NLA Assignment 1

Name: Abhigyan Ghosh
Roll Number: 20171089

IBM Model 1:

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)
     set total_s(e) = 0 for all e
     for all words e in e_s
       for all words f in f_s
         total_s(e) += t(e|f)
     for all words e in e_s
       for all words f in f_s
         count(e|f) += t(e|f) / total_s(e)
         total(f)   += t(e|f) / total_s(e)
   for all f
     for all e
       t(e|f) = count(e|f) / total(f)
```

- I have used `defaultdict` library to store all counts.
- All initial counts default to zero except t(e|f) which is initialized to 0.1.
- The code runs for 10 iterations as more iterations take a lot more time to train
- Instead of storing a matrix for t, we are storing a dictionary with a tuple key of (e,f)

Problems in the IBM model 1
I did not make the entire pipeline for SMT. Only translation probabilities are found using this model. Manual inspection reveals a lot of errors especially for words which occur rarely. For example, proper nouns are very rare and are aligned to function words in most contexts.

Outputs:
Link to output data for train: t_out.pickle

HMM:

Did not implement.