DBL Data Challenge

Group 10

Ver 0.6
04/July/2018

Zhen Tian 1288423
Charchika Sinha 0987861
Leon Willems 1288563
Maren Buermann 1287230
Ilona Toikka 1251805
David Sebastiaan Dietzenbacher 1235266
Yuqing Zeng 1285843

Backup full sample link

If any of the content is not working, the following link should provide a workable test sample.
https://drive.google.com/file/d/1U6GptcdqjCIJnNmANKUpfRw5Kf0Edygz/view?usp=sharing

Table of Content

File Structure

/zip_folder

Contains required test json zips

/extract (Automatically generated)

Contains extracted json file

/obj (Automatically generated)

Contains csv file and pickle file

/R code (Automatically generated)

Contains R code to run t test, see /R code/RcodeReadMe.txt

./

Contains all jupyter notebooks

Please execute by index sequence

Contains core.db and sqlite3 database

has twitter, user, conversation tables. (Automatically generated)

Contains README file

With PDF and word doc

Jupyter notebooks

Environment setup

0_setup_env.ipynb

Build database

1_unzip.ipynb
2_load_database.ipynb

Create conversation list pickle and insert into database

3_make_conversations_list.ipynb
4_conversation_table.ipynb

Functional jupyter notebooks

- Word frequency count
    - 5_wordcount.ipynb
    - 6_word_sentiment_change.ipynb
- Visualization
    - 7_hex_plot_with_R_code.ipynb
    - 8_violinplots.ipynb
    - 9_sentiment_barplot.ipynb
    - 10_sentiment_heatmap.ipynb
    - 11_tweet_volume_heatmap.ipynb

Required tools & Modules (0_setup_env.ipynb will help you to install packages)

- Anaconda 3
    - Jupyter Notebook
    - Python 3.6
        - os
        - sys
        - pandas
        - zipfile
        - sqlite3
        - json
        - time
        - matplotlib : version 2.0.2 required
        - seaborns
        - numpy
        - msgpack
        - TextBlob
        - pickle
        - vaderSentiment
        - nltk
            - Package ('punkt')
            - Package ('stopwords')
- R and RStudio
    - R package "BSDA"
- Windows 10

Accessing data

1. Please put the required zip files in the 'zip_folder'.
2. Run jupyter notebooks from 0-11(Kernel -> Restart Kernel and Run All Cells), when 'Done' is printed, the running of one jupyter notebook is completed.
3. R code is used to run t test
4. If any of the code is not working, please download sample code with the link below.

https://drive.google.com/file/d/1U6GptcdqjCIJnNmANKUpfRw5Kf0Edygz/view?usp=sharing