In [1]:

```
import os
import json
from pathlib import Path
import zipfile
import email
from email.policy import default
from email.parser import Parser
from datetime import timezone
from collections import namedtuple
import pathlib

import pandas as pd
import s3fs
from bs4 import BeautifulSoup
from dateutil.parser import parse
from chardet.universaldetector import UniversalDetector

from pyspark.ml import Pipeline
from pyspark.ml.feature import CountVectorizer
from pyspark.ml.feature import HashingTF, Tokenizer
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.ml.pipeline import Transformer
from pyspark.sql.functions import udf
from pyspark.sql.types import StructType, StringType

import pandas as pd

current_dir = Path(os.getcwd()).absolute()
results_dir = current_dir.joinpath('results')
results_dir.mkdir(parents=True, exist_ok=True)
data_dir = current_dir.joinpath('data')
data_dir.mkdir(parents=True, exist_ok=True)
enron_data_dir = data_dir.joinpath('enron')

output_columns = [
        'payload',
        'text',
        'Message_D',
        'Date',
        'From',
        'To',
        'Subject',
        'Mime-Version',
        'Content-Type',
        'Content-Transfer-Encoding',
        'X-From',
        'X-To',
        'X-cc',
        'X-bcc',
        'X-Folder',
        'X-Origin',
        'X-FileName',
        'Cc',
```

```
        'Bcc'
]

columns = [column.replace('-', '_') for column in output_columns]

ParsedEmail = namedtuple('ParsedEmail', columns)

spark = SparkSession\
    .builder\
    .appName("Assignment04")\
    .getOrCreate()
```

The following code loads data to your local JupyterHub instance. You only need to run this once.

In [ ]:
```
def copy_data_to_local():
    dst_data_path = data_dir.joinpath('enron.zip')
    endpoint_url='https://storage.budsc.midwest-datascience.com'
    enron_data_path = 'data/external/enron.zip'

    s3 = s3fs.S3FileSystem(
        anon=True,
        client_kwargs={
            'endpoint_url': endpoint_url
        }
    )


    s3.get(enron_data_path, str(dst_data_path))

    with zipfile.ZipFile(dst_data_path) as f_zip:
        f_zip.extractall(path=data_dir)

copy_data_to_local()
```

This code reads emails and creates a Spark dataframe with three columns.

Assignment 4.1

In [2]:
```
def read_raw_email(email_path):
    detector = UniversalDetector()

    try:
        with open(email_path) as f:
            original_msg = f.read()
    except UnicodeDecodeError:
        detector.reset()
        with open(email_path, 'rb') as f:
            for line in f.readlines():
                detector.feed(line)
                if detector.done:
                    break
        detector.close()
```

```
        encoding = detector.result['encoding']
        with open(email_path, encoding=encoding) as f:
            original_msg = f.read()

    return original_msg

def make_spark_df():
    records = []
    for root, dirs, files in os.walk(enron_data_dir):
        for file_path in files:
            current_path = Path(root).joinpath(file_path)
            rel_path = os.path.relpath(current_path, enron_data_dir)
            username = current_path.parts[12]
            id = rel_path
            original_msg = read_raw_email(current_path)

            records.append((username, id, original_msg))

    col = ['username', 'id', 'original_msg']
    return spark.createDataFrame(data=records, schema=col)

df = make_spark_df()
```

In [6]:
```
df.show()
```

```
+----------+--------------------+--------------------+
|   username|                  id|        original_msg|
+----------+--------------------+--------------------+
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <2789...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <9989...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <5326...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <4139...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <3025...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <9701...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1644...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1135...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <6203...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <9658...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1770...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <2719...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1667...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <2905...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <5664...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1448...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1045...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1836...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <1327...|
|reitmeyer-j|reitmeyer-j/delet...|Message-ID: <2495...|
+----------+--------------------+--------------------+
only showing top 20 rows
```

In [7]:
```
df.printSchema()
```

```
root
 |-- username: string (nullable = true)
 |-- id: string (nullable = true)
 |-- original_msg: string (nullable = true)
```

Assignment 4.2

Use `plain_msg_example` and `html_msg_example` to create a function that parses an email message.

