Northeastern University, Khoury College of Computer Science

CS 6120 Natural Language Processing
Final Project Template
Due: April 24, 2025(100 points)

Qinyuan Shen
PROJECT GIT REPOSITORY

1 Executive Summary and Abstract

This project analyzes thematic and linguistic evolution between Arthur Conan Doyle's original Sherlock Holmes stories (1887-1927) and CBS's Elementary (2012-2019) using modern NLP techniques. By combining semantic search with ChromaDB, topic modeling via BERTopic, and Phi-3-mini-4k-instruct for context-aware Q&A, we quantify how 130 years of cultural shifts reshape detective fiction while preserving core narrative DNA.

2 Background and Related Work

This project leverages state-of-the-art NLP technologies and methodological advancements to analyze thematic and stylistic differences across Sherlock Holmes adaptations. Below are key components of the technical framework:

1. BERTopic for Thematic Analysis The project employs **BERTopic** Grootendorst [2022], a topic modeling framework that leverages transformer-based embeddings. Unlike traditional LDA, BERTopic integrates SBERT embeddings to capture semantic nuances, enabling cross-corpus comparisons between 19^{th}-century novels and modern scripts. Its hierarchical clustering and dynamic topic reduction capabilities (configured to 20 topics) ensure interpretable themes, such as distinguishing *forensic science* in *Elementary* from Doyle's *Victorian morality*.

2. Sentence-BERT for Semantic Embeddings **Sentence-BERT** Reimers and Gurevych [2019] generates domain-invariant embeddings using the `all-mpnet-base-v2` model. This approach ensures robust cross-domain semantic similarity calculations between Doyle's prose and modern dialogue, addressing challenges like lexical shifts (e.g., *telegram* vs. *text message*).

3. UMAP and HDBSCAN for Clustering **UMAP** McInnes et al. [2018] reduces embedding dimensions while preserving global semantic structure (`n_components=5`, `cosine` metric). **HDBSCAN** McInnes et al. [2017] identifies dense topic clusters (`min_cluster_size=15`), avoiding rigid assumptions about cluster count. This combination optimizes topic coherence in heterogeneous corpora.

4. spaCy for Text Preprocessing The pipeline uses **spaCy** Explosion AI [2023] for speaker-tag extraction, tokenization, and noise removal (e.g., stage directions). Rule-based matching ensures Sherlock-specific dialogue isolation, critical for character-centric analysis.

5. Scikit-learn and Chroma for Feature Engineering Scikit-learn's **CountVectorizer** Pedregosa [2011] extracts bi-grams with stopword filtering (`min_df=5`), capturing stylistic markers like Holmes' signature deductive phrasing (*My dear Watson...*). **ChromaDB** Chroma [2023] manages embeddings and metadata, enabling efficient retrieval of novel-script parallels.

3 Methodology

Decoding the Detective leverages modern NLP techniques to analyze thematic and stylistic fidelity between Arthur Conan Doyle's original Sherlock Holmes novels and CBS's *Elementary* TV adaptation. The system combines three core capabilities:

a) **Semantic Search Engine**: Built using Chroma vector databases Chroma [2023] and Sentence-BERT embeddings Reimers and Gurevych [2019], enabling retrieval of semantically similar passages across:
   - 9 original novels (Project Gutenberg)
   - 154 *Elementary* episodes (fan-curated transcripts)

b) **Adaptation Fidelity Metrics**:
   - BERTopic modeling Grootendorst [2022] for thematic distribution comparison
   - spaCy-driven linguistic analysis Explosion AI [2023] of language evolution
   - Network analysis of character dialogue dynamics

c) **RAG-Powered Q&A System**:
   - Microsoft's Phi-3-mini-4k-instruct model
   - Answer grounding in retrieved passages
   - Confidence checks to prevent hallucination

Repository: https://github.com/mahamayashen/Decoding-the-Detective
Live Demo: Dockerized Streamlit interface with pre-built vector indices

4 Data and Data Analysis

4.1 Data Source(s)

Original Sherlock Holmes Novels

A Study in Scarlet (1887) Doyle [1887]

The Sign of the Four (1890) Doyle [1890]

The Adventures of Sherlock Holmes (1892) Doyle [1892]

The Memoirs of Sherlock Holmes (1894) Doyle [1894]

The Hound of the Baskervilles (1902) Doyle [1902]

The Return of Sherlock Holmes (1905) Doyle [1905]

The Valley of Fear (1915) Doyle [1915]

His Last Bow (1917) Doyle [1917]

The Case-Book of Sherlock Holmes (1927) Doyle [1927]

Modern Adaptation Scripts

Elementary (CBS 2012-2019) 154 episodes retrieved from Forever Dreaming [2012-2019] under Fair Use doctrine

