Shangzhi LOU
Bastien CHEREL

TRIP ADVISOR RECOMMENDATION CHALLENGE
Beating BM25

Objective

The project aims to **outperform the BM25 algorithm** by applying advanced **Natural Language Processing (NLP) techniques** and knowledge acquired during the course. Using the Kaggle TripAdvisor review dataset, the goal is to build a robust semantic ranking model that leverages embeddings from SentenceTransformer (all-mpnet-base-v2) to capture deeper contextual meanings.

Through preprocessing, tokenization, embedding generation, and similarity computation, the project showcases how modern NLP tools and methods can deliver **more accurate, contextually relevant recommendations** compared to traditional keyword-based retrieval approaches. The performance comparison is quantified using **Mean Squared Error (MSE)** between the rankings produced by the two models.

Pipeline

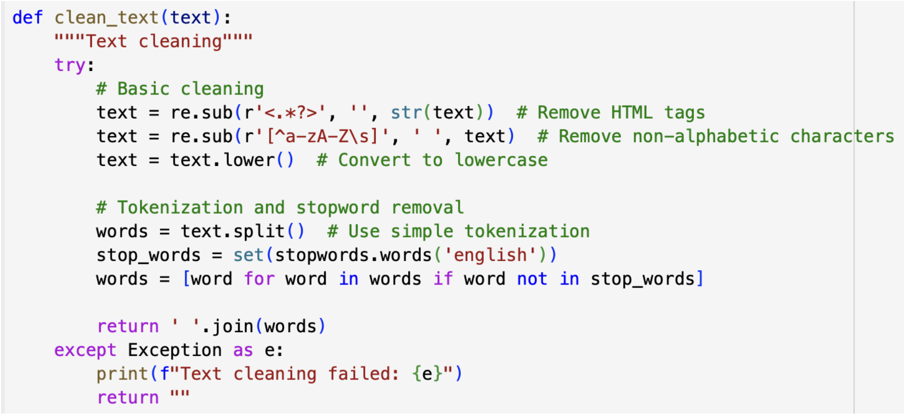
1. Dataset: Kaggle TripAdvisor Reviews

The project utilizes the Kaggle TripAdvisor review dataset, which contains a large collection of user reviews and their associated ratings. However, due to computational constraints, training the semantic model on the full dataset would take approximately four hours on a T4 GPU in Google Colab.

The dataset was downsized to 100 samples. This downsizing enables quicker iteration and testing while still providing a meaningful comparison between the semantic model and the BM25 algorithm. The smaller dataset retains the core features required for evaluation, allowing the project to focus on demonstrating the effectiveness of the techniques rather than being bottlenecked by processing time.

2. Preprocessing Steps

In the preprocessing step, we focused on cleaning the reviews to make them suitable for analysis. This involved removing unnecessary elements like, special characters, and stopwords, as well as converting the text to lowercase to ensure consistency across the dataset. We then tokenized the cleaned text into individual words to prepare it for techniques like BM25 and embedding generation. Finally, we saved the processed data to a file, allowing us to reuse it without repeating the cleaning process, which saved time during multiple experiments.



```python
def clean_text(text):
    """Text cleaning"""
    try:
        # Basic cleaning
        text = re.sub(r'<.*?>', '', str(text))  # Remove HTML tags
        text = re.sub(r'[^a-zA-Z\s]', ' ', text)  # Remove non-alphabetic characters
        text = text.lower()  # Convert to lowercase

        # Tokenization and stopword removal
        words = text.split()  # Use simple tokenization
        stop_words = set(stopwords.words('english'))
        words = [word for word in words if word not in stop_words]

        return ' '.join(words)
    except Exception as e:
        print(f"Text cleaning failed: {e}")
        return ""
```

3. BM25 vs. Our Model

BM25 is a keyword-based retrieval method that ranks reviews by calculating term frequency-inverse document frequency (TF-IDF) metrics. It works by measuring how often query terms appear in a document while considering their importance in the overall dataset. BM25 relies heavily on exact term matches and struggles with synonyms, paraphrases, or deeper contextual understanding.

In contrast, our semantic similarity model uses embeddings generated by the all-mpnet-base-v2 transformer to represent the meaning of reviews as dense numerical vectors. Relevance is determined by computing cosine similarity between the query embedding and the embeddings of the reviews, allowing the model to capture context, synonyms, and nuanced language patterns.

