Think Play Hack – World Views

Eric Godat, Ph.D.
Rob Kalescky, Ph.D.
Aren Cambre, D.Eng.

Taos 2019

Meet the Data Team

Eric Godat

- Ph.D. in Theoretical Particle Physics
- Data Science Research Applications Developer

Rob Kalescky

- Ph.D in Theoretical and Computational Chemistry
- High Performance Computing Applications Scientist

Aren Cambre

- Doctorate of Engineering in Engineering Management
- Director of Web Development and Adjunct Professor of Economics

Meet the Data Team

Eric Godat

- Ph.D. in Theoretical Particle Physics
- Data Science Research Applications Developer

Rob Kalescky

- Ph.D in Theoretical and Computational Chemistry
- High Performance Computing Applications Scientist

Aren Cambre

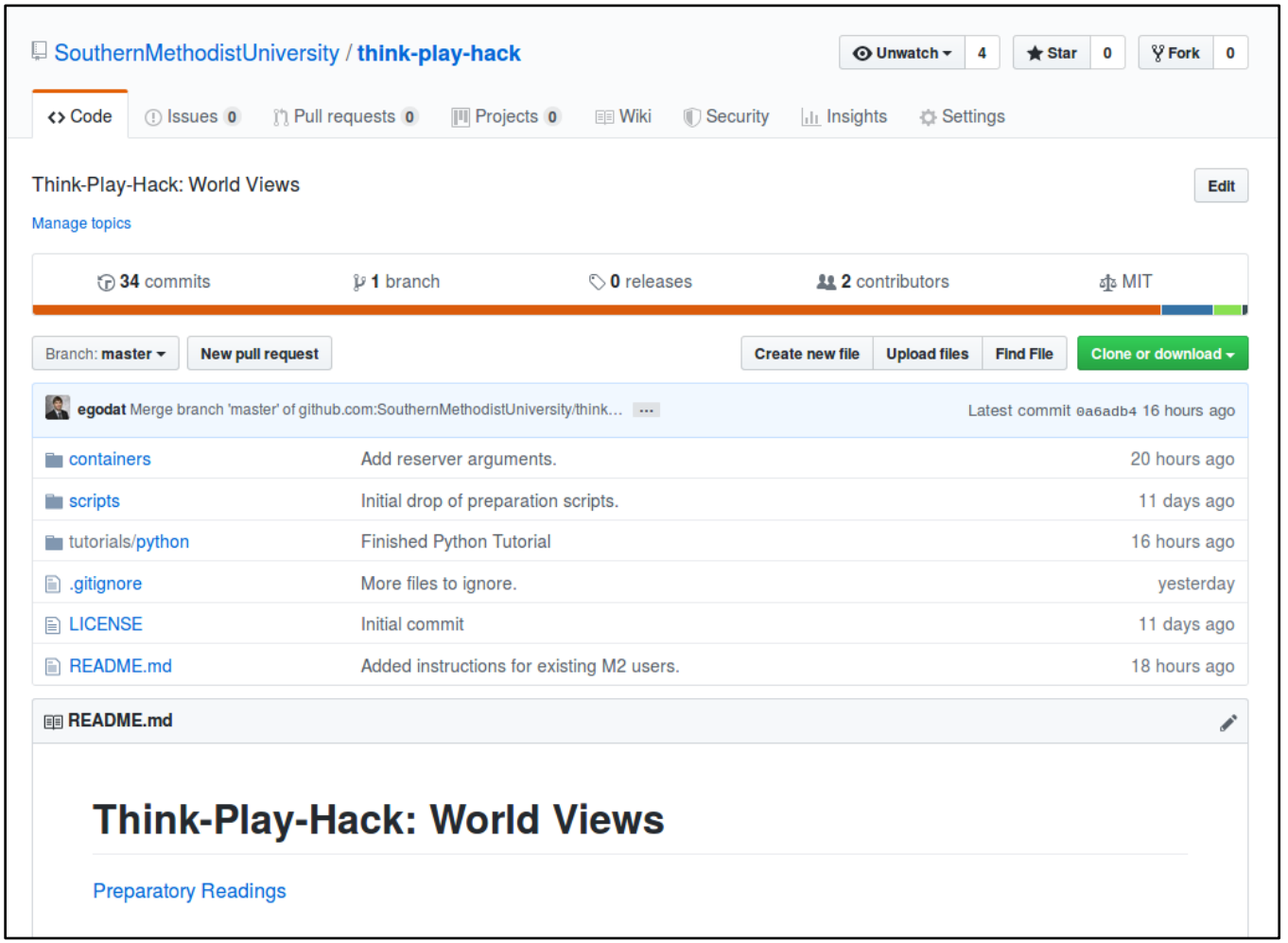
- Doctorate of Engineering in Engineering Management
- Director of Web Development and Adjunct Professor of Economics

