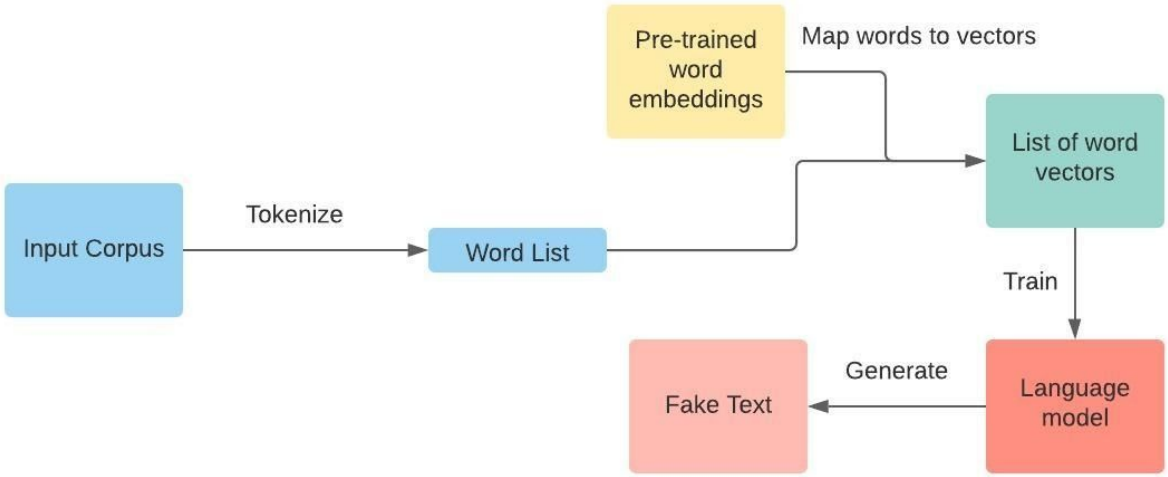
CSE583: Natural Language Processing
Zhan Shi, Neilly Herrera Tan, Angel Burr

Procedural design

- The main use case for this program is for generating fake text. Our program takes an input of some text, and outputs a wall of fake text with similar sentiment to the input text.
    - User input: A few paragraphs of text such as fake speeches, a Wikipedia article, a poem, etc.
    - Output: Computer generated fake text that mimics the style of the input text.

- This text generator program is meant for fun. With a more advanced program, a person may be able to generate greeting cards or fortune cookies based on larger or smaller text inputs. These examples are implications for future work. For class purposes, our program may not be advanced enough to generate meaningful text in this manner.

Component design



- **<u>Preprocessing</u>**
  Preprocess the input text with 4 main functions in mind:
  1. Tokenize the input.
  2. Indexing
  3. Generate <EndOfSentence> tokens
  4. Break token list into pieces based on <EOS> tokens for batch training

- **<u>Training</u>**
  Train the language model with the input given, containing two parts:

1. Initializing the embedding matrix with pre-trained vectors when available, otherwise start at random.
2. Train the LSTM model

- **Decoding**
  Decode the language model to generate fake text. With both greedy method and top k random sampling method in mind.