Semantic Search in Articles Using NLP

Table of Contents:

1. Introduction
2. Project Overview and Goals
    - Path 1: Traditional Search with Semantic Enhancements
    - Path 2: Word2Vec-Based Semantic Search
    - Path 3: Transformer-Powered Semantic Search
3. Data Description
    - Dataset Characteristics
    - Data Utilization
4. Baseline experiment
    - Traditional Search with Semantic Enhancements
5. Second Experiment
    - Word2Vec-Based Semantic Search
6. Third Experiment
    - Transformer-Powered Semantic Search
7. Challenges
8. Overall Conclusion

Introduction:

The project sets out to revolutionize search capabilities through the strategic implementation of Natural Language Processing (NLP) techniques. Traditional search engines rely on lexical matches, whereas semantic search transcends this limitation by comprehending the content of search queries, including synonym identification. The primary goal is to implement semantic search within articles and concurrently extract significant keywords, contributing to more precise and context-aware information retrieval.

Project Overview and Goals:

The project unfolds in three distinct paths, each constituting a unique step towards crafting a fully-fledged semantic search system. It is crucial to emphasize that these paths are not hierarchical; rather, each path represents a standalone approach, progressively moving from traditional search methods to more advanced semantic understanding techniques.

Path 1: Traditional Search with Semantic Enhancements

The first path involves augmenting traditional search methods with some level of semantic understanding. This might include incorporating synonym recognition and contextual analysis into the search algorithm while still relying on conventional keyword matching. This serves as a foundational step toward more sophisticated semantic search.

Path 2: Word2Vec-Based Semantic Search

In the second path, the project aims to implement a fully semantic search using Word2Vec embeddings. This involves training or utilizing pre-trained Word2Vec models to represent words in a semantically meaningful way. The search algorithm will leverage these embeddings to capture semantic relationships and enhance the accuracy of search results.

Path 3: Transformer-Powered Semantic Search

The third and most advanced path explores the use of powerful transformer models, for semantic search. Transformers have demonstrated remarkable capabilities in capturing complex semantic structures and context. The project will investigate integrating transformer architectures into the search process to achieve a highly sophisticated and context-aware semantic search system.

Data Description:

The dataset used for this project is sourced from the IMDb movies dataset, containing information related to 10,055 movies. The dataset comprises the following columns:

1.Movie Title:

- Description: Title of the movie.

- Type: Categorical.

2. Year:

- Description: The year in which the movie was released.

- Type: Numerical (Discrete).

3. Genre:

- Description: The style or category of the movie.

-Type: Categorical

4. Synopsis:

- Description: A concise summary outlining the basic plot of the movie as available in IMDb.

- Type: Textual.

5. Cast:

- Description: List of individuals involved in the movie.

- Type: Categorical.

Dataset Characteristics:

- Number of Instances (Rows): 10,055

- Number of Features (Columns): 5

Data Utilization:

The IMDb movies dataset, with its diverse movie-related information, provides a rich source for conducting semantic search and keyword extraction. Leveraging the "Synopsis" column as English articles for search purposes, coupled with careful cleaning and preprocessing, ensures that the data is optimized for semantic analysis. This comprehensive approach enhances the dataset's utility in achieving the project's goals of semantic search and extracting relevant keywords from movie synopses.

Baseline experiment:

Traditional Search with Semantic Enhancements:

Goal:

Build a comprehensive foundation for semantic search by incorporating enhancements in vocabulary representation, feature extraction, and semantic modeling.

Steps:

1. Word Dictionary Building:

- Utilize Gensim: Leverage the Gensim library to create a word dictionary, assigning unique IDs to each word, and recording their frequency counts.

2. Bag of Words Technique:

- Apply Bag of Words Model: Implement the doc2bow method to iterate through words in the text, updating frequency counts for existing words and introducing new words into the corpus.

3. Tf-Idf and LSI Model Building:

- Build Tf-Idf Models: Create Tf-Idf Models: Utilize the Gensim Tf-idf Model constructor to build Tf-Idf models, evaluating word importance within each document. This model is created based on the Bag of Words (BoW) representation of my documents and the dictionary built in the previous step.

- Pass to LSI Model: Feed the Tf-Idf models into the Latent Semantic Indexing (LSI) model, specifying the desired number of features.

4. Create Similarity Index for LSI Corpus:

- Utilize Matrix Similarity: Implement Matrix Similarity from Gensim to create a similarity index for the LSI corpus.

- Specify Features: Set the number of features for the similarity index, ensuring consistency with the number specified for the LSI model.

