Assignment : Application of TF-IDF for spam detection
Name : Anuvrat Tiku
ID : 010822084
CMPE 239 Data Mining

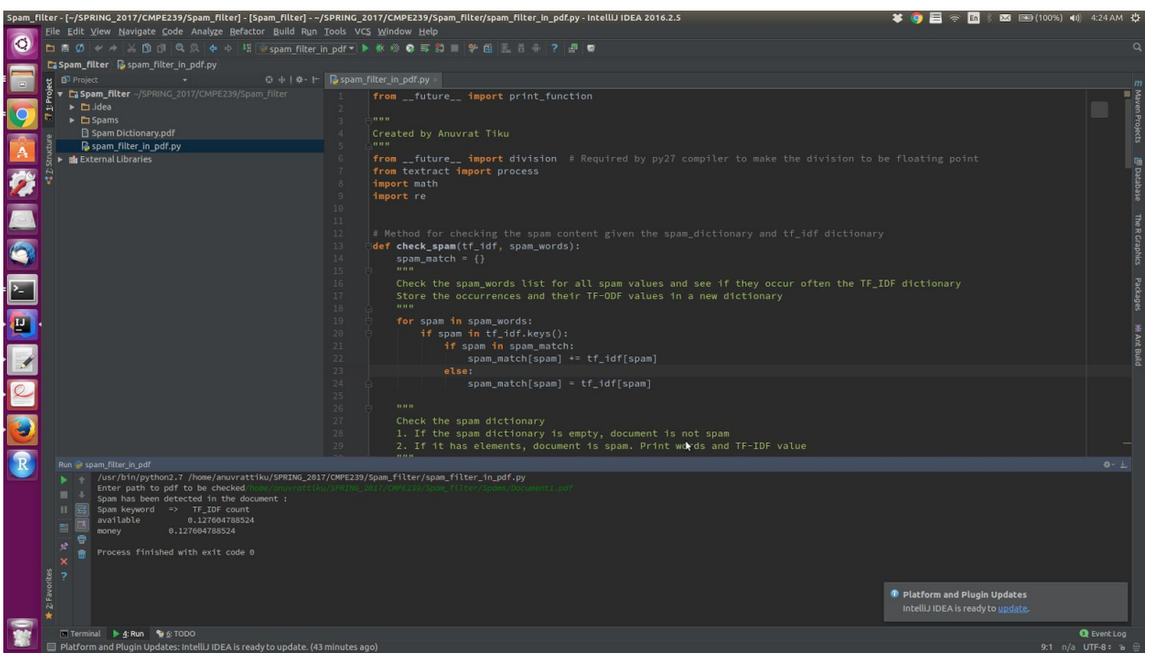
I created a program that takes input as a set of paths to build a corpus from. All the files which have to be scanned for spam are fed to the program.
The program created a corpus and computed IDF for the corpus with respect to the file. The program also computed TF with respect to the file. TF*IDF = TF-IDF. The product of these two values gives us the TF-IDF value.

Every file to be tested is passed into the program and it detect if the file is spam based on the dictionary. If a word in the spam dictionary is also found in the TFIDF dictionary and the file, the file is declared spam.

My program detected all files as spam. Below are the spam keywords found in each file and the TF_IDF value of each spam.

