Assignment(EX08)

Column 4 i.e. MetadataFrom had list of all senders. When I tried to extract all senders and checked the number of times each sender have send the email, I found that there were senders with just one word like H, nan(this is a blank space in MetadataFrom) this both are irrelevant data which later I removed from my list.

After cleaning that list I found top 10 sender as 'McHale, Judith A','PIR','Verma, Richard R','Slaughter, Anne-Marie','Valmoro, Lona J','Jiloty, Lauren C','sbwhoeop','Sullivan, Jacob J','Mills, Cheryl D','Abedin, Huma'.

From all of the above senders the McHale, Judit A send most email i.e. 1380 times.

Later I represented all the top 10 senders using matplotlib library and created histogram.

Below is the histogram-

A screenshot of a cell phone

Description automatically generatedFurthermore, I extracted all the data which was asked in question point 4.

That was mostly from column A, I extracted that data, cleaned it by removing numeric values, punctuations, blank spaces.

After that I used that words and perform wordcloud of that clean data. After doing that I was able to find most used word in entire text like- case, unclassified, department, data, doc, etc.

Following is the wordcloud.

A screenshot of a cell phone

Description automatically generated

Bonus (Google findings)

When we compare the analysis done above from .csv file provided by the professor, it is clear that Huma Abedin has send maximum emails.

According to FBI agents there were 340,000 emails send by Huma Abedin.

This emails were found on Weiner’s ipad, iphone and laptop devices. The reports also say that these mails were send from Clintons private email.

Reference- https://www.dailymail.co.uk/news/article-7553413/FBI-agents-thousands-Clinton-emails-described-oh-s-t-moment-new-book-claims.html