CSCI 5015 Assignment 6
Due April 3, 11:59 PM.
50 points

Objective

Working with messy data. Create more graphs, including Word Clouds, with matplotlib and seaborn.

Data set

You've given been a ZIP file, **data.zip** that contains archives of five Usenet groups. These files come from the Internet Archive, https://archive.org. The Internet Archive makes many data collections available for download and often preserves historical data that is often lost when a website disappears. Our data comes from the Usenet Historical Collection, https://archive.org/details/usenethistorical, which contains Gigabytes worth of Usenet data from the 80's, 90's and early 2000's.

If you haven't heard of Usenet before, it was a large discussion board that allowed people to gather online and discuss topics of interest, for example, Linux, movies, cooking, whatever your interest was. Think of it as a completely text-based Reddit. For more information, read about its history at the usual place, Wikipedia, https://en.wikipedia.org/wiki/Usenet.

Once you get your homework done, you may want to check out the Archive's Internet Arcade collection, https://archive.org/details/internetar which is a library of arcade (coin-operated) video games from the 1970s through to the 1990s, and playable in your browser. Remember, **after** you complete your homework.

The archives you been given come from the following groups

- comp.org.acm - announcements from the Association for Computing Machinery, the world's largest scientific and educational computing society
- rec.arts.origami - discussions about folding paper
- sci.fractals - discussions about fractals, the king of mathematical eye candy
- linux.dev.kernel - discussions between the developers who created Linux
- rec.food.chocolate - mmmmmmm, chocolate

The data is in the MBOX format (MBOX stands for MailBOX). MBOX is a text-based format that contains a set of messages. Messages are divided into headers and the content of the message. Headers for a message come first and have an identifier, a colon, and then some data after it. For example,

```
From 3072346469177122312
X-Google-Thread: 1160c2,91073f25a01ab25d
Path: g2news2.google.com!news1.google.com!news3.google.com!newshub.sdsu.edu!tethys.csu.net!okeanos.csu.net!53ab2750!not-for-mail
From: Tyler Spaeth
Newsgroups: alt.arts.origami,rec.arts.origami,
Message-ID: <20070114121111116807-cadix@maccom>
References:
MIME-Version: 1.0
Content-Type: text/plain; charset=ISO-8859-1; format=flowed
Content-Transfer-Encoding: 8bit
Subject: Re: trying to locate stores
User-Agent: Unison/1.7.7
Lines: 25
```

The one exception, *I think*, is the first header item, the first **From**, which does not have a colon and is followed by a large positive or negative integer. I think that line starts each message, but don't count on it. Also don't rely on every message having all the same headers. Depending on when the message was sent and where it was sent from, a message may get a different set of headers than other messages in the same archive.

Even the values for headers are inconsistent. For example, here are several examples of the various date-time formats that can be found in the files for the **Date:** header.

- Date: 1996/06/21
- Date: 21 Jun 1996 11:49:45 +0200
- Date: Mon, 5 Oct 1992 22:26:07 GMT
- Date: 23 Sep 92 17:05:49 GMT

After the headers, is the message body. Any text there is pretty much valid since a person can type whatever they want.

Warning: Usenet could be used by anyone, much like Reddit. People talked about their opinions and could often get fired up, using not-so-polite language. I have tried to pick archives of groups that would normally avoid over-the-top conversations and comments, but given the amount of the data here, you may see some inappropriate language. Such is the nature of data.

Coding instructions

Create a Jupyter notebook file named **assignment6.ipynb**. Unzip the file, **data.zip** and place the five TXT files in the same folder as your notebook.

Make the first cell in your notebook a Markdown cell and create a header that says "Usenet Analysis" and then on the next line, have a plain paragraph that says "Written by John Nicholson" (hopefully, it is obvious that you should replace my name with your name). It should look roughly like this after you render the Markdown:

Usenet Analysis

Written by John Nicholson

Use your notebook to answer the questions below. Your notebook should have both code cells and Markdown cells. The code cells should read and process the data, and the Markdown cell should explain what the associated code cell does and why you wrote the code the way you did. If you wish, add extra cells to make your main code easier to read and use.

Hints

Before jumping into making a graph, print the data to make sure you are capturing the right thing and to help you understand what you are working with. This is especially important when you need to do some data clean-up.

Using regular expressions may or may not help on some questions. Sometimes, it is easier to make more than one regular expression rather one giant regular expression that tries to do everything.

