WikiBot: A Chatbot for Summarizing and Recommending Wikipedia Articles

Group 30: Ayush Patel, Shivansh Verma, Spandan Maaheshwari

Description:

WikiBot is a conversational AI tool that summarizes and recommends Wikipedia articles to users based on their input. The bot scrapes data from Wikipedia to generate summaries and recommend related articles to users who ask for a particular topic or question. This tool can help users quickly access relevant information and discover new articles related to their interests.

Dataset:

The Wikipedia Vital Level 4 dataset is a curated collection of over 10,000 articles that cover essential topics in human knowledge. These articles meet certain criteria, such as being notable, widely recognized, and having a significant impact on human knowledge. The dataset is organized into categories, such as "Art and Architecture," "Philosophy and Religion," "Science and Technology," and "Social Sciences." The articles come with metadata, such as their length, date of last edit, and pageviews, which can be used to understand their popularity and relevance.

Link: Wikipedia: Vital articles/Level/4 - Wikipedia

Approach to solution:

WikiBot allows us to quickly access important information from relevant Wikipedia articles. When the user input a question or topic of interest, the chatbot retrieves the most relevant articles using web scraping techniques. It then applies NLP techniques such as NLTK and GENSIM to extract the most important information from the articles and generate a concise summary for the user. We are going to use t-SNE for the process of feature selection for our recommendation model.

In addition to summarizing articles, WikiBot also recommends related articles based on my topic or question input. This is done using a recommendation algorithm that utilizes cosine distance and User-based collaborative filtering to identify similar articles. The algorithm employs Laplace smoothing to avoid the problem of zero probabilities. We can easily interact with WikiBot by typing in the questions or topics of interest. The chatbot then responds with a summarized version of the relevant Wikipedia article, as well as related articles for the user to explore further. They can also ask relevant questions related to the articles and WikiBot provides informative answers.

Its ability to retrieve and summarize vast amounts of information in real-time makes it an invaluable tool for me as a researcher, student, or anyone in need of quick access to information on a given topic. We are also going to use user-based collaborative filtering in which we will use unsupervised machine learning algorithms such as K-means, Agglomerative clustering, Hierarchical clustering, etc. The recommendations will be generated using user-based collaborative filtering.

Related Work:

Paper 1: Yamada, I., Tamaki, R., Shindo, H., & Takefuji, Y. (2016). Studio Ousia's Quiz Bowl Question Answering System. arXiv preprint arXiv:1603.07042.

Some of the key findings from the paper include:

- The system generated high-quality candidate answers and was able to learn from large amounts of textual data, including Wikipedia articles, to improve its performance. It successfully answered a wide range of question types in areas such as literature, science, history, and fine arts.

Some of the drawbacks or limitations of the paper include:

- The Quiz Bowl Question Answering System's performance is heavily dependent on the quality and quantity of the training data and may struggle with answering questions outside of the training data. The system's complex architecture may also be difficult to replicate or implement for other researchers or developers.

Paper 2: "Dialogue Generation Using Wikipedia as Knowledge Base" by Peng et al. (2018)

Some of the key findings from the paper include:

- The proposed approach in the paper showed superior performance to existing state-of-the-art dialogue generation models across various evaluation metrics. It was able to generate more informative and diverse responses by incorporating external knowledge from Wikipedia and using a latent variable model. The model was versatile enough to handle various types of dialogue tasks, including chit-chat, question answering, and knowledge-grounded dialogue.

However, there are also some limitations or drawbacks of the proposed approach, such as:

- The model's performance is affected by the quality and bias of the available Wikipedia articles and may generate factually correct but socially inappropriate responses. The model's training and reliance on external knowledge sources require a large amount of data and computational resources, which may limit its applicability and raise ethical and privacy concerns.

Methodology, Tools and Techniques:

Extracting data from Wikipedia will be accomplished using Python libraries such as BeautifulSoup and Gensim, while pre-processing, summarizing, and calculating similarity will be done using unsupervised machine learning algorithms such as clustering (e.g., K-means, Hierarchical clustering) and generate recommendations using (e.g., User- based collaborative filtering, Content-based filtering) algorithms, with TensorFlow and PyTorch as our chosen frameworks for training the model. We are going to use t-SNE for the process of feature selection for our recommendation

model. For text summarization, we will test both extractive and abstractive methods and evaluate their respective performance and training time. Although, abstractive methods produce more human-like responses, they are more challenging to create. Therefore, we will compare the performance of both extractive and abstractive methods using our dataset.

References:

1. Yamada, I., Tamaki, R., Shindo, H., & Takefuji, Y. (2016). Studio Ousia's Quiz Bowl Question Answering System. arXiv preprint arXiv:1603.07042.
2. "Dialogue Generation Using Wikipedia as Knowledge Base" by Peng et al. (2018)
3. "Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia" by Yamada et al. (2018)