techniques revealed significant insights into how subtle linguistic changes can profoundly affect poetic interpretation. For instance, modifying a single word like "worm" to "unknown worm" in Blake's "The Sick Rose" introduced a deeper level of ambiguity and complexity, demonstrating the powerful role of word choice in shaping a poem's meaning and emotional impact. This project highlighted not only the delicate interplay between language and literary expression but also the effectiveness of NLP in uncovering nuanced interpretations of classical poetry.

Furthermore, the quantitative assessment using cosine similarity provided a valuable metric for gauging the semantic shifts resulting from POS substitutions. Despite these linguistic alterations, the core themes of nature, mortality, and human experience remained strikingly resilient in both the original and transformed texts. This resilience, as evidenced through topic modeling and text summarization, underscores the enduring power of thematic elements in poetry. This experiment thus offered a compelling example of how NLP can be utilized to deepen our understanding and appreciation of literary art, reaffirming the intricate bond between language and poetic creativity.

2. Comparative Analysis with Gold Standard Texts

Introduction

The project aimed to critically analyze a range of poems against the benchmark of Pushcart Prize-nominated poems, using advanced Natural Language Processing (NLP) techniques. This involved a granular examination of POS distribution, thematic coherence, and emotional expression within the poems.

A. Statistical Analysis and POS Tagging

Detailed POS Distributions:

- Plots were created to illustrate the distributions of POS tags such as nouns (NN), verbs (VBG, VBN), adjectives (JJ), and adverbs (RB) across different poems.
- One notable pattern was the dominance of nouns (NN) and adjectives (JJ) in nominated poems. For example, in Christopher Hunter's "LOUISA, AGE 6, AT REST", nouns and adjectives accounted for a significant portion of the POS distribution.

Comparative POS Ratio Analysis:

- The analysis revealed varying POS ratios across poems. For instance, John Mitchell's "NICHOLAS II" displayed a high frequency of nouns (NN: 41), indicating a strong focus on subjects and entities.
- The POS ratios were then superimposed to identify stylistic trends common in nominated poems, such as a balanced use of verbs (VBG, VBN) and adjectives (JJ).

B. Topic Modeling

Analysis of Prize-Winning Poems:

- LDA revealed distinct thematic clusters in nominated poems. Themes like "nature's transience" and "personal reflection" were prominent.
- Specific topics identified included existential themes (Topic 1: "crash," "silver," "tiptoe") and introspective musings (Topic 2: "stay," "trying," "speed").

Comparative Topic Modeling:

- Topics in non-nominated poems, such as Nandini Sethi's "The Things I Never Said," were also analyzed. These poems often presented a wider thematic variety but with less thematic density compared to the nominated poems.

C. Sentiment Analysis

Gold Standard Sentiment Analysis:

- The Pushcart-nominated poems often showed a balanced sentiment profile. For example, "LOUISA, AGE 6, AT REST" by Christopher Hunter exhibited a neutral sentiment predominance ('neu': 0.952).

Comparative Sentiment Analysis:

- Non-nominated poems sometimes displayed more polarized sentiments. For instance, Dave Muddy's "Pride’s Wrath" had a complex sentiment distribution with notable positive ('pos': 0.342) and negative ('neg': 0.222) elements.

D. Comparative Ranking and Evaluation

Ranking System:

- Developed a cosine similarity-based ranking system to compare new poems to the nominated ones. This method identified thematic and stylistic resemblances.
- For example, "GOODWILL" by an unknown author was closely aligned with "THE BURNING OF THE WHALES" (Similarity Score: 1.00), indicating similar thematic and stylistic elements.

Data Scraping and POS Analysis:

- Comprehensive POS data was scraped and analyzed for each poem. This analysis provided insights into the linguistic characteristics that might align more closely with the gold standard. For instance, "Unseen disability" by Katanya Marks was rich in nouns (NN: 29) and adjectives (JJ: 9), resembling the structure seen in nominated poems.

Conclusions

This detailed analysis elucidates the complex interplay of linguistic, thematic, and emotional elements in poetry, particularly in relation to the standards set by Pushcart Prize nominations. Key findings suggest that nominated poems tend to exhibit a balanced mix of POS tags, coherent thematic expression, and a nuanced blend of sentiments. In contrast, non-nominated poems offer a broader spectrum of emotional and thematic diversity, sometimes deviating from the more nuanced balance observed in the nominated works. This investigation not only enhances our understanding of poetic compositions but also provides a framework for aspiring poets to gauge their work against recognized standards in contemporary poetry.

