

Applied Machine Learning

Lab 3 – Working with Text Data

Overview

In this lab, you will use R or Python to work with text data. Specifically, you will use code to clean text, remove stop words, and apply Porter stemming to the remaining words.

What You'll Need

To complete this lab, you will need the following:

- An Azure ML account
- The files for this lab

Note: To set up the required environment for the lab, follow the instructions in the [Setup Guide](#) for this course.

Exploring Text Data

The text data used in this lab consists of a collection of tweets that have been categorized as positive or negative.

Explore the Tweets Dataset

1. In the folder where you extracted the lab files for this module (for example, C:\DAT203.3x\Lab03), open the **tweets.csv** file, using either a spreadsheet application such as Microsoft Excel, or a text editor such as Microsoft Windows Notepad.
2. View the contents of the **tweets.csv** file, noting that it contains tweets with a numeric indication of sentiment in which 4 indicates a positive tweet, and 0 indicates a negative tweet:

	sentiment_label	tweet_text
1		
2	4	@elephantbird Hey dear, Happy Friday to You Already had your rice's bowl for lunch ?
3	4	Ughhh layin downnnn Waiting for zeina to cook breakfast
4	0	@greeniebach I reckon he'll play, even if he's not 100%...but i know nothing!! ;) It won't be the same without him.
5	0	@vaLewee I know! Saw it on the news!
6	0	very sad that http://www.fabchannel.com/ has closed down. One of the few web services that I've used for over 5 years
7	0	@Fearnecotton who sings 'I Remember'? i alwaysss hear it on Radio 1 but never catch the artist
8	4	With God on ur side anything is possible....
9	0	@LoveSmrs why being stupid?
10	0	Having dived back into the guts of Expression Engine, its a flexible CMS if you have to use it as a dev, not great for client
11	0	@emoskank awww take him with you!
12	4	the video on VH1 is much better than the u tube one
13	0	i ran out of champagne..... i wonder if i call my brother i could convince him to bring me up a bottle....
14	0	@carolinefjones I wish I was going to the show tonight.
15	0	Doing homework..then bed..waking up at 4 is gonna be awful
16	4	@AshleyTMSYF Hey Ashley when will the Hush Hush; Hush Hush video be out? Can't wait)
17	4	Scratch that I enjoy seein people that left for college and came back for summer it makes me chuckle
18	4	@demi_superfan1 hey im good sorry i took so long 2 reply and im just chillin listening 2 music wbu?
19	4	@Lydiajohn13 Good morning!!! You're up early... Chilling and chomping so far.. How is yours? And how was training?
20	4	@Joni2281 omggg no lol! i saw it was a trending topic, and i think its been released! i have to get it soon hehe
21	0	@danielak Who will say "Good Morning" when I head to bed over the next month Seriously have a marvelous
22	4	@katyperry enjoy your time with your family
23	0	Okay so it's quarter past midnight on a school night and I'm still awake! Gotta be up in like 6 hours
24	4	@yelyahwilliams http://twitpic.com/6u375 - Add velcro for more fun

3. Close the data file without saving any changes.

Explore the Stopwords Dataset

1. Open the **stopwords.csv** file and review its contents. Note that this file contains a list of common words such as “a”, “the”, “it”, and so on, which are generally not helpful in determining the meaning or sentiment of a sentence or paragraph.
2. Close the file without saving any changes.

Upload the Datasets to Azure Machine Learning

1. Browse to <https://studio.azureml.net> and sign in using the Microsoft account associated with your free Azure ML account.
2. If the **Welcome** page is displayed, close it by clicking the **OK** icon (which looks like a checkmark). Then, if the **New** page (containing a collection of Microsoft samples) is displayed, close it by clicking the **Close** icon (which looks like an X).
3. At the bottom left, click **NEW**; and in the **NEW** dialog box, in the **DATASET** tab, click **FROM LOCAL FILE**. Then in the **Upload a new dataset** dialog box, browse to select the **tweets.csv** file from the folder where you extracted the lab files on your local computer. Enter the following details, and then click the ✓ icon.
 - **This is a new version of an existing dataset:** Unselected
 - **Enter a name for the new dataset:** tweets.csv
 - **Select a type for the new dataset:** Generic CSV file with a header (.csv)
 - **Provide an optional description:** Tweets.
4. Wait for the upload of the dataset to complete, then click **OK** on the status bar at the bottom of the Azure ML Studio page.

5. Repeat the previous steps to upload the stopwords.csv file as a new dataset with the following properties:
 - **This is a new version of an existing dataset:** Unselected
 - **Enter a name for the new dataset:** stopwords.csv
 - **Select a type for the new dataset:** Generic CSV file with a header (.csv)
 - **Provide an optional description:** Stopwords.

Working with Text Data in Jupyter

Now you are ready to use R or Python code in a Jupyter notebook to work with the text data.

Upload a Jupyter Notebook

1. In Azure ML Studio, click **NEW**; and in the **NEW** dialog box, in the **NOTEBOOK** tab, click **Upload**. Then in the **Upload a new notebook** dialog box, browse to select the notebook file for your preferred language (R or Python) from the folder where you extracted the lab files on your local computer – the R version of the notebook is named **TextPrep_R.ipynb**, and the Python version is named **TextPrep_Py.ipynb**. Enter the following details, and then click the ✓ icon.
 - **Enter a name for the new notebook:** TextPrep_R or Text_Prep_Py
 - **Select a language for the new notebook:** R or Python 2
2. Wait for the upload of the notebook to complete, then click **OK** on the status bar at the bottom of the Azure ML Studio page.

Use Code to Work with the Time Series Data

1. In Azure ML Studio, on the Notebooks tab, open the **TextPrep_R** or **Text_Prep_Py** notebook you uploaded in the previous procedure.
2. Follow the instructions in the notebook to work with the time series data.
3. When you have completed all of the coding tasks in the notebook, save your changes and then close and halt the notebook.

Summary

In this lab, you have used R or Python in a Jupyter notebook to work with text data.