# Assignment 04 Outputs and Code

For Assignment 4, I started the assignment locally but ran into issues with Assignment 4.3 and SPARK errors popping up. As a result, I finished the assignment using Google Colab with the following outputs and code.

## Assignment 4.1 Code Output:

Table

Description automatically generated

## Assignment 4.2 Code Output:

Graphical user interface, text, application, email

Description automatically generated

Outputs from the parsed email message in html and plain text are shown below:

Andy,

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

Thanks and Best Regards,

Jeff Hammad

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 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 provider. Click here for

a list of tested and supported phones that you can use with the MIK. Cable not provided.

------------------

To modify your profile or unsubscribe from Palm newsletters, click here.

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, PalmModem, PalmPoint, 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 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 owners.

## Assignment 4.3 Code Output:Table Description automatically generated with medium confidence

+------------------+--------------------+--------------------+

| id| words| features|

+------------------+--------------------+--------------------+

|davis-d/2\_trash/1\_|[, >, , , , , >, ...|(3,[0,1,2],[697.0...|

|davis-d/2\_trash/3\_|[----------------...|(3,[0,1,2],[118.0...|

|davis-d/2\_trash/4\_|[-----original, m...|(3,[0,2],[17.0,1.0])|

|davis-d/2\_trash/2\_|[fyi..., thanks.,...|(3,[0,1,2],[57.0,...|

|2\_trash/candis/12\_|[----------------...| (3,[0],[41.0])|

|2\_trash/candis/10\_|[hi, mommy!, , ye...|(3,[0,1,2],[4.0,1...|

|2\_trash/candis/11\_|[hey, sweetie,, ,...| (3,[0,1],[8.0,1.0])|

|2\_trash/candis/13\_|[----------------...|(3,[0,1,2],[65.0,...|

|2\_trash/candis/14\_|[----------------...|(3,[0,2],[58.0,1.0])|

|2\_trash/candis/15\_|[----------------...|(3,[0,1,2],[29.0,...|

|2\_trash/candis/17\_|[----------------...| (3,[0],[43.0])|

|2\_trash/candis/16\_|[----------------...|(3,[0,1,2],[18.0,...|

|2\_trash/candis/19\_|[----------------...| (3,[0],[14.0])|

| 2\_trash/candis/1\_|[are, you, on, th...| (3,[0,1],[1.0,1.0])|

|2\_trash/candis/18\_|[, , , , --------...| (3,[0],[11.0])|

| 2\_trash/candis/3\_|[candis, all, you...|(3,[0,1,2],[16.0,...|

| 2\_trash/candis/2\_|[listen, girly!, ...| (3,[0,2],[9.0,1.0])|

| 2\_trash/candis/4\_|[what, is, your, ...| (3,[0,2],[1.0,1.0])|

| 2\_trash/candis/5\_|[candis, -, , why...| (3,[0],[3.0])|

| 2\_trash/candis/6\_|[----------------...| (3,[0],[16.0])|

+------------------+--------------------+--------------------+

only showing top 20 rows

## Assignment 4 Code:

*## Mounting Drive*

**import** os

**from** google.colab **import** drive

drive**.**mount('/content/drive', force\_remount **=** **True**)

os**.**chdir(r"/content/drive/My Drive/dsc650/dsc650/assignments/assignment04")

**!**pwd

Mounted at /content/drive

/content/drive/My Drive/dsc650/dsc650/assignments/assignment04

In [ ]:

*## Needed to install openjdk to eliminate pysaprk error message.*

**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

*## Install chardet*

**!**pip install chardet

**from** chardet.universaldetector **import** UniversalDetector

*## install pyspark*

**!**pip install pyspark

**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 **=** '/content/drive/My Drive/dsc650/data/external/enron'

*## os.environ["PYSPARK\_PYTHON"] = "C:/Users/jkmey/anaconda3/envs/dsc650/python.exe"*

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]

*## Show the output for column names to prove code executed.*

print(columns)

ParsedEmail **=** namedtuple('ParsedEmail', columns)

*## Print the ParsedEmail to confirm the code executed.*

print(ParsedEmail)

*## Import findspark*

*## import findspark*

*## findspark.init()*

spark **=** SparkSession\

**.**builder\

**.**appName("Assignment04")\

**.**getOrCreate()

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/

Requirement already satisfied: chardet in /usr/local/lib/python3.9/dist-packages (4.0.0)

