NATURAL LANGUAGE PROCESSING: AUTHOR STYLOMETRY 06

Hello guys, see in this video we will prepare the data set. So as the instructor said we will go to the project Gutenberg, you can see the link here. If you go to this web page you will see there are many E books which you can download for free as a text version so we are going to use these E books for our analysis so for you what we have done we have created an Archive which contains all the which contains selected papers which we are going to use for our analysis. So you need to download an unzip the Archive or some papers containing eighty five documents that we will use for our analysis, this eighty five documents have been extracted from this project Gutenberg E book version. When you unzip the Archive it will create a directory called data, so you will have a folder name data, this will be our working directory so what do you have to do? You have to whatever python code you are going to create just create it in this data folder, we have the data will be using nltk library which comes with the anaconda cloud so you can just try importing nltk can your python console and you can see that it is working. Make sure you download all the collections of nltk using nltk download so before we can proceed with the analysis we need to load the files containing all the eighty five papers into convenient data structures ok, the first step is to assign each of the eighty five papers to the proper set since all the files have extend a format we can easily assign each paper to its corresponding authors for this we will use the dictionary data structure so the idea with the dictionary data structure is comes with the key and the value payer so with each key you can store, store its value and you can easily access this particular value using the key so if some of you know the basic data structure then you must know hashing also, in hashing what we do we create a hash of the particular data and whenever we want to extend the data we use that hash value to extend the data so the dictionary data contains the key value payers in Archive key would be the authors name and the list of paper numbers will be the values associated with these keys so let’s see the data world.