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 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()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(ne
wLevel).
23/07/03 03:19:55 WARN NativeCodeLoader: Unable to load native-hadoop library for yo
ur platform... using builtin-java classes where applicable
```

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

Note: I did not run the following code and instead manually copied the enron folder in to the data folder

In [2]:
```
# 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 [3]:
```
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 is now the file path to the current email.
            current_path = Path(root).joinpath(file_path)
            record = {}

            ## Use this path to read the following information
            ## original_msg
            record['original_msg'] = read_raw_email(current_path)

            ## username (Hint: It is the root folder)
            record['username'] = Path(root).name

            ## id (The relative path of the email message)
            record['id'] = os.path.relpath(current_path, os.path.dirname(os.path.di

            records.append(record)

    ##  Create the Spark dataframe
    schema = ["original_msg", "username", "id"]

    return spark.createDataFrame(records, schema)

df = make_spark_df()
```

In [4]:
```
df.show()
```

```
[Stage 0:>                                                          (0 + 1) / 1]
```

```
+--------------------+--------------------+-----------+
|        original_msg|            username|         id|
+--------------------+--------------------+-----------+
| davis-d/vacation/7_|Message-ID: <1645...|   vacation|
| davis-d/vacation/5_|Message-ID: <1159...|   vacation|
|davis-d/vacation/17_|Message-ID: <6829...|   vacation|
| davis-d/vacation/8_|Message-ID: <6790...|   vacation|
| davis-d/vacation/4_|Message-ID: <6053...|   vacation|
| davis-d/vacation/1_|Message-ID: <3258...|   vacation|
|davis-d/vacation/18_|Message-ID: <3016...|   vacation|
| davis-d/vacation/6_|Message-ID: <2285...|   vacation|
| davis-d/vacation/3_|Message-ID: <2308...|   vacation|
| davis-d/vacation/9_|Message-ID: <3059...|   vacation|
| davis-d/vacation/2_|Message-ID: <2401...|   vacation|
|davis-d/notes_inb...|Message-ID: <3091...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <3292...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <7892...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <1445...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <9976...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <3052...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <7428...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <1204...|notes_inbox|
|davis-d/notes_inb...|Message-ID: <5931...|notes_inbox|
+--------------------+--------------------+-----------+
only showing top 20 rows
```

In [5]:
```
df.printSchema()
```

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

Assignment 4.2

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

In [6]:
```
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=CBBE377A-24F5
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/ Associate

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="#ff990
<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.d?BERea
</tr>
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-im
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?LkReaQA5eczXL=21"><
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?BkReaQA5eczXO=31"><
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?JkReaQA5eczXRs=41">
  <td width="19"><img src="http://images4.postdirect.com/master-images/404707/clear
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-im
</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/master-im
<td width="574"><br>
  <table border="0" cellpadding="0" cellspacing="0" width="574" bgcolor="#99ccff"
  <tr>
    <td width="50"><img src="http://images4.postdirect.com/master-images/404707/c
    <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 Mobile C
      extending the special offer of 25% off through November 30, 2001. So there'
      increase the functionality of your Palm&#153; III, IIIx, IIIxe, IIIc, V or
      new Palm OS v4.1 through this extended special offer. You'll receive the br
      <b>for just $29.95 when you use Promo Code <font color="#FF0000">OS41WAVE</
      <b>$10 savings</b> off the list price.
      <br><br>
      <a href="http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51">Click here to
      <br><br>
      <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61"><img src="http
      <br><br>
      You can do a lot more with your Palm&#153; handheld when you upgrade to the
      favorite features just got even better and there are some terrific new addi
      <br><br>
      <LI> Handwrite notes and even draw pictures right on your Palm&#153 handhel
      <LI> Tap letters with your stylus and use Graffiti&#174; at the same time w
      <LI> Improved Date Book functionality lets you view, snooze or clear multip
      <LI> You can easily change time-zone settings</LI>

      <br><br>
      <a href="http://insync-online.p04.com/u.d?WkReaQA5eczXRb=71"><img src="http
      <br><br>
      <LI> <nobr>Mask/unmask</nobr> private records or hide/unhide directly withi
      <LI> Lock your device automatically at a designated time using the new Auto
      <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 src="http
      <br><br>
      <LI> Use your GSM compatible mobile phone or modem to get online and access
      <LI> Stay connected with email, instant messaging and text messaging to GSM
      <LI> Send applications or records through your cell phone to schedule meeti
           important information to others</LI>

      <br><br>
      All this comes in a new operating system that can be yours for just $29.95!
      upgrade to the new Palm&#153; OS v4.1</a> and you'll also get the latest Pa
      <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 absolutely en
      and is only available through the Palm Store when you use Promo Code <b><fo
      <br><br>
      <img src="http://images4.postdirect.com/master-images/404707/bottom_button.
      <br><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
```

```
        </font></td>
      <td width="50"><img src="http://images4.postdirect.com/master-images/404707/c
    </tr>
    </table></td>
    <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-
  </tr>
  <tr>
  <td colspan="3"><img src="http://images4.postdirect.com/master-images/404707/bott
  </tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
  <tr>
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/cle
    <td width="474"><font face="arial, verdana" size="-2" color="#000000"><br>
    * This feature is available on the Palm&#153; IIIx, Palm&#153; IIIxe, and Palm&
    ** Note: To use the MIK functionality, you need either a Palm OS&#174; compatib
    with  <nobr>built-in</nobr> modem or data capability that has either an infrare
    are using a phone, you must have data services from your mobile service provide
    a list of tested and supported phones that you can use with the MIK. Cable not
    <br><br>
    ------------------<br>
    To modify your profile or unsubscribe from Palm newsletters, <a href="http://in
    Or, unsubscribe by replying to this message, with "unsubscribe" as the subject
    <br><br>
    ------------------<br>
    Copyright&#169; 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP, Ha
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem, Pa
    and the Palm Platform Compatible Logo are registered trademarks of Palm, Inc. P
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, PalmPix, Pa
    trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of Palm, Inc
    product names may be trademarks or registered trademarks of their respective ow
    <img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/cle
  </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></body>
</html>

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

In [7]:

