In [1]:

```
import os
import json
import sys
import re
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
from pyspark.sql import functions as F
from pyspark.sql.types import 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)
```

```
os.environ['PYSPARK_PYTHON'] = sys.executable
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable

spark = SparkSession \
    .builder \
    .appName("Assignment04") \
    .config("spark.executor.memory", "2g") \
    .config("spark.driver.memory", "2g") \
    .getOrCreate()
```

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

In [2]:

```
def copy_data_to_local():
    dst_data_path = data_dir.joinpath('enron.zip')
    enron_data_path = '../../../data/external/enron/enron.zip'

    with zipfile.ZipFile(enron_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]:

```
# update the enron_data_dir to use the proper directory now
enron_data_dir = "data"

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.
            ## Use this path to read the following information
            ## original_msg
            ## username (Hint: It is the root folder)
            ## id (The relative path of the email message)
            current_path = Path(root).joinpath(file_path)

            # Add a list of records with the username, id, and original message to the r
            original_msg = read_raw_email(current_path)
            id = os.path.relpath(current_path, "data/")
            username = id.split("\\")[0]

            records.append({"id": id, "username": username, "original_msg": original_msg

    # Add the records to the dataframe
    return spark.createDataFrame(records)

df = make_spark_df()
```

In [4]:

```
df.show()
```

```
+--------------------+--------------------+--------+
|                  id|        original_msg|username|
+--------------------+--------------------+--------+
|  davis-d\2_trash\1_|Message-ID: <1774...| davis-d|
|  davis-d\2_trash\2_|Message-ID: <2467...| davis-d|
|  davis-d\2_trash\3_|Message-ID: <2833...| davis-d|
|  davis-d\2_trash\4_|Message-ID: <1972...| davis-d|
|davis-d\2_trash\c...|Message-ID: <1964...| davis-d|
|davis-d\2_trash\c...|Message-ID: <7345...| davis-d|
|davis-d\2_trash\c...|Message-ID: <5686...| davis-d|
|davis-d\2_trash\c...|Message-ID: <7218...| davis-d|
|davis-d\2_trash\c...|Message-ID: <3016...| davis-d|
|davis-d\2_trash\c...|Message-ID: <1233...| davis-d|
|davis-d\2_trash\c...|Message-ID: <2215...| davis-d|
|davis-d\2_trash\c...|Message-ID: <1365...| davis-d|
|davis-d\2_trash\c...|Message-ID: <2251...| davis-d|
|davis-d\2_trash\c...|Message-ID: <8556...| davis-d|
|davis-d\2_trash\c...|Message-ID: <1807...| davis-d|
|davis-d\2_trash\c...|Message-ID: <2705...| davis-d|
|davis-d\2_trash\c...|Message-ID: <2977...| davis-d|
|davis-d\2_trash\c...|Message-ID: <3065...| davis-d|
|davis-d\2_trash\c...|Message-ID: <2798...| davis-d|
|davis-d\2_trash\c...|Message-ID: <3108...| davis-d|
+--------------------+--------------------+--------+
only showing top 20 rows
```

In [5]:

```
df.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- original_msg: string (nullable = true)
 |-- username: 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-24F58854-
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 progr

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="#ff9900">
<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?BEReaQA5ec
</tr>
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?LkReaQA5eczXL=21"><img s
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?BkReaQA5eczXO=31"><img s
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?JkReaQA5eczXRs=41"><img
  <td width="19"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/
</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-images/
  <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/clear.
      <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 Connec
        extending the special offer of 25% off through November 30, 2001. So there's sti
        increase the functionality of your Palm&#153; III, IIIx, IIIxe, IIIc, V or Vx ha
        new Palm OS v4.1 through this extended special offer. You'll receive the brand n
        <b>for just $29.95 when you use Promo Code <font color="#FF0000">OS41WAVE</font>
        <b>$10 savings</b> off the list price.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51">Click here to view 
        <br><br>
        <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61"><img src="http://im
        <br><br>
        You can do a lot more with your Palm&#153; handheld when you upgrade to the Palm
        favorite features just got even better and there are some terrific new additions
        <br><br>
        <LI> Handwrite notes and even draw pictures right on your Palm&#153 handheld</LI
        <LI> Tap letters with your stylus and use Graffiti&#174; at the same time with t
        <LI> Improved Date Book functionality lets you view, snooze or clear multiple al
        <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://im
        <br><br>
        <LI> <nobr>Mask/unmask</nobr> private records or hide/unhide directly within the
        <LI> Lock your device automatically at a designated time using the new Autolocki
        <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://im
        <br><br>
        <LI> Use your GSM compatible mobile phone or modem to get online and access the 
        <LI> Stay connected with email, instant messaging and text messaging to GSM mobi
        <LI> Send applications or records through your cell phone to schedule meetings a
             important information to others</LI>