In [8]:

```
plain_msg_example = """
Message-ID: <6742786.1075845426893.JavaMail.evans@thyme>
Date: Thu, 7 Jun 2001 11:05:33 -0700 (PDT)
From: jeffrey.hammad@enron.com
To: andy.zipper@enron.com
Subject: Thanks for the interview
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Hammad, Jeffrey </O=ENRON/OU=NA/CN=RECIPIENTS/CN=NOTESADDR/CN=CBBE37
X-To: Zipper, Andy </O=ENRON/OU=NA/CN=RECIPIENTS/CN=AZIPPER>
X-cc:
X-bcc:
X-Folder: \Zipper, Andy\Zipper, Andy\Inbox
X-Origin: ZIPPER-A
X-FileName: Zipper, Andy.pst

Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Ass

Thanks and Best Regards,

Jeff Hammad
"""

html_msg_example = """
Message-ID: <21013632.1075862392611.JavaMail.evans@thyme>
Date: Mon, 19 Nov 2001 12:15:44 -0800 (PST)
From: insynconline.6jy5ympb.d@insync-palm.com
To: tstaab@enron.com
Subject: Last chance for special offer on Palm OS Upgrade!
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: InSync Online <InSyncOnline.6jy5ympb.d@insync-palm.com>
X-To: THERESA STAAB <tstaab@enron.com>
X-cc:
X-bcc:
X-Folder: \TSTAAB (Non-Privileged)\Staab, Theresa\Deleted Items
X-Origin: Staab-T
X-FileName: TSTAAB (Non-Privileged).pst
```

```
<html>

<html>
<head>
<title>Paprika</title>
<meta http-equiv="Content-Type" content="text/html;">
</head>
<body bgcolor="#FFFFFF" TEXT="#333333" LINK="#336699" VLINK="#6699cc" ALINK=
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
  <td width="582" colspan="9"><nobr><a href="http://insync-online.p04.com/u.
</tr>
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/ma
  <td width="20"><img src="http://images4.postdirect.com/master-images/40470
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?LkReaQA5eczX
  <td width="20"><img src="http://images4.postdirect.com/master-images/40470
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?BkReaQA5eczX
  <td width="20"><img src="http://images4.postdirect.com/master-images/40470
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?JkReaQA5eczX
  <td width="19"><img src="http://images4.postdirect.com/master-images/40470
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/ma
</tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/ma
  <td width="574"><br>
    <table border="0" cellpadding="0" cellspacing="0" width="574" bgcolor="#
    <tr>
      <td width="50"><img src="http://images4.postdirect.com/master-images/4
      <td width="474"><font face="verdana, arial" size="-2"color="#000000">
        <br>
        Dear THERESA,
        <br><br>
        Due to overwhelming demand for the Palm OS&#174; v4.1 Upgrade with M
        extending the special offer of 25% off through November 30, 2001. So
        increase the functionality of your Palm&#153; III, IIIx, IIIxe, IIIc
        new Palm OS v4.1 through this extended special offer. You'll receive
        <b>for just $29.95 when you use Promo Code <font color="#FF0000">OS4
        <b>$10 savings</b> off the list price.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51">Click h
        <br><br>
        <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61"><img sr
        <br><br>
        You can do a lot more with your Palm&#153; handheld when you upgrade
        favorite features just got even better and there are some terrific n
        <br><br>
        <LI> Handwrite notes and even draw pictures right on your Palm&#153
        <LI> Tap letters with your stylus and use Graffiti&#174; at the same
        <LI> Improved Date Book functionality lets you view, snooze or clear
        <LI> You can easily change time-zone settings</LI>

        <br><br>
        <a href="http://insync-online.p04.com/u.d?WkReaQA5eczXRb=71"><img sr
```

```
        <br><br>
        <LI> <nobr>Mask/unmask</nobr> private records or hide/unhide directl
        <LI> Lock your device automatically at a designated time using the n
        <LI> Always remember your password with our new Hint feature*</LI>

        <br><br>
        <a href="http://insync-online.p04.com/u.d?VEReaQA5eczXRQ=81"><img sr
        <br><br>
        <LI> Use your GSM compatible mobile phone or modem to get online and
        <LI> Stay connected with email, instant messaging and text messaging
        <LI> Send applications or records through your cell phone to schedul
             important information to others</LI>

        <br><br>
        All this comes in a new operating system that can be yours for just
        upgrade to the new Palm&#153; OS v4.1</a> and you'll also get the la
        <nobr>1-800-881-7256</nobr> to order via phone.
        <br><br>
        Sincerely,<br>
        The Palm Team
        <br><br>
        P.S. Remember, this extended offer opportunity of 25% savings absolu
        and is only available through the Palm Store when you use Promo Code
        <br><br>
        <img src="http://images4.postdirect.com/master-images/404707/bottom_
        <br><img src="http://images4.postdirect.com/master-images/404707/cle
        </font></td>
      <td width="50"><img src="http://images4.postdirect.com/master-images/4
    </tr>
    </table></td>
    <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/
  </tr>
  <tr>
  <td colspan="3"><img src="http://images4.postdirect.com/master-images/4047
  </tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
  <tr>
    <td width="54"><img src="http://images4.postdirect.com/master-images/404
    <td width="474"><font face="arial, verdana" size="-2" color="#000000"><b
    * This feature is available on the Palm&#153; IIIx, Palm&#153; IIIxe, an
    ** Note: To use the MIK functionality, you need either a Palm OS&#174; c
    with  <nobr>built-in</nobr> modem or data capability that has either an
    are using a phone, you must have data services from your mobile service
    a list of tested and supported phones that you can use with the MIK. Cab
    <br><br>
    ------------------<br>
    To modify your profile or unsubscribe from Palm newsletters, <a href="ht
    Or, unsubscribe by replying to this message, with "unsubscribe" as the s
    <br><br>
    ------------------<br>
    Copyright&#169; 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandST
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmMo
    and the Palm Platform Compatible Logo are registered trademarks of Palm,
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, Palm
    trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of Pa
```

```
    product names may be trademarks or registered trademarks of their respec
    <img src="http://images4.postdirect.com/master-images/404707/clear.gif"
    <td width="54"><img src="http://images4.postdirect.com/master-images/404
  </tr>
</table><br><br><br><br>
<!-- The following image is included for message detection -->
<img src="http://p04.com/1x1.dyn" border="0" alt="" width="1" height="1">
<img src="http://p04.com/1x1.dyn?0vEGou8Ig30ba2L2bLn" width=1 height=1></boc
</html>

</html>
"""
plain_msg_example = plain_msg_example.strip()
html_msg_example = html_msg_example.strip()
```

