CS 245 - Assignment 2

Usage

Method `main` of class `A2` prints spelling suggestions for "onomatopoeia" if called with command line arguments:

- the path to the token frequency file ("unigram_freq.csv")
- the number of suggested tokens for each prefix of "onomatopoeia".

Method `main` of class `MisspellingMain` prints spelling suggestions for misspelled tokens in a misspelings file if called with command line arguments:

- the path to the token frequency file ("unigram_freq.csv")
- the path to the misspelings file ("misspelling.csv").

This method processes the misspelings file for each value of the number of suggested tokens (for each prefix) from `3` and `7`.

Improvements

The analysis of suggestions for common misspelings in "misspelling.csv" showed that this spelling suggestion algorithm performs poorly if a spelling mistake occurs in a character at the beginning of a token. If a mistake occurs in the 6th character or later, the algorithm almost always finds the correct suggestion. If a mistake occurs in the 2nd character or earlier, the algorithm almost always doesn't find the correct suggestion.

If a mistake occurs in the `i`th (one-based) character in an input token, this algorithm considers all types in "unigram_freq.csv" having a common prefix of length `i-1` with the input token. Characters after the misspelled character aren't used to narrow the search. In order to improve the quality of suggestions, the algorithm should consider all characters in the input token equally.

- One idea is to search in "unigram_freq.csv" for types having a specific character counts, for example, two "p"s, one "a", one "l", one "e" if the input token is "apple". This search may be approximate. In this case, character order doesn't matter.
- Another idea is to search in "unigram_freq.csv" for types having characters in specific places approximately, for example, (if the input token is "apple"):
  - "a" is the 1st or the 2nd
  - "p" is the 1st or the 2nd or the 3rd
  - "l" is the 3rd or the 4th or the 5th
  - and so on.