Project 3 - Build A Language Model

Tasks

1. Build a language model using trigrams
2. Find out how many words in the vocabulary are starting words.
3. Pick a word from this list at random and generate a sentence using the trigram model built in step 1
4. Build a language model using 4-grams
5. Generate a sentence using the 4-gram model - use step 2 to pick a random word
6. Which one of the language models (trigram or 4-gram) performed better? Why?

Note: Use 90% of the corpus (sentences) to learn the model parameters and the remaining to test your model. Do not forget to construct the vocabulary :)