**Pattern recognition**

**Assignment #2**

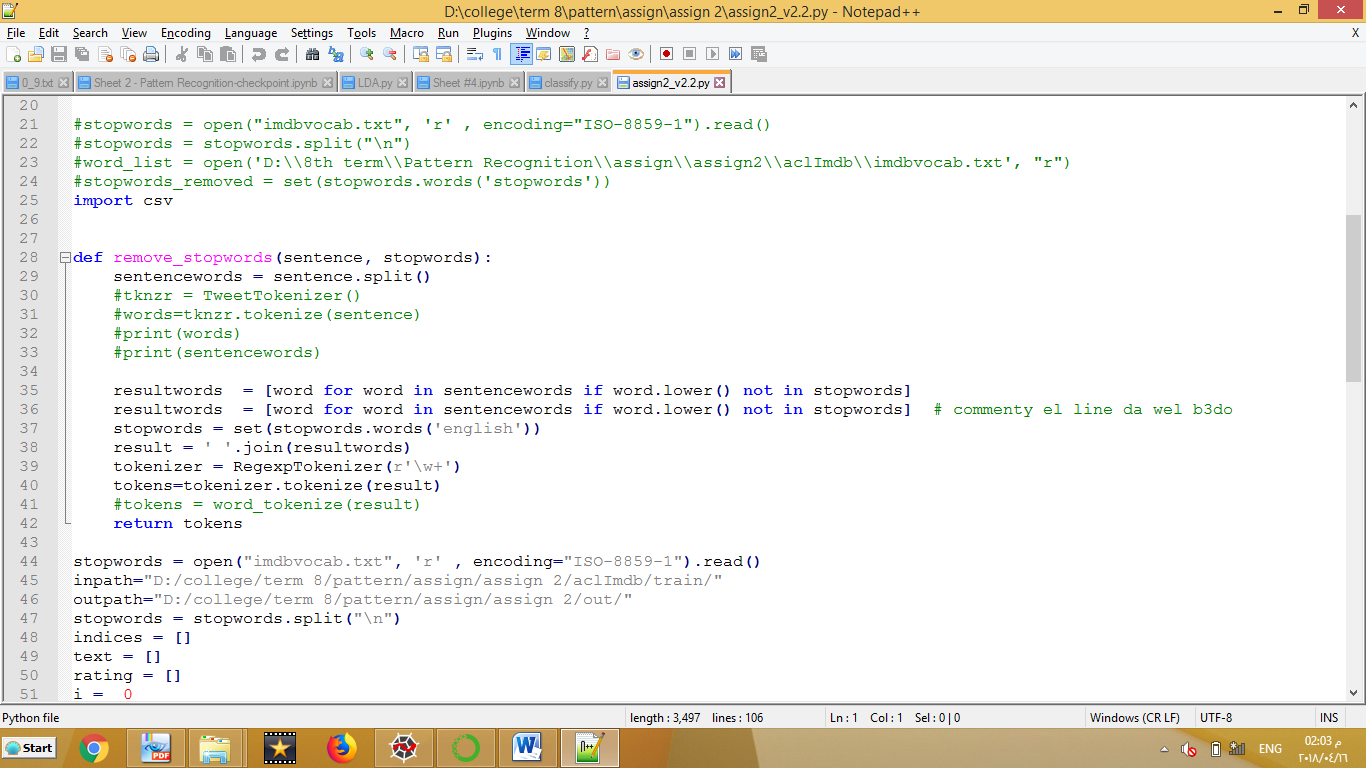
**NLP**

**Nada Hazem**

**Mohamed Samy**

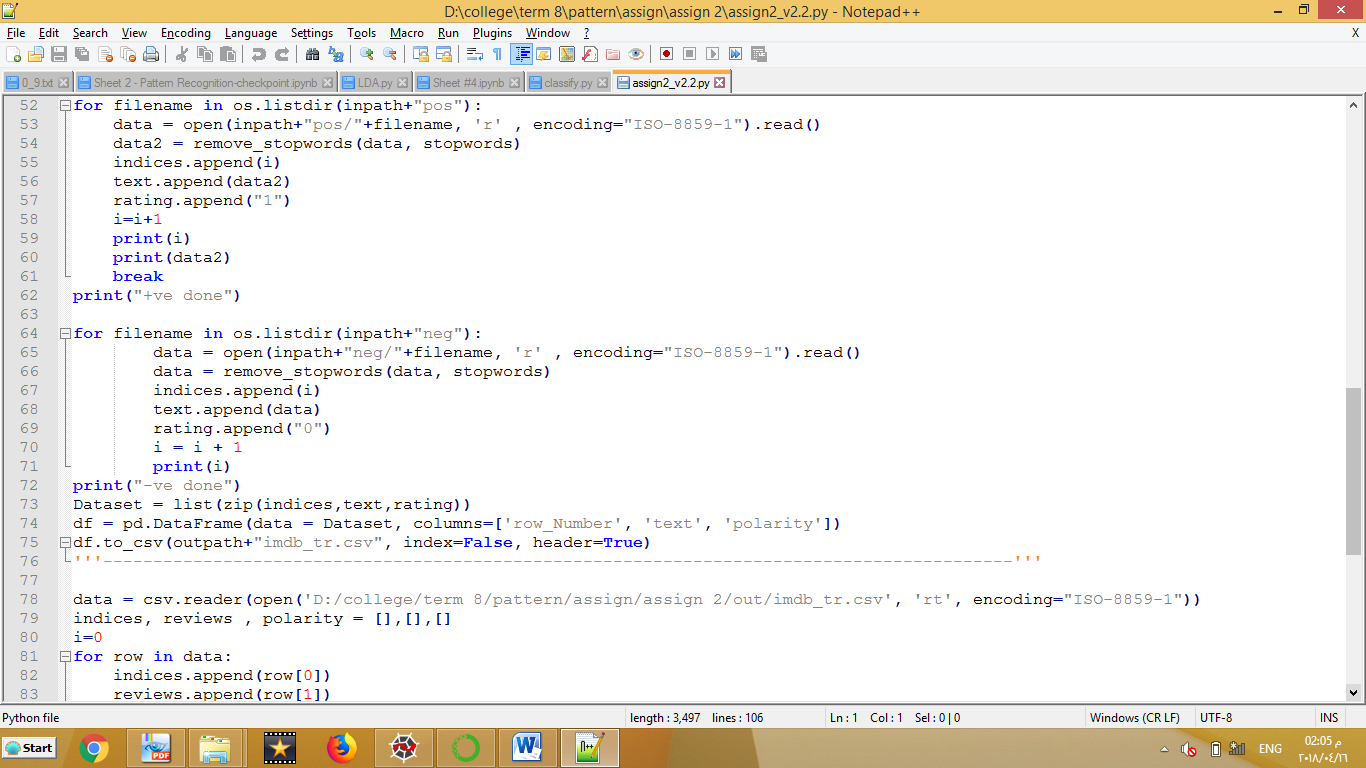
**Seif Eldin Samer**

**Codes used:**



***Stopwords*** variable