3. Language Model Fine-Tuning for Poetic Style Mimicking

Objective

The project's primary goal was to fine-tune a language model to emulate the poetic style of William Blake. By training the model on a curated collection of Blake's poems, the aim was to enable it to generate new poetry that aligns closely with Blake's thematic and stylistic characteristics.

Methodology

Poet Selection: William Blake was chosen for his distinctive poetic style. Data Collection: 30-40 of Blake's poems were scraped and compiled into a single JSONL file, William_Blake.jsonl.

Model Selection: The Mistral 7B model was chosen, and configured for 4-bit quantization to enhance processing efficiency.

Tokenization and Dataset Preparation: Poems were tokenized using a custom tokenizer designed for compatibility with the model's architecture.

Model Training: The Mistral 7B model underwent fine-tuning with the tokenized Blake poems dataset.

Experiments and Results

Base Model Evaluation: Before fine-tuning, the base model generated Blake-style poems with varying degrees of success, some capturing the essence of Blake's themes and style.

Chunking Strategy Comparison: The [poem] and [line] chunking options were evaluated to determine their effectiveness in generating coherent language. The [poem] chunking was found to produce more coherent and stylistically consistent outputs.

Cosine Similarity Analysis: Generated poems were compared with the original Blake poems using cosine similarity. High similarity scores, particularly with "Auguries of Innocence," indicated the model’s success in capturing Blake's style.

Generated Poems Analysis: Five poems were generated post-training, revealing a significant improvement in coherence and stylistic alignment with Blake's work.

Key Findings

High Fidelity to Blake’s Style: The fine-tuned model effectively emulated Blake's unique poetic style, both in thematic depth and stylistic elements.

Optimal Chunking Approach: The [poem] chunking method was found to be

superior in maintaining coherence and thematic continuity.
Quantitative Validation: The cosine similarity scores demonstrated a strong correlation between the model-generated poems and Blake's originals, particularly with "Auguries of Innocence," highlighting the model's nuanced understanding of Blake's work.
Qualitative Confirmation: Human assessment of the generated poems confirmed their resemblance to Blake's style, affirming the model’s effectiveness.

Conclusion

This project successfully demonstrates the capacity of advanced language models to mimic the poetic style of a specific author, in this case, William Blake. The fine-tuned model displayed a remarkable ability to generate new poetry that resonated with Blake's thematic richness and stylistic idiosyncrasies. This achievement underscores the potential of AI in creative literary endeavors, presenting innovative avenues for exploring and expanding upon the literary legacies of historic poets. The model's proficiency in capturing Blake's complex thematic elements and stylistic nuances paves the way for future explorations in AI-assisted literary creation and analysis.

4. Vector DB Integration and Retrieval-Augmented Generation (RAG)

Objective

This project focused on integrating Weaviate, a vector database, with a local Large Language Model (LLM), specifically targeting enhanced NLP tasks around the works and historical context of William Blake. It aimed to leverage the synergies between database-driven context retrieval and the generative capabilities of LLMs for nuanced natural language processing.

Methodology

Setup and Dependencies: Initiation involved installing necessary packages and importing dependencies, including HuggingFace's sentence transformer for embedding and the 'zephyr-7b-alpha-sharded' model for the LLM.
Weaviate Configuration: The embedding model 'sentence-transformers/all-mpnet-base-v2' was specified for Weaviate, with model arguments set for CUDA optimization, ensuring efficient processing. Data Ingestion into Weaviate: Over 30 poems by William Blake and multiple articles detailing significant historical events during his lifetime were ingested into Weaviate's vector database for semantic searches and contextual retrievals. Local LLM Setup: The LLM was loaded with 4-bit quantization to balance performance and resource usage. A custom tokenizer was initialized to align with the specific model requirements, including the setting of special tokens. Conversation Chain Implementation: Two approaches were employed for generating responses:

- Direct LLM Prompting: The model was queried directly with prompts related to Blake's life and works.
- RAG Augmentation: The Retrieval-Augmented Generation approach combined direct LLM responses with contextually relevant information pulled from Weaviate's database.