In [9]:
```
def parse_html_payload(payload):
    """
    This function uses Beautiful Soup to read HTML data
    and return the text.  If the payload is plain text, then
    Beautiful Soup will return the original content
    """
    soup = BeautifulSoup(payload, 'html.parser')
    return str(soup.get_text()).encode('utf-8').decode('utf-8')

def parse_email(original_msg):
    result = {}
    html = parse_html_payload(original_msg)
    msg = Parser(policy=default).parsestr(html)

    body = msg.get_payload()

    tuple_result = tuple([str(result.get(column, None)) for column in columr
    return body
```

In [10]:
```
parsed_msg = parse_email(plain_msg_example)
```

In [11]:
```
print(parsed_msg)
```

```
Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ As
sociate program.  I enjoyed talking to you, and look forward to contributin
g to the success that the program has enjoyed.

Thanks and Best Regards,

Jeff Hammad
```

In [12]:
```
parsed_html_msg = parse_email(html_msg_example)
```

In [13]:
```
print(parsed_html_msg)
```

Paprika

Dear THERESA,

Due to overwhelming demand for the Palm OS® v4.1 Upgrade with Mobile Connectivity, we are extending the special offer of 25% off through November 30, 2001. So there's still time to significantly increase the functionality of your Palm™ III, IIIx, IIIxe, IIIc, V or Vx handheld. Step up to the new Palm OS v4.1 through this extended special offer. You'll receive the brand new Palm OS v4.1 for just $29.95 when you use Promo Code OS41WAVE. That's a $10 savings off the list price.