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/

Collecting pyspark

Downloading pyspark-3.3.2.tar.gz (281.4 MB)

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 281.4/281.4 MB 4.8 MB/s eta 0:00:00

Preparing metadata (setup.py) ... done

Collecting py4j==0.10.9.5

Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 199.7/199.7 KB 17.8 MB/s eta 0:00:00

Building wheels for collected packages: pyspark

Building wheel for pyspark (setup.py) ... done

Created wheel for pyspark: filename=pyspark-3.3.2-py2.py3-none-any.whl size=281824028 sha256=406a931238bd1e589ea63d91393e122c77ac4f3b36cb121f4dfc47b7eed31a53

Stored in directory: /root/.cache/pip/wheels/6c/e3/9b/0525ce8a69478916513509d43693511463c6468db0de237c86

Successfully built pyspark

Installing collected packages: py4j, pyspark

Attempting uninstall: py4j

Found existing installation: py4j 0.10.9.7

Uninstalling py4j-0.10.9.7:

Successfully uninstalled py4j-0.10.9.7

Successfully installed py4j-0.10.9.5 pyspark-3.3.2

['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']

<class '\_\_main\_\_.ParsedEmail'>

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

In [ ]:

'''

Commenting out this entire cell and manually placed the files into the Assignment 4 data folder.

The files were pulled from C:/Users/jkmey/Documents/Github/DSC650\_Course\_Assignments/dsc650/data/external/enron/

and moved to C:/Users/jkmey/Documents/Github/DSC650\_Course\_Assignments/dsc650/dsc650/assignments/assignment04/data/enron.

I did not zip the files because I was running into issues when executing the code with enron.zip. Left files uncompressed.

'''

'''

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()

'''

Out[ ]:

"\ndef copy\_data\_to\_local():\n dst\_data\_path = data\_dir.joinpath('enron.zip')\n endpoint\_url='https://storage.budsc.midwest-datascience.com'\n enron\_data\_path = 'data/external/enron.zip'\n\n s3 = s3fs.S3FileSystem(\n anon=True,\n client\_kwargs={\n 'endpoint\_url': endpoint\_url\n }\n )\n\n \n s3.get(enron\_data\_path, str(dst\_data\_path))\n \n with zipfile.ZipFile(dst\_data\_path) as f\_zip:\n f\_zip.extractall(path=data\_dir)\n \ncopy\_data\_to\_local()\n"

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

## Assignment 4.1

In [ ]:

*## Import additional packages from pyspark*

**from** pyspark.sql.types **import** StructField

**from** pyspark.sql.types **import** StructType

**from** pyspark.sql.types **import** StringType

**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 **=** []

print("enron\_data\_dir", enron\_data\_dir)

**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)

record **=** {}

username **=** os**.**path**.**basename(os**.**path**.**dirname(root))

id **=** username**+**"/"**+**os**.**path**.**basename(root)**+**"/"**+**file\_path

record["id"] **=** id

record["username"] **=** username

record["original\_msg"] **=** read\_raw\_email(current\_path)

records**.**append(record)

*## Complete the code to create the Spark dataframe*

schemaString **=** "id username original\_msg"

fields **=** [StructField(field\_name, StringType(), **True**) **for** field\_name **in** schemaString**.**split()]

schema **=** StructType(fields)

**return** spark**.**createDataFrame(records, schema)

df **=** make\_spark\_df()

enron\_data\_dir /content/drive/My Drive/dsc650/data/external/enron

In [ ]:

df**.**show()

