NATURAL LANGUAGE PROCESSING: AUTHOR STYLOMETRY 09

So you can see here i have downloaded all the collections for nltk you can see its written whatever i needed to download everything has been installed here and will go to our program let me just zoom it we can see let me ok yeah everything is visible now i define what my function is exactly and what this paper dictionary is doing, i will just save it just run it again so everything looks perfect by running it i have created the papers dictionary so by if i just type in my let me just zoom this one ipython console here you can see the ipython console. If i type papers you can see the dictionary Madison contains a list of all his papers Hamilton contains list of all his fifty two papers jay contains four papers, shared are three and these are twelve disputed papers so our dictionary is ok let me clear it ok so next what you have to do, you have to call this function so that we create the actual dictionary that we wanted to create our main dictionary is dictionary key value will be the authors name and the list i mean the value the key will be the authors name and the value will be the string of words i mean all the sentence all the words in a single string so that we can analyse it easily. So that what we are going to do we will create a dictionary out of authors so let me write here let me give the name of the dictionary here is federal list by author ok so this is the dictionary and let me write the loop for author comma files in papers dot items ok so let me just try this once what exactly it does so let me let's see the let see my ipython console. I will just try this console for i comma j and i already created this papers dictionary so papers dot i items ok i will just print i and k ok you can see it is printing all the values Madison and in his list, Hamilton and his list and jay and his list shared all the list, the list disputed and the list so its work perfectly i wanted this only so i will go to the next line what i am going to do so authors will give me so what i is? I will give me the name of the author and j is giving me the list so author here the variable author will give me the name of the author and files, files here will give me the list so files contain list author contains authors name so what i am going to do the dictionary which the entire dictionary which is just created i will use this federal list by author and key is what? Key is the authors name here it is author, key is author so i will create a author through which i am looking through and what will be the value, value is the list. So what i will do? What i am going to do? I am going to call this function now read files and i will i am going to call this function by giving the list as a parameter so the file name will be my will be the list so files what it will do? The list will go here ok then all the numbers, numbers in this contain numbers so for all the numbers it will read all the files for that particular author and then contacting it all the words strings whatever it is into a single list. Ok and that list i for a particular author i will use it to analyse. What i am what my idea is? I will use that list i will count all the words of a particular length and i will plot the distribution that's it ok so you will now you know what i am exactly doing why i am calling this function here, so this will give me the list of all the words for a particular author so my dictionary will contain everything here ok i am getting some errors so what is the problem is i put here file name rather than file ok lets run this cool its working so i will just try to print this so that everything looks perfect so for author in papers print this my federal list by author the author author is my variable here i will just print something some sliced hundred character so that lets see everything is

working or not? I am getting something it means the dictionary is perfect. I will just delete this. Ok my dictionary has been created, now the only thing i have to do is on this dictionary i have the data i have created my i have done my pre processing i will do the analysis i will count the word and their length and i will plot the distribution and i will see it whether there is difference or not. Let’s see this in next video. Thank you.