Assignment 1

(Due: Sept 27, 11:59pm)

Zoie uses English in her everyday work. She finds it sometimes difficult to decide which **preposition** she should use in some cases. For example, should she say "schedule a meeting in 3pm" or "schedule a meeting at 3pm"? One day she visited HKU and came across a student who happens to study COMP3361 this semester.

"You must be an expert in NLP!", said Zoie.
"Well, I only got to know what is **language model** last week", said the student.
"That's *more* than enough", said Zoie with her eyes shining like stars.
"Could you please **write a program** for me that tells me what **preposition** I should use in a sentence?"
"Sure! My pleasure!", said the student.

And here comes the **assignment 1**.

Download the training corpus from https://nlp.cs.hku.hk/comp3361/hw1/train.fo1
It is *lowercased white-space tokenized* and contains many examples sentences of prepositions, specifically "at, in, of, for on".

We also provide a validation set at https://nlp.cs.hku.hk/comp3361/hw1/valid.fo1 where these 5 prepositions are replaced with a special token **GW**.

Your task is to **build a model** that trains on (and only on) the training corpus and output for each special token **GW** in the validation (and soon the test) set, which preposition is the proper one there. The correct answers for the validation set are listed in https://nlp.cs.hku.hk/comp3361/hw1/valid.fo2. Please follow that format for your output as well.

We will release the real test set (text.fo1) for the task on Sept 20 (a week before the deadline). Please submit to moodle a zip file that contains (1) your code, (2) a write-up (pdf) that explains your model, and (3) your model's predictions (strictly following the format of the output of the validation set) text.fo2.

A simple baseline is to build a BIGRAM model for this task. Models better than that will be scored more than 60/100. The score consists of two parts, the write-up for your model (40%) and the final performance of your model on the test set (60%). Late submissions will not be graded.

Good luck, Zoie!