+------------------+--------+--------------------+

| id|username| original\_msg|

+------------------+--------+--------------------+

|davis-d/2\_trash/1\_| davis-d|Message-ID: <1774...|

|davis-d/2\_trash/3\_| davis-d|Message-ID: <2833...|

|davis-d/2\_trash/4\_| davis-d|Message-ID: <1972...|

|davis-d/2\_trash/2\_| davis-d|Message-ID: <2467...|

|2\_trash/candis/12\_| 2\_trash|Message-ID: <5686...|

|2\_trash/candis/10\_| 2\_trash|Message-ID: <1964...|

|2\_trash/candis/11\_| 2\_trash|Message-ID: <7345...|

|2\_trash/candis/13\_| 2\_trash|Message-ID: <7218...|

|2\_trash/candis/14\_| 2\_trash|Message-ID: <3016...|

|2\_trash/candis/15\_| 2\_trash|Message-ID: <1233...|

|2\_trash/candis/17\_| 2\_trash|Message-ID: <1365...|

|2\_trash/candis/16\_| 2\_trash|Message-ID: <2215...|

|2\_trash/candis/19\_| 2\_trash|Message-ID: <8556...|

| 2\_trash/candis/1\_| 2\_trash|Message-ID: <1807...|

|2\_trash/candis/18\_| 2\_trash|Message-ID: <2251...|

| 2\_trash/candis/3\_| 2\_trash|Message-ID: <2977...|

| 2\_trash/candis/2\_| 2\_trash|Message-ID: <2705...|

| 2\_trash/candis/4\_| 2\_trash|Message-ID: <3065...|

| 2\_trash/candis/5\_| 2\_trash|Message-ID: <2798...|

| 2\_trash/candis/6\_| 2\_trash|Message-ID: <3108...|

+------------------+--------+--------------------+

only showing top 20 rows

In [ ]:

df**.**printSchema()

root

|-- id: string (nullable = true)

|-- username: 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 [ ]:

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-24F58854-862567DD-591AE7>

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 program. I enjoyed talking to you, and look forward to contributing to the success that the program has enjoyed.

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?BEReaQA5eczXB=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-images/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 src="http://images4.postdirect.com/master-images/404707/screen1.gif" alt="" width="165" height="159" hspace="0" vspace="0" border="0"></a><br><img src="http://images4.postdirect.com/master-images/404707/screen1\_text.gif" alt="" width="93" height="26" hspace="0" vspace="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 src="http://images4.postdirect.com/master-images/404707/screen2.gif" alt="" width="165" height="159" hspace="0" vspace="0" border="0"></a><br><img src="http://images4.postdirect.com/master-images/404707/screen2\_text.gif" alt="" width="93" height="26" hspace="0" vspace="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.postdirect.com/master-images/404707/screen3\_text.gif" alt="" width="93" height="26" hspace="0" vspace="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 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&#153; 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

<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://images4.postdirect.com/master-images/404707/title1.gif" alt="" width="336" height="20" hspace="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 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 the enhanced onscreen keyboard</LI>

<LI> Improved Date Book functionality lets you view, snooze or clear multiple alarms 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://images4.postdirect.com/master-images/404707/title2.gif" alt="" width="460" height="20" hspace="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 Autolocking 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://images4.postdirect.com/master-images/404707/title3.gif" alt="" width="461" height="31" hspace="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 mobile phones</LI>

<LI> Send applications or records through your cell phone to schedule meetings and 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 href="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 desktop 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 color="#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" width="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-images/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.gif" 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.gif" 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 modem or a phone

with <nobr>built-in</nobr> 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 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 provided.

<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, HandWEB, Graffiti,

HotSync, iMessenger, MultiMail, Palm.Net, PalmConnect, PalmGlove, PalmModem, PalmPoint, 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 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 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.gif" 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>

"""

plain\_msg\_example **=** plain\_msg\_example**.**strip()

html\_msg\_example **=** html\_msg\_example**.**strip()

In [ ]:

*## Test out the email reader for the html path.*

email\_path\_html **=** '/content/drive/My Drive/dsc650/dsc650/assignments/assignment04/examples/html//message-0.txt'

print(read\_raw\_email(email\_path\_html))

Message-ID: <9603832.1075862117623.JavaMail.evans@thyme>

Date: Mon, 26 Nov 2001 09:43:37 -0800 (PST)

From: irene@m-ul.com

To: pmims@enron.com

Subject: 4 FREE Airline Tickets!!

