Word Sense Disambiguation

The purpose of this exercise is to learn to classify word usages in context into senses. The data consist of six sense-annotated English words: TWA.tar.gz. Each word occurrence in context (instance) is labeled with one of two possible senses.

Naïve Bayes for WSD

Implement the supervised Naïve Bayes algorithm for WSD. Train and evaluate your classifier using the leave-one-out method, as follows:

- Each of the six word types is a separate binary classification problem.
- For each word type
  - Choose one instance as a test item. The rest of the instances will be training examples.
  - Extract features (use words in surrounding context) and compute feature counts from the training examples.
  - Apply the model to the test instance. For smoothing you can use “add epsilon”.
  - Choose the next instance as a test item and repeat, until every instance was chosen once.
- Compute and report aggregate accuracy for each word type.

Your submission should include a brief report on your solution, as well as the source code you wrote to solve the exercise. If possible, avoid using nonstandard libraries not included in the distribution of the programming language you used. Make sure to include brief instruction on how to run your code. Send your solutions to snlp@lsv.uni-saarland.de by Friday, 1 July. Include [Tue] or [Wed] in the subject of the email to indicate which session you are attending. Do NOT send you solutions to the course mailing list.

Important: Please use PDF as a document format. If you need to compress files, use ZIP or GZIP.