Experiments and Results

Direct LLM Responses: These responses, while informative, generally lacked depth in historical details and specific connections to Blake’s work. RAG-Enhanced Responses: This approach yielded more detailed and contextually rich responses, integrating specific historical events and their implications on Blake's poetry.

Comparison and Evaluation:

- Without RAG: The responses were broader, touching upon general themes in Blake's poetry but missing specific historical context.
- With RAG: Demonstrated a remarkable improvement in specificity and relevance, showcasing the efficiency of integrating vector database retrieval with LLMs.

Example

Query: "How is William Blake's 'Songs of Innocence and of Experience' related to 'The Slave Trade'?"

- Without RAG: Response generalized Blake's themes in poetry, veering away from the specific query.
- With RAG: Response specifically linked "The Little Black Boy" from Blake's collection to the anti-slavery campaign, highlighting Blake's stance against slavery and racism.

Key Findings

- Enhanced Contextual Understanding: RAG's integration with Weaviate significantly improved the LLM's ability to provide context-specific information, especially concerning historical events and literary analysis.
- Accuracy and Relevance: RAG-augmented responses were more aligned with the query’s intent, offering a nuanced understanding not achievable by the LLM alone.
- Efficiency in Information Retrieval: The combination of Weaviate's vector database and the LLM facilitated efficient information retrieval, showcasing the potential for complex query handling in literary and historical research.

Conclusion

The fusion of Weaviate's vector database capabilities with a locally hosted LLM creates a powerful tool for advanced NLP tasks. This hybrid system excels in providing detailed, contextually relevant responses, especially in fields requiring deep understanding, such as literature and history. The project underscores the potential of this integrated approach in enhancing the quality of responses for complex queries, offering a promising direction for future research and applications in AI-driven historical and literary analysis. The success in addressing queries related to William Blake demonstrates the model's applicability in educational and research settings, where accuracy and depth of context are paramount.

5. **Knowledge Graph Assembly and News Article Mining** Objective

The primary aim was to construct and analyze Knowledge Graphs (KG) derived from William Blake's poetry and related historical articles, focusing on their structure, content, and thematic interconnections.

Methodology

Data Preparation:

- Poems by William Blake and articles on pertinent historical events were selected.
- Texts were processed using Spacy's en_core_web_sm pipeline, facilitating natural language processing and analysis.

Knowledge Graph Construction:

- Employed Textacy for Subject-Verb-Object (SVO) extraction from Blake's texts and related articles.
- Generated nodes and relations for the KGs based on the extracted SVO triples.
- Created Directed Graphs (DiGraph) using NetworkX for both poem and article texts.
- Nodes represented entities or concepts, while edges represented their relational actions or interactions.

Graph Visualization:

- Utilized matplotlib and NetworkX to visualize the KGs, providing a graphical representation of the thematic and relational structures within the texts.
- Edge labels indicated the nature of the actions connecting the nodes.

Experiments and Results

KG for Poem:

- Nodes: Entities like 'I', 'joy', 'day', 'year', 'sit', 'hour', etc., were extracted from the poem.
- Relations: Actions like 'love', 'drives', 'spend', 'can_take', etc., were used to connect the nodes, depicting the poem's thematic and narrative structure. ● Visualization: The graph showed a complex web of interactions, illuminating the poem's thematic depth and structure.

KG for Article:

- Similar to the poem KG, but derived from a historical article, providing a visual representation of the article's thematic elements and their interconnections.

Topic Modeling:

- Applied LDA topic modeling to the articles, identifying major themes like land, tenants, acts, and Inclosure.
- These themes informed the construction of micro KGs, further dissecting the thematic elements.

Key Findings

Thematic Interconnection:

- The KGs revealed a deep interplay of themes and concepts within and between Blake's poetry and the historical context.
- The poem KG captured the emotional and philosophical nuances, while the article KG provided a factual, historical perspective.

Insightful Visualizations:

- The graphs provided an intuitive visual understanding of complex thematic structures, making abstract concepts more tangible.

Comparative Analysis:

- By comparing the poem and article KGs, similarities and differences in thematic elements and narrative structures were identified.
- This comparison allowed for a nuanced understanding of how Blake's poetry might reflect or diverge from the contemporaneous historical narrative.

