Sentiment Analysis of Twitter and How it Relates to Illini Football

STAT 385 FA2018 - Team Poisseidon

Joe Broton
Braden Gebavi
Brad Gibbons
Tim Soo

November 13, 2018

Abstract

The purpose of this project is to explore sentiment analysis of twitter and how it relates to the illini football team. We will be analyzing the tweets regarding the last game against Northwestern on 11/24. This is in order to see how fans and students view the illini football team and their opinions on Lovie Smith. Lovie Smith has a 6-year contract with the Illini and after 3 years we have failed to make a bowl game.

Contents

1 Instructions

This document will walk you through some of the necessary steps of formatting your report. Do not mistake the length of this document as an example of the length of a proper report. Length is not important. Communicating your idea in a concise but complete manner is important. The goal of the proposal is to capture details found in Figure 1.

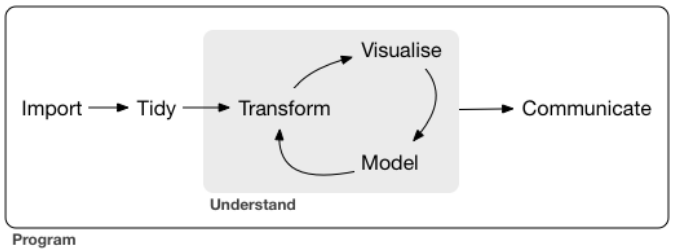

Figure 1: Data science workflow emphasized in R4DS. The photo has been reduced using chunk size options.

2 Introduction

The **introduction** section provides a preview of the project's focus. Within this section, provide an overview on the selected topic for the consumption of a manager. In essence, the manager must be able to understand what the project is and why they should support the endeavor. You are allowed to make the assumption that the manager is knowledgeable in base R concepts. Make sure to answer the following questions:

- What problem or topic are you addressing?
- Why is it interesting or important? In particular, what evidence supports this conclusion?
- Cite papers or reputable sources that back up this claim. (You may want to find - material using Google Scholar.)
- Where did the problem or topic come from?
- What is your idea for addressing the problem or topic?
- How does your idea match with the course's focus on statistical programming?

Consider adding subsections in this section. For example, consider adding a **data** subsection. The data subsection would describe your data. What is it? Where did it come from? How will it be useful in answering your problem? Provide references to information about the data, but explain enough that your reader does not need to utilize them.

Keep in mind, you may use or construct any dataset of your choice under two conditions:

1. there is a minimum of **500** observations and **10** variables;
2. data is *not* from either UC Irvine Machine Learning Repository or Kaggle, *not* used as a data set in the course, and *not* found in textbooks.

The dataset may be relevant to research outside of this course, another field, or some other interest of the groups. If you have any questions about whether your data is appropriate, do not hesitate to ask. If you plan to use data from either a research project or current job be sure to gain permission from the data controller.

3 Related Work

The **Related Work** section must provide an overview of pre-existing solutions. In essence, please credit those who enabled you to consider embarking on this project, or as Issac Newton in a letter to Robert Hooke on 15 February 1676 more aptly put it:

If I have seen further it is by standing on the shoulders of Giants.

Address the following questions:

- What other ideas have been attempted?
- Why is your team's idea original compared to prior work?

4 Methods

The **methods** section should discuss how you plan to solve your problem. The overall details of the project including any preliminary work. In particular, the implementation details behind the approach should be explained at length here. The more details you can provide, the better feedback your group can receive. As a result, the section serves as a roadmap of what features are going to be developed and any external dependencies that are required. **The majority of your code should be *suppressed* from the displaying in this section**. Please refer to code and figures placed in the appendix. The latter can be referenced using:

`Figure \\ref{fig:code-chunk-name-here}.`

For example, the figure of the data science workflow is accessible via Figure 1.

To satisfy this section, provide detailed responses for the following:

- What packages will you use in your implementation?
- What code will the group need to write for the project?
- Provide low-fidelity prototypes (e.g. *sketches* on paper) in the **Appendix** of:
  - Visualisations
    - What kinds of graphs will you use?
    - Label axes, provide a title, and mention any interactivity.
  - Interface
    - All projects need a Shiny Application.
    - Sketch how a user will work with the shiny application.
- What have you done or learned so far for the project?

We are primarily wanting to ensure that your project has met the criterion of the data science pipeline. In essence, we want to see evidence that your project has:

- Reading data into *R* or accessing data via an API.
- Data transformations (e.g. Tidying (`tidyr`), Summarizing (`dplyr`), et cetera.)
- Data visualization (e.g. `ggplot2`, `plotly`, `gganimate`)
- R functions either in external packages or included in a new *R* package
- Interactive Interface (e.g. `shiny`)
- Reproducibility

5 Feasibility

The **Feasibility** section is meant to act as a way to reflect upon the proposal. Generally speaking, there will be three weeks of heavy development time afforded to the group. Building a detailed ecosystem or heavily scripting in a different language will likely not lead your team to success. Hence, please provide a project management overview of *who* on your team will be doing *what* and *when* by answering:

- Is this project able to be completed before the end of the semester?
- What steps must occur to complete the project before the end of the semester?
- What is the work plan to accomplish the necessary tasks before the end of the semester?
  - Specify who is doing what and when.
  - Consider making a Gantt chart to highlight each stage of the project.

