Teaching Social Media Analytics in R

Huaxia Rui

Simon Business School, University of Rochester

2017 IT Teaching Workshop, UC Irvine
Agenda

1. Design
2. Sample Labs
3. Sample Case Project
4. Lessons
Agenda

1. Design
   - Background
   - Infrastructure
   - Course Content

2. Sample Labs

3. Sample Case Project

4. Lessons
In 2014, I started developing a new course, **social media analytics**, in an effort to help Simon students understand and analyze social media in the business context.

### Timeline

- **2014-2015**: offered the first time as an MBA elective course with a total enrollment of 25 students in 1 section.
- **2015-2016**: 2 sections with a total enrollment of 77, including 37 in the MSBA section and 40 in the MBA section.
- **2016-2017**: 2 sections with a total enrollment of 133, including 99 in the MSBA/MSMA section and 34 in the MBA section.
Course Highlights

- Very hands-on, with a mixture of lectures and labs.
- I chose R as the programming language for all hands-on exercises to be consistent with other courses at Simon.
- I used a cloud-based R environment to give students a homogeneous platform for all class activities.

  http://r1.twittersensor.com
  ...
  http://r9.twittersensor.com

  Username: IT2017    Password: UCIrvine
Hardware
### Software

#### Servers

- **Public Frontend Server** (1): direct students to different nodes
- **Internal RStudio Server** (9): R interface for students to interact with data
- **Internal Data Server** (1): central database storing all data used in the course

- I used CentOS 6.5/7.0 as the OS for servers.
- I run a web server on the frontend server which is public facing. The rest of the servers hide in a private network.
- Potential monitoring of user activities in the cloud for teaching analytics.
Software

Administrative Programs

- **account management**: create/delete user accounts based on student enrollment data
- **cluster management**: distribute student data in the cluster
- **maintenance**: backup, cleaning, restoration
Topics

Social Media Marketing

- offensive marketing: brand campaigns on social media
- defensive marketing: customer service on social media

Mining Social Media Content

- managing text data in R
- syntactical analysis
- vector semantics and latent semantic analysis (LSA)
- clustering and topic modeling: pLSA, LDA, sLDA, ...
- text classification: naive bayesian, maximum entropy, SVM, ...

Deep Learning for Textual Analytics

- feedforward neural network language model, CBOW/skip-gram
- recurrent neural network (RNN) and long short-term memory (LSTM)
Structure

Lecture, Lab, & Case Project

- Half of each class for lecturing theories & algorithms or case discussion.
- Half of each class for lab during which students are given some data and are expected to complete some task individually or in groups.
- TA sits in the lab part of each class.

Homework & Exam

- Individual assignments are similar to lab exercises but are typically simpler.
- Final exam is either a take-home project or a computer-based test, both individualized to prevent cheating.
In-Class Exam

Each student is given an individualized csv file containing some tweets sent by customers to airlines. Most of these tweets are complaints, but some are not. The task is to find out those tweets that are NOT complaints.

Grade for this task will depend on the precision, the recall, and the scalability of the proposed approach.
Teaching Social Media Analytics in R

Design

Course Content

Sample Exam

Take-home Project

- Each student selects a brand he/she is interested in. Selections are mutually exclusive.
- Students then submit brand keywords through a Web interface.
- Each student will receive a link to download 1% of all tweets containing those keywords from 2011 until last month.
- Each student will then analyze how the topics and sentiment of brand-related tweets change over time and submit a report along with the code.
Agenda

1 Design

2 Sample Labs
   - Sample Lab: Clustering Physician Reviews
   - Sample Lab: Discovering Topics on Twitter
   - Sample Lab: Deep Learning with Text

3 Sample Case Project

4 Lessons
Latent Semantic Analysis

Task

- Students are given 300 documents.
- Each document contains an online review submitted by a patient on a doctor.
- Use LSA to analyze the corpus and construct its 2-dimensional semantic representation.
- Cluster the corpus into 3 clusters and compare the cluster with the specialty of the doctor being reviewed.
3 Clusters
Ebola

- 09/25/2014: Duncan seeks care at Texas Health Presbyterian Hospital for a fever, sharp headache and other symptoms. He is sent home with antibiotics.
- 09/28/2014: Duncan returns to the hospital in an ambulance and is isolated.
- 09/30/2014: Tests performed by the CDC and a Texas laboratory confirm that Duncan has Ebola, the first case diagnosed on U.S. soil.
Tweets on Ebola

Figure: The data is from a random sample of tweets mentioning the word *ebola* from 10/01/2014 to 10/20/2014. This figure shows the word cloud for 10/14/2014.
NIH director blames budget cuts for lack of Ebola vaccine, but why did they waste money on this stuff?! The Right http://t.co/pcVSdXir0D

Ad blames GOP for Ebola Budget Cuts; wait, what, we cut the budget? when? How much? #TCOT

NIH Director: We’d Probably Have An Ebola Vaccine If Not For Budget Cuts http://t.co/4rAEH6nrX3 #uniteblue #PDMFN
Heathrow to start screening for Ebola: Heathrow airport is to start Ebola screening for passengers flying into... http://t.co/sHOj3tt3ig

#travel #tourism Ebola crisis: ’huge disruption’ expected at Heathrow as screening begins: Experts have warned... http://t.co/j0pfV4tpnS
Sample Topic from LDA

HAIR OF THE DOG, BABY! "@WGRZ: Texas nurse infected with Ebola receives blood transfusion from Ebola survivor."

