Checkpoint 1 Documentation

CSE 447 | Group 35: Jasmine Zhang, Caitlyn Widjaja, Theresa Tran

Dataset

We obtained the WikiText-2 dataset using the Hugging Face ‘datasets’ library, and exported the cleaned training split to a local UTF-8 text file. This pipeline operates entirely offline for reproducibility and efficiency. This is well-suited for character-level prediction tasks because it contains natural and diverse text with varied vocabulary and punctuation structures.

Method

We are using a character-level n-gram model that utilizes trigram, bigram, and unigram statistics. To do this, we train our model by counting character-level unigrams, bigrams, and trigrams across the entire training corpus, then use this to precompute the top-3 most frequent next characters for each observed context. With this information, we can use the longest available context (trigram → bigram → unigram fallbacks) to build accurate predictions. A 3-character prediction is returned following a standard procedure:

- Trigram Predictions → for input strings with more than 2 characters
- Bigram Predictions → for input strings with 1 character
- Unigram Predictions → for empty strings and punctuation (unseen contexts)