Web Scraping with rvest

Anisha BharathSingh

R-Ladies NYC Lightning Talks
September 1st, 2020

Twitter: @thesparemoments
Blog: https://journeytodatascientist.blog

What is Web Scraping?

- Web scraping, or web data extraction, is the process of extracting data or information from websites.
  - Can be done manually, but is typically refers to the automated process using web scrapers.

A