Doctor gives blood for Ebola-infected Dallas nurse: Sections. SFGatevisit|Home|navigation-www|1 SFGatevisit|... http://t.co/OBBuPsUbJo

YOU CAN BEAT IT: Ebola survivor donates plasma to sick Dallas nurse... http://t.co/W5n6i5XjEo
Sample Topic from LDA

Business #Economies #US_News #CDC #Ebola #Facebook Zuckerberg and Wife Give CDC $25 Million to Fight Ebola: NEW http://t.co/fyz700ueG0

These are the real men of God "@ye2ndeh: Mark Zuckerberg Donates $25 Million To Combat Ebola http://t.co/70HBIe7DPN #NGTrends"

@zaynmalik donate money to help fight Ebola
Teaching Social Media Analytics in R

Sample Labs

Sample Lab: Discovering Topics on Twitter

**Term Frequency: 15:00 – 16:00**

Figure: One hour of random tweets in 2014
Sample Topic: 15:00 – 16:00
Sample Topic: 15:00 – 16:00

Keywords:
- airlines
- people
- follow
- come
- right
- photo
- malaysia
- good
- new
- back
- make
- much
- adviser
- automatically
- still
- another
- hate
- one
- plane
- love
- shot
- ukraine
- love
- passenger
- morning
- breaking
- thing
- on
- way
- guys
- crash
- check
- know
- carrying
- twitter
- want
- flight
- help
- know
- anyone
- say
- make
- reports
- still
- much
- still
- don't
- back
Routes of MH17

MH17

13:00

Dnipropetrovsk

13:10

Donetsk

13:20

Hrabove

13:10

Torez

13:20

Russia

Sample Lab: Discovering Topics on Twitter

Sample Labs

Teaching Social Media Analytics in R
model=word2vec(train_file = "data.txt",
               output_file = "vec.bin",binary=1)

bin_to_txt("vec.bin","vector.txt")
anarchism originated as a term of abuse first used against early working class radicals including the diggers of the English revolution and the sans culottes of the French revolution whilst the term is still used in a pejorative way to describe any act that used violent means to destroy the organization of society it has also been taken up as a positive label by self defined anarchists the word anarchism is derived from the greek without archons ruler chief king anarchism as a political philosophy is the belief that rulers are unnecessary and should be abolished although there are differing interpretations of what this means anarchism also fits related social movements that advocate the elimination of authoritarian institutions particularly the state the word anarchy as most anarchists use it does not imply chaos nihilism or anomic but rather a harmonious anti authoritarian society in place of what are regarded as authoritarian political structures and coercive economic institutions anarchists
- Similar countries next to each other, and capital cities arranged in parallel.

\[ V(Paris) - V(France) + V(German) \approx V(Berlin) \]

\[ V(King) - V(Man) + V(Woman) \approx V(Queen) \]

- Abstract concepts can be learnt just by reading lots of news articles.
Agenda

1. Design
2. Sample Labs
3. Sample Case Project
4. Lessons
Teaching Social Media Analytics in R

Sample Case Project

Social Media Customer Service

@Spotify @SpotifyCares Hey spotify! I hate you need wifi before use offline mode. Why i just cant do it with my phone internet connection?
11:40 PM - 12 Nov 2011

@SpotifyCares @SpotifyCares - Nov 20
@AlaraPerimm 2: to Settings > Music Quality and turn off the 'Sync Over Cellular' option on NW

Alara Perimm @AlaraPerimm - Nov 20

@AlaraPerimm @SpotifyCares Okay thanks to your reply, finally i found it. Now i love you more

View other replies

Ali Desidero @AliDesidero21 - Nov 20
@AlaraPerimm @SpotifyCares hey dude I she is my girl and you're dead

@AliDesidero21 Hey Ali! This seems appropriate: spoti.fi/1WcDRXJ 3/J

SpotifyCares @SpotifyCares - Nov 20

@AliDesidero21 Ali Desidero21 Hi Ali Desidero21 Hey Ali! This seems appropriate: spoti.fi/1WcDRXJ 3/J

Ali Desidero @AliDesidero21 - Nov 20
@SpotifyCares last chance fancy pants. You wanna fight with me?
open.spotify.com/track/47ZqWjW... pic.twitter.com/pvcXORUJ3u

Huaxia Rui @huaxia

@UPS Package shown delivered but couldn't find it. Tracking: 1Z21E11A0337255638
11:54 PM - 15 Feb 2016

Reply to @UPS

UPS Customer Support @UPSHelp - Feb 16
@huaxia Oh no! Sorry about that. Please email us details, tracking # and contact # to help@ups.com. We will check on it. ^KD

