CSCI - 5832
NAMED ENTITY RECOGNITION

December 5, 2017

Approaches explored for NER with human gene corpus provided, and corresponding F-1 score. Libraries Used: NLTK, Pandas

- **HMM Based NER Using Viterbi Algorithm:**
  - Vocabulary Size: All the tokens with count > 1
  - Achieved a F-1 score of 0.472 on previously unseen test data

- **Maximum Entropy Based NER using custom features**:
  - Vocabulary Size: All the tokens with count > 1
  - Feature Set finalized after multiple iterations and incrementally adding to the list
    - Current Word Index
    - Word: Current, Previous, Next
    - Word Shape: Current, Previous, Next
    - Word Length: Current, Previous, Next
    - Prefix: Current, Previous, Next
    - Suffix: Current, Previous, Next
    - Lemma: Current, Previous, Next
    - UNK attribute: To account for out-of-vocabulary words
    - Prefix Pairs: Previous + Current, Current + Next, Previous + Current + Next
    - Suffix Pairs: Previous + Current, Current + Next, Previous + Current + Next
    - Word Shape Pairs: Previous + Current, Current + Next, Previous + Current + Next
    - Lemma Pairs: Previous + Current, Current + Next, Previous + Current + Next
    - Previously three assigned tags to account for tag transition trend
  - All these features account for tag transition trend, word information, and word context information which helped model to classify IOB tags for giving word and its context.
  - Achieved a F-1 score of around 0.60 on previously unseen test data for all the runs