S20CS7.401 Introduction to NLP **(Due Date: 4th February, 2020)**

Assignment 1: Language Modeling and Smoothing

Instructor: Manish Shrivastava *TA:* Alok Debnath

1 Question

We have provided a training corpus `corpus.txt`. Use it to create an n-gram language model, where n can be provided as a parameter. Perform smoothing on the language model using:

- Witten Bell Smoothing, and
- Kneyser Ney Smoothing

Answer in the README: Compare the models of smoothing and explain in which cases the outputs of the two smoothing mechanisms differ and why.

2 Submission Format and Other Instructions

Submit a single file of the format `<roll_number>.zip`, which contains:

- `README`: Which provides instructions on how to generate the language model. Also answer the descriptive question as instructed.
- `language_model.py`: Runs the language model given the following:

  ```
  $ python3 language_model.py <value of n> <smoothing type> <path to input corpus>
  ```

  where n can be between 1 and 3, and smoothing type can be `k` for Kneyser Ney or `w` for Witten Bell. On running the file, the expected output is a prompt, which asks for a sentence and provides the probability of that sentence using the two smoothing mechanisms.

  Therefore an example would be:

  ```
  $ python3 language_model.py 2 k ./corpus.txt
  input sentence: I am a man.
  0.899742021
  ```

Please follow the submission instructions carefully. Plagiarism shall not go unpunished.