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

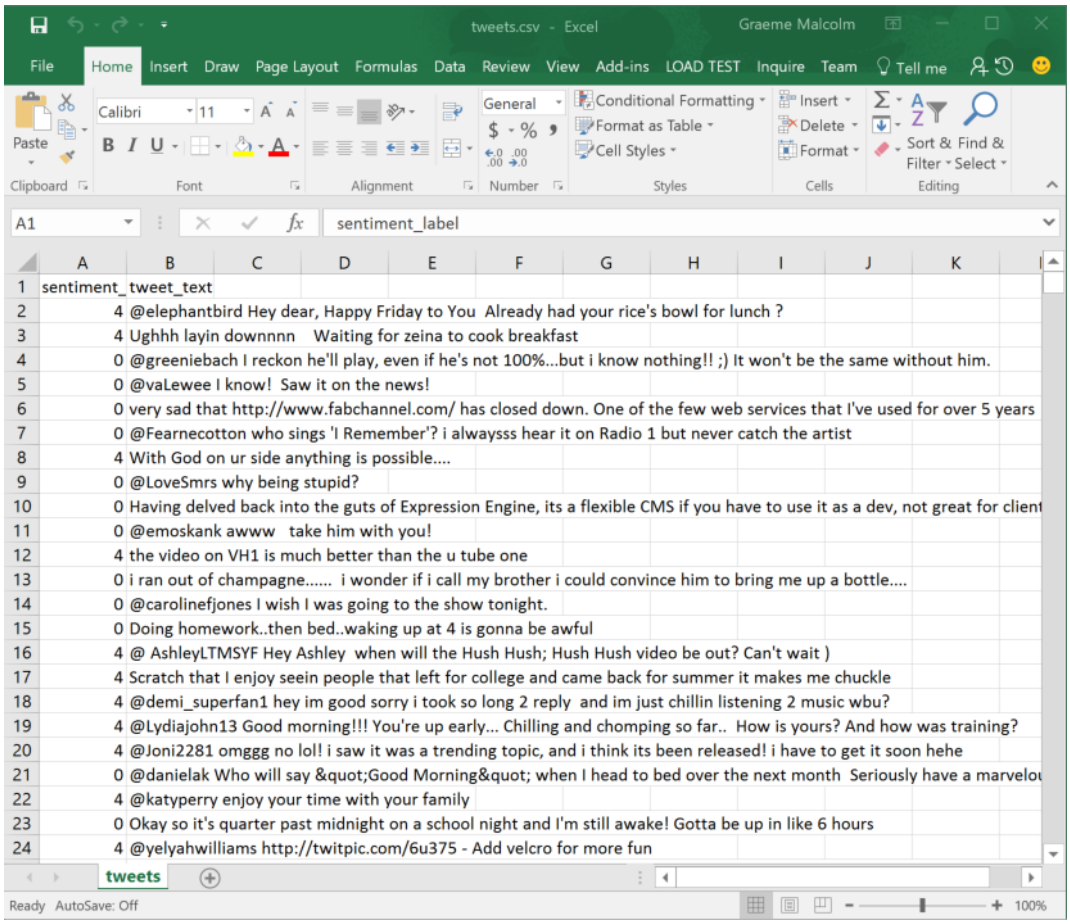
Note: To set up the required environment for the lab, follow the instructions in the Setup Guide for this course.

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:



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.ipnyb**, 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.