



Natural Language Processing CS 1462

Assignment 1: Text generation using N-gram -

Individual work - deadline: 12 April mid-night marks: 5 marks

Use the corpus (plain text) to generate N-gram models where n : 2-6. The user should enter number of words (tokens) in the desired sentence, for example 10 words and should enter one word. The output should have 10 words/tokens. Evaluate 8 test samples by human expert (yourself or your friend).

Submit:

- 1) the code and the test sample (input + output)
- 2) the 10 high frequent (trigram) in the corpus in a text file

You can use python/java and any NLP library.

You can use the corpus **KSUCCA Files** from here: <https://sourceforge.net/projects/ksucca-corpus/files/>

Or

Any Arabic plain corpus for any domain.