        <br><br>
        All this comes in a new operating system that can be yours for just $29.95! <a h
        upgrade to the new Palm&#153; OS v4.1</a> and you'll also get the latest Palm de
        <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 ends on
        and is only available through the Palm Store when you use Promo Code <b><font co
        <br><br>
        <img src="http://images4.postdirect.com/master-images/404707/bottom_button.gif" 
        <br><img src="http://images4.postdirect.com/master-images/404707/clear.gif" widt
        </font></td>
      <td width="50"><img src="http://images4.postdirect.com/master-images/404707/clear.
    </tr>
    </table></td>
    <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-image
  </tr>
  <tr>
  <td colspan="3"><img src="http://images4.postdirect.com/master-images/404707/bottom.gi
  </tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
  <tr>
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/clear.gi
    <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&#153;
    ** Note: To use the MIK functionality, you need either a Palm OS&#174; compatible mo
    with  <nobr>built-in</nobr> modem or data capability that has either an infrared por
    are using a phone, you must have data services from your mobile service provider.  <
    a list of tested and supported phones that you can use with the MIK. Cable not provi
    <br><br>
    ------------------<br>
    To modify your profile or unsubscribe from Palm newsletters, <a href="http://insync-
    Or, unsubscribe by replying to this message, with "unsubscribe" as the subject line
    <br><br>
    ------------------<br>
    Copyright&#169; 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP, HandWEB
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem, PalmPoi
    and the Palm Platform Compatible Logo are registered trademarks of Palm, Inc. Palm,
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, PalmPix, Palm Po
    trade dress, PalmSource, Smartcode, and Simply Palm are trademarks of Palm, Inc. All
    product names may be trademarks or registered trademarks of their respective owners.
    <img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="474"
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/clear.gi
  </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)

    ## TODO: Use Python's email library to read the payload and the headers
    ## https://docs.python.org/3/library/email.examples.html
    # Set the result by using the output column from the message.
    for output_column in output_columns:
        result[output_column] = msg[output_column]

    # Add payload and text to the results.
    result['Payload'] = original_msg
    result['text'] = msg.get_payload()

    # Add a space to ensure the results match the expected outcome.
    print()

    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 progr
```

```
am.  I enjoyed talking to you, and look forward to contributing to the success that 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)
```