6 Conclusion

The **Conclusion** section provides a summary of the entire proposal. This acts as the final paragraph that can be used to justify the work being proposed. In general, this means you should make one last push to identify the problem, potential solution, and its novelty.

If a group's project is well written, uses thoughtful and creative approaches, and is sufficiently interesting you may be asked to have your work "published" as an example for future students. **All group members will have to agree to publication.** You may also be asked to make edits before publication, but you should be sure to **proofread** and **spellcheck** your work before your initial submission.

7 Appendix

The **Appendix** section contains figures, sample data, and other miscellaneous entries. Generally, this sketch seeks to contain all of your *planning* information.

- Provide the sketches of visualisations and the shiny application.
- Provide an overview on the desired functions.
  - What is a function's input? Output? How are functions related to each other.
  - For example, `read_data("hospital_data.csv")` must be called before `tidy_hospital()`, et cetera.
- Provide a sample of the data set you intend to use (~10 observations).

If you used previous code chunks within the document, this information can be dynamically retrieved and embedded.

```
# Sets default chunk options
knitr::opts_chunk$set(
  # Figures/Images will be centered
  fig.align = "center",
  # Code will not be displayed unless `echo = TRUE` is set for a chunk
  echo = FALSE,
  # Messages are suppressed
  message = FALSE,
  # Warnings are suppressed
  warning = FALSE
)
# All packages needed should be loaded in this chunk
pkg_list = c('knitr', 'kableExtra', 'magrittr')

# Determine what packages are NOT installed already.
to_install_pkgs = pkg_list[!(pkg_list %in% installed.packages()[,"Package"])]

# Install the missing packages
if(length(to_install_pkgs)) {
  install.packages(to_install_pkgs, repos = "https://cloud.r-project.org")
}

# Load all packages
sapply(pkg_list, require, character.only = TRUE)
knitr::include_graphics("images/data-science.png")
kable(
  head(mtcars, 20),
  format = "latex",
  caption = "This is an example of a table in the Appendix. Notice that it is way too big, and has way t
  booktabs = TRUE
) %>%
  kable_styling(latex_options = c("striped", "scale_down"))
kable(
  head(mtcars, 20),
  format = "latex",
  caption = "This is another example of a ridiculous table. Notice that it is automatically numbered.",
  booktabs = TRUE
) %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

Table 1: This is an example of a table in the Appendix. Notice that it is way too big, and has way too much information. We use the `kableExtra` package to shrink it down, but even then, no one would actually read this table.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1

Table 2: This is another example of a ridiculous table. Notice that it is automatically numbered.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1

7.1 Formatting Notes

7.1.1 `R` Code and `rmarkdown`

An important part of the report is communicating results in a well-formatted manner. This template document should help a lot with that task. Some thoughts on using `R` and `rmarkdown`:

- Chunks are set to not echo by default in this document.
- Consider naming your chunks. This will be necessary for referencing chunks that create tables or figures.
- One chunk per table or figure!
- Tables should be created using `knitr::kable()`.
- Consider using `kableExtra()` for better presentation of tables. (Examples in this document.)
- Caption all figures and tables. (Examples in this document.)
- Use the `img/` sub-directory for any external images.
- Use the `data/` sub-directory for any external data.

7.1.2 LaTeX

While you will not directly work with LaTeX, you may wish to have some details on working with TeX can be found in this guide by UIUC Mathematics Professor A.J. Hildebrand.

With `rmarkdown`, LaTeX can be used inline, like this, $a^2 + b^2 = c^2$, or using display mode,

$$\mathbb{E}_{X,Y}\left[(Y - f(X))^2\right] = \mathbb{E}_X \mathbb{E}_{Y|X}\left[(Y - f(X))^2 \mid X = x\right]$$

You **are** required to use BibTeX for references. With BibTeX, we could reference the `rmarkdown` paper (Allaire et al. 2015) or the tidy data paper. (Wickham and others 2014) Some details can be found in the `bookdown` book. Also, hint, Google Scholar makes obtaining BibTeX reference extremely easy. For more details, see the next section...

8 References

The **References** section acts as a bibliography for all papers referenced in the **Introduction**, **Related Works**, and **Method** sections. The references should be formated in Chicago author-date format, which is the default for RMarkdown.

- Provide a list (5+) of papers or items you have read to write this proposal.
- Please list all *R* packages or software referenced.

To acquire software citation information, *R* has a built-in command that creates a BibTex and in-line text citation. To generate the citation of an installed *R* package, type:

```
# In R
citation(package="pkg_name")
```

For example, to cite `dplyr`, one would generate the BibTex entry from:

```
citation(package="dplyr")
```

```
@Manual{dplyr:2018,
    title = {dplyr: A Grammar of Data Manipulation},
    author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller},
    year = {2018},
    note = {R package version 0.7.7},
    url = {https://CRAN.R-project.org/package=dplyr},
}
```

Note, we added a "name" to the autogenerated citation of `dplyr:2018`. Using this name, we can reference the work within the paper via (Wickham et al. 2018) or Wickham et al. (2018).

Allaire, JJ, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, and Rob Hyndman. 2015. "Rmarkdown: Dynamic Documents for R." *R Package Version 0.5*.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2018. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and others. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). Foundation for Open Access Statistics: 1–23.