Preet Raut (51)
Dipanshu Vartak (62)
Shubham Warik (66)

Text Summarization on Amazon Reviews

Abstract:

This project seeks to address the challenge of information overload in the context of fine food reviews on Amazon, where the dataset contains an extensive collection of over 500,000 reviews hosted on Kaggle. The primary objective is to develop a robust and effective sequence-to-sequence (seq2seq) model for creating concise and contextually relevant summaries of these reviews. To achieve this, a two-layered bidirectional Recurrent Neural Network (RNN) architecture, featuring Long Short-Term Memory (LSTM) units, is employed for the encoding of the input data, while the target data is processed through two LSTM layers.

Problem Statement:

The proliferation of e-commerce platforms like Amazon has revolutionized the way consumers access and evaluate products through the vast repository of user-generated product reviews. However, this abundance of information poses a significant challenge: customers are often inundated with an overwhelming volume of fine food reviews. This plethora of data can be a hindrance to users seeking to extract meaningful insights about products. As such, the research project is designed to tackle the need for automated text summarization, specifically tailored to fine

food reviews on Amazon. By developing a model capable of distilling the essence of these reviews into concise, informative summaries, the project aims to empower consumers, save their valuable time, and enhance the quality of their decision-making when purchasing fine foods.

Methodology:

1. Data Preparation:
   - The initial phase of the project involves gathering a substantial dataset of fine food reviews from Amazon and performing data preprocessing to ensure that it is suitable for training a seq2seq model.
   - The data is tokenized and cleaned, removing any superfluous information or noise.

2. Seq2seq Architecture:
   - A two-layered bidirectional RNN architecture, incorporating LSTMs, is adopted for encoding the input data, which encompasses the reviews.
   - The target data, which is the generation of summaries, is managed through a two-layered structure, with each layer employing LSTMs, enhanced by the Bahdanau attention mechanism. This architecture is chosen to facilitate the understanding and summarization of the review content.

3. Training and Validation:
   - The model is trained on the prepared dataset, with a focus on optimizing accuracy in generating summaries.
   - The model's performance is subsequently validated, using relevant metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), to assess the quality of the generated summaries.

4. Hosting Infrastructure:
   - Deployment of the model necessitates a hosting infrastructure capable of managing the large dataset effectively. This infrastructure should accommodate both training and inference processes to ensure practical usability.

5. User Testing:
   - A crucial aspect of this research involves user testing to assess the utility of the generated summaries for consumers who rely on fine food reviews when making purchase decisions.
   - Feedback is gathered from users to evaluate the quality, coherence, and relevance of the summaries.

Conclusion:

This project represents a significant step forward in addressing the information overload issue surrounding fine food reviews on Amazon. By developing an automated summarization system that can distill essential insights from this abundance of data, the project contributes to enhancing the accessibility and usability of product information for consumers, ultimately elevating their online shopping experiences.