```
<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="#ff9900">
<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?BEReaQA5ec
zXB=1"><img src="http://images4.postdirect.com/master-images/404707/upper_left.gif" alt
="" width="103" height="110" hspace="0" vspace="0" border="0"></a><a href="http://insync
-online.p04.com/u.d?AkReaQA5eczXE=11"><img src="http://images4.postdirect.com/master-ima
ges/404707/upper_right.gif" alt="" width="479" height="110" hspace="0" vspace="0" border
="0"></a></nobr></td>
</tr>
<tr valign="top">
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/
404707/clear.gif" width="4" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
width="20" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?LkReaQA5eczXL=21"><img s
rc="http://images4.postdirect.com/master-images/404707/screen1.gif" alt="" width="165" h
eight="159" hspace="0" vspace="0" border="0"></a><br><img src="http://images4.postdirec
t.com/master-images/404707/screen1_text.gif" alt="" width="93" height="26" hspace="0" vs
pace="0" border="0"></td>
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
width="20" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?BkReaQA5eczXO=31"><img s
rc="http://images4.postdirect.com/master-images/404707/screen2.gif" alt="" width="165" h
eight="159" hspace="0" vspace="0" border="0"></a><br><img src="http://images4.postdirec
t.com/master-images/404707/screen2_text.gif" alt="" width="93" height="26" hspace="0" vs
pace="0" border="0"></td>
  <td width="20"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
width="20" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  <td width="165"><br><a href="http://insync-online.p04.com/u.d?JkReaQA5eczXRs=41"><img
src="http://images4.postdirect.com/master-images/404707/screen3.gif" alt="" width="165"
height="159" hspace="0" vspace="0" border="0"></a><br><img src="http://images4.postdirec
t.com/master-images/404707/screen3_text.gif" alt="" width="93" height="26" hspace="0" vs
pace="0" border="0"></td>
  <td width="19"><img src="http://images4.postdirect.com/master-images/404707/clear.gif"
width="19" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-images/
404707/clear.gif" width="4" height="1" hspace="0" vspace="0" border="0" alt=""></td>
</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-images/
404707/clear.gif" width="4" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  <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/clear.
gif" width="50" height="1" hspace="0" vspace="0" border="0" alt=""></td>
      <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 Connec
tivity, we are
        extending the special offer of 25% off through November 30, 2001. So there's sti
ll time to significantly
        increase the functionality of your Palm&#153; III, IIIx, IIIxe, IIIc, V or Vx ha
ndheld. Step up to the
        new Palm OS v4.1 through this extended special offer. You'll receive the brand n
ew Palm OS v4.1
        <b>for just $29.95 when you use Promo Code <font color="#FF0000">OS41WAVE</font>
</b>. That's a
        <b>$10 savings</b> off the list price.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?NkReaQA5eczXRh=51">Click here to view
a full product demo now</a>.
        <br><br>
        <a href="http://insync-online.p04.com/u.d?MkReaQA5eczXRm=61"><img src="http://im
ages4.postdirect.com/master-images/404707/title1.gif" alt="" width="336" height="20" hsp
ace="0" vspace="0" border="0"></a>
        <br><br>
        You can do a lot more with your Palm&#153; handheld when you upgrade to the Palm
OS v4.1. All your
        favorite features just got even better and there are some terrific new addition
s:
        <br><br>
        <LI> Handwrite notes and even draw pictures right on your Palm&#153 handheld</LI
>
        <LI> Tap letters with your stylus and use Graffiti&#174; at the same time with t
he enhanced onscreen keyboard</LI>
        <LI> Improved Date Book functionality lets you view, snooze or clear multiple al
arms all with a single tap </LI>
        <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://im
ages4.postdirect.com/master-images/404707/title2.gif" alt="" width="460" height="20" hsp
ace="0" vspace="0" border="0"></a>
        <br><br>
        <LI> <nobr>Mask/unmask</nobr> private records or hide/unhide directly within the
application</LI>
        <LI> Lock your device automatically at a designated time using the new Autolocki
ng feature</LI>
        <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://im
ages4.postdirect.com/master-images/404707/title3.gif" alt="" width="461" height="31" hsp
ace="0" vspace="0" border="0"></a>
        <br><br>
        <LI> Use your GSM compatible mobile phone or modem to get online and access the
web</LI>
        <LI> Stay connected with email, instant messaging and text messaging to GSM mobi
le phones</LI>
        <LI> Send applications or records through your cell phone to schedule meetings a
nd even "beam"
             important information to others</LI>

        <br><br>
        All this comes in a new operating system that can be yours for just $29.95! <a h
ref="http://insync-online.p04.com/u.d?MkReaQA5eczXRV=91">Click here to
```

```
        upgrade to the new Palm&#153; OS v4.1</a> and you'll also get the latest Palm de
sktop software. Or call
        <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 ends on
November 30, 2001
        and is only available through the Palm Store when you use Promo Code <b><font co
lor="#FF0000">OS41WAVE</font></b>.
        <br><br>
        <img src="http://images4.postdirect.com/master-images/404707/bottom_button.gif"
align="right" alt="" width="295" height="60" hspace="0" vspace="0" border="0">
        <br><img src="http://images4.postdirect.com/master-images/404707/clear.gif" widt
h="474" height="1" hspace="0" vspace="0" border="0" alt="">
        </font></td>
      <td width="50"><img src="http://images4.postdirect.com/master-images/404707/clear.
gif" width="50" height="1" hspace="0" vspace="0" border="0" alt=""></td>
    </tr>
    </table></td>
    <td width="4" bgcolor="#CCCCCC"><img src="http://images4.postdirect.com/master-image
s/404707/clear.gif" width="4" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  </tr>
  <tr>
  <td colspan="3"><img src="http://images4.postdirect.com/master-images/404707/bottom.gi
f" width="582" height="67" hspace="0" vspace="0" border="0"></td>
  </tr>
</table>
<table border="0" cellpadding="0" cellspacing="0" width="582">
  <tr>
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/clear.gi
f" width="54" height="1" hspace="0" vspace="0" border="0" alt=""></td>
    <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&#153;
Vx. <br><br>
    ** Note: To use the MIK functionality, you need either a Palm OS&#174; compatible mo
dem or a phone
    with  <nobr>built-in</nobr> 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.  <
a href="http://insync-online.p04.com/u.d?RkReaQA5eczXRK=101">Click here</a> for
    a list of tested and supported phones that you can use with the MIK. Cable not provi
ded.
    <br><br>
    ------------------<br>
    To modify your profile or unsubscribe from Palm newsletters, <a href="http://insync-
online.p04.com/u.d?KkReaQA5eczXRE=121">click here</a>.
    Or, unsubscribe by replying to this message, with "unsubscribe" as the subject line
of the message.
    <br><br>
    ------------------<br>
    Copyright&#169; 2001 Palm, Inc. Palm OS, Palm Computing, HandFAX, HandSTAMP, HandWE
B, Graffiti,
    HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem, PalmPoi
nt, PalmPrint,
    and the Palm Platform Compatible Logo are registered trademarks of Palm, Inc. Palm,
the Palm logo,
    AnyDay, EventClub, HandMAIL, the HotSync Logo, PalmGear, PalmGlove, PalmPix, Palm Po
wered, 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 owners.
</font>
    <img src="http://images4.postdirect.com/master-images/404707/clear.gif" width="474"
height="1" hspace="0" vspace="0" border="0" alt=""></td>
```