Mime-Version: 1.0

Content-Type: text/plain; charset=us-ascii

Content-Transfer-Encoding: 7bit

X-From: Travel Agent <irene@m-ul.com>@ENRON

X-To: pmims@enron.com

X-cc:

X-bcc:

X-Folder: \PMIMS (Non-Privileged)\Mims, Patrice L.\Deleted Items

X-Origin: Mims-Thurston-P

X-FileName: PMIMS (Non-Privileged).pst

<html>

<head>

<title>Untitled Document</title>

<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">

</head>

<body bgcolor="#FFFFFF"><center>

<table width="540" border="1" cellspacing="0" cellpadding="3" bordercolor="#999999">

<tr bgcolor="#0099FF">

<td bgcolor="#FFFFFF" height="48">

<div align="center"><font face="Verdana, Arial, Helvetica, sans-serif" size="2">If

you do not wish to receive email from Bargain Bazaar, please Click Here:<br>

<a href="http://www.m-ul.com/e/unsub.cgi?l=9994&m=1736394363">UNSUBSCRIBE</a>

</font></div>

</td>

</tr>

</table>

<br>

<table width="540" border="1" cellspacing="0" cellpadding="0" bordercolor="#666600">

<tr>

<td>

<table width="100%" border="1" cellspacing="0" cellpadding="0" bordercolor="#CCFFFF">

<tr>

<td>

<table width="100%" border="0" cellspacing="0" cellpadding="5">

<tr align="center">

<td><img src="http://www.m-ul.com/email/images/accessories/travelagent-header.jpg" width="525" height="80"></td>

</tr>

<tr align="center">

<td>

<table cellspacing="0" cellpadding="0" border="0" align="center" width="601">

<tr>

<td width="18" height="159" valign="top"></td>

<td width="607" height="159" valign="top"><a href="http://www.m-ul.com/e/c.cgi?j=20011125\_62&e=1736394363&r=u424&d=6&p=1"><img src="http://www.m-ul.com/email/images/html102401/header1.gif" width="235" height="149" border="0"><img src="http://www.m-ul.com/email/images/html102401/header2.gif" width="147" height="149" border="0"><img src="http://www.m-ul.com/email/images/html102401/header3.gif" width="223" height="149" border="0"></a></td>

<td width="4" height="159" valign="top"></td>

</tr>

<tr>

<td height="2" colspan="3" valign="top">

<p align="center"><font size="2" face="Times New Roman, Times, serif"><b><font color="#663399" face="Arial, Helvetica, sans-serif" size="5">HERE'S

THE DEAL!</font></b></font></p>

<p align="center"><font size="2" face="Times New Roman, Times, serif"><font face="Arial, Helvetica, sans-serif" size="3">

We've just <b>DOUBLED</b> the value of the Web's hottest

<b>FREE OFFER</b>. Not 2 - but<br>

<b>4 ROUND-TRIP AIR TICKETS - ABSOLUTELY FREE</b>

to exciting destinations across the U.S., Caribbean,

Hawaii or Mexico. Take the family or take two trips!

And except for your hotel stay, there's <b>NO OBLIGATION</b>

and <b>NO PURCHASE REQUIRED</b> for your trip. Imagine

flying your family <b>FREE</b> to Florida, California,

Las Vegas, Jamaica, Aruba, Bahamas, or other exciting

vacation paradise! Plus, you get a 30-day<b> FREE</b>

trial in Preferred Traveller, where you'll save big

whenever and wherever you travel, <br>

dine or entertain!<br>

<b>Act now! This FREE offer is good only for a limited

time! </b></font></font></p>

<div align="center">

<table width="80%" border="1" cellpadding="10" cellspacing="0" bordercolordark="#FFFFFF" bordercolorlight="#663399" align="center">

<tr>

<td>

<div align="center"><font face="Arial, Helvetica, sans-serif"><b><font color="#FF0033">

The 4 FREE AIRLINE TICKETS are yours to keep,<br>

just for registering! Plus, your FREE TICKETS

are<br>

<u>GOOD FOR ONE FULL YEAR</u>! </font></b></font></div>

</td>

</tr>

</table>

</div>

<p align="center"><font size="3" face="Arial, Helvetica, sans-serif"><b><font color="#663399" size="5">HERE'S

ALL YOU NEED TO DO!</font></b> </font></p>

<p align="center"><font face="Arial, Helvetica, sans-serif"><b>What

do you have to lose?<br>

<font color="#FF0033">4 FREE AIRLINE TICKETS!<br>

GREAT TRAVEL SAVINGS!<br>

100% SATISFACTION GUARANTEE! </font></b></font></p>

<p align="center"><b><font face="Arial, Helvetica, sans-serif">Act

now and you'll soon be jetting off to the<br>

vacation destination of your choice! </font></b></p>

</td>

</tr>

</table>

</td>

</tr>

</table>

</td>

</tr>

</table>

</td>

</tr>

</table>

</center>

<img height="1" width="1" src="http://www.m-ul.com/e/e\_open.cgi?j=20011125\_62&mail\_id=1736394363"></body>

