Unsupervised Learning

This assignment is **optional**. If you choose to do it, it can only help your grade (i.e., I'll apply a *max* operator.) But, it can be fun.

Details

Use unsupervised learning on the 1149 text files in [data.tgz] to find some interesting structure. You can build an HMM and model them and generate new files, you can do NMF on a document-word frequency matrix and find topics, you can use clustering and try to find clusters, etc. Half the fun will be appropriate preprocessing.

Turn in, on moodle:

1. code that can be used to replicate your results, don't include the data but assume it's in a `data/` subdir; this can be in the form of the URL of a git repo if you'd like
2. a brief report explaining what you did and what interesting structure (if any) you found. If you're turning in a git repo, this can just be a file there; if you do that please name it `report.*`