**Assignment 1 Report**

**Name: Syiti Liviani M.**

**SID: 1301174515**

**Class: IF-41-INT**

In this assignment, I applied Bigram Language Model for 100 articles in Bahasa Indonesia with keyword “korupsi”. The model then will be evaluated by using perplexity from a sample test, then the result will be analyzed.

**Scraping Article**

First, the webpage link for articles that contains keyword “korupsi” is collected using BeautifulSoup library. The source that I choose is detik.com because all the articles are using Bahasa Indonesia. From the links that have been gathered then I use BeautifulSoup again to extract the articles from each page. The result of 100 articles then is saved into file article.txt.

**Bigram Language Model**

For building the model, I’m using NLTK library to tokenize each words (unigram). The tokens are turned into lowercase characters and the special characters are removed. Then, I create the unigram frequency table by calculating how many times a word has appeared. Then, by using the unigram frequency table, I create the probability table by dividing the frequency of a word with the total words. After that, I calculate the bigram frequency table by using the tokens from before and calculate how many times does a 2 ordered token (bigram) has appeared in the articles. Then, the probability of each bigram is calculated by dividing the frequency of the bigram with the frequency of the 2nd token. For testing, the model will count the probability from each bigram for total probability, and 1 will dive the total probability. The result then calculated with root of 6 to find the perplexity.

**Test Sample**

The model is evaluated by using one sentence that closely related to the topic and one sentence that’s unrelated to the topic.

|  |  |
| --- | --- |
| **Sentence** | **Perplexity** |
| Related to “Korupsi” | 10.12245824169573 |
| Unrelated to “Korupsi” | 1948.6138337906568 |

Here, we can see the difference from the perplexity. The sentence that is related to “Korupsi” gets lower perplexity if compared to sentence that’s unrelated to “korupsi”. High perplexity means that unrelated sentence has lower probability rather than related sentence, which also means that the model can do better with related sentences that is in the corpus rather than sentences with unknown words.