**ABSTRACT**

The online shopping is becoming popular day by day . One of the most difficult tasks for online retailers is to retain the customers. For this they deploy diffrent innovative techniques. One is they recommend different products to their customers. The other way to improve their productivity is to do the sentimental analysis of their customers. It could be done through the sentimental analysis product reviews.

We will do sentimental analysis for different customers of e-commerce for better marketing strategies.

**FUNCTIONAL REQUIREMENTS**

* Retriving and processing training data from social media.
* Retrieving and storing reviews from ecommerce sites namely.
* Predicting the sentiment.
* Comparing different classification models.

**NON FUNCTIONAL REQUIREMENTS**

* Python
* Reviews of particular language are taken i.e. English.

**MODULES**

**Data Retrieval**-This module is responsible for crawling data from web using beautiful soup library.

**Data Storage**-This module stores the reviews in database.

**Feature Extraction**-This extracts features from reviews, updates the word count of all the reviews from time to time, saves the dataset in original form.

**Data Analysis-**This module creates testing and training dataset and update files.

**Classification techniques**-This module classifies whether the product has positive sentiment or not. This also keeps a tap on the nature of reviews being developed on a whole i.e. positive, negative or neutral.

For data analysis and classification, Sci-kit Learn library, K-Neighbor, Naive Bayes , SVMs is used.

**Brief description of work-flow:**

#### **Tokenization:**[Tokenization](https://en.wikipedia.org/wiki/Tokenization_(lexical_analysis))is the name given to the process of chopping up sentences into smaller pieces (words or tokens). The segmentation into tokens can be done with decision trees, which contains information to correctly solve the issues you might encounter.

**Word Filtering:** It is the reduction of each word to its base/stem form (by chopping of the affixes).

1. Capital letters should be normalized to lowercase, unless it occurs in the middle of a sentence; this could indicate the name of a writer, place, brand etc.

2.What should be done with the apostrophe (‘); “George’s phone” should obviously be tokenized as “George” and “phone”, but I’m, we’re, they’re should be translated as I am, we are and they are. To make it even more difficult; it can also be used as a quotation mark.

**Bag-of-Words model:** after the text has been segmented into sentences, each sentence has been segmented into words, the words have been tokenized and normalized, we can make a simple [bag-of-words](https://en.wikipedia.org/wiki/Bag-of-words_model) model of the text. In this bag-of-words representation you only take individual words into account and give each word a specific subjectivity score. This subjectivity score can be looked up in a sentiment lexicon.

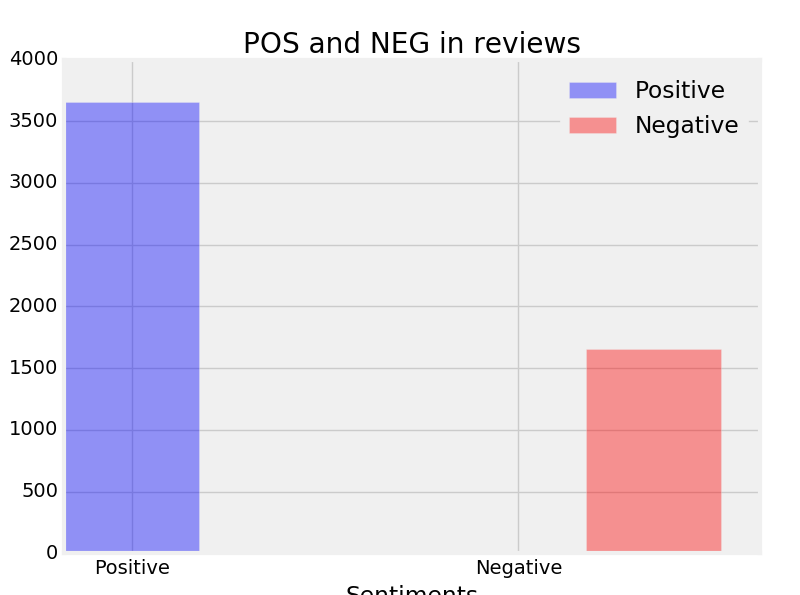
After all of these pre-processing steps:

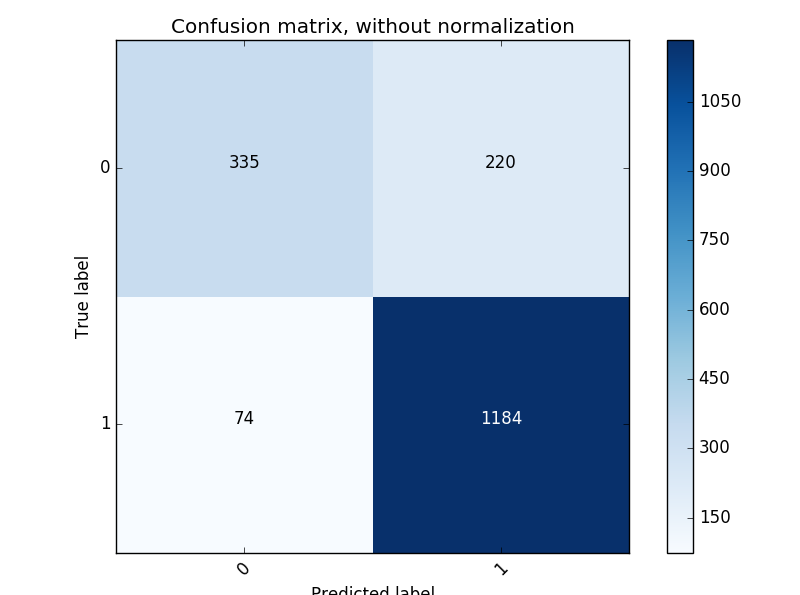
* Training of model is done on the basis of three different classification techniques namely
* Naive-Bayes , SVMs(support vector machines), k-neighbor.
* The accuracy and confusion matrix is calculated .
* After this a compound classifier made which gives the final result on the basis of majority of the above classifiers.