</html>

In [ ]:

*## Test out the email reader for the html path.*

email\_path\_plain **=** '/content/drive/My Drive/dsc650/dsc650/assignments/assignment04/examples/plain/message-0.txt'

print(read\_raw\_email(email\_path\_plain))

Message-ID: <14562086.1075853934677.JavaMail.evans@thyme>

Date: Mon, 24 Jan 2000 01:37:00 -0800 (PST)

From: office.chairman@enron.com

To: all.downtown@enron.com

Subject: Over $50 -- You made it happen!

Mime-Version: 1.0

Content-Type: text/plain; charset=ANSI\_X3.4-1968

Content-Transfer-Encoding: quoted-printable

X-From: Office of the Chairman

X-To: All Enron Downtown

X-cc:

X-bcc:

X-Folder: \Dana\_Davis\_Dec2000\Notes Folders\Discussion threads

X-Origin: Davis-D

X-FileName: ddavis2.nsf

On Wall Street, people are talking about Enron. At Enron, we're talking=20

about people...our people. You are the driving force behind every success=

=20

that our company has experienced, including our high-performing stock price=

,=20

which surpassed the $50 mark only a few days ago. You made it happen!

To show our appreciation for your hard work and commitment to Enron=01,s=20

continued success, eligible regular full-time and regular part-time employe=

es=20

who were on the payroll of a wholly-owned Enron company at Dec. 31, 1999 wi=

ll=20

receive 50 Enron stock options. A special stock option award certificate a=

nd=20

a copy of the stock plan will be sent to you in the next few weeks.

About the stock options grant:

The grant was effective Jan. 18, 2000, therefore, the option grant price is=

=20

$55.50.

Options will have a seven-year term, which means you must exercise these=20

options before seven years have passed.

Options will vest 25 percent on Feb. 28, 2000 and 25 percent each subsequen=

t=20

January 18th thereafter until fully vested, as long as you are an employee =

of=20

Enron. This means you may exercise 25 percent of these options as early as=

=20

Feb. 28.

This grant is made in accordance with the terms and provisions of the Enron=

=20

Corp. Stock Plans and the award documents, which you will receive in the=20

coming weeks.=20

Due to varying international restrictions and legalities, Enron employees i=

n=20

certain international locations will receive some other form of recognition=

. =20

Your local management and human resources representative will communicate=

=20

further details.

Visit home.enron.com for additional information. There you can access a li=

st=20

of frequently asked questions regarding stock options. And you can view a=

=20

special message from Ken Lay. To access the special message, you must have=

=20

IP-TV and a sound card. If you are not able to view Ken=01,s message, a=20

transcript is available. If you encounter any technical difficulties, plea=

se=20

contact your PC Help Desk.

To help celebrate this occasion Plaza Java located in the Enron Building wi=

ll=20

offer 50-cent Coffee of the Day and 50-cent Krispy Kreme donuts today only,=

=20

while quantities last.

Again, thank you for all you do to make Enron a successful global energy an=

d=20

broadband services company and a great place to work.=20

In [ ]:

**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*

*## https://docs.python.org/3/library/email.examples.html*

result['payload'] **=** msg**.**get\_payload()

result['text'] **=** parse\_html\_payload(result['payload'])

*## Use error handling to setup key,value within msg.items()*

**try**:

**for** key, value **in** msg**.**items():

result[key**.**replace('-', '\_')] **=** value

**except** Exception **as** e:

print(f"Problem parsing email: email\_path e")

*## use error handling to parse date and convert date with isoformat().*

**try**:

result["Date"] **=** parse(result["Date"], ignoretz **=** **False**)**.**isoformat()

**except** Exception **as** e:

print(f"Problem converting date: {result**.**get('date')} {e}")

tuple\_result **=** tuple([str(result**.**get(column, **None**)) **for** column **in** columns])

