

Assignment 4

L665/B659: Applying Machine Learning Techniques in CL; Sandra Kuebler

- For this assignment, you will go back to the English lexical sample data from the SensEval-3 competition.
- For the current assignment, we will also concentrate on 3 words out of the set: `arm.n`, `difficulty.n`, and `interest.n`, i.e., the same train/test sets as in assignment 3.
- But this time, we will use Scikit-learn and a different feature set.
- We use the features set from assignment 3 plus the context features (see Escudero et al. (2000). Naive Bayes and Exemplar-Based Approaches to Word Sense Disambiguation Revisited). This means, you need to extract a word list from the training data and then add a binary feature per word. The same list is used to model the training and test data.
- The data format for Scikit-learn is the following: You need two files, one containing your features, one the correct class. In both files, the order of instances needs to be the same. One line per instance.
- Then train two different classifier for each of the 3 words and run them on your test sets.
- Report the timbl and the two new results. How do they differ?
- Work in groups of 2-3 people.
- **DUE DATE:** Sunday, March 13, 11:59pm. Please submit your report via canvas.