Resources

Data Website

smu.edu/think-play-hack

- ReadMe
  - Readings
  - Prompts
  - Instructions
- Containers
- Slides
- Tutorials
- Scripts



Data Menu

Data Menu

Data	Size	Feature
Reddit	374 GB	Did you know the first message posted to Reddit was “test”?
Usenet	405 GB	30 years of the Internet from before it was the Internet
Pizzagate	146 MB	Remember that conspiracy from 2016? We do.
Folktales	6.1 MB	Nearly 1,000 English translations of folktales from around the globe and lots of wolves.
Project Gutenberg	N/A	Over 59,000 free books
Icelandic Sagas	2.8 MB	The tales of 16 guys named Bjorn and many more
Danish Legends	1.1 MB	“If the eyebrows have grown together over a man's nose, then it's a werewolf,” now you know.
Wikiplots	227 MB	112,000+ plot summaries from books, movies and video games
Biblical Versions	1 GB	Did you know there was a special bible just for slaves?

Data Menu

Da		
Reddit		eddit was
Usenet		he Internet
Pizzagate		do.
Folktales		es from
Project Gute		
Icelandic Sa		more
Danish Lege		a man's
Wikiplots		vies and
Biblical Vers		for slaves?

WARNING

Some of the sets provided on this menu contain adult and objectionable material.

Proceed with caution.

Data Details: Reddit

- A collection of all comments posted to Reddit.com
  - October 2007 until May 2015
- Torrented as 1 TB of raw JSON files
- Restructured into PostGres Database and tsv files
  - NSFW content removed
  - Active links encrypted
- Accessible on Box, from M2 or from external hard drives

Fields	Description
ID	Unique identifier for each comment
Subreddit	Title of the subreddit
Subreddit ID	Unique identifier of the subreddit
Link ID	ID of the link this comment is in
Created UTC	Time stamp when the comment was posted
Parent ID	ID of the parent comment or link
body	Text of the comment
ups	Number of up votes
downs	Number of down votes

Data Details: Usenet

- Usenet is a distributed messaging service that allowed users to subscribe to "newsgroups" and send messages to each other
  - Started in 1980, still active today
- Archive of 30 years of raw messages obtained by webscraping archive.org
  - Originally 2 TB compressed
  - 405 GB cleaned and structured
    - Only messages in English
    - Removed attachments
    - Encrypted active links
    - Limited content to Usenet's Big 8 + alt
      - comp, misc, news, rec, sci, soc, talk, humanities, alt
- Available on Box, from M2 or external hard drives

Fields	Description
Date	Date message was sent
X-Google-Language	Language of the message (should be English,ASCII-7-bit)
Subject	Subject of the message
Newsgroup	Newsgroup the message was sent to
Organization	Organization the sender is a member of
Body	Body text of the message

Data Details: Pizzagate

- Pizzagate is a conspiracy theory that demonstrates how fake news can go viral and continuously become more violent until causing real, physical threats to others. Starting as an unfounded rumor that a pizza shop owner was a pedophile, Pizzagate theories continued to grow, implicating the Democratic party as assisting this hole-in-the-wall pizzeria run a child sex ring. The scandal ended when a shooter walked into the pizzeria with an AR-15 to rescue the children.
- The set consists of:
  - Webscraped data
    - Twitter
    - Reddit
    - Voat
    - Major news outlets
      - Including 'Fake News' Outlets
    - Wikipedia
  - OCR'd Images
    - Memes
    - Screenshots
    - Photos
- This set was provided to us by Tim Tangherlini
  - Cleaned by Steph Buongiorno and Josh Hernandez
- Available on Box, from M2 or external hard drives