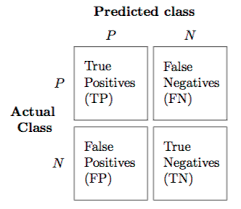
**REFERENCES**

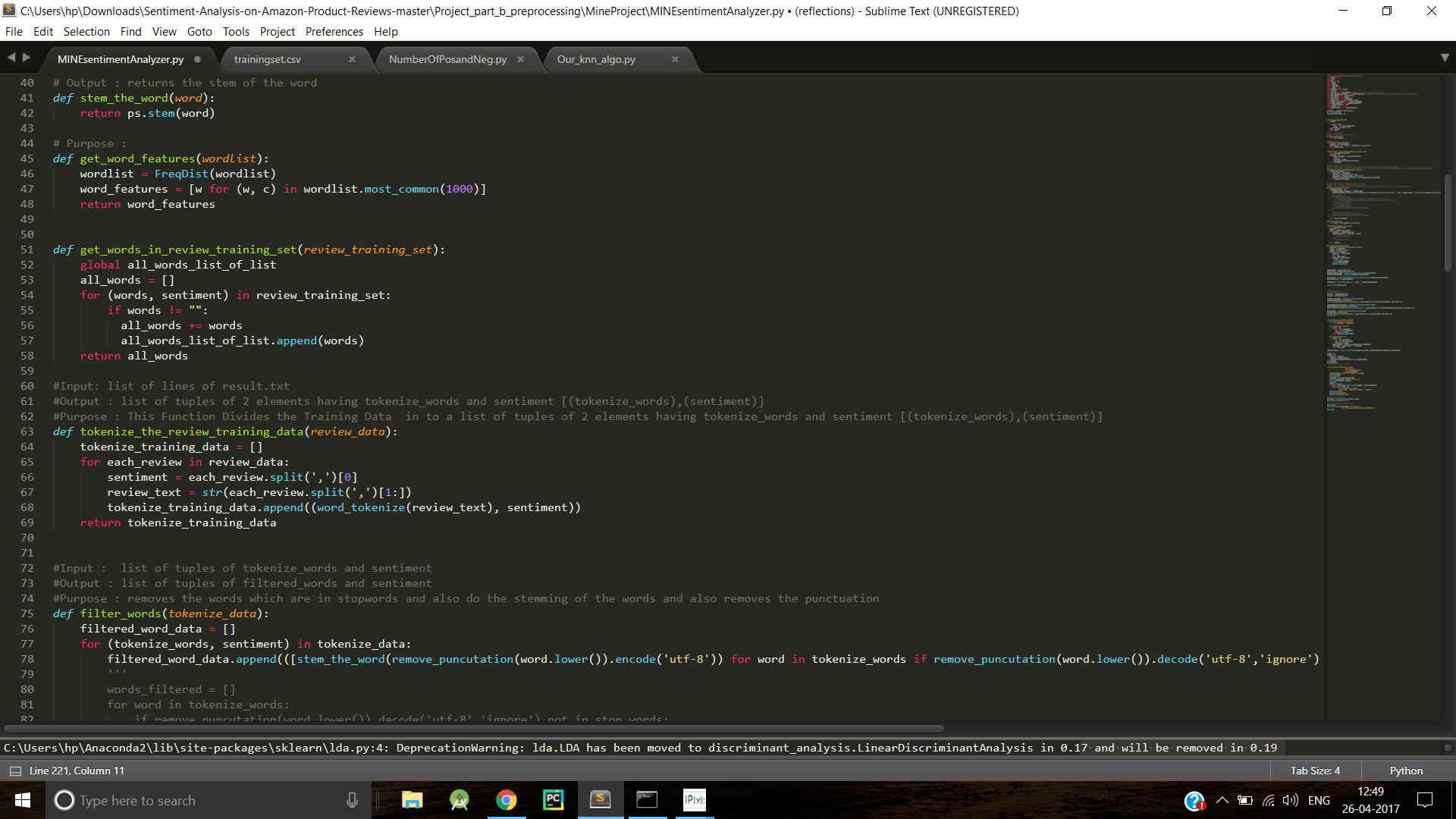
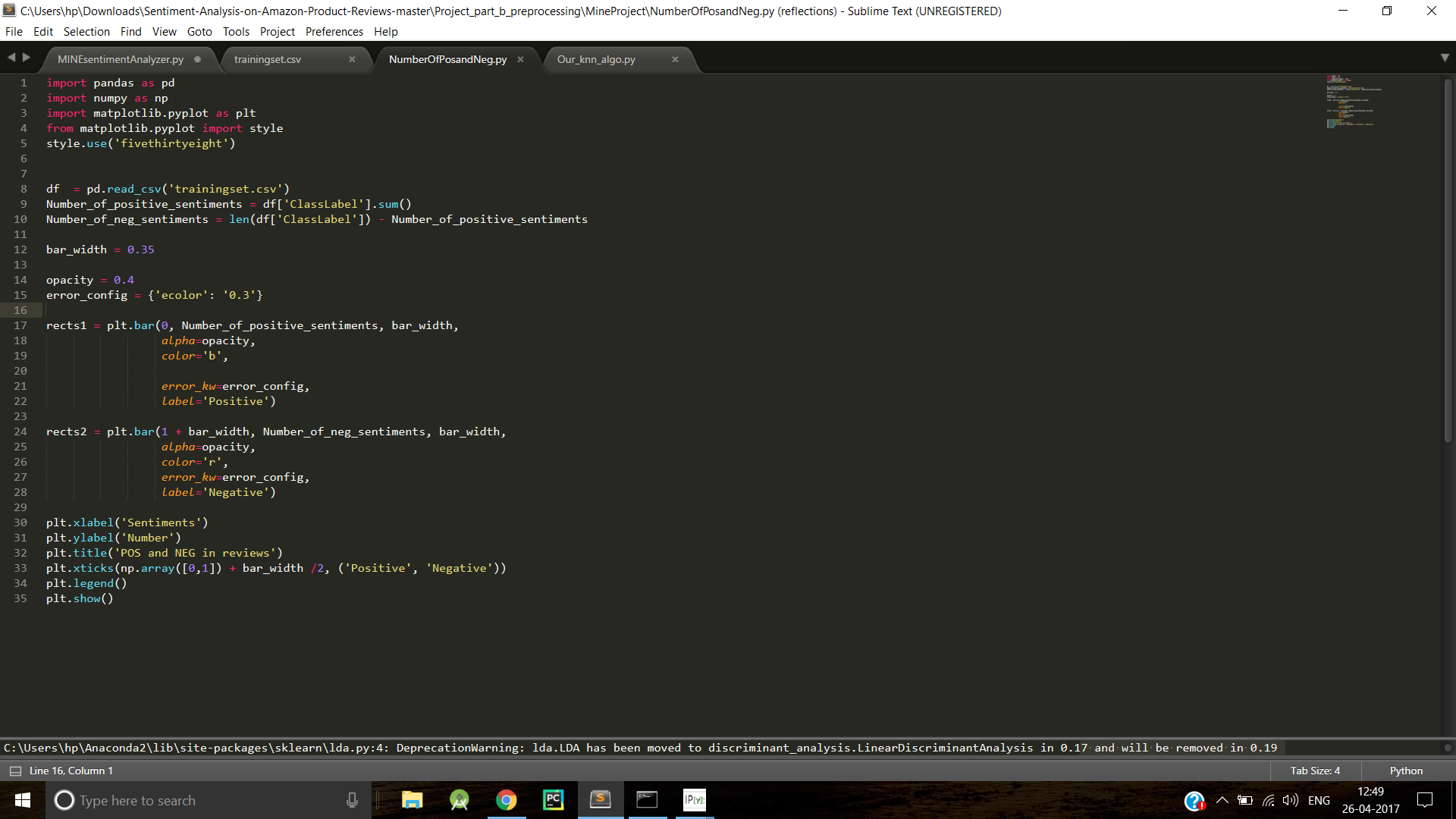
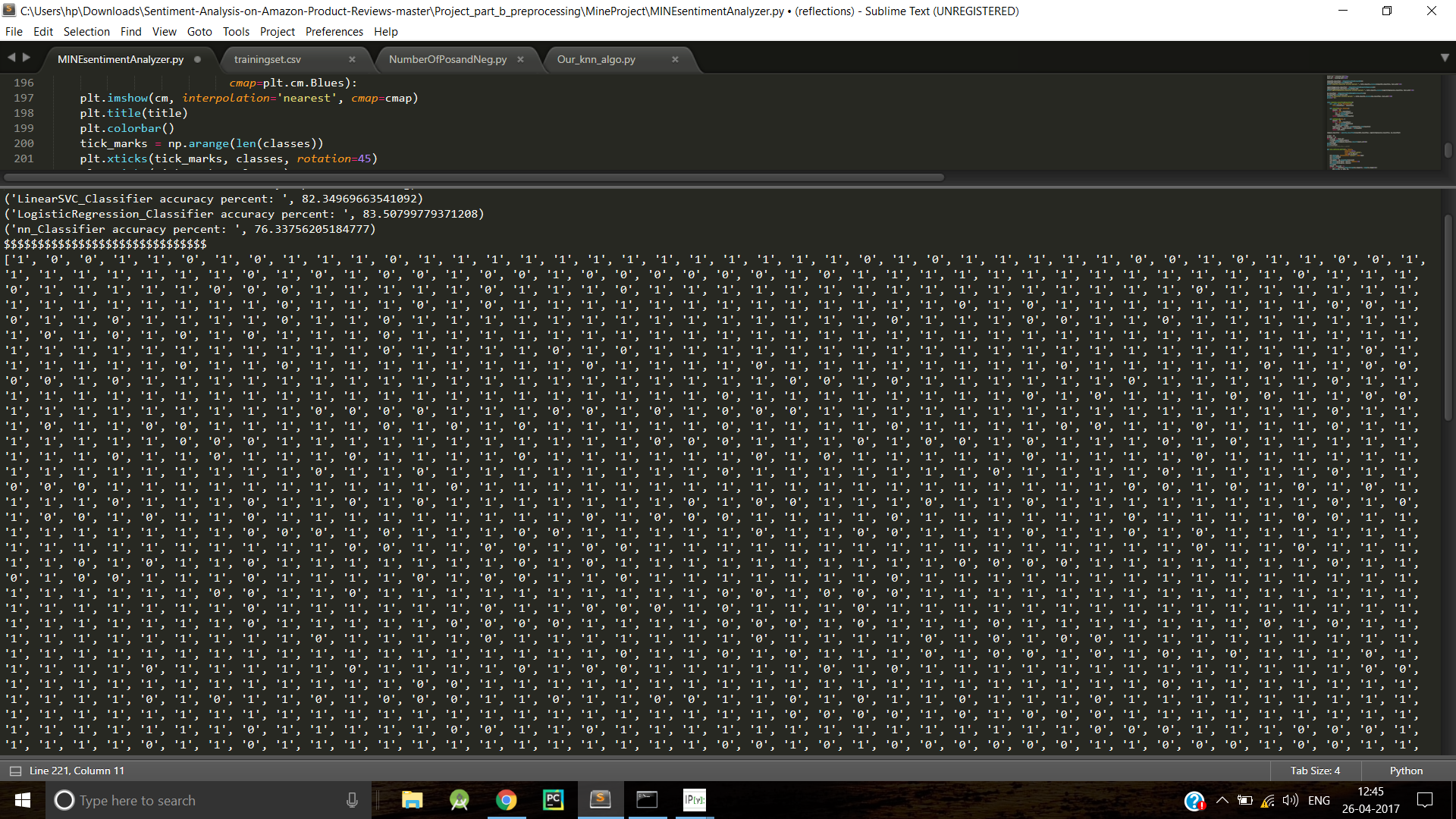
* <http://ataspinar.com/2016/02/01/sentiment-analysis-with-bag-of-words-part-2/>
* <https://en.wikipedia.org/wiki/Hyperplane>
* <http://ataspinar.com/2016/01/21/sentiment-analysis-with-bag-of-words/>
* <https://archive.ics.uci.edu/ml/datasets/Amazon+book+reviews>
* <http://scikit-learn.org/stable/>
* <http://scikit-learn.org/stable/supervised_learning.html#supervised-learning>

**GRAPH PLOTS**

****

****

****

****