Requirements

- No single code cell should produce all the output for more than one question. At a minimum, you will need 6 code cells, one for each question.
- If you make additional code cells, they do not have to display data. For example, you may want to write a helper function in a code cell. You still need a Markdown cell describing the function and why you wrote the function in the first place.
- All code cells need a Markdown cell explaining what the code does and why you wrote it the way you did. Explanations should be complete sentences and paragraphs, use good grammar, and be spell-checked.
- Before turning in your notebook, ensure that all code cells that produce output have their output displayed. I should not have to run your notebook in order to see the output. You should all render all Markdown cells. That is, I should not see the special Markdown characters like the `#`. I should see nicely rendered text.
- For all questions with a plot or graph, you must
    - plot a graph in addition to the required Markdown cell that explains why you wrote the code the way you did
    - Use a different set of colors for each graph
    - Give each graph an appropriate title. For seaborn assign the plot to a variable and then use that variable to call **`set_title`**

      ```
      ax = sns.countplot(...
      ax.set_title('My awesome title')
      ```

      For matplotlib, give the plot function a title argument, e.g. **`title='My awesome title'`**

Questions

1. How many messages are in each archive?

   Make a countplot that shows how many messages are in each archive. We will assume the **From** with no colon followed by a large positive or negative integer can be used to identify the start of a message, so count those lines.

   In addition to making the countplot, print the message counts for each archive below the graph.

2. For each archive, what is the average number of lines per message?

   Make a countplot that shows the average number of lines per message in each archive. Use the header **Line:** to find the number of lines. When computing the average, divide by the number of **Line:** headers you found, not the **From** header we used above.

3. Who were the top 5 posters in each group?

   You do not need to plot this. Create a nice, printed report for each archive listing the user, and the number of postings they made.

   Use the **From:** header to identify message posters. Some people changed email addresses over time or changed their email software so that their **From:** value changed. That means you may see the same person listed more than once in your top 5. For example,

   - From: "David C. Ullrich" <ullrich@math.okstate.edu>
   - From: David Ullrich <ullrich@math.okstate.edu>
   - From: David C. Ullrich <david_ullrich@my-deja.com>

   These are probably the same person. The first two definitely are, but one has quotes and one doesn't. You do not have to fix this. Just count each unique value after **From:** header label. If your top 5 for an archive lists what is probably the same person more than once, as long as the name difference similar to what you see above, we will count them as unique people.

4. For all messages, which mail programs (or user agents) were the most popular for sending messages to Usenet?

   Make a countplot of the top 15 user agents. Sort the counts so that they appear from the highest count to the lowest.

   There are two headers **User-Agent:** and **X-Http-User-Agent** that can be used to identify what mail program was used to send a message. The values for these are all over the place, but there is some consistency. You get two basic formats

   - User-Agent: Mozilla/5.0 (X11; U; Linux i686; en-US; rv:0.9.9) Gecko/20020513
   - User-Agent: Mailgate Web Server

   In the first case there is a basic identifier followed by a slash with lots of specifics about the user agent. In those cases, just extract the user agent between the colon and the first forward slash, which would be **Mozilla** above. In the second case, there is no slash, so just extract everything after the colon, which would be **Mailgate Web Server** above.

5. As the World Wide Web grew, use of Usenet fell. Does our data support this statement?

   Extract the year from each **Date:** header. Remember, there are many formats in the fields you you will need to clean up the values and make them consistent. Count how many times the year appears for each archive.

   When you plot, create one graph that has years in the x-axis and counts in the y-axis. Draw five lines one for each archive's count. Make sure to have a legend on the graph so that you know which line represents which archive. You can start with the single basic plot from matplotlib documentation as a starting point, https://matplotlib.org/gallery/lines_bars_and_markers/simple_plot.html.

   Be aware that you are being asked to graph something and then use that graph to answer the question. Make sure to add a write-up below the graph explaining your interpretation of the graph.

6. Word clouds can act as a type of "fingerprint" for a set of data. Does our data support that statement?

   Create a word cloud for each archive. You should not pass the header lines to the word cloud library, make sure to only pass message data. Create five word clouds, one for each archive.

   Be aware, you need to discuss your interpretation of the word clouds in relationship to the statement above.

Tips and help

Remember, we talked about files and listing directories in a previous lecture.

When you are running code, you may not see results right away. You should look to the left of the code cell for the **`In [ ]:`** marker. If you see a number, for example, **`In [8]:`**, the code in the cell is not currently running. But if you see an asterisk, for example, **`In [*]:`**, the code in the cell is running on the kernel, and results may not be available yet. The asterisk should change to a number when the code is done.

Remember, you can stop a running cell by choosing the "stop" button in the menu. It is the one that looks like a black square.

You can always come see me in my office or send me email. If you send me email, you should send your code as an attachment. Don't copy/paste your code into the message because that will make it harder for me to debug your code. Send your file as a **`.ipynb`** file

Submission

When your program is correct, log into D2L and locate the Dropbox for assignment 6. Upload the file

- assignment6.ipynb

into the D2L dropbox.

Contact me if you have any problems.