Data Details: Folktales

- From Aesop to The Brother's Grimm, folktales from around the globe (mostly Europe)
- Extracted from the Multilingual Folk Tale Database (www.mftd.org)
  - 908 English Translations
- Available on Box, from M2 or external hard drives

Field	Description
ID	ID number from mftd
ATU Code	Code correcponding to ATU Classification of Folk Tales
Author	Original Author
Country of Origin	Country from which the folktale originated
Original Title	Untranslated title
Source	Source location, often times the original book publication
Story	The text of the story
Story Type	The classification corresponding to the ATU code
Title	The translated title
Translated	Name of the translator
Year Translated	Year the text was translated
Year Written	Year the story was written

Data Details: Project Gutenberg

- A digital repository of 59,000 public domain books
  - https://www.gutenberg.org/
- Books are available via:
  - R – gutenbergr
  - Python - Gutenberg
  - Direct download
    - Plain Text
    - EBook

Data Details: Icelandic Sagas and Danish Legends

- Icelandic Sagas
  - Eyrbyggja Saga
  - Orkney Saga
  - Sagas of the Icelanders
- Plain Text format
- Provided to us by Tim Tangherlini
- Available on Box, from M2 or external hard drives

- Danish Legends from the Danish Folklore Database
  - 2000+ English Translations
- This set was provided to us by Tim Tangherlini
- Available on Box, from M2 or external hard drives

Data Details: WikiPlots

- A collection 112,000+ of summaries from books, movies, shows and video games from Wikipedia
  - Must contain a "Plot" or "Plot Summary" subsection
- Previously collected by David Robinson from Data Camp
  - Available on GitHub here
  - Some example analyses using the Tidyverse:
    - Examining Story Arcs: http://varianceexplained.org/r/tidytext-plots/
    - Gender and Verbs: http://varianceexplained.org/r/tidytext-gender-plots/
- Only 2 fields:
  - Title – name of the media
  - Plot – Extracted plot summary
- Available on Box, from M2 or external hard drives

Data Details: Bible Versions

- Nearly 200 different versions and translations of the Bible
  - Collected from ph4.org and archive.org by Steph Buongiorno
  - Includes a version censored for slaves in the British Colonies
- Each version comes in a SQLite database
  - Multiple tables
    - Verses table contains text
- Available on Box, from M2 or external hard drives

Data Access

How do I get access to this data?

- Internet Options:
  - From the website:
    - Links to Box
  - Connect to ManeFrame2
    - More on this to follow
- Offline Options:
  - Ask the Data Team
    - External hard drives loaded with data
      - All sets
    - Flash drives loaded with the smaller sets only

Compute

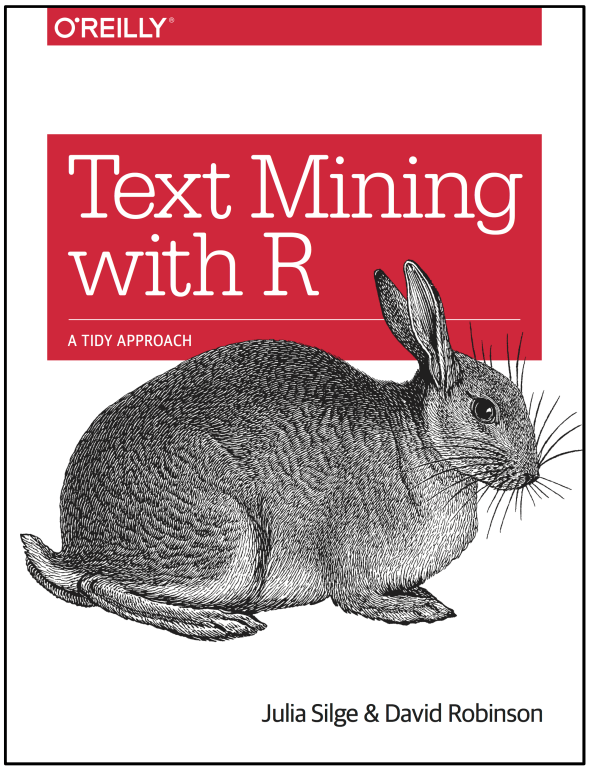
How do I do things with the data?