```
    <td width="54"><img src="http://images4.postdirect.com/master-images/404707/clear.gi
f" width="54" height="1" hspace="0" vspace="0" border="0" alt=""></td>
  </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>
```

Assignment 4.3

In [12]:

```python
## This creates a schema for the email data
email_struct = StructType()

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

In [16]:

```python
## 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):

    def clean_email_text(text):
        # Remove any special characters
        return re.sub(r'[^\w\s]', '', text)

    new_df = input_df.select(
        'username', 'id', 'original_msg', parse_email_func('original_msg').alias('parsed
    )

    # Define the UDF with a return type
    clean_email_text_udf = F.udf(clean_email_text, StringType())

    # Clean the email text
    new_df = new_df.withColumn('cleaned_text', clean_email_text_udf(new_df['parsed_email

    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.sql.Da
        :returns: transformed dataset
        """
        return dataset.transform(parse_emails)

## Use the custom ParseEmailsTransformer, Tokenizer, and CountVectorizer
## to create a spark pipeline; Use the libraries already added at the top and the class
custom_transformer = ParseEmailsTransformer()
tokenizer = Tokenizer(inputCol = "cleaned_text", outputCol = "words")
count_vectorizer = CountVectorizer(inputCol = "words", outputCol = "features")

# Add the new items to the pipeline.
email_pipeline = Pipeline(stages = [
        custom_transformer,
        tokenizer,
        count_vectorizer
    ]
)
model = email_pipeline.fit(df)
result = model.transform(df)
```

In [17]:

```
result.select('id', 'words', 'features').show()
```

```
+--------------------+--------------------+--------------------+
|                  id|               words|            features|
+--------------------+--------------------+--------------------+
|  davis-d\2_trash\1_|[, , , , , , , , ...|(81402,[0,1,2,3,4...|
|  davis-d\2_trash\2_|[fyi, thanks, , f...|(81402,[0,1,2,3,4...|
|  davis-d\2_trash\3_|[, forwarded, by,...|(81402,[0,1,2,4,6...|
|  davis-d\2_trash\4_|[original, messag...|(81402,[0,2,4,6,7...|
|davis-d\2_trash\c...|[hi, mommy, , yes...|(81402,[0,1,2,4,6...|
|davis-d\2_trash\c...|[hey, sweetie, , ...|(81402,[0,1,6,9,1...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,2,9,18,...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,1,2,3,4...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,2,3,4,6...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,1,2,3,6...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,1,2,3,5...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,2,9,18,...|
|davis-d\2_trash\c...|[, , , , , forwar...|(81402,[0,2,6,9,1...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,2,9,24,...|
|davis-d\2_trash\c...|[are, you, on, th...|(81402,[0,1,6,9,2...|
|davis-d\2_trash\c...|[listen, girly, ,...|(81402,[0,2,3,6,9...|
|davis-d\2_trash\c...|[candis, all, you...|(81402,[0,1,2,3,6...|
|davis-d\2_trash\c...|[what, is, your, ...|(81402,[0,2,6,10,...|
|davis-d\2_trash\c...|[candis, , , why,...|(81402,[0,6,11,37...|
|davis-d\2_trash\c...|[, forwarded, by,...|(81402,[0,2,9,12,...|
+--------------------+--------------------+--------------------+
only showing top 20 rows
```