```
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 = {}
```

```
    msg = Parser(policy=default).parsestr(original_msg)

    # Use Python's email library to read the payload and the headers

    # Extract headers
    headers = ['Message-ID', 'Date', 'From', 'To', 'Subject', 'Mime-Version', 'Cont
    for header in headers:
        result[header] = msg[header]

    # Read Payload
    # The following code can decide whether or not the email contains the string "h
    if "html" in str(msg):
        result['text'] = parse_html_payload(msg.get_payload(decode=True))
    else:
        result['text'] = msg.get_payload(decode=True).decode('utf-8')


    ## https://docs.python.org/3/library/email.examples.html
    tuple_result = tuple([str(result.get(column, None)) for column in columns])
    return ParsedEmail(*tuple_result)
```

In [8]:
```
parsed_msg = parse_email(plain_msg_example)
```

In [9]:
```
print(parsed_msg.text)
```

```
Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associate p
rogram.  I enjoyed talking to you, and look forward to contributing to the success t
hat the program has enjoyed.

Thanks and Best Regards,

Jeff Hammad
```

In [10]:
```
parsed_html_msg = parse_email(html_msg_example)
```

In [11]:
```
print(parsed_html_msg.text)
```

Paprika

Dear THERESA,

Due to overwhelming demand for the Palm OS® v4.1 Upgrade with Mobile Connect
ivity, 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 han
dheld. Step up to the
new Palm OS v4.1 through this extended special offer. You'll receive the bra
nd 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 addit
ions:

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 o
nscreen keyboard

Improved Date Book functionality lets you view, snooze or clear multiple alarms 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 Autolocking 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 mobile 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 Palm 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 absolutely ends on November 30, 2001
and is only available through the Palm Store when you use Promo Code OS41WAVE.

* This feature is available on the Palm™ IIIx, Palm™ IIIxe, and Palm™ Vx.
** Note: To use the MIK functionality, you need either a Palm OS® compatible modem or a phone
with built-in modem or data capability that has either an infrared port or cable exits. If you
are using a phone, you must have data services from your mobile service provide

```
r.  Click here for
    a list of tested and supported phones that you can use with the MIK. Cable not p
rovided.

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

    ------------------
    Copyright© 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP, HandWEB,
Graffiti,
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem, Pal
mPoint, PalmPrint,
    and the Palm Platform Compatible Logo are registered trademarks of Palm, Inc. Pa
lm, the Palm logo,
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, PalmPix, Pal
m Powered, the Palm
    trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of Palm, Inc.
All other brands and
    product names may be trademarks or registered trademarks of their respective own
ers.
```

Assignment 4.3

In [12]:
```
## This creates a schema for the email data
email_struct = StructType()

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

In [13]:
```
## 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').alias('p
    )
    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:`pyspark.s
        :returns: transformed dataset
        """
        return dataset.transform(parse_emails)

# Use the custom ParseEmailsTransformer, Tokenizer, and CountVectorizer
# to create a spark pipeline
email_pipeline = Pipeline(
    ## TODO: Complete code
    stages=[
        ParseEmailsTransformer(),
        Tokenizer(inputCol='text', outputCol='words'),
        CountVectorizer(inputCol='words', outputCol='features')
    ]
)

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

```
23/07/03 03:20:49 WARN TaskSetManager: Stage 1 contains a task of very large size (1
289 KiB). The maximum recommended task size is 1000 KiB.
```

In [14]:
```
result.select('id', 'words', 'features').show()
```

```
+----------+--------------------+--------------------+
|        id|               words|            features|
+----------+--------------------+--------------------+
|   vacation|[davis-d/vacation...|(13725,[8343],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[9483],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[9263],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[11097],[1...|
|   vacation|[davis-d/vacation...|(13725,[9495],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[6465],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[13489],[1...|
|   vacation|[davis-d/vacation...| (13725,[632],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[7470],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[8679],[1.0])|
|   vacation|[davis-d/vacation...|(13725,[4298],[1.0])|
|notes_inbox|[davis-d/notes_in...|(13725,[5463],[1.0])|
|notes_inbox|[davis-d/notes_in...|(13725,[1435],[1.0])|
|notes_inbox|[davis-d/notes_in...| (13725,[522],[1.0])|
|notes_inbox|[davis-d/notes_in...|(13725,[12674],[1...|
|notes_inbox|[davis-d/notes_in...|(13725,[8733],[1.0])|
|notes_inbox|[davis-d/notes_in...|(13725,[11234],[1...|
|notes_inbox|[davis-d/notes_in...|(13725,[4235],[1.0])|
|notes_inbox|[davis-d/notes_in...|(13725,[2788],[1.0])|
|notes_inbox|[davis-d/notes_in...|(13725,[3547],[1.0])|
+----------+--------------------+--------------------+
only showing top 20 rows
```

note: I am not convinced that the "words" parsed correctly from the "original message"

In [ ]: