CSC 247/447 Programming Assignment #1 Due Feb 25th, 12:30PM

Your first programming assignment will be to implement a program to do the "guess the color" task we discussed in class. Specifically, you will be given a set of input, each consisting of a two word fragment of a noun phrase, such as "R cat" or "B dog". The task is to read a file containing these inputs and output a file that provides you answers, e.g., "red cat" "brown dog".

As a resource for this assignment, you can use the Brown corpus as outlined in Hannah's email and lecture on Python. Here are the steps she mentioned to install the corpus and some NLTK tools.

1. Install NLTK: run `pip install -U nltk`
2. Download Brown Corpus: run `python3 -c "import nltk; nltk.download('brown')"`

You also may find it useful to stem the words to improve your statistics from the Brown corpus. You can use WordNet's built-in lemmatizer. You can download the data for the module with

```
import nltk
nltk.download('wordnet')
```

and import this module with

```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
```

and then run on a word with lemmatizer.lemmatize('word').
As an example, there are three cases of "black hairs" and 10 "black hair" in the Brown corpus. If you lemmatize the words you'll end up with 13 "black hair".

You may use whatever techniques you wish, combining hand-built rules as well as corpus statistics. As an example of using statistics, say the input is "B ribbon". Looking at the brown corpus we see three cases of "blue ribbon" and one case of "black ribbon". So using this evidence you might guess "blue".

Testing

Once you submit you program, we will run it on two tests. The first will consist of cases that occur in the Brown corpus, such as the "B ribbon" example above. You can get full credit on the first test portion of the grade by getting all these right. The second test set will be examples that do not occur in the Brown corpus, but are cases that a human can accurately guess the correct. For instance, it might contain "G frog" and the answer is Green. This is a significant challenge to do well on as you don't have a source of commonsense knowledge. But see what you can come up with to earn bonus points. (At a minimum you could at least use the most common color found for a letter in the Brown corpus).

Details for Submission of your Program

Submit your program and the documentation on Blackboard. Your program should be defined so that it can be run with the command line

```
python3 [yourprogram].py [input-file] [output-file]
```

and input file we will provide would look like
“b ribbon”
“g bottle”
and your output should be like
“blue ribbon”
“green bottle”

In addition to you program, you should write up a document that outlines the strategies you used, especially how you deal with hard cases where there are not any exact matches in the Brown corpus.