Redwan Khan and Alex Barganier

University at Buffalo

LAB3 DOCOMENTATION

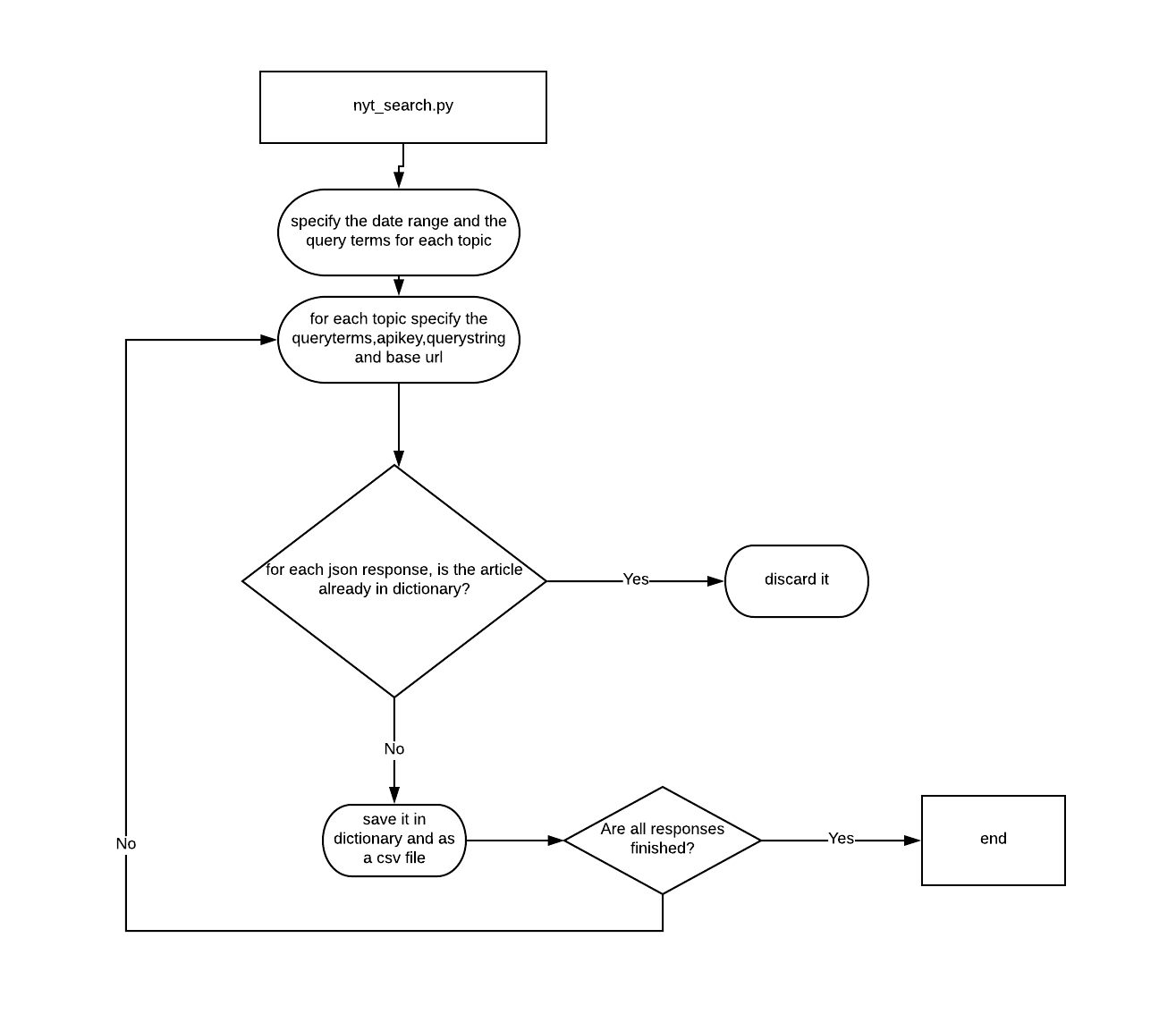
**nyt\_search.py**

Code used for collecting news articles from NYtimes using NYtimes API

**topics** - structure that contains the keywords we used for generating articles related to Sports,Business, Politics and Entertainment

**daterange(start\_date, end\_date)** - specifies a range of dates between which we want to collect news articles

**get\_article\_body(url)** - collects the article text from the url specified



**filter\_data.py**

Code used to filter the collected data by removing stop words and saving each collected article in a seperate file. Even after specifying keywords some of the articles collected for politics might have actually appeared in the business section of Nytimes. So we needed to filter the articles based on their url.

