

**Faculty of Engineering and Technology**

**Computer Science Department**

**COMP338**

**Project 3 2st - 2021/2022**

**Project Description**

1. **Build a Tri-gram language model**

Each student needs to collect an **Arabic** corpus of at least 100,000 words, but the more is better. A bonus will be given if the corpus contains any Arabic dialect.

Students cannot use the same corpus, fully or partially.

Write a program to tokenize the corpus into tokens/words, then build a tri-gram model for this corpus. That is, your language model is a table that contains: the token, the token counts, and the token probability.

The language model should be saved in CSV format.

1. **Develop Word Substitution interface**

Develop a JAVA program that uses your language model to substitute any word in a given sentence. In other words, the user can write a sentence in Arabic, and when clicking on any word in this sentence, the program will show the top candidate words that can be used to replace this word. The program will show each candidate word with its probability. Only the top 10 candidate words will be shown order by probability. The candidate words should be retrieved using the language model, by calculating their probabilities based on the previous words (up to 6 words) – using the chain rule and Markov assumption..

Example:

The user wrote (كل عام وانت بألف خير) then clicked on (وانت). The program showed the candidate words with propablities.

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | كل عام **وانت** بألف خير |  |  |  | | --- | --- | | وأنتم  وانت  زي  مثل  غير | 0.8  0.7  0.3  0.16  0.02 | |

**Submission:**corpus language model.csv, source code, and all files used to run the project.

During the discussion, students will be also asked theoretical questions related NLP.

**Deadline:**7/06/2022