5. Apply Semantic Search and Extract Hot Keywords:

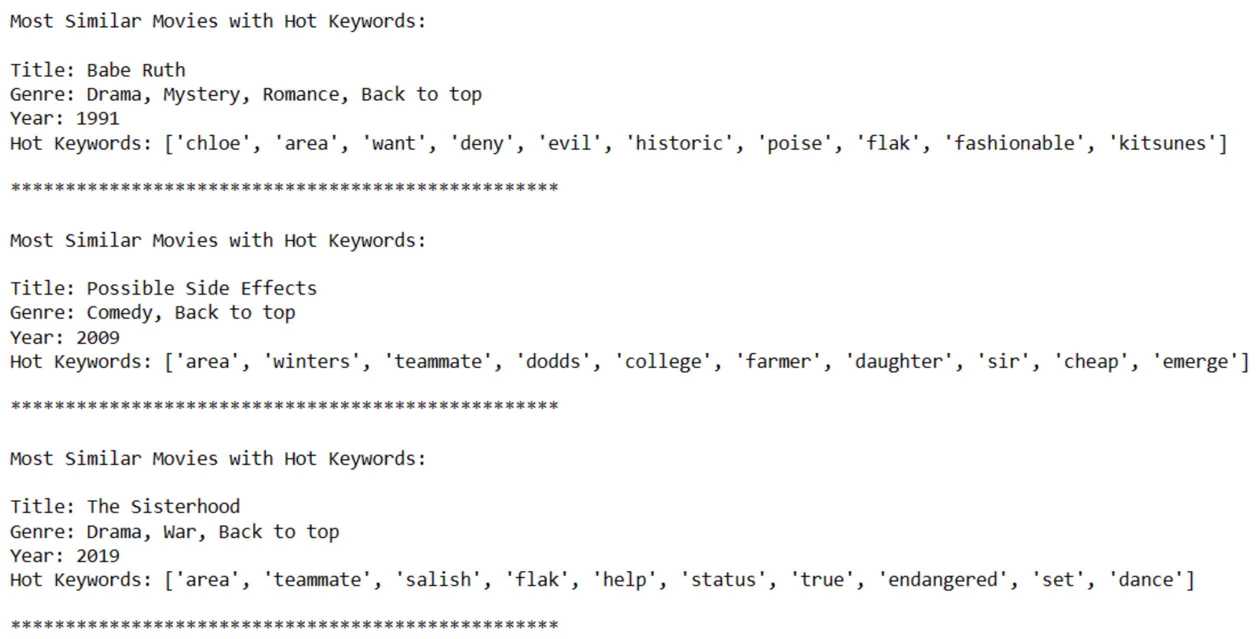
- Use Python Script: Utilize the Python script named "basic_search" for applying semantic search and extracting hot keywords.

- semantic search: Leverage the semantic enhancements developed through the word dictionary, bag of words, Tf-Idf, and LSI model to perform more accurate and context-aware searches.

- Extract Keywords: Implement keyword extraction methods within the script to identify and highlight significant terms.

Results:

The culmination of the implemented steps is reflected in the search output and extracted keywords. The search results:



```
Most Similar Movies with Hot Keywords:

Title: Babe Ruth
Genre: Drama, Mystery, Romance, Back to top
Year: 1991
Hot Keywords: ['chloe', 'area', 'want', 'deny', 'evil', 'historic', 'poise', 'flak', 'fashionable', 'kitsunes']

**************************************************

Most Similar Movies with Hot Keywords:

Title: Possible Side Effects
Genre: Comedy, Back to top
Year: 2009
Hot Keywords: ['area', 'winters', 'teammate', 'dodds', 'college', 'farmer', 'daughter', 'sir', 'cheap', 'emerge']

**************************************************

Most Similar Movies with Hot Keywords:

Title: The Sisterhood
Genre: Drama, War, Back to top
Year: 2019
Hot Keywords: ['area', 'teammate', 'salish', 'flak', 'help', 'status', 'true', 'endangered', 'set', 'dance']

**************************************************
```

Baseline Experiment Conclusion:

The comprehensive approach, including vocabulary representation, feature extraction, semantic modeling, similarity indexing, and application of semantic search and keyword extraction using the "basic_search" script, establishes a solid foundation for semantic search capabilities. This holistic framework ensures a structured and optimized system ready for subsequent experiments and advancements in semantic search.

Second Experiment:

Word2Vec-Based Semantic Search

Goal: I aimed to develop a fully-fledged semantic search using Word2Vec.

Steps:

1. Word2Vec Model Training:

- Trained the tokenized movie synopses on a Word2Vec model.

2. Semantic and General Search with Movie Embeddings:

- Checked if a specific movie is in our dataset.

  - If the movie is present:

  - Obtained embeddings for both its synopsis and the name.

  - Calculated the similarity between the synopsis embeddings and the embeddings of the rest of the movie synopses.

  - Enriched the semantic search process by considering the similarity between movie synopses.

  - Ranked the movies based on their similarity scores.

