**Case study – Enron email classification**

The Enron corpus is a large database of over 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse.

We have cleaned the data and provided you with only sent messages. Please see the sample below

|  |  |  |
| --- | --- | --- |
|  | file | body |
| 0 | allen-p/\_sent\_mail/1. | Here is our forecast\n\n |
| 1 | allen-p/\_sent\_mail/10. | Traveling to have a business meeting takes the... |
| 2 | allen-p/\_sent\_mail/100. | test successful.  way to go!!! |
| 3 | allen-p/\_sent\_mail/1000. | Randy,\n\n Can you send me a schedule of the s... |
| 4 | allen-p/\_sent\_mail/1001. | Let's shoot for Tuesday at 11:45. |

**Fields**

1) file : file name under mailbox directory and the hierarchy is constructed as <sender\_name>/<mailbox\_type>/<file\_number>

2) body : email body of sent messages

**Download link to dataset** - <https://drive.google.com/file/d/1yNMKT2-DoLCZMLlrAdqy6iNWQVCYq1OS/view?usp=sharing>

**Instructions**

Download the data from the link.

1) Analyse the corpus

2) Build a text processing pipeline to clean the data

3) Identify and extract top 10 senders

4) Build a classification model to predict sender given email content

5) Document your solution as

                - Solution developed

                - Process diagram (explaining all the steps you took to arrive at the solution)

                - Explain model performance

                - Visualisation to support your solution