**filter\_keywords** **:** consists of the following fields

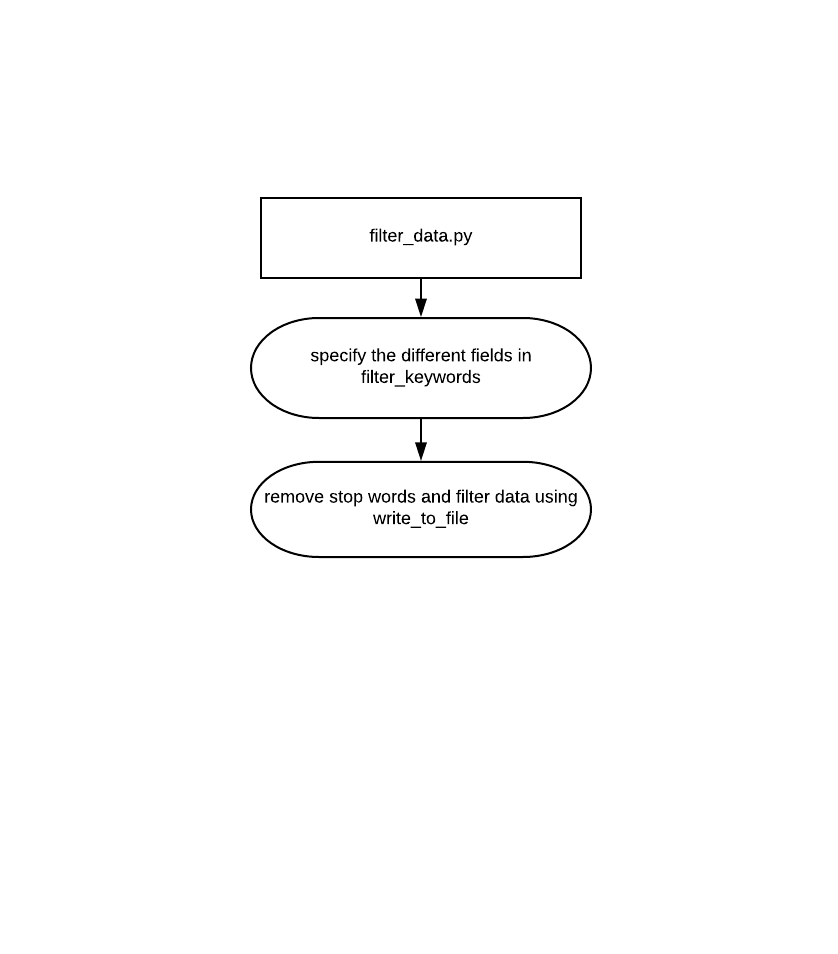
'**filename\_prefix':** prefix of the filename,

**'train\_dir':** the place where the training data files should be saved,

'**test\_dir':** the place where testing data files should be saved

**'source\_file':** which one is the source file to look into for saving the files,

**'url\_keywords':** what keywords to look for in an url for filtering the articles.

**write\_to\_file(fn\_prefix, dict):** gets the prefix name of file and the dictionary consisting of url and content. Opens up a file with the prefix and puts the content of news articles in that file after removing stop words.

**tf\_idf.py**

Training and Testing paths are specified for the different topics of articles. Random articles from Washington Post have also been collected and the testing paths for those articles have also been specified.The labels of the different articles have also been given.

**buildTextRDD(directory, label\_id):** return a dataframe of label ids and articles content.

**buildTfIdfRddAllTopics(business, sports, politics, entertainment):** builds a dataframe for each topic and combines all of the dataframes. Returns a rescaled data of labels and features.

**runTestForModel(model, test\_rdd):** makes a tuple of predicted label and actual label after running a machine learning model. Also gives the accuracy of the results.

Outputs saved in **EvaluationMetrics.txt, featureengineeringtestrdd.txt, featureengineeringtrainrdd.txt,PredictionNYtimes.txt, PredictionWashington.txt**