Click here to view a full product demo now.

You can do a lot more with your Palm™ handheld when you upgrade to the Palm OS v4.1. All your favorite features just got even better and there are some terrific new additions:

Handwrite notes and even draw pictures right on your Palm™ handheld
Tap letters with your stylus and use Graffiti® at the same time with the enhanced onscreen keyboard

Improved Date Book functionality lets you view, snooze or clear multiple a larms all with a single tap
You can easily change time-zone settings

Mask/unmask private records or hide/unhide directly within the application
Lock your device automatically at a designated time using the new Autolock ing feature
Always remember your password with our new Hint feature*

Use your GSM compatible mobile phone or modem to get online and access the web
Stay connected with email, instant messaging and text messaging to GSM mob ile phones
Send applications or records through your cell phone to schedule meetings and even "beam"
important information to others

All this comes in a new operating system that can be yours for just $29.95! Click here to
upgrade to the new Palm™ OS v4.1 and you'll also get the latest Pal m desktop software. Or call
1-800-881-7256 to order via phone.

Sincerely,
The Palm Team

P.S. Remember, this extended offer opportunity of 25% savings absol utely ends on November 30, 2001
and is only available through the Palm Store when you use Promo Cod e OS41WAVE.

* This feature is available on the Palm™ IIIx, Palm™ IIIxe, and Palm™ V x.

** Note: To use the MIK functionality, you need either a Palm OS® compa tible modem or a phone
with built-in modem or data capability that has either an infrared por

```
t or cable exits.  If you
    are using a phone, you must have data services from your mobile service
provider.  Click here for
    a list of tested and supported phones that you can use with the MIK. Ca
ble not provided.

    ------------------
    To modify your profile or unsubscribe from Palm newsletters, click her
e.
    Or, unsubscribe by replying to this message, with "unsubscribe" as the
subject line of the message.

    ------------------
    Copyright© 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP,
HandWEB, Graffiti,
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmM
odem, PalmPoint, PalmPrint,
    and the Palm Platform Compatible Logo are registered trademarks of Pal
m, Inc. Palm, the Palm logo,
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, Pal
mPix, Palm Powered, the Palm
    trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of P
alm, Inc. All other brands and
    product names may be trademarks or registered trademarks of their respe
ctive owners.
```

Assignment 4.3

In [14]:
```
## This creates a schema for the email data
email_struct = StructType()

for column in columns:
    email_struct.add(column, StringType(), True)
```

In [15]:
```
def make_spark_df2():
    records = []
    for root, dirs, files in os.walk(enron_data_dir):
        for file_path in files:
            current_path = Path(root).joinpath(file_path)
            rel_path = os.path.relpath(current_path, enron_data_dir)
            username = current_path.parts[12]
            id = rel_path
            original_msg = read_raw_email(current_path)
            parsed_msg = parse_email(original_msg)
```

```
            records.append((username, id, parsed_msg))

    col = ['username', 'id', 'parsed_msg']
    return spark.createDataFrame(data=records, schema=col)

df2 = make_spark_df2()
```