4. Evaluation

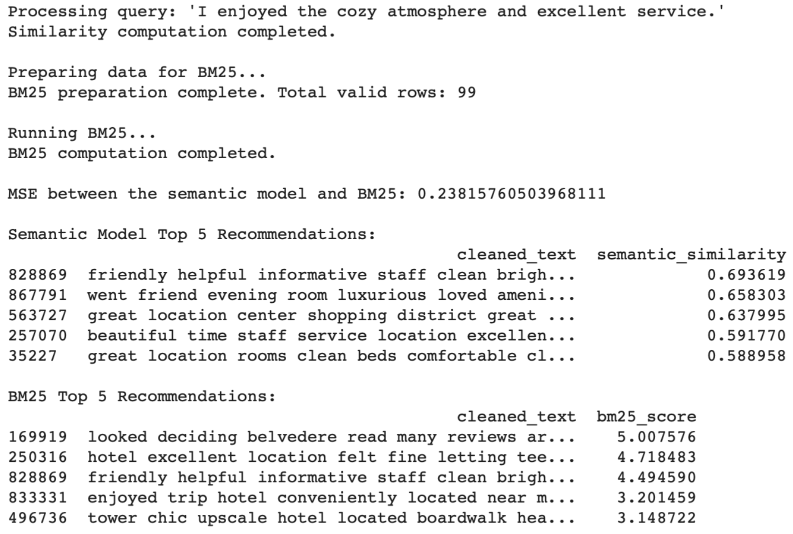
To evaluate performance, we used **Mean Squared Error (MSE)** to compare the similarity scores generated by both models. This metric quantifies the alignment between the rankings produced by BM25 and our semantic similarity model. A lower MSE indicates that the semantic model closely mimics BM25's ranking structure. Furthermore, by analyzing the top-ranked reviews from each method, we assessed how well the models met user intent, emphasizing the practical advantage of a deeper, context-aware approach in real-world scenarios. This comprehensive evaluation highlights the superiority of modern NLP techniques for delivering more meaningful and relevant recommendations.

5. Results

The Semantic Similarity Model has several strengths. It is more accurate in understanding the user's intent and captures the contextual meaning of the text, allowing it to rank reviews based on their relevance to the query. This enables the model to deliver more contextually appropriate recommendations. However, the model is not without its weaknesses. It still requires further tuning and may not always produce perfect rankings. For example, a review about "great location" in the top 5 might not be ideal if the user is more interested in aspects like service or atmosphere, which the model may not prioritize as effectively in every case.

On the other hand, BM25 has its own set of strengths. It is efficient and interpretable, especially for keyword-based matching. This makes it a solid choice for traditional search tasks where exact term matching is important. However, BM25's major weakness lies in its limited ability to understand the broader context or meaning behind the words. It struggles with recognizing the semantic relationships between terms, leading to less relevant recommendations when the query is context-heavy or nuanced.

In conclusion, the semantic model outperforms BM25 by delivering results that better reflect the user's intent and the specific sentiments expressed in the query. The semantic model is able to capture deeper nuances and context, offering a more relevant ranking of reviews. However, further refinement of the model could enhance its accuracy, ensuring it ranks reviews even more closely in line with user-specific preferences.



```
Processing query: 'I enjoyed the cozy atmosphere and excellent service.'
Similarity computation completed.

Preparing data for BM25...
BM25 preparation complete. Total valid rows: 99

Running BM25...
BM25 computation completed.

MSE between the semantic model and BM25: 0.23815760503968111

Semantic Model Top 5 Recommendations:
                                              cleaned_text  semantic_similarity
828869  friendly helpful informative staff clean brigh...             0.693619
867791  went friend evening room luxurious loved ameni...             0.658303
563727  great location center shopping district great ...             0.637995
257070  beautiful time staff service location excellen...             0.591770
35227   great location rooms clean beds comfortable cl...             0.588958

BM25 Top 5 Recommendations:
                                              cleaned_text  bm25_score
169919  looked deciding belvedere read many reviews ar...    5.007576
250316  hotel excellent location felt fine letting tee...    4.718483
828869  friendly helpful informative staff clean brigh...    4.494590
833331  enjoyed trip hotel conveniently located near m...    3.201459
496736  tower chic upscale hotel located boardwalk hea...    3.148722
```