Programming Assignment

Word Sense Disambiguation (WSD) with Decision Lists

Hints and Tips

1. You can use **BeautifulSoup** to process the XML files.

2. Please preprocess the train and test data accordingly. You can convert them to lower case, standardize plural lines, and remove special characters.

3. You can create a dictionary with id (Ex: line-n.w8_060:3604:) sense (Ex: phone), and the list of tokens (['atlanta', 'based', 'bellsouth', …]) and use this dictionary for further operations.

4. You may use the ConditionalFreqDist() from NLTK. Conditional frequency distributions are used for recording the number of times each sample occurred, given the condition under which the experiment was run.

You can find more information here: https://lost-contact.mit.edu/afs/cs.pitt.edu/projects/nltk/docs/tutorial/probability/conditionalfreqdist.html

5. The size of the window can be from two words to the left to one word to the right. The window should not include 0 (the word itself).

6. You may use ELEProbDist from NLTK to generate/to use as input to the ConditionalProbDist function from NLTK. More about ConditionalProbDist can be found here: https://www.cs.bgu.ac.il/~elhadad/nlpproj/hocr/doc/project.external.nltk_probability.ConditionalProbDist-class.html

7. If the ratio of conditional probabilities between phone and product equals zero, you can assign the likelihood to zero since log(0) is undefined.

8. You can sort the decision list according to probabilities to decrease the execution time.

9. For the purpose of logs, you can use the logger module as it is a professional way to do it (https://docs.python.org/3/library/logging.html). This is not mandatory; you can use simple File I/O for achieving the same.

10. Mention in the comments if any additional features are incorporated into the program.