- How many brought laptops? How many have experience using R/Python?
  - If not, don't worry!
- Access to ManeFrame2
  - SMU's High Performance Cluster
    - A handful of accounts have been generated for the event
  - Data and Environment are ready to go
  - Interested? Ask us how to get access
- We have setup:
  - A Docker Image with R/Python
  - Data Loading Packages in Python and R to run locally
  - Some starter scripts to help you get going
- Check the Website for code and documentation

Tutorials

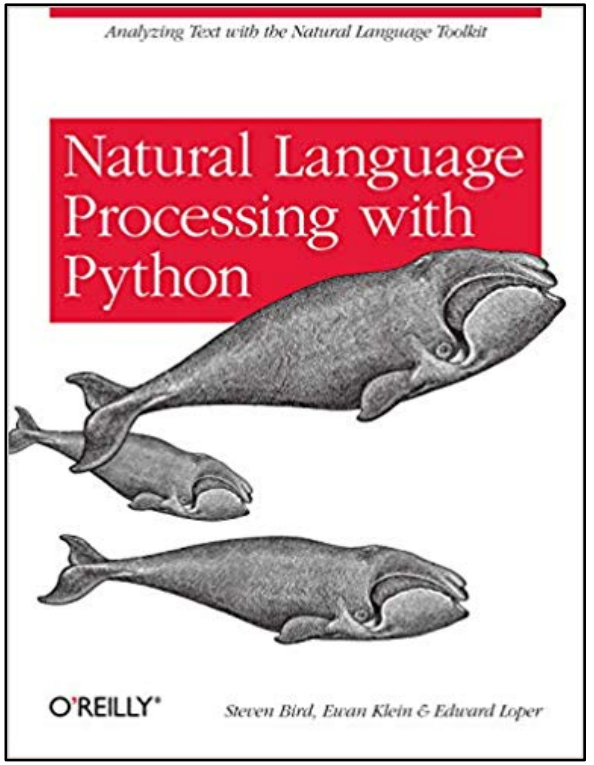
Text Mining in R

- One hour overview of how to use the Tidyverse to do Text Mining in R
  - This afternoon at 2:30
  - Parallel to Python Tutorial
- As much of the following that we get to:
  - R Basics for text
  - Tokenization
  - Word Frequencies
  - Sentiments
  - Wordclouds
- Working on text from Project Gutenberg
- Resources:
  - GitHub with Tutorial Code
  - https://www.tidytextmining.com/



Natural Language Processing with Python

- One hour overview of how to use Natural Language Processing Text Mining in Python
  - This afternoon at 2:30
  - Parallel to the Python Tutorial
- As much of the following that we get to:
  - Python Basics for text
  - Counting words and characters
  - Cleaning text data
  - Advanced Text Processing
  - Visualizations
- Working on text from Folktales and Wikiplots
- Resources:
  - GitHub with Tutorial Code
  - http://www.nltk.org/book/



Getting Help

How to reach us:

- We have a Slack channel
  - TPH World Views 2019
  - tphworldviews2019.slack.com
- Find a mistake?
  - Submit a Git Issue on GitHub
- Need help getting started?
  - Come to tutorial
- We will be here all week so don't be afraid to ask

External Resources

- Here are some links to tutorials for some of the languages and tools we use:
  - GitHub: https://guides.github.com/activities/hello-world/
  - Python: https://mode.com/resources/python-tutorial/
  - R: https://www.statmethods.net/r-tutorial/index.html
  - Linux: https://www.digitalocean.com/community/tutorials/
  - SQL: https://mode.com/sql-tutorial/
- By no means is this an exhaustive list