ITCS 4111/5111 Introduction to Natural Language Processing

Assignment 1 Due: September 15th, 2019 11:59 pm

Your submissions should be in a single python notebook. Make sure code is documented well. All submissions should be uploaded to Canvas. Email submissions will not be considered for grading.

The problems marked with ** are optional for students enrolled in the ITCS 4111 section of this course, all other problems are mandatory. ALL problems are mandatory for the students enrolled in ITCS 5111.

Problem 1 (Language Model Creation) (80 points).

In this exercise, you will train probabilistic language models to distinguish between words in different languages. Rather than looking up whole words in a dictionary, you will build models of character sequences so you can make a guess about the language of unseen words. You will need to use NLTK and the Universal Declaration of Human Rights corpus.

We will compare across different languages from the Universal Declaration of Human Rights documents. Use the following code to load the corpus and create sets of four languages.

```
import nltk
from nltk.corpus import udhr

english = udhr.raw('English-Latin1')
french = udhr.raw('French_Francais-Latin1')
italian = udhr.raw('Italian_Italiano-Latin1')
spanish = udhr.raw('Spanish_Espanol-Latin1')
```

If you do not have the UDHR dataset already installed with your version of NLTK, use nltk.download() to download the corpus.

Create training, development and test samples for English, French, Italian, and Spanish from these sets.

```
english_train, english_dev = english[0:1000], english[1000:1100]
french_train, french_dev = french[0:1000], french[1000:1100]
italian_train, italian_dev = italian[0:1000], italian[1000:1100]
spanish_train, spanish_dev = spanish[0:1000], spanish[1000:1100]
```

```
english_test = udhr.words('English-Latin1')[0:1000]
french_test = udhr.words('French_Francais-Latin1')[0:1000]
italian_test = udhr.words('Italian_Italiano-Latin1')[0:1000]
spanish_test = udhr.words('Spanish_Espanol-Latin1')[0:1000]
```

Build unigram, bigram, and trigram character models for all four languages. You may find it convenient to use the NLTK classes FreqDist and ConditionalFreqDist, described in chapter 2 of the NLTK book (http://www.nltk.org/book/).

For each word in the English test sets, calculate the probability assigned to that string by English vs. French unigram models, English vs. French bigram models, and English vs. French trigram models. Use the test set to report accuracy of your models. You should report the accuracies of the uni-, bi-, and tri-gram models.

Problem 2 (Language Model Comparison) (20 points).**

Perform the same experiment as above for Spanish vs. Italian. Which language pair is harder to distinguish?

Note: ITCS 5111 students will be graded out of a 100 points total for all problems in this assignment. ITCS 4111 students will be graded out of 80 points total for all problems in this assignment, except for those marked with **.