NLP HW1

**Algorithm:**

Word correction – generating all the candidates according to the error model, filter all the out of vocabulary words, multiply the probability of a candidate word after correction and the probability of it in the vocabulary. Then, choosing the word with the highest probability overall.

Sentence correction – generating all the possible candidate sentences that has only words in the vocabulary. Then, for each candidate sentence, iterating over all his words and multiplying the probability of these words with each other. Their probability determined by alpha if the word is the same as in the original sentence, otherwise (1-alpha)/number\_of\_candidates. Just like in the presentation. Then, multiple the aggregated probability of the words in the sentence with the probability of the entire sentence being part of the language model. I used that formula and not using the error model because it just did not correct the sentences correctly with reasonable alpha

**Text generation** – given a context, choosing the next word randomly according to the probabilities of the candidates. If there are no candidates, continue by choosing random word from the starting sentence candidates according to their probabilities.

**Additional packages:**

* Numpy – I used it in the text generation function in order to pick random word according to a specific distribution.

**Additional functions:**

* \_sentence\_tokenizer – tokenizing text into sentences
* \_normalize\_text – replacing newlines with spaces and removing any non-letter or white space characters
* \_choose\_given\_context – choosing the next word given a context and a language model
* \_generate\_candidates\_with\_proba – create a map of all the candidates of a given word with the probability for that candidate
* \_generate\_sentence\_candicate – generate all the candidate sentences up to a given number of mistakes
* \_p - calculate the probability of a given word w is supposed to be x
* \_context\_freq – caching the frequency of a given context in the language model
* \_filter\_word\_candidates - filter words out of vocab and duplicates