4.2 Data Analysis and Exploration

Corpus Statistics Analysis

Basic Corpus Statistics

Metric	Novels	Scripts
Total Chunks	3,972	12,077
Avg. Chunk Length	935.47	458.46
Victorian Terms	16,462	24,538
Modern Terms	1,884	5,902

Key Entity Analysis

Character Mentions (PERSON)

Novels		Scripts	
Holmes	3,052	Watson	1,236
Watson	1,047	Holmes	793
Lestrade	260	Joan	304
McMurdo	207	Gregson	204
Mycroft	37	Bell	138

Locations (GPE)

Novels		Scripts	
London	368	New York	377
England	139	London	170
America	47	U.S.	60
Chicago	41	Marcus	294
India	27	Queens	61

Organizations (ORG)

Novels		Scripts	
Scotland Yard	59	NYPD	198
Gregson	38	FBI	159
McGinty	31	Bell	199
Agra	22	CSU	86
Times	17	DEA	30

Temporal References (DATE)

Novels		Scripts	
yesterday	93	today	346
years	47	yesterday	293
Monday	46	last week	80
a week	34	months	57
two days	28	years	95

Cultural Term Analysis

Category	Novels	Scripts	Difference
Victorian Institutions	59 (Scotland Yard)	198 (NYPD)	+236%
Forensic Terms	0	112	+%
Deductive Markers	164 (”therefore”)	47 (”DNA”)	-71%

Key Findings

- **Modernization Paradox**: Scripts use **49% more Victorian terms** absolutely but have **21% lower relative ratio** compared to modern terms
- **Forensic Shift**: Complete transition from deductive markers (”therefore”) to technical terms (”DNA”)
- **Character Dynamics**: Watson's mentions increase from **1:3 ratio** (novels) to **1:1.5** in scripts

5 Results and Evaluation

5.1 Model Configuration

Component	Parameters
Embedding Model	all-MiniLM-L6-v2 (384D)
UMAP	n_neighbors=15, n_components=5
HDBSCAN	min_cluster_size=15, min_samples=5
Vectorizer	ngram_range=(1,2), min_df=5
Topic Reduction	nr_topics=20

5.2 Top 5 Topics

Topic	Count	Key Terms	Representative Document Excerpt
-1	8,518	holmes, hes, little, night	*"Sherlock Holmes," said Peters... we found in the Brixton Workhouse Infirmary...*
0	3,900	hes, klled, mrder, ive	*"Instead, Tim Bledsoe gets shot and stuffed in a wall... why not call the cops?"*
1	1,833	holmes, sir, face, little	*"There are forces here which may be more dangerous than those he has escaped..."*
2	446	father, ive, oh, addict	*"Sorry, but I can't let that slide. Not after everything I've been through..."*
3	310	patients, dr, eric, hospital	*"They're all dead now. We believe they were all... m*rder*d..."*

5.3 Notable Topics

Topic	Theme	Representative Terms
7	Environmental Crisis	collapse, bees, trees, silk, worm
13	Species Conservation	animals, species, extinct, truck, birth
12	Technology	bella, computer, virus, program, AI
5	Historical Mysteries	map, book, treasure, seal, unger

5.4 evaluation

6 Evaluation

6.1 Implementation-Specific Metrics

- **Retrieval Precision**
  - Code proof:

```
# app.py
results = chroma_search(query, collection,
                        corpus="both", top_k=3)
st.write(f"Score: {result['score']:.4f}")
```

- **Answer Validation**
  - 3 forbidden terms enforced via regex:

```
# answer_generation.py
forbidden_terms = ["BBC", "Benedict Cumberbatch",
                   "Sherlock BBC"]
```

6. **Answer Faithfulness**:
   - *Metric*: % of claims directly supported by retrieved contexts.
   - *Why*: Prevents hallucination by grounding answers in evidence.
   - *Code*:

```
# answer_generation.py prompt structure
prompt = "Base answers mainly on these contexts:\n{context_chunks}"
```

7. **Source Attribution Accuracy**:
   - *Metric*: Correct novel/episode identification from metadata.
   - *Why*: Critical for literary analysis validity.
   - *Code*:

     ```
     # Metadata handling (app.py)
     st.write(f"Novel: {metadata.get('novel', 'Unknown')}")
     ```

6.2 Evaluation: Contextual Accuracy

To assess the system's ability to ground answers in textual evidence, we tested the query: *"In which novel does Sherlock Holmes mention his cocaine use?"*
Answer: Sherlock Holmes' cocaine use is famously mentioned in *the_memoirs_of_sherlock_holmes*, a collection of short stories by Sir Arthur Conan Doyle. In these tales, Dr. Watson observes that Holmes occasionally turns to a "seven-percent solution" of cocaine to alleviate boredom during periods of inactivity, when challenging cases are scarce. Watson describes this habit as one of Holmes' few vices, contrasting it with his otherwise austere and disciplined lifestyle.
Top Retrieved Passage:

```
Result 1 | Source: Novel
Score: 0.6915
Novel: the_memoirs_of_sherlock_holmes
". Sherlock Holmes was a man who seldom took exercise for exercise's sake. Few m
```

