**Enron Email Dataset Coding Exercise**

<https://www.cs.cmu.edu/~./enron/>

1. How many emails did each person receive each day?
2. Let's label an email as "direct" if there is exactly one recipient and "broadcast" if it has multiple recipients. Identify the person (or people) who received the largest number of direct emails and the person (or people) who sent the largest number of broadcast emails.
3. Find the five emails with the fastest response times. (A *response* is defined as a message from one of the recipients to the original sender whose subject line contains all of the words from the subject of the original email, and the response time should be measured as the difference between when the original email was sent and when the response was sent.)

**Solution：**

1. **Python program to extract data from raw files to CSV file.** (***Email.py***)

Fields needed to answer questions: Date, From, To, Cc, Bcc and Subject

Read lines from raw text files for different fields. Total 1703 Emails.

Another program to find the raw files based on key words like Subject. (***FindEmail.py***)

1. **SSIS to load raw data into database as dbo.EmailRawData**
2. **SQL queries (*EmailCode.sql*)**

1). Processed raw data:

***Note: Some emails don't have ‘To’ recipients, so I have removed them from our dataset. 1557 emails left.***

Email Date to remove Prefix days and the ending PDT/PST; Replaced unqualified xml letters in the recipients (To, Cc, Bcc); Made label of ‘Direct’ and ‘Broadcast’.

2). Split ‘To’, ‘Cc’ and ‘Bcc’ to recipients rows based on ‘From’.

3). Formatted Data to remove duplicate from the splitting and also extra space

1. **Answers (Attached the result as excel files as well)**

1). Attached excel from the result.

2). [steven.kean@enron.com](mailto:steven.kean@enron.com) sent the largest number of broadcast emails

[maureen.mcvicker@enron.com](mailto:maureen.mcvicker@enron.com) received the largest number of direct emails

3).Five emails with fastest response times:

**Note: Removed empty subject emails.**

**Removed the case when original sender also put the same email addess into recipients.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| OriginalSender | Replier | OriginalEmailDate | ResponseEmailDate | OriginalSubject | ResponseTimeMins |
| karen.denne@enron.com | jeff.dasovich@enron.com | 4/13/01 5:30 | 4/13/2001 5:34 | Re: CONFIDENTIAL - Residential in CA | 4 |
| michelle.cash@enron.com | lizzette.palmer@enron.com | 10/26/01 9:18 | 10/26/2001 9:22 | RE: CONFIDENTIAL Personnel issue | 4 |
| paul.kaufman@enron.com | jeff.dasovich@enron.com | 5/10/01 6:51 | 5/10/2001 6:55 | RE: Eeegads... | 4 |
| lizzette.palmer@enron.com | michelle.cash@enron.com | 10/26/01 9:13 | 10/26/2001 9:18 | RE: CONFIDENTIAL Personnel issue | 5 |
| lizzette.palmer@enron.com | michelle.cash@enron.com | 10/26/01 9:22 | 10/26/2001 9:28 | RE: CONFIDENTIAL Personnel issue | 6 |