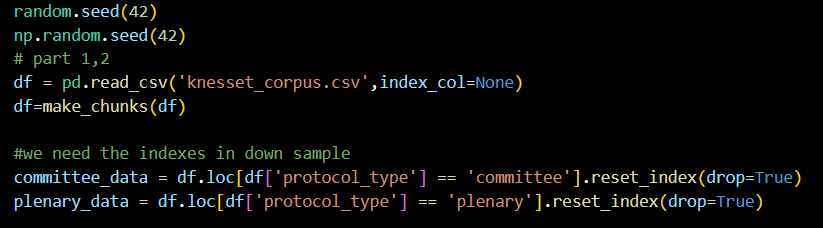
Natural Language processing – HW 3

Mias Ghantous – 213461692

Faisal Omari

Section 1,2.

As we see we split the data according to his type after making the chunks



Make\_chunks:

1. we divide the data into groups by using the function which takes the name of the columns as a parameter and then make a group for every possible compensation of value from those columns.

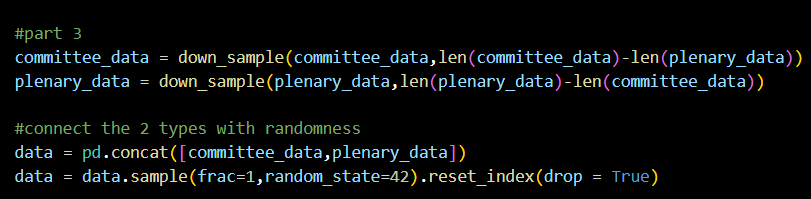
2. second we use which takes a function as a parameter and do this function for each group, we sent which extract the relevant data from the group and put them into a data frame.

3. apply concatenate the result by herself.

4. in the code we group by 'protocol\_type' and 'protocol\_name' because we want the keenest number in the classification

Section 3:

As we see we call down sample for both of the data and then connect them and then randomize it:



For chunk size 5 we get that:

The size of the committee data before the down sample was 5872 and also after.

The size of the plenary data 14095 and after the down sample we have 5872

Section 4.

1. We used TFIDF Victimizer because he can give more accurate predictions:Take a photo of the prints for both of the vectors
2. for our vector we used a 7 words:



Because those words have a lot of occurrences in plenary documents compared to committee.

Also we used the helped the accuracy a lot:

Show to pictures one before and one after

Also we tried to include the average length to the feature vector but it made the accuracy worst

Show pictures

Section 5:

1. in knn also we used linear kernel for the SVM.
2. And here are our results:

Show the picture of the results

Section 6:

We predicted using the SVM that belongs to the BoW because he has the best accuracy.

Section 7 (questions):