-python vs java

-feature sets to use

* libraries to implement HMM and CRF:
  + problems:
    - data provided pre-processed, libraries (MALLET) seems to expect a string
    - suggestions/workarounds/otherlibraries?
* If we have a lib in that format:
  + Have a library that we know will take in this data, having processed data to be BIO
  + that's fed into the model--> to train
  + run that model on the test data that outputs tagged data
  + the tagged data is what we submit to kaggle
* Feature Sets
  + Only with CRFs🡪 HMM not a feature based model
  + Can use either weasel and hedge words
    - Weasel words= words which imply uncertainty?
    - Hedge words = words which confine emotions ex. a bit etc.
    - Identifying hedges and their scope
    - Look into papers for feature set
      * Look at references in webpage
      * Hedge = weasel
      * Sample data
      * Sample data
      * Look at methods
      * Paper on SEQUENCE LABELING
      * SVM AND CRF
      * FEATURES ABOUT HEDGE CUES
    - Using hedge or nothing
    - Explain why you are doing this; why sentence level relates to token level
    - Hedge word: IOB
    - Don’t do dependency parsing
    - Lemma: strip suffixes
    - Dp = dependency parser Don’t use
    - \*Look at hedge cue dictionary and scope indicator
    - Look online of hedge cue diciotnaries, create own dictionary from hedge word list from only
    - Start with word property feature (ex. Simply word itself, POS, \*\*dictionary of hedge words)
* Algorithm:

Stanford NLP:

NLTK

HMM example

Nltk hmm github repo example