  - Selected the most similar movies as the search results.

  - If the movie is not present:

  - Obtained embeddings for the name of the movie.

  - Compared these embeddings with the embeddings of the synopses of other movies in the dataset.

  - Enabled a more general search by utilizing the semantic representation of the movie name.

  - Ranked the movies based on their similarity scores.

  - Selected the most similar movies as the search results.

3. Extracting Hot Keywords:

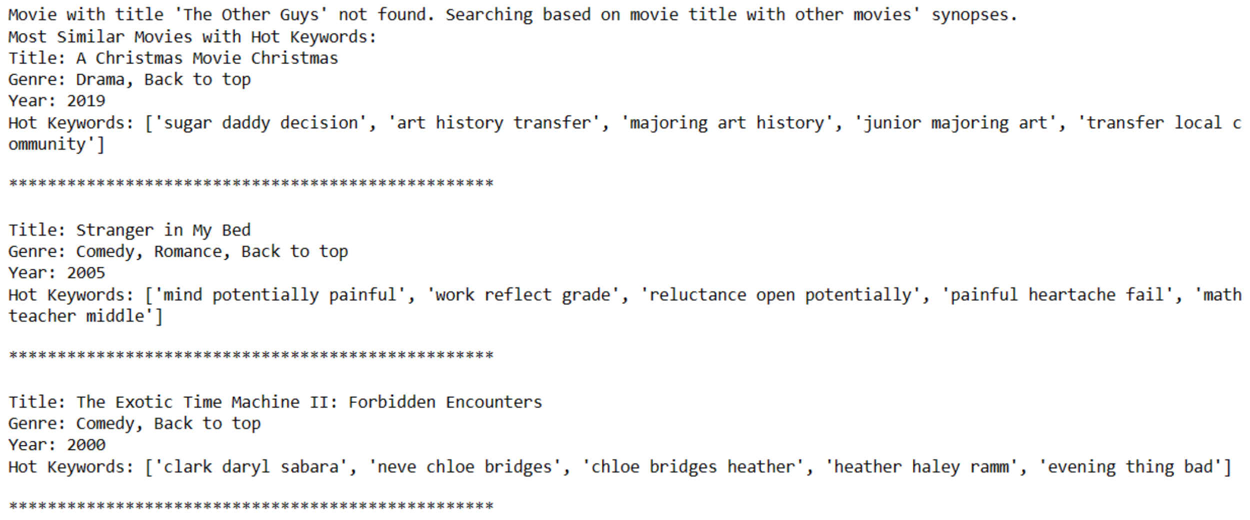
- Implemented two methods for extracting hot keywords:

  - First method: Calculated the embedding for each word in the synopsis, measuring similarity to the mean embedding of other words, and selecting highly similar words.

  - Second method: Utilized the YAKE (Yet Another Keyword Extractor) algorithm for automated extraction of key phrases and keywords from the movie synopsis.

- Enhanced the variety and coverage of extracted keywords, contributing to a more comprehensive understanding of movie content

Results: The culmination of the implemented steps is reflected in the search output and extracted keywords. The search results:



```
Movie with title 'The Other Guys' not found. Searching based on movie title with other movies' synopses.
Most Similar Movies with Hot Keywords:
Title: A Christmas Movie Christmas
Genre: Drama, Back to top
Year: 2019
Hot Keywords: ['sugar daddy decision', 'art history transfer', 'majoring art history', 'junior majoring art', 'transfer local c
ommunity']

**************************************************

Title: Stranger in My Bed
Genre: Comedy, Romance, Back to top
Year: 2005
Hot Keywords: ['mind potentially painful', 'work reflect grade', 'reluctance open potentially', 'painful heartache fail', 'math
teacher middle']

**************************************************

Title: The Exotic Time Machine II: Forbidden Encounters
Genre: Comedy, Back to top
Year: 2000
Hot Keywords: ['clark daryl sabara', 'neve chloe bridges', 'chloe bridges heather', 'heather haley ramm', 'evening thing bad']

**************************************************
```

Second Experiment Conclusion:

In this experiment, the integration of Word2Vec embeddings for semantic and general search significantly advances search capabilities. By considering movie embeddings and incorporating ranking based on semantic similarity, the search results are refined for both specific and general queries. Additionally, the implementation of advanced keyword extraction methods contributes to a nuanced understanding of movie content. This experiment establishes the foundation for a more sophisticated and effective semantic search system for movie synopses.

Third Experiment:

Transformer-Powered Semantic Search

Goal:

To leverage Transformer-based embeddings for an advanced semantic search system, enriching the search process and extending its capabilities beyond traditional methods.