Analysis:

- **Strengths**:
  - The system correctly identified both the source novel and the thematic context (cocaine as a response to boredom).
  - Semantic search retrieved a passage containing the key phrases "occasional use of cocaine" and "monotony of existence," aligning with the answer's claims.
- **Limitations**:
  - **Knowledge Gap bridging**: The missing "seven-percent solution" detail in retrieved text reveals a *context chasm*—the system patches canonical knowledge without explicitly flagging inferences. Future versions could:
    - Integrate human feedback loops for annotating such gaps
    - Add " Prior Knowledge" tags when answers exceed retrieved evidence
  - **Threshold Roulette**: The 0.6915 similarity score sits in a confidence gray zone. Without:
    - Dynamic thresholds
    - User-tunable sliders ("Strict vs. Creative" modes)

    we risk either over-filtering niche references or permitting hallucination creep.

- **Silent Hallucinations**: While the core answer was correct, the system could *over-confidently* invent details (e.g., wrongly specifying cocaine concentration as 5%). Current architecture lacks:
  - A "Flag Uncertain" button for questionable claims
  - Confidence intervals per factual assertion (e.g., "London mention: 95% certainty vs. dosage: 60%")

This case demonstrates the system's ability to surface **thematically relevant evidence** even when exact phrasal matches (e.g., "seven-percent") are absent. However, it highlights the need for clearer confidence calibration when distinguishing between *retrieved knowledge* and *model prior knowledge* in hybrid QA systems.

7 Conclusions

So, What Did We Learn? Turns out, Sherlock Holmes is the ultimate shapeshifter of literature. By applying NLP to Doyle's original stories and *Elementary* scripts, three key insights emerged:

- **Culture Leaves Fingerprints**: While *Elementary* uses 49% more Victorian terms than the originals (surprise!), they're overshadowed by modern tech like DNA analysis. Think antique furniture in a smart home—old vibes, new tools.
- **Watson's Glow-Up**: The novels framed Watson as Sherlock's sidekick. Modern scripts? Partners-in-crime-solving.
- **Core DNA Survives**: Despite surface changes (NYPD replacing Scotland Yard), Sherlock's essence—obsession with puzzles, moral gray areas—remained intact.

What's Next? With more time (and GPU power):

- **Theme Timelines**: Map topic evolution episode-by-episode rather than across centuries
- **AI Creativity**: Fine-tune models to rewrite Victorian mysteries as Gen-Z social media dramas

The methodology could extend beyond detective fiction—researchers already use similar approaches to track shifts in war reporting or slang evolution. Our toolkit might crack those cases too.

References

Chroma. Chroma: Ai-native open-source embedding database, 2023. URL https://docs.trychroma.com/.

Arthur Conan Doyle. *A Study in Scarlet*. Project Gutenberg, 1887. URL https://www.gutenberg.org/files/244/244-0.txt.

Arthur Conan Doyle. *The Sign of the Four*. Project Gutenberg, 1890. URL https://www.gutenberg.org/files/2097/2097-0.txt.

Arthur Conan Doyle. *The Adventures of Sherlock Holmes*. Project Gutenberg, 1892. URL https://www.gutenberg.org/files/1661/1661-0.txt.

Arthur Conan Doyle. *The Memoirs of Sherlock Holmes*. Project Gutenberg, 1894. URL https://www.gutenberg.org/files/834/834-0.txt.

Arthur Conan Doyle. *The Hound of the Baskervilles*. Project Gutenberg, 1902. URL https://www.gutenberg.org/files/2852/2852-0.txt.

Arthur Conan Doyle. *The Return of Sherlock Holmes*. Project Gutenberg, 1905. URL https://www.gutenberg.org/files/108/108-0.txt.

Arthur Conan Doyle. *The Valley of Fear*. Project Gutenberg, 1915. URL https://www.gutenberg.org/files/3289/3289-0.txt.

Arthur Conan Doyle. *His Last Bow*. Project Gutenberg, 1917. URL https://www.gutenberg.org/files/2350/2350-0.txt.

Arthur Conan Doyle. *The Case-Book of Sherlock Holmes*. Project Gutenberg, 1927. URL https://www.gutenberg.org/files/69700/69700-0.txt.

Explosion AI. spaCy: Industrial-strength nlp in python, 2023. URL https://spacy.io/.

Forever Dreaming. Elementary (cbs) episode transcripts, 2012-2019. URL https://transcripts.foreverdreaming.org/viewforum.php?f=12. 154 episodes retrieved under Fair Use.

Maarten Grootendorst. Bertopic: Leveraging bert embeddings for topic modeling. *arXiv preprint arXiv:2203.05794*, 2022.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Fabian et al. Pedregosa. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.