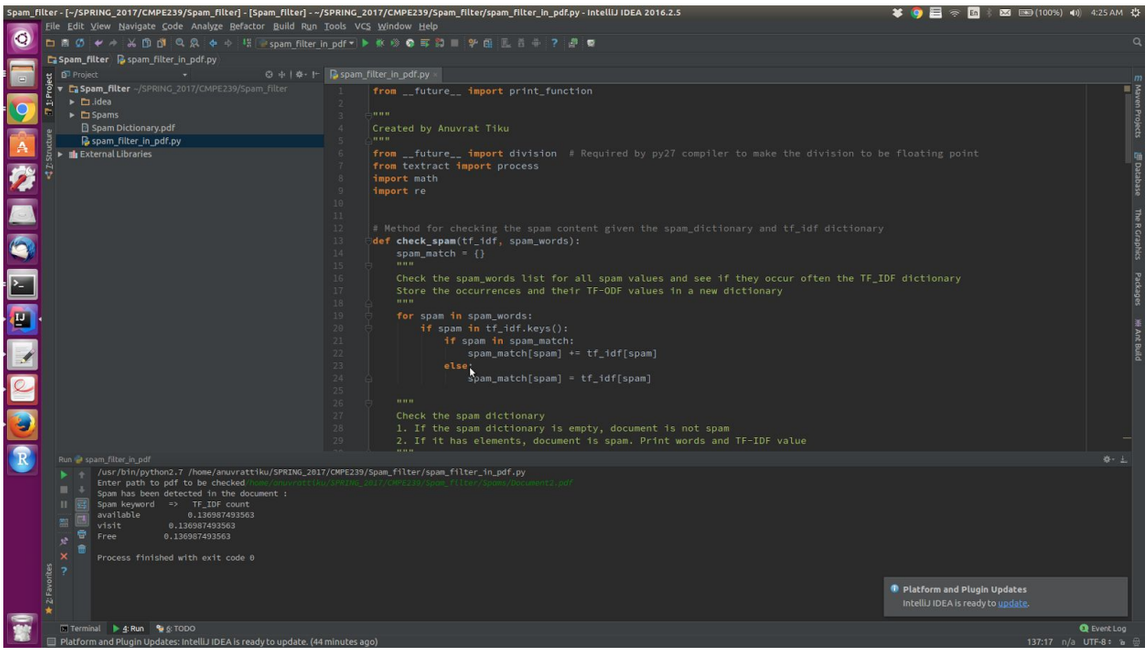
Document1 : Spam

Spam keyword => TF_IDF count
available 0.127604788524
money 0.127604788524



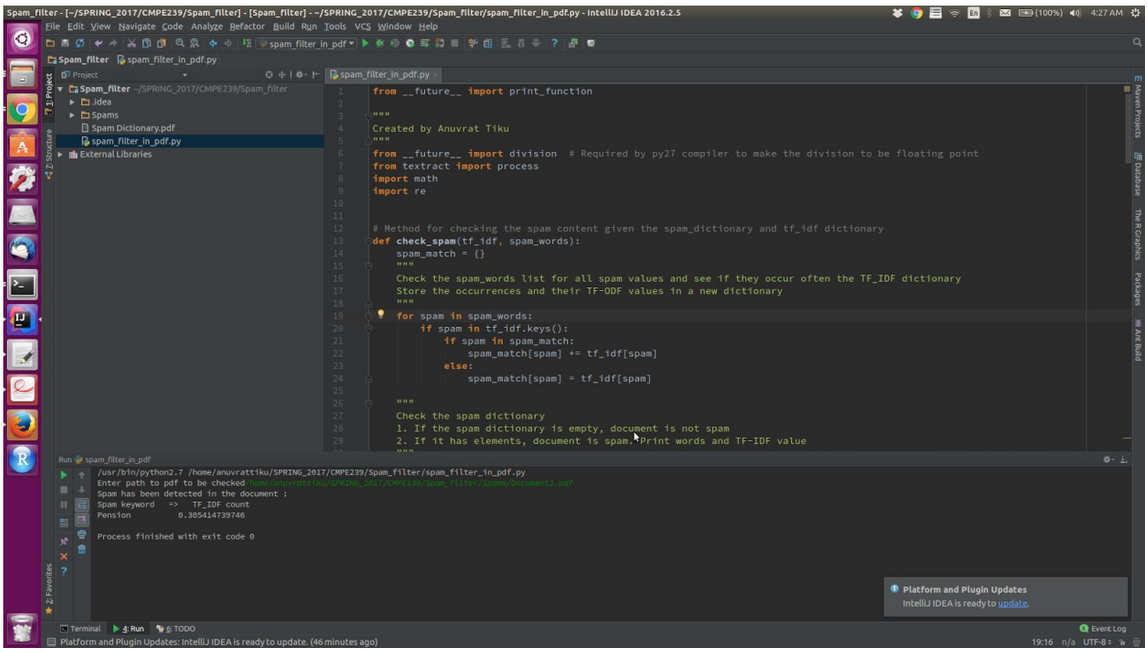
Document2 : Spam

Spam keyword => TF_IDF count
available 0.136987493563
visit 0.136987493563
Free 0.136987493563