reads the file of unwanted words in the reviews

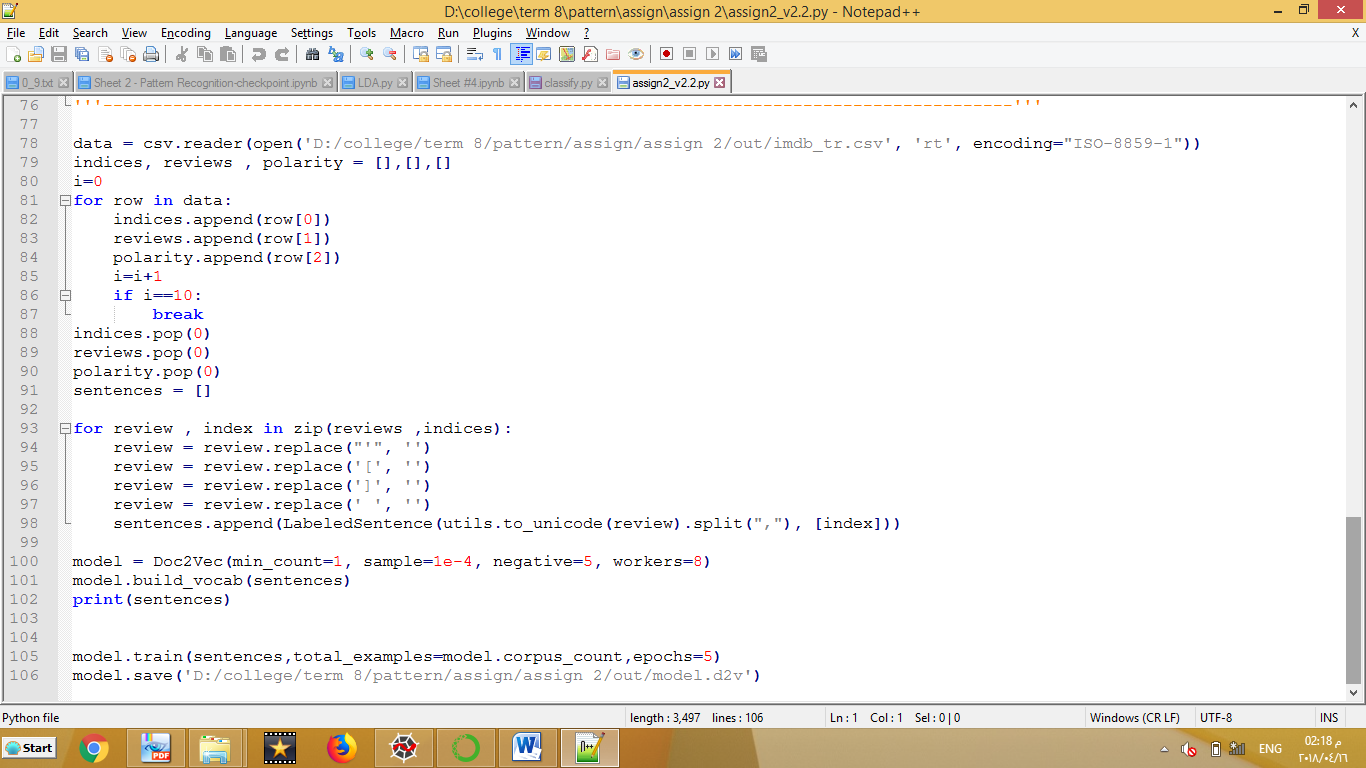
The ***remove\_stopwords*** function removes stop words from the reviews and tokenize the review into tokens of words and return tokens

