

## **NATURAL LANGUAGE PROCESSING: AUTHOR STYLOMETRY 07**

So i am on this web page you can click on the link given in the description to download the data so if you click you will get to this page, you can just click on the download button to download the data, i have already downloaded it, you will get a archive version of the data just click on it so to unzip it and you will get a folder name data so i have this folder name data i will just open it and i can see all the texts so name of the files are federal lists underscore one two three four and so on, these are the files which have been written by some of the authors and our key idea is to understand their style of writing so that we can get some ideas of how we can distinguish different authors using their style of writing. So our analysis we will use these files as input and we will plot the word length distribution and through that we will try to make some sense that whether different people write in a different way or not so these are key idea and these are input file please download it, see it once and let's start with the code so we are splitting the codes into five categories so these categories are if these files for example are we are going to split categorise these files into five categories, first there are some fifty one papers known to have been written by Alexander Hamilton and there are in these files there are fourteen papers to have been written by James Madison, four of the five papers are known to have written by john j and three papers were co-written by Hamilton and Madison and there are twelve papers disputed between Hamilton and Madison which we don't know who exactly was the author of these twelve papers so our idea is to understand the style of writing of these papers and then try to analyse this twelve papers whether they were written by Hamilton or Madison so we will use our key idea which we discussed previous videos that we will try to plot the word distribution i mean word length distribution and with this we will try to see whether whether it is really that people people have different way of writing people have different way of writing papers such that they use different length of words differently so in next video will try to write a code for this and will plot the graph of this length word distribution will try to analyse the data. Thank you.