ReviewSpeak – Trigram-Based Word Prediction Using Movie Reviews Corpus

# 1. Introduction

Natural Language Processing (NLP) enables machines to understand and generate human language. One intriguing application is next-word prediction, where a model predicts the most probable word that follows a given sequence. This project demonstrates a simple but powerful trigram-based model built from the NLTK movie reviews corpus, using Python and Gradio for a live, interactive interface.

# 2. Objective

The main objective of this project is to:

- Build a Trigram Language Model from real-world text data (movie reviews),

- Predict the next most probable word given two input words,

- Provide a user-friendly interface for experimentation using Gradio.

# 3. Dataset Used

- Corpus: nltk.corpus.movie\_reviews

- Source: NLTK (Natural Language Toolkit)

- Content: 2000 movie reviews categorized into positive and negative sentiments.

The corpus provides rich linguistic patterns suitable for building an n-gram language model.

# 4. Methodology

✅ Preprocessing:

- Tokenization using nltk.word\_tokenize.

- Lowercasing of all tokens.

- Removal of punctuation and non-alphabetic tokens.

🔗 Trigram Model Creation:

- Trigram Definition: A sequence of three consecutive words (w1, w2, w3).

- The model records how often a word w3 follows the pair (w1, w2).

🔮 Prediction Function:

- Accepts a pair of words (w1, w2).

- Looks up the most frequently observed third word w3 that followed (w1, w2).

🌐 Gradio Interface:

- Simple textbox for user input.

- Displays the predicted next word based on trigram model.

# 5. Tools & Libraries

- Python: Primary programming language.

- NLTK: For natural language processing and corpus access.

- Gradio: For creating a web-based interface.

- Collections: Counter and defaultdict for counting word occurrences.

- Tokenization: To process raw text into usable tokens.

# 6. Output and Interpretation

The application returns the most probable next word after the two provided words.

Example:

Input: this is

Output: great (if "this is great" is common in the corpus)

The output reflects common word patterns found in movie reviews, useful for:

- Text auto-completion

- Language generation

- Improving human-computer interaction

# 7. Applications and Future Scope

🔍 Applications:

- Text editors and smart keyboards

- Chatbots and virtual assistants

- Educational tools to understand grammar and context

- Predictive typing systems

🚀 Future Improvements:

- Extend to 4-grams or 5-grams for more context.

- Use neural language models like GPT or LSTM.

- Allow dynamic corpus uploads.

- Support sentence completion or full-text generation.

# 8. Conclusion

This project showcases a basic but effective NLP technique — trigram modeling — and demonstrates its potential in word prediction tasks. It balances simplicity and utility by combining classic NLP with modern interactive tools like Gradio.

By using real-world data from movie reviews, the model offers a meaningful insight into common English language usage in opinions and critiques. This forms a strong foundation for more advanced NLP applications such as chatbots, writing assistants, or personalized content generation tools.

Name:Olivia Nereparambil

College:Agnel Polytechnic,Vashi