In [16]:
```
df2.show()
```

```
+----------+--------------------+--------------------+
|   username|                  id|          parsed_msg|
+----------+--------------------+--------------------+
|reitmeyer-j|reitmeyer-j/delet...|CNNSI.com  =09  =...|
|reitmeyer-j|reitmeyer-j/delet...|i couldn't have s...|
|reitmeyer-j|reitmeyer-j/delet...|\n\n-------------...|
|reitmeyer-j|reitmeyer-j/delet...|Howdy!\n\nWe need...|
|reitmeyer-j|reitmeyer-j/delet...|\nLate on October...|
|reitmeyer-j|reitmeyer-j/delet...|\n\nYahoo! Direct...|
|reitmeyer-j|reitmeyer-j/delet...|Dear Jay,\n\nWith...|
|reitmeyer-j|reitmeyer-j/delet...|\n---------------...|
|reitmeyer-j|reitmeyer-j/delet...|25% Off 10 Great ...|
|reitmeyer-j|reitmeyer-j/delet...|Jay,\n\nHow would...|
|reitmeyer-j|reitmeyer-j/delet...|\nUntitled Docume...|
|reitmeyer-j|reitmeyer-j/delet...|The All-Employee ...|
|reitmeyer-j|reitmeyer-j/delet...|\n\n\n\n\n\n\n\n\...|
|reitmeyer-j|reitmeyer-j/delet...|\n\nA Dialogue wi...|
|reitmeyer-j|reitmeyer-j/delet...|Are you late buyi...|
|reitmeyer-j|reitmeyer-j/delet...|\n\n2 Motorola Ta...|
|reitmeyer-j|reitmeyer-j/delet...|Hi All,\n\nWe've ...|
|reitmeyer-j|reitmeyer-j/delet...| ezboard Tips & T...|
|reitmeyer-j|reitmeyer-j/delet...|\nJust a note to ...|
|reitmeyer-j|reitmeyer-j/delet...|Thank you.\n\n --...|
+----------+--------------------+--------------------+
only showing top 20 rows
```

In [18]:
```
## This creates a user-defined function which can be used in Spark
parse_email_func = udf(lambda z: parse_email(z), email_struct)

def parse_emails(input_df):
    new_df = input_df.select(
        'username', 'id', 'original_msg', parse_email_func('original_msg').a
    )
    for column in columns:
        new_df = new_df.withColumn(column, new_df.parsed_email[column])

    new_df = new_df.drop('parsed_email')
    return new_df

class ParseEmailsTransformer(Transformer):
    def _transform(self, dataset):
        """
        Transforms the input dataset.

        :param dataset: input dataset, which is an instance of :py:class:`py
        :returns: transformed dataset
```

```
        """
        return dataset.transform(parse_emails)

## Use the custom ParseEmailsTransformer, Tokenizer, and CountVectorizer
## to create a spark pipeline

parse = ParseEmailsTransformer()
tokenizer = Tokenizer(inputCol='parsed_msg', outputCol='words')
vectorizer = CountVectorizer(inputCol=tokenizer.getOutputCol(), outputCol='f

stages = [tokenizer, vectorizer]

email_pipeline = Pipeline(stages=stages)

model = email_pipeline.fit(df2)
result = model.transform(df2)
```

In [21]:
```
result.select('id', 'words', 'features').show()
```

```
+------------------+------------------+------------------+
|                id|             words|          features|
+------------------+------------------+------------------+
|reitmeyer-j/delet...|[cnnsi.com, , =09...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[i, couldn't, hav...|(108640,[0,2,3,8,...|
|reitmeyer-j/delet...|[, , ----------...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[howdy!, , we, ne...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, late, on, octo...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, , yahoo!, dire...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[dear, jay,, , wi...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, ----------...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[25%, off, 10, gr...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[jay,, , how, wou...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, untitled, docu...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[the, all-employe...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, , , , , , , ...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, , a, dialogue,...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[are, you, late, ...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, , 2, motorola,...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[hi, all,, , we'v...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, ezboard, tips,...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[, just, a, note,...|(108640,[0,1,2,3,...|
|reitmeyer-j/delet...|[thank, you., , ,...|(108640,[0,1,2,3,...|
+------------------+------------------+------------------+
only showing top 20 rows
```