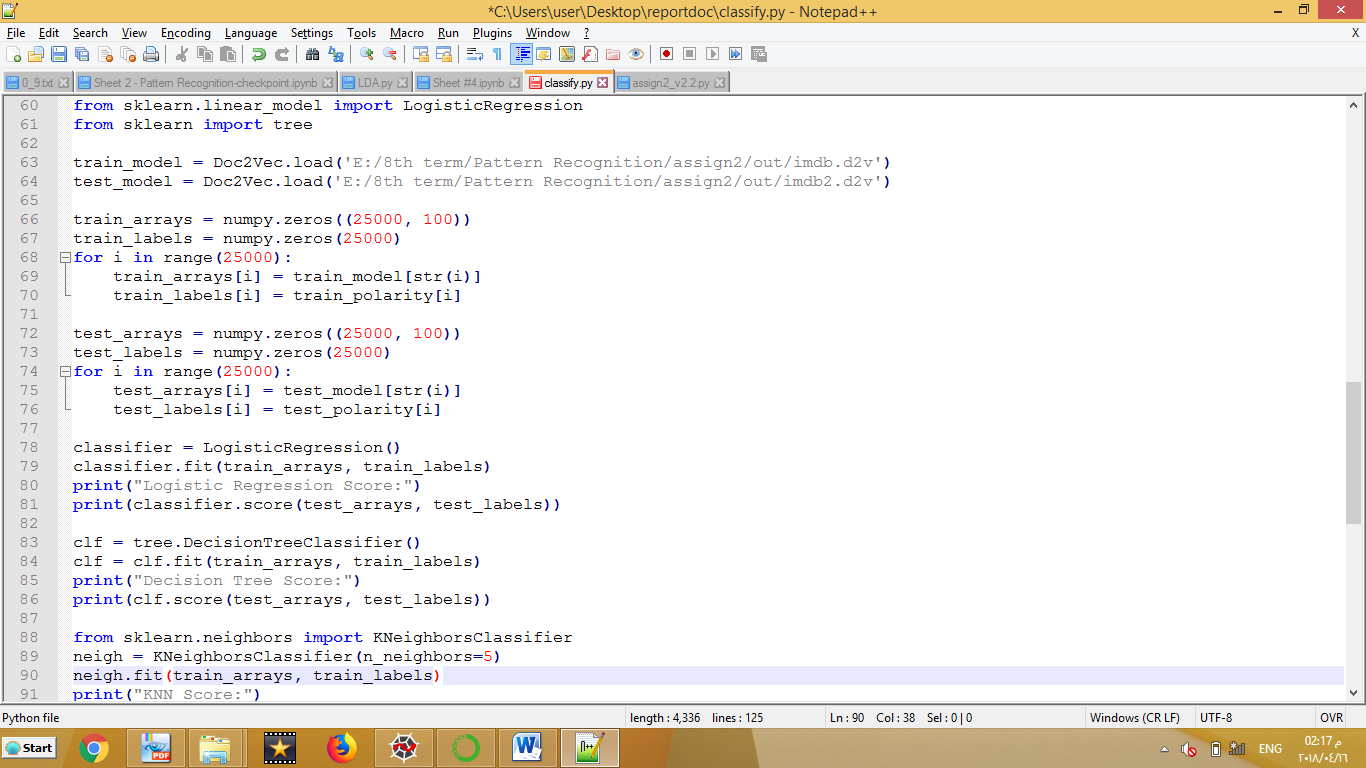


The first for loop reads the positive folder then the negative reviews. Once for the test data and once for the train data.

Then writes the data back into files to easily read it again.

**Review to vector code:**





**Preprocessing effect:**

We only used tokenizer, we removed unwanted words from the reviews

**Feature choice:**

-Doc2vec

-Word2Vec

**Classifier choice:**

we used :

-knn

-random forrest

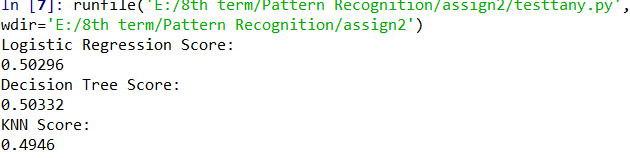
-decision trees

-naïve bayes

-linear regression

-SVM

***Results***



At 3 neighbors

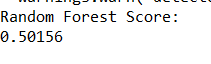
C:\Users\user\AppData\Local\Microsoft\Windows\INetCache\Content.Word\knn5.png

At 5 neighbors



C:\Users\user\AppData\Local\Microsoft\Windows\INetCache\Content.Word\rf1.png

Random forest with depth=2



Random forest with depth 4 (max depth)

**Same Tests with word2vec**