Topic-Specific Graphs:

- The micro KGs based on LDA topic modeling offered focused insights into specific themes, adding depth to the overall analysis.

Conclusion

The construction and analysis of Knowledge Graphs from William Blake's poetry and related historical articles provided a novel approach to literary analysis. The KGs elucidated thematic interconnections, narrative structures, and the poet's engagement with contemporary events. This method proved valuable in uncovering the subtle interplay between literary works and their historical context, offering insights that traditional textual analysis might overlook. The visual aspect of KGs added an intuitive dimension to understanding complex literary and historical relationships. This project underscores the potential of KGs in enhancing literary studies, particularly in exploring the connections between literature and historical contexts.

These experiments collectively contribute to a comprehensive understanding of the application of NLP and AI in poetry analysis and creation. The results from each experiment offer unique insights into the capabilities of computational methods in literary studies, providing a multi-dimensional view of the project's objectives.

Discussion

The results of the NLP Poets Project offer intriguing insights into the intersection of Natural Language Processing and poetry. The experimentations, particularly the generation of new poems using a fine-tuned Mistral 7B model, highlight the significant potential of AI in creative literary endeavors. The comparison between the generated poems and William Blake's original work reveals that while the AI can mimic stylistic elements to a considerable degree, there are nuances in thematic depth and emotional resonance that remain uniquely human. This observation aligns with existing literature that suggests AI's capabilities in language generation are rapidly advancing, but the subtleties of human creativity and emotional expression are more challenging to replicate.

The Pushcart Probability Assessment experiment sheds light on the quantifiable aspects of poetry that contribute to literary recognition. The results indicate a gap between AI-generated works and award-winning poetry, suggesting that certain qualitative aspects might not yet be fully captured by computational methods. This finding is significant as it underscores the complex interplay between form, content, and emotional appeal in literature, a domain still largely governed by human judgment.

The NLP Statistical Analysis and Text File Analysis experiments provide a detailed view of linguistic patterns and structures, contributing to our understanding of poetic composition. The identification of a 'golden cluster' demonstrates that certain linguistic and thematic elements are more prevalent in recognized works, offering a potential roadmap for aspiring poets.

The integration of Knowledge Graphs and Vector Databases showcases the potential of these technologies in enhancing literary research. The ability to contextually link poems with historical articles presents a novel approach to studying literature, opening new avenues for interdisciplinary research.

Conclusion

The key findings of this research include:

- Advanced NLP and AI technologies can effectively emulate certain stylistic aspects of poetry, but the complete replication of the emotional depth and thematic nuances of human-created poetry remains challenging.
- Quantitative methods, such as the Pushcart Probability Assessment, can provide valuable insights into the characteristics of recognized poetry, although they may not fully capture the essence of literary excellence.
- NLP Statistical Analysis reveals discernible patterns in acclaimed poetry, offering valuable insights into the mechanics of poetic writing.
- The application of Knowledge Graphs and Vector Databases in literary studies represents an innovative approach, enhancing the depth and breadth of literary analysis.

From these findings, it is clear that the intersection of NLP and poetry offers fertile ground for further research. Future studies could explore the development of more advanced AI models capable of capturing the emotional and thematic depth of poetry. Additionally, there is scope for expanding the use of Knowledge Graphs and Vector Databases to include a broader range of literary works and historical contexts.

In conclusion, the NLP Poets Project not only demonstrates the current capabilities and limitations of AI in the realm of poetry but also opens up new possibilities for the use of NLP in literary studies. The project sets a foundation for future explorations in this evolving field, suggesting a future where technology and literature continue to intersect in novel and meaningful ways.

References

[1] .https://arxiv.org/pdf/1708.03310.pdf

[2].https://www.researchgate.net/publication/313874773_NLP_based_Poetry_Analysis_and_Generation

[3]. https://www.jsr.org/hs/index.php/path/article/download/4418/2107/28475

[4].https://www.researchgate.net/publication/348563721_Knowledge_Graphs_and_Natural-Language_Processing

[5].https://medium.com/@ashminipw27/the-beginners-guide-to-vector-databases and-natural-language-models-a04c4f253d71