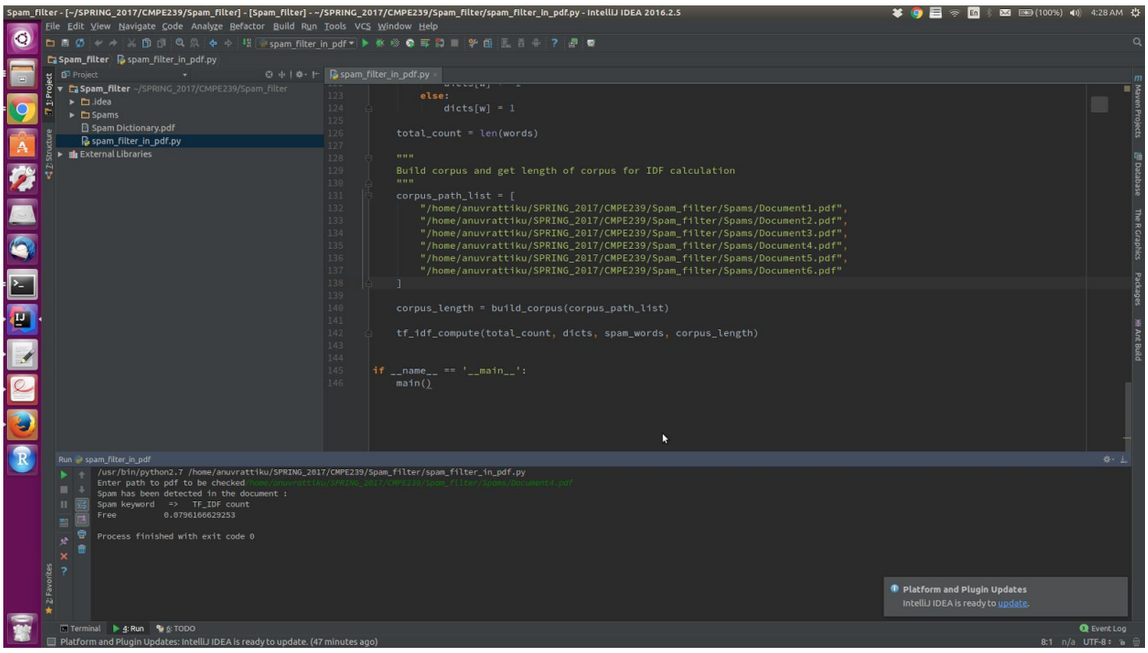
Document3 : Spam

Spam keyword	**=>**	**TF_IDF count**
Pension		0.305414739746



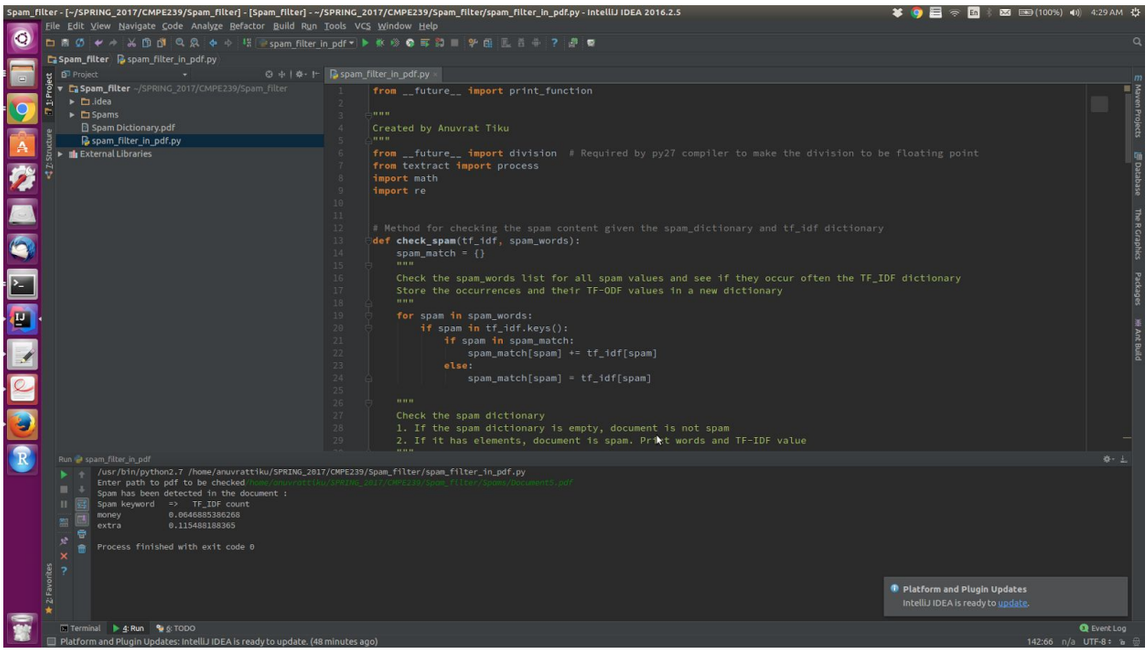
Document4 : Spam

Spam keyword	**=>**	**TF_IDF count**
Free		0.0796166629253



