NATUTRAL LANGUAGE PROCESSING: AUTHOR STYLOMETRY 10

So everything is ready we have created the dictionary federal list by authors which contain all the data where key s the authors name and the value is the authors data. Now what we have to do, we have to create words distribution word length distribution exactly so will count the word with the particular length will count them i will create a graph out of it. So let me create a tuple name author authors and the first one is Hamilton second is Madison third is disputed then we have jay and the shared documents ok my tuple is there for authors, i will create two dictionaries author tokens ok let me explain this dictionaries let me create first and length distribution so what i am going to do i go through all these authors and whatever dictionary i created in this dictionary the key value is the authors name even see all this authors name are the key values here i will take the i will take the strings the data contains everything and i will remove all the punctuated things other letters only the words will be there i will take only the words because what we want is the word and their length so first i will do i will create a loop for author in authors is my authors list yeah tokens is equal to i will create a token tokens is equal to nltk dot there is a method for it word tokenize so what it does? It takes a string and create tokens out of it i mean each word will be each string each token tokenize so everything for example punctuated letters words everything will be tokenized, tokenize what i will take the data, so data is present in this dictionary federal list by author and i am taking for each author this import nltk ok so this is done lets will after will filter out the punctuations for that there is an nltk library for it so i will take the author tokens ok so the dictionary i created ok and the key will be key value will be the author and now here is the tricky part i too remove all the punctuations here so i will do it in a very small code so let me let's see this first for token token in tokens you know what tokens is here ok if any here it is c dot alpha this is a function for c in token so what i am doing for each token in tokens i am checking if it is and alpha alphabet it is made of all the characters only or not any other special characters or numbers something if it is only alphabets then i am i will create a list of it otherwise i will discard it. So for token for token in tokens put the token there so put that tokens in this list that's it so this line is nothing else you can do it in other way also i just wanted to put it in a single line so for token and tokens if that it's an alpha correct for c in token whatever token i am getting here is a string so for c in token means all the words in this string if it is a made up of only alphabets i will put this token inside the list otherwise i will discard it so then i got the words out of the for each authors now i just have to calculate the length of it so token length which i created token lengths ok token lengths will be again let me just write it again for token in author tokens of what authors, for token in author tokens what token put what length of it. So i got the i am iterating over all the words which i got for each authors and i am just getting the list of lengths that's it great i got the list after that i want to get the length distribution of author so i will create a list out of it length distributions authors ok and i call nltk dot frequency distribution this is the function name and i will pass the token lengths and i will just plot length distributions of author let just plot it so maximum length i will put as fifteen up t fifteen words the length of the fifteen words sorry title is equal to author, author means what even it will plot it will show for which author this plot is so this dictionary is

length distribution yeah that's right it was throwing the warning error now its fine lets let me just make it little bit big and just run it ok i am getting some error str object has not attribute alpha sorry just should be is alpha because i am checking whether it is an alphabet or not. Ok let me sorry let me run it again ok you can see the plots here you can see there are plots for Hamilton Madison disputed and jay and shared so there are exactly five plots, there are three authors Hamilton, Madison and jay the shared plot is for documents which were shared by Madison and Hamilton and disputed are those which we don't know. Now see this see the plot it looks exactly almost same but it's not exactly if you see the plot of jay the samples are the word count i mean they are there are two words two length words there are twelve hundred words around twelve hundred words around of length two and so on it goes distribution goes like this for jay it's definitely not what this disputed plot is it means that whatever the documents are there in the disputed are definitely not written by jay you can see there the beginning and the tale is definitely different for jay and disputed but if you see and also if you see the shared one its look similar to Hamilton more as compare to the Madison one but if you see the disputed one you cannot you cannot say who exactly wrote the disputed one either it is Madison or Hamilton but you can definitely say that jay didn't write you can see the plots are different the beginning part is little bit similar to the Hamilton one you can see the beginning of this, this plot of disputed but the tale is more or less similar to Madison or even Hamilton so we cannot even generalise with this but this is a very fun way of seeing what exactly is going on or is it any way related with this at least we were able to say that this jay definitely not didn't write these disputed. This analysis was given when it was given it was it didn't account for the authors way of writing things how they choose words what kind of words they choose they are different there are so many research on this and still not idea so people are still working on this stylometry and now they now there are so many tools through which you can analyse the authors signature and you can predict or you categorised authors documents by analysing their styles. So it has, so this what the program we did was just to see just to motivate you this idea of stylometry you can exactly actually given a document you can do some analysis and get the authors name if you if you have previous data of it. So this is exactly known as data analysis, now a days we have a lot of data you analyse the data and you whenever you get a new data you try to infer from the past data whether this new data is something or not this was the very very main example of doing it because this is the only this was the quickest and simple way of doing it. If you go little further it will become very complex so since this is the very basic course we dint want you to go through the complex one so we did this basic analysis to understand the data at least and we can see the plots and we can at least infer something through this. There are so many methods of doing stylometry you can go to the web and you can just type stylometry and you will get a lot of lot of way please see this nltk library this is this is a very good library for analysing text, documents actually it is known as nlpk analysis so please see this libraries if you have any problem post it on discussion form. Thank you.