**Iteration 1: September 29, 2021**

Query : Dataset 7 -- Query Correction.

1. Background.txt, if being used for final evaluation of the project, resembles real world queries a lot more i.e they have spelling mistakes and other errors whereas in the test.tsv file, all the queries just need auto-completion and not spelling correction. So we are getting confused between both the files: test.tsv and background.txt because they are quite different as we are training the model based on the tsv file.
2. Exact same inputs in training and test files have different outputs.

Example:

Input Test File: united Output Test File: united states map vietnamese american

Input Train File: united Output Train File: united states crime rate compared

Remarkably, the train contains only the given output and not the other one and vice-versa, this is to clarify the case of multiple possible outputs.

1. In this project do we have to autocomplete the queries like Google does, when we start entering something it shows us relevant search results or do we have to correct the queries by removing the mistakes made by the user. Also what do we mean by “validating the search results” here, “improving this correction by validating the search results”.

**Reply:**

**1. Background files are a bunch of queries whose mapping to the auto-completed data is not given, you can use it for your modelling optionally.**

**2. This is fine because there is no one-one mapping for queries, for example if you google any text you get multiple recommendations, you can choose amongst these.**

**3. For this part you have to correct the mistakes in the queries.**

Query: Dataset 3 - Syntax Analysis in Source Code

1. The data is split like

>> bcb\_reduced

>> 2

>> default

>> sample

>> selected

>> 3

>> default

>> sample

>> selected

>> …..and so on

Could you please explain more about the structure of this and which one should be used

for training and testing?

1. How should we use `bcb.h2` and `bcb.trace` for the project? And what kind of information is included in these two files?
2. `default` folder of folder `8` and `9` is empty.

**Reply: You can use the codes in the ‘default’ folder for training**

Query: Dataset 5 - Sentiment based aspect analysis

1. In order to extract aspect terms, can we use [spacy’s POS tagging](https://spacy.io/usage/linguistic-features#pos-tagging) (which gives the POS tagging and some relations between the words)

**Reply: Yes. You can use POS tagging from libraries directly for aspect extraction.**

Query: Dataset 6 - POS tagging

1. The question asks us to compare the performance of RNNs and HMMs. However, the dataset given is not really labelled and it just appears to be a large block of text. How do we train RNNs without these labels?

**Reply: New labeled dataset has been uploaded on the drive**

**12/10/21**

Query: Dataset 5 - Sentiment based aspect analysis

1. Can we use pre trained Word2Vec vectors to find similarity between words in order to distinguish between positive and negative words?
2. Can we also use Naive Bayes Classifier for classification?

**Reply: Try and not use pretrained models. You can use the classification method of your choice.**

Query: Dataset 3 - Syntax Analysis in Source Code

1. Are all the files inside the default folders free from syntax errors?

**Reply: As per my knowledge, yes. You can induce syntax errors if you want to test your model**

Query: Dataset 7 -- Query Completion

1. There is an issue in converting all tokens with frequency = 1 into <UNK> as mentioned in the tut because in the training file, there are around 5600 out of 7500 with frequency=1. Also we cannot reduce the frequency threshold further as it is the lowest already. So we are at a roadblock.