Huaxia Rui @huaxia - Feb 17
Got the package from neighbor. @UPSHelp
“We started on Twitter as a one-way dialogue...lo and behold, what was most surprising is that our customers wanted to speak to us!” — Brooks Brothers

“Twitter, unlike anything else, afforded us an opportunity to honor our vision of hospitality, at any time around the globe.” — Hilton
In contrast with traditional call centers, airlines respond to less than half of the tweets directed at them by complaining customers!
Not All Customers Are Created Equal

Forbes

- Genesys, a global omni-channel customer experience and contact center solution provider for business clients including major airlines, banks and telecommunications companies, integrated Klout score into its solutions.
- LiveOps, a cloud call center company, was implementing ways of prioritizing which customers to contact first, based on the number of Twitter followers customers have.

MIT Technology Review

Citibank, McDonald’s, Delta Airlines, and Coca-Cola are among the companies that can pull up a Klout score, according to Jesse Engle, the CEO and cofounder of CoTweet, which incorporates Klout into its CRM software and counts those four companies as customers. "Everyone engaging in social media is constantly making judgments about who to engage with and how to engage," Engle says.
Prioritization Based on Influence?

VP of Air France-KLM

The audience we have is very suspicious towards famous people getting treated differently. This is not our policy, because we know it will backfire on us.

Hello @AlaskaAir! I heard customers with more followers are more likely to be responded and more likely to be responded faster. True?

3:14 PM - 21 Feb 2016

@Alaska Airlines @AlaskaAir - Feb 21
@yinglu No, that is not true. We respond in the order we receive messages
-Rachel
Prioritization Based on Influence?

Hello @united! I heard customers with more followers are more likely to be responded and more likely to be responded faster. Is this true?

3:09 PM - 21 Feb 2016

Not responded as of this date

S&P500 closed at 2,047.6 on April 8, 2016
Two Polls

As a customer, do you think it's fair to be prioritized based on your social media influence?

As a customer service manager, do you want to prioritize based on customers' social media influence?
Two Fundamental Questions

The Normative Question

Should companies prioritize customers based on their influence?

The Positive Question

Do companies actually prioritize customers based on their influence?
The Normative Question

Whichever prioritization criteria you choose, be sure they pass the “sniff test.” If these criteria became public, would they embarrass you? If so, they probably need some reworking. Remember, because of the public nature of Twitter, and the fact that anybody can set up a Twitter account and give it any username, analysts, journalists, bloggers and consumer advocates can test your Twitter customer service response, and “reverse engineer” reconstruct your prioritization policies.
The absence of considerations of fairness and loyalty from standard economic theory is one of the most striking contrasts between this body of theory and other social sciences – and also between economic theory and lay intuitions about human behavior.

Firms might sometimes forgo exploiting a legal but "unfair" profit opportunity either because their owners and managers prefer acting fairly or because customers may be willing to punish an offending firm by reducing business transactions with that firm.
The Normative Question

- Service fairness is a customer’s perception of the degree of *justice* in a firm’s behavior.
- Differential customer treatment involves the distribution of a social media team’s attention and time among customers.

Distributive justice is concerned with the distribution of the conditions and goods that affect individual wellbeing.
The Normative Question

The Three Principles of Distributive Justice (Deutsch 1985)

- **Equity**: in cooperative relations in which economic productivity is a primary goal, equity rather than equality or need will be the dominant principle of distributive justice.

- **Equality**: In cooperative relations in which the fostering or maintenance of enjoyable social relations is the common goal, equality will be the dominant principle of distributive justice.

- **Need**: In cooperative relations in which the fostering of personal development and personal welfare is the common goal, need will be the dominant principle of distributive justice.
The Positive Question

- 1,228,415 tweets sent by customers to seven major U.S. airlines from September 2014 to May 2015.
- Develop a sentiment classifier to extract all complaining tweets.
- 173,662 conversations from the complaining tweets.
- Use latent semantic analysis to reduce the dimensionality.
- Cluster the conversations into different topics and styles.
- Estimate an econometric model to reveal the influence-based prioritization.
The Positive Question

An airline is more likely to respond to a complaint sent to it by a customer with a higher number of followers.

An airline is more likely to respond faster to a complaint sent to it by a customer with a higher number of followers.
Agenda

1. Design
2. Sample Labs
3. Sample Case Project
4. Lessons
Lessons with Solutions

- Cloud setup isolates students and faculty from many OS environment issues.
- Just use one public facing node and be prepared with security and school IT regulations.
- Do NOT use RStudio Server Professional if you don’t have a very powerful machine.
- Node will fail, hardware or software. Have redundancy.
- Automatic cleaning of all R instances after midnight.
- Give code snippet instead of asking students to code from scratch.
Lessons with Solutions?

- To improve their marketability, students need to develop deep understanding of analytics, which requires technical and abstract treatment and comprehensive training, yet they are so heterogeneous in their backgrounds and tastes.
- *Keeping students happy no matter what* is a constant pressure.

**Strategies to help our kids eat less sugar...**

...and still enjoy the season!