ITCS 4111/5111 Introduction to Natural Language Processing

Assignment 3 **Due: October 27th, 2019 11:59 pm**

Total points: 100

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.

The learning goals of this assignment are:

- Understand the fundamentals of statistical machine translation
- Implement the IBM model 1

Corpora **(uploaded in the Assignment 3 folder)**

1. German – English
2. French – English

Machine Translation

Problem 1 (30 points): First pick one of the aligned corpus from the above two choices (German-English or French-English). You will implement a function that will output a table containing the word translation probabilities that were learned (note: think of an efficient data structure for such a sparse matrix).

Problem 2 (50 points): You will implement a function that outputs the alignment for each sentence pair in the training data based on the IBM Model 1 discussed in class (this function should work on the corpus you have chosen for Problem 1).
Here is a potential pseudo-code for IBM Model 1:

```
initialize t(e|f) uniformly
do until convergence
  set count(e,f) to 0 for all e,f
  set total(f) to 0 for all f
  for all sentence pairs (e_s,f_s)
    for all words e in e_s
      set sum = 0
      for all words f in f_s
        sum += t(e|f)
      for all words f in f_s
        count(e,f) += t(e|f) / sum
        total(f)   += t(e|f) / sum
  for all f
    for all e
      t(e|f) = count(e,f) / total(f)
```

Problem 3 (20 points):** Repeat Problem 2 for the other language pair. This problem is optional for ITCS 4111 students and mandatory for ITCS 5111 students.