Steps:

1. Semantic Search with Transformer Embeddings:

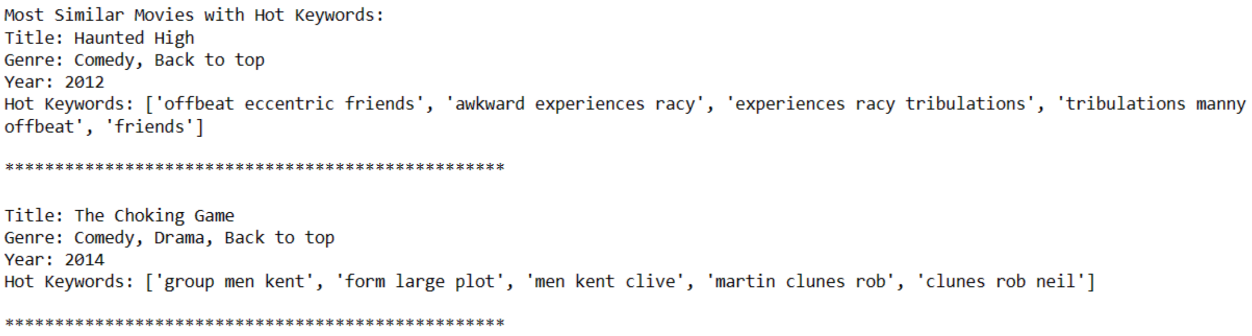
- Checked if a specific movie is in our dataset.
  - If the movie is present:
    - Obtained Transformer embeddings for its synopsis.
    - Calculated the similarity between the Transformer embeddings and the embeddings of the rest of the movie synopses.
    - Enriched the semantic search process by considering the similarity between movie synopses.
  - If the movie is not present:
    - Obtained embeddings for the name of the movie.
    - Compared these embeddings with the embeddings of the synopses of other movies in the dataset.
    - Enabled a more general search by utilizing the semantic representation of the movie name.

2. Extracting Hot Keywords:

- Developed two methods for extracting hot keywords:
  - First method: Calculated the embedding for each word in the synopsis, measuring similarity to the mean embedding of other words, and selecting highly similar words.
  - Second method: Utilized the YAKE (Yet Another Keyword Extractor) algorithm for automated extraction of key phrases and keywords from the movie synopsis.
- Enhanced the variety and coverage of extracted keywords, contributing to a more comprehensive understanding of movie content.

Results:

The culmination of the implemented steps is reflected in the search output and extracted keywords. The search results:



```
Most Similar Movies with Hot Keywords:
Title: Haunted High
Genre: Comedy, Back to top
Year: 2012
Hot Keywords: ['offbeat eccentric friends', 'awkward experiences racy', 'experiences racy tribulations', 'tribulations manny
offbeat', 'friends']

**************************************************

Title: The Choking Game
Genre: Comedy, Drama, Back to top
Year: 2014
Hot Keywords: ['group men kent', 'form large plot', 'men kent clive', 'martin clunes rob', 'clunes rob neil']

**************************************************
```

Third Experiment Conclusion:

The implementation of Transformer-powered embeddings in semantic search demonstrates an advanced and effective approach to information retrieval. The refined search results, combined with sophisticated keyword extraction methods, contribute to a more nuanced understanding of movie synopses. This experiment establishes the Transformer-powered semantic search system as a powerful tool for extracting meaningful information from movie datasets, showcasing its potential for applications in diverse domains.

Challenges:

This project, conducted in Jupyter Notebook, not only honed technical skills in natural language processing but also highlighted the iterative and evolving nature of developing sophisticated semantic search systems. Challenges, such as combining semantic and traditional approaches and refining keyword extraction, underscored the importance of adaptability and continuous refinement. These experiences contributed to a deeper understanding of information retrieval methodologies.

Overall Conclusion:

The journey from the initial experiment to the advanced Transformer-powered semantic search system marks a significant progression in the development of a robust semantic search system.

1. Baseline Experiment:

In this initial phase, traditional search methods were augmented with semantic enhancements.

2. Word2Vec Semantic Search:

- The second experiment introduced Word2Vec embeddings, enhancing both specific and general search capabilities. Keyword extraction methods were refined to provide a more comprehensive understanding of movie content.

3. Transformer-Powered Semantic Search:

- The third experiment marked a substantial leap with the incorporation of Transformer-based embeddings, elevating semantic search to new heights. This advanced approach not only refined the search process but also extended its capabilities, showcasing the potential for sophisticated information retrieval.

Throughout these experiments, the goal was to progress from traditional methods to more advanced techniques, culminating in a Transformer-powered semantic search system. The stepwise evolution demonstrates the system's adaptability and effectiveness, offering users refined search results and enriched keyword extraction for a more comprehensive and context-aware information retrieval experience. This overall journey underscores the continual enhancement and sophistication achieved in the pursuit of an optimal semantic search solution.