**return** ParsedEmail(**\***tuple\_result)

In [ ]:

parsed\_msg **=** parse\_email(plain\_msg\_example)

<ipython-input-11-6a57cbc2b755>:7: MarkupResemblesLocatorWarning: The input looks more like a filename than markup. You may want to open this file and pass the filehandle into Beautiful Soup.

soup = BeautifulSoup(payload, 'html.parser')

In [ ]:

print(parsed\_msg**.**text)

Andy,

Thanks for giving me the opportunity to meet with you about the Analyst/ Associate program. 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 [ ]:

parsed\_html\_msg **=** parse\_email(html\_msg\_example)

In [ ]:

print(parsed\_html\_msg**.**text)

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 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 provider. Click here for

a list of tested and supported phones that you can use with the MIK. Cable not provided.

------------------

To modify your profile or unsubscribe from Palm newsletters, click here.

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, PalmModem, PalmPoint, 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 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 owners.

## Assignment 4.3

In [ ]:

*## This creates a schema for the email data*

email\_struct **=** StructType()

**for** column **in** columns:

email\_struct**.**add(column, StringType(), **True**)

In [ ]:

*## Import bs4*

**import** bs4

In [ ]:

*## 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('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.DataFrame`

:returns: transformed dataset

"""

**return** dataset**.**transform(parse\_emails)

*## Use the custom ParseEmailsTransformer, Tokenizer, and CountVectorizer*

*## to create a spark pipeline*

*## Setup tokenizer and vectorizer to put into pipeline.*

tokenizer **=** Tokenizer(inputCol **=**"text", outputCol**=**"words")

vectorizer **=** CountVectorizer(inputCol **=** tokenizer**.**getOutputCol(), outputCol **=** "features", vocabSize **=** 3)

*## Setup Stages to include tokenizer and vectorizer created above.*

stages **=** [tokenizer, vectorizer]

email\_pipeline **=** Pipeline(stages**=**stages)

*## Complete code*

new\_df **=** parse\_emails(df)

model **=** email\_pipeline**.**fit(new\_df)

result **=** model**.**transform(new\_df)

In [ ]:

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

+------------------+--------------------+--------------------+

| id| words| features|

+------------------+--------------------+--------------------+

|davis-d/2\_trash/1\_|[, >, , , , , >, ...|(3,[0,1,2],[697.0...|

|davis-d/2\_trash/3\_|[----------------...|(3,[0,1,2],[118.0...|

|davis-d/2\_trash/4\_|[-----original, m...|(3,[0,2],[17.0,1.0])|

|davis-d/2\_trash/2\_|[fyi..., thanks.,...|(3,[0,1,2],[57.0,...|

|2\_trash/candis/12\_|[----------------...| (3,[0],[41.0])|

|2\_trash/candis/10\_|[hi, mommy!, , ye...|(3,[0,1,2],[4.0,1...|

|2\_trash/candis/11\_|[hey, sweetie,, ,...| (3,[0,1],[8.0,1.0])|

|2\_trash/candis/13\_|[----------------...|(3,[0,1,2],[65.0,...|

|2\_trash/candis/14\_|[----------------...|(3,[0,2],[58.0,1.0])|

|2\_trash/candis/15\_|[----------------...|(3,[0,1,2],[29.0,...|

|2\_trash/candis/17\_|[----------------...| (3,[0],[43.0])|

|2\_trash/candis/16\_|[----------------...|(3,[0,1,2],[18.0,...|

|2\_trash/candis/19\_|[----------------...| (3,[0],[14.0])|

| 2\_trash/candis/1\_|[are, you, on, th...| (3,[0,1],[1.0,1.0])|

|2\_trash/candis/18\_|[, , , , --------...| (3,[0],[11.0])|

| 2\_trash/candis/3\_|[candis, all, you...|(3,[0,1,2],[16.0,...|

| 2\_trash/candis/2\_|[listen, girly!, ...| (3,[0,2],[9.0,1.0])|

| 2\_trash/candis/4\_|[what, is, your, ...| (3,[0,2],[1.0,1.0])|

| 2\_trash/candis/5\_|[candis, -, , why...| (3,[0],[3.0])|

| 2\_trash/candis/6\_|[----------------...| (3,[0],[16.0])|

+------------------+--------------------+--------------------+

only showing top 20 rows