Yitong Wang

CSCI-GA-2590

Natural Language Processing

Homework6

Feature set includes:

Feature set 1: ‘current’ and ‘pos’

Feature set 2: ‘current’, ‘pos’ and ‘initCaptilized’, which indicates whether the word is capitalized.

Feature set 3: ‘prev’, ‘prevPOS’, ‘current’, ‘pos’, ‘next’, ‘nextPOS’, ‘initCaptilized’

Feature set 4: ‘prev’, ‘prevPOS’, ‘current’, ‘pos’, ‘next’, ‘nextPOS’, ‘initCaptilized’, ‘suffixIng’ (which indicates whether the word has a suffix of ‘ing’)

For the precision, recall and the f-measure:

For NPGroup1:

Precision is 0.8182861514919664, Recall is 0.8404715127701375, F-Measure is 0.8292304710215158

here

For NPGroup2:

Precision is 0.7626280466266337, Recall is 0.8483300589390963, F-Measure is 0.8031994047619048

here

For NPGroup3:

Precision is 0.8755690440060698, Recall is 0.906876227897839, F-Measure is 0.8909476934954642

here

For NPGroup4:

Precision is 0.8718726307808946, Recall is 0.9037328094302554, F-Measure is 0.8875168821146056

For the number of correct tags:

Feature set1: Correct tags 9186, accuracy is 0.9186

Feature set2: Correct tags 8272, accuracy is 0.8272

Feature set3: Correct tags 9280, accuracy is 0.928

Feature set4: Correct tags 9277, accuracy is 0.9277

List of codes: (Please run them in the order written here)

**readfile.py**: a python program that reads the training file into different 4 files based on 4 feature sets (named event1, event2, event3, event4)

**Tagger.java**: The file that contains the given code, which transform each ‘event’ into its corresponding model file (named ‘model1’, ‘model2’, ‘model3’, ‘model4’).

**TestTagger.java**: The program that transform the model into results and test the results against the test file.