# CS5213 Project Guides

* LDA
  + Download the LDA package from <http://www.cs.princeton.edu/~blei/lda-c/lda-c-dist.tgz>
  + Detailed guides and data format required are described in **readme.txt** in the package
* HLTAs
  + Go to <https://github.com/kmpoon/hlta> and detailed guides can be found.
* nHDP
  + The nHDP package offered here is different from the one described in paper. The main change is that we let nHDP to calculate per-document loglikelihood instead of per-word loglikelihood in the original codes.
  + Data : Two example data are given: Nips1000\_binary\_train.mat and Nips1000\_binary\_test.mat.

Three variables can be found in each data:

1. vocab: The vocabulary set. The word id is the index of the corresponding cell it is in.
2. Xid: Each cell represents a document, the numbers in the cell are the id’s of the words appearing in the document. The word ids are in the same order as their positions in the document.
3. Xcnt: Each cell represents a document, the numbers in the cell are the numbers of times of a particular word appearing in the document. This particular word is the one in the same position as indicated by Xid. (Here we only use binary data so each word will only be counted once if they appear, so the number is 1.)
   * script\_nHDP\_Nips1k.m: To run nHDP to build a tree. (SampleTree.mat)
   * script\_test.m: To test the obtained tree with test data and give per-document loglikelihood.