MIS381N (Fall 2018) Homework6

Due on 11/05 before 12 pm

Created an inverted index as described in class (and in the textbook), where the output is a file of key/value pairs, with the key being an email address and the reference type (To, From, Cc, Bcc), and the value being a list of email IDs in which that address was referenced in the email.

For this assignment you will need to parse the emails in the input file (one email per line in the input file, separated by newlines). You should only be concerned with parsing the header of the email (fields like To, From, Cc, ...), and you need not do anything with the email content or embedded/forwarded emails other than skip over them. In the data set (two files) the emails have these fields in the header:

- Message-ID:
- Date:
- From:
- To:
- Subject:
- Cc: [optional]
- Mime-Version:
- Content-Type:
- Content-Transfer-Encoding
- Bcc: [optional]
- X-From:
- .....

Input files:

dataSet3a.txt, dataSet3b.txt

Artifacts to submit:

- Assignment3Code.zip - Python code
- Assignment3Output.txt - output generated by your solution

Example line for the output file for input data file dataSet3Tiny.txt

```
Cc:kimberly.olinger@enron.com['<6571521.1075857500090.JavaMail.evans@thyme>']

From:jane.tholt@enron.com     ['<6571521.1075857500090.JavaMail.evans@thyme>']

To:chris.foster@enron.com     ['<6571521.1075857500090.JavaMail.evans@thyme>']
```