CSE 5525: HOMEWORK 3 | PART-OF-SPEECH TAGGING

-Shiuli Das, 500043947

OBJECTIVE: Implement structured perceptron and Viterbi algorithm for part-of-speech tagging. Explore performance of part-of-speech taggers when applied to different domains.

RESULTS:

1. World's most frequent tag baseline
   Accuracy : 72.88%

2. Tagging using Viterbi algorithm and Structured Perceptron weight update:< Tested on twitter >
   Using parameter averaging, trained on-
   a. Twitter
      Accuracy: 87.07%
   b. Penn Tree Bank
      Accuracy: 75.37%
   c. IRC chat data
      Accuracy: 82.47%
   d. Twitter + Penn Tree Bank + IRC chat data
      Accuracy: 89.23%

3. Named Entity Recognition
   a. Without additional features
   processed 11570 tokens with 356 phrases; found: 265 phrases; correct: 153.
   accuracy: 96.27%; precision: 57.74%; recall: 42.98%; FB1: 49.28
        ENTITY: precision: 57.74%; recall: 42.98%; FB1: 49.28 265

   b. Adding a "NAMES" feature which contains set of names extracted from the following -
   - http://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html
   - https://en.wikipedia.org/wiki/List_of_companies_of_the_United_States
   - https://simple.wikipedia.org/wiki/List_of_United_States_cities_by_population
   - https://en.wikipedia.org/wiki/List_of_restaurant_chains_in_the_United_States
   - names of months of the year

   processed 11570 tokens with 356 phrases; found: 274 phrases; correct: 157.
   accuracy: 96.31%; precision: 57.30%; recall: 44.10%; FB1: 49.84
        ENTITY: precision: 57.30%; recall: 44.10%; FB1: 49.84 274