**Project Report 12**

PLSA: Text Document Clustering

**NAME: DHIVAKAR.R**

**COURSE: AI and ML**

Question:

Perform topic modelling using the 20 Newsgroup dataset (the dataset is also available in sklearn datasets sub-module). Perform the required data cleaning steps using NLP and then model the topics

1. Using Latent Dirichlet Allocation (LDA).

2. Using Probabilistic Latent Semantic Analysis (PLSA)

**Prerequisites**

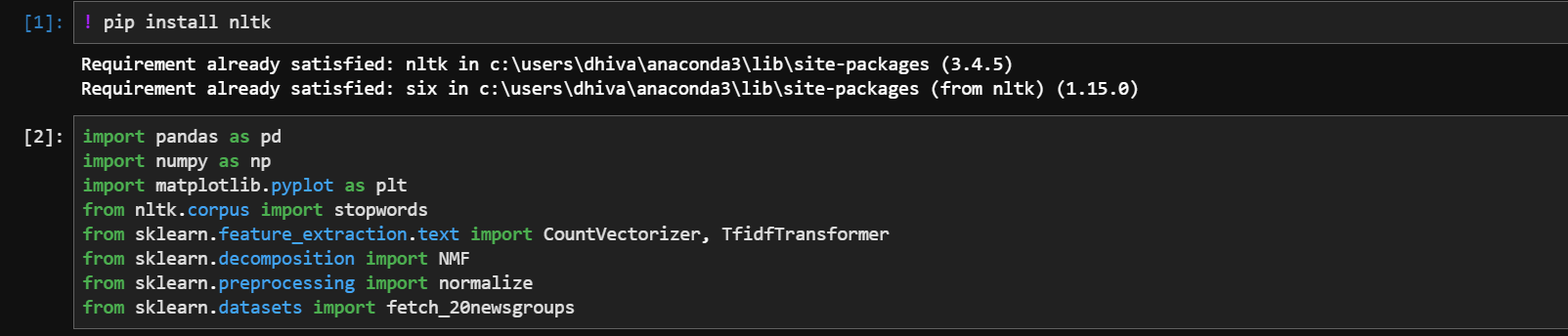
What things you need to install the software and how to install them:

Python 3.6 This setup requires that your machine has latest version of python. The following url <https://www.python.org/downloads/> can be referred to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: <https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-asan-internal-or-externalcommand/> . Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.

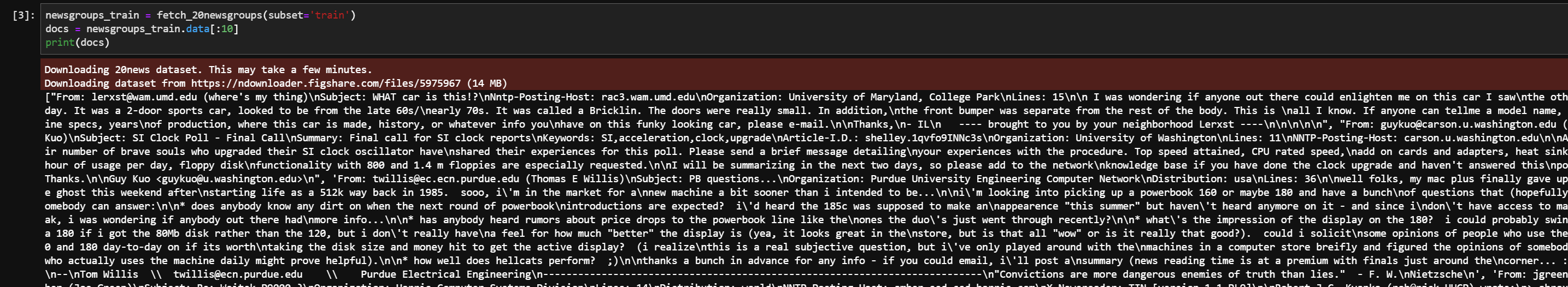
Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url <https://www.anaconda.com/download/> You will also need to download and install below 3 packages after you install either python or anaconda from the steps above Sklearn (scikit-learn) numpy scipy if you have chosen to install python 3.6

Dataset Link: <https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html>

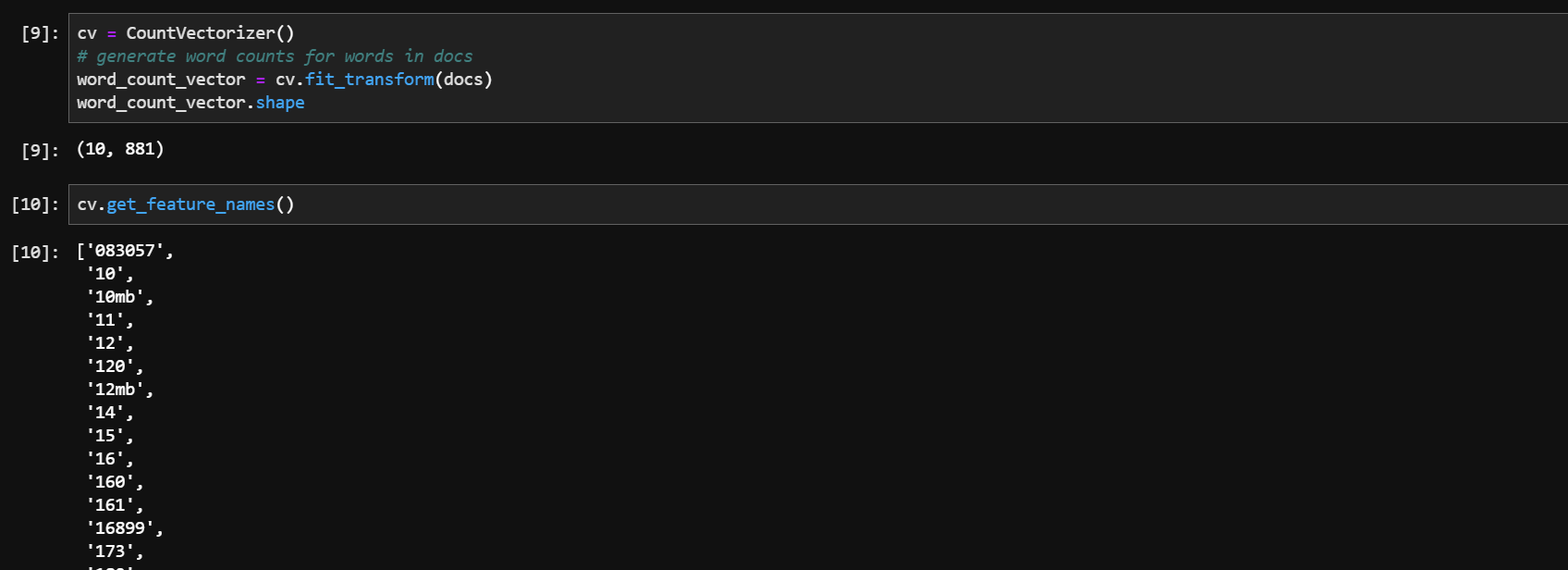
Load all required libraries and Dateset



Downloading dataset



generate word counts for words in docs



get tfidf vector for first document

