Feature Selection

1. Assume that you want to detect *collocations* consisting of two words in a text using the $\chi^2$-test.

   What are the two events whose independence you want to measure?

   Create manually a table similar to that on Slide 54 of Chapter 6.

2. Implement *Information Gain* and *Mutual Information* feature selection.

3. Using the *Congressional Voting Records Data Set** get a ranking of the 16 features by IG and PMI respectively.

*Collocations are sequences of words that are used very often together and may obtain a specific meaning, e.g. San Francisco, Sherlock Holmes, DVD player, hand made, bar of soap, etc.

**Congressional Voting Records

Classification

1. Implement a *Naive Bayes classifier* for categorial data

2. Using the above data, train your Bayes Classifier on the best 2, 3...16 features and plot the performance in terms of #misclassifications.

   This makes two plots with 2, 3...16 on the $x$-axis and #misclassifications on the $y$-axis.

3. BIG BONUS: Now that you have implemented the methodology, train and test it on some appropriate natural language data.

---

**Note on Submission** Please use PDF as a document format. If you need to compress files, use ZIP or GZIP.

If you attend the Mon. tutorial, the deadline is Saturday 28th June, 23:59. In that case send the solutions to s9tshbck@stud.uni-saarland.de.
If you attend the **Wed. tutorial**, the **deadline** is **Sunday 29th June, 23:59**. In that case send the solutions to *kalofali@ccid.upatras.gr*.

Please indicate each group member's name in the filename:

*Ex08-MemberName1(MN)-MemberName2(MN)-MemberName3(MN).zip*