

PhD Student Workshop
Web Mining for Communication Research
April 22-25, 2014

Web Data Analysis

Instruction

Lectures: Cheng-Jun Wang

Introduction: There are five demonstrations in this file. The five demonstrations include (1) installing software; (2) simulating networks; (3) describe networks with NodeXL and R; (4) Exponential random graph modeling (ERGM) of US. Congressmen's retweet network; and (5) Sentiment analysis with supervised machine learning.

To start your practices, download the **R scripts** and **dataset** first

1. <http://weblab.com.cityu.edu.hk/workshops/com2014/CongData.zip>
2. <http://weblab.com.cityu.edu.hk/workshops/com2014/RDemo.zip>

Save them in any folder, for example: **D:/demos/**

You can inspect every step the demos by checking each subfolder in the demos:

1. D:/demos/demo1_install_softwares/R_demos.html
2. D:/demos/demo2_simulate_networks/simulate_networks.html
3. D:/demos/demo3_describe_the_network/describe_the_network_with_igraph.html
4. D:/demos/ demo4_ergm_analysis/ergm of retweet network.html
5. D:/demos/ demo5_sentiment_analysis/sentiment_analysis.html

You can also view them online:

1. http://chengjun.github.io/web_data_analysis/demo1_install_softwares/
2. http://chengjun.github.io/web_data_analysis/demo2_simulate_networks/
3. http://chengjun.github.io/web_data_analysis/demo3_describe_the_network/
4. http://chengjun.github.io/web_data_analysis/demo4_ergm_analysis/
5. http://chengjun.github.io/web_data_analysis/demo5_sentiment_analysis/

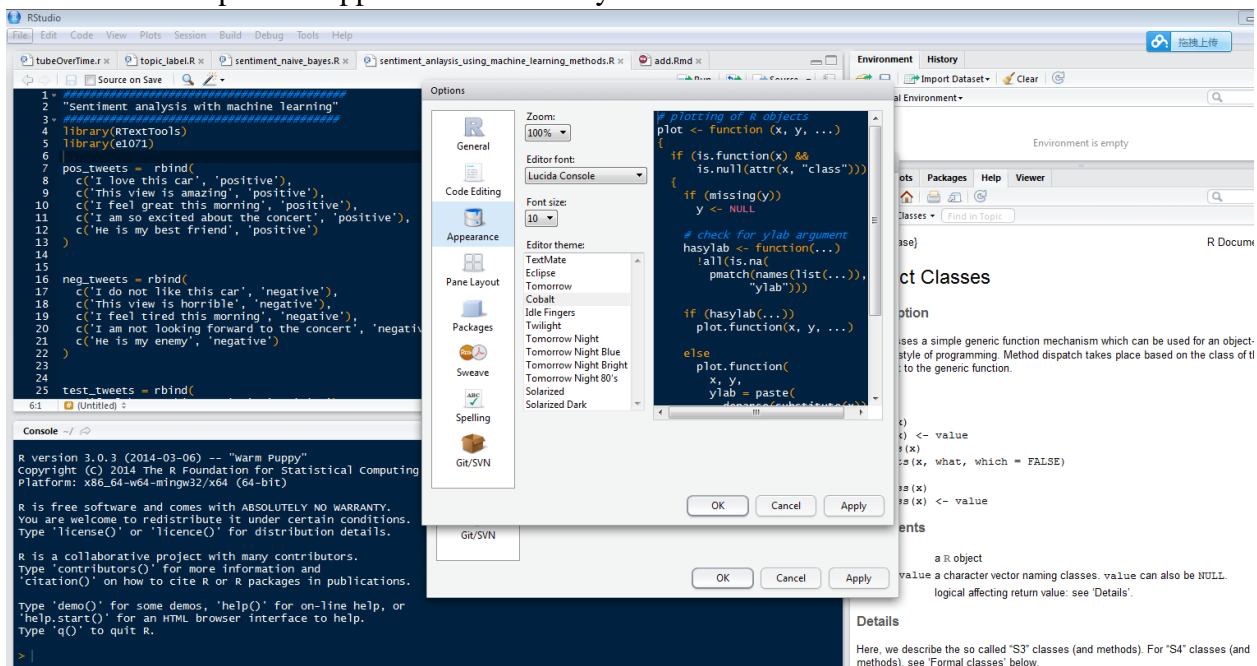
Demo 1: installing software

Step 1. Download and install R, Rstudio, and NodeXL

- <http://cran.r-project.org/>
- <https://www.rstudio.com/ide/>
- <http://nodexl.codeplex.com/>

Step 2. Customizing Rstudio

Tools->Global options->appearance->choose your favorite editor theme



Step 3. Following the **Demo1 R_Basics.R** in the subfolder of D: /demos/R scripts

Secret: From this moment on, you become a hacker!

Homework 1:

Learn the basics of R <http://tryr.codeschool.com/>

More information on R: <https://www.rstudio.com/training/online.html>

Demo2 Simulate the networks

Step 1: following the **Demo2 simulate_networks.R** script

Step 2. Installing **igraph** in Rstudio

Step 3. Simulating the networks with igraph

Step 4. Plot degree distribution

Step 5. Fit power law

Demo3 Describe the Network

Step 1. following the **Demo3 describe_network_with_igraph.R** script to set work directory and load in data

Step 2. Build up network object and set network attributes

Step 3. Plot networks

Step 4. Calculate Graph Statistics

Step 5. Calculate Centrality Measures

Step 6. Calculate the Shortest path

Step 7. Plot the network diameter

Demo 4. ERGM of U.S. congressmen's retweet networks

Step 1. following the **Demo4 ERGM.R** script to Set work directory and load in data

Step 2. Build up network object and set network attributes

- Detach igraph and install statnet

Step 3. Plot networks with statnet

Step 4. Model specification of ERGM

Step 5. Mcmc diagnostics

Demo 5. Sentiment analysis with supervised machine learning

Step 1. install RTextTools and e1071 [**Exercise**]

Step 2. following the **Demo5 sentiment_analysis.R** script and Read data

Step 3. Build up document-term matrix

Step 4. Train the model

Step 5. Test the training model

Step 6. The other machine learning methods

Homework 2

Dance with the process again with a new data:

Download Data Zip from here: <https://github.com/victorneo/Twitter-Sentimental-Analysis>

Compare your results with the given R script [Try your best first]