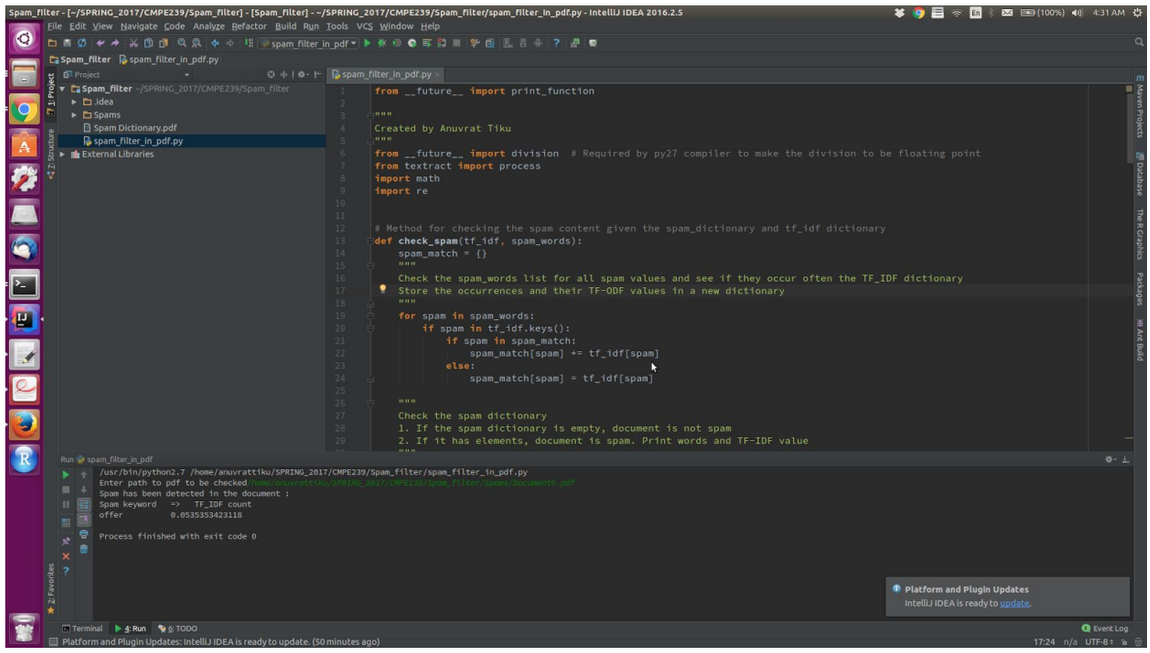
Document5 : Spam

Spam keyword	**=>**	**TF_IDF count**
money		0.0646885386268
extra		0.115488188365



Document6 : Spam

Spam keyword	**=>**	**TF_IDF count**
offer		0.0535353423118



All 6 docs are spam.