Name: _____
CS159: Natural Language Processing
PReview #3
Due in class Feb 9, 2015

**Review**

Finish the in-class lab. Describe what you found for Parts 3 and 4. Note that Unicode errors are ubiquitous - there was one in the lab writeup! In Part 4, the title of the document to retrieve should be Château.

■

For the Wikipedia article of your choice, download and parse both the English and the Simple English version of the article. You can find the Simple English version by selecting "Simple English" in the list of languages on the left. Note that not every article has a Simple English "translation," so you may want to start by finding a good simple article and then finding the English version.
Using what you've seen and/or read about in nltk, identify which words show up more often in the Simple English article and which words show up more often in the English article relative to the size of the article. Comment on what, if anything, do you find interesting.

■

**Preview**

Explain what is meant by the Noisy Channel Model. How does it apply to machine translation?

■

**MS Exercise 2.4**: Are $X$ and $Y$ as defined in the following table independently distributed?

| | | | | |
|---|---|---|---|---|
| $x$ | 0 | 0 | 1 | 1 |
| $y$ | 0 | 1 | 0 | 1 |
| $p(X=x,Y=y)$ | 0.32 | 0.08 | 0.48 | 0.12 |

■