For Task 2a, we must first execute parser.py on yelp dataset. The resulting trec file is then used by generateIndexTask2.java to create indexes. The indexed data is then accessed by TrainingAndTestingData.java which will generate Training and testing text files. We need to run this file separately with different values to generate train and test data [details described in java file]. The output of this java file is then used by prediction.py python script.

**parser.py** – this python script extracts dataset, where city is Charlotte/Las Vegas, and reviews; between 2001 to 2014, for businesses in Bakeries category.

**generateIndexTask2.java** – This java file creates index for dataset using Lucene API. Consists of following functions: createIndex() [creates the lucene index for dataset] and displayStats() [displays index stats].

**TrainingAndTestingData.java** – Using the indexed data, retrieves and categorizes review texts into rating categories and generates a text file. All code within main function, except printing to textfile handled by fileReviews() method. This file needs to be run separately for train and test data with different input values to generate respective train and test data.

**prediction.py** – We have written a linear coding approach. The file extracts reviews from retrieval step and performs POS tagging followed by stemming and lowercase. This is repeated for all 5 categories of review. We train the classifier and test the classifier in this step. The prediction of rating is also done in this script. More details within script as comments.