AM 207: Homework 5

Verena Kaynig-Fittkau and Pavlos Protopapas
Due: 11.59 P.M. Thursday April 14th, 2016

Note: This homework is only for one week

Instructions:

- Upload your answers in an ipython notebook to Canvas.
- We will provide you imports for your ipython notebook. Please do not import additional libraries.
- Your individual submissions should use the following filenames: AM207_YOURNAME_HW5.ipynb
- Your code should be in code cells as part of your ipython notebook. Do not use a different language (or format).
- **Do not just send your code. The homework solutions should be in a report style. Be sure to add comments to your code as well as markdown cells where you describe your approach and discuss your results.**
- Please submit your notebook in an executed status, so that we can see all the results you computed. However, we will still run your code and all cells should reproduce the output when executed.
- If you have multiple files (e.g. you've added code files or images) create a tarball for all files in a single file and name it: AM207_YOURNAME_HW5.tar.gz or AM207_YOURNAME_HW5.zip

Have Fun!

In [3]:
```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

import seaborn as sns
sns.set_style("white")

import time
import timeit

import scipy.stats
import pandas as pd
import pymc as pm

import re
import numpy as np

import string
```

Problem 1: HMM... I Think Your Text Got Corrupted!

In this problem you should use a Hidden Markov Model to correct typos in a text without using a dictionary. Your data is in two different text files:

- `Shakespeare_correct.txt` contains the words of some sonnets from Shakespeare
- `Shakespeare_typos.txt` contains the same text, but now some of the characters are corrupted

For convenience both text files only contain lower case letters a-z and spaces.

First build a first order HMM:

- What are the hidden states and what are the observed states?

- What should you do to generate your HMM probability matrices?
- For some of the HMM parameters, you won't have enough training data to get representative probabilities. For example, some of your probabilites might be 0. You should address this problem by adding a small pseudocount, similar to the motif finding problem from a previous assignment.
- Implement the Viterbi algorithm and run it on a test portion that contains errors. Show that your Viterbi implementation can improve text of length 100, 500, 1000, and 2000. Note: To do this correctly you would have to withhold the part of the text that you use for testing when you estimate the parameters for you HMM. For the sake of this homework it is ok though to report training error instead of test error. Just be aware that the correction rate you are reporting most likely is a very optimistic estimate.
- What correction rate do you get?

Important: Wikipedia has a nice article on Viterbi (https://en.wikipedia.org/wiki/Viterbi_algorithm). **Please do not use the python implementation from this article!** (The lecture notebook also has the version from Wikipedia). Using dictionaries for Viterbi is really not intuitive and using numpy is typically faster. The article has very nice pseudo code that should enable you to easily program Viterbi by yourself. Please also refrain for this problem from using any other third party implementations.

Now for a second order HMM: By using a second order HMM, you should be able to get a better correction rate.

- Give an intuitive explanation why a second order HMM should give better results.
- Implement your second order text correction. Hint: If you think a bit about the model you won't even have to change your Viterbi implementation.
- Compare your correction rates against the first order model for text length of 100 and 500, (you can do 1000 as well if your computer is fast enough).
- How well would your implementation scale to HMMs of even higher order?

Extra Problem 2: Final Project Review

You will be contacted shortly by a TF to meet and discuss your final project proposal. Be sure to take advantage of this feedback option. Review meetings should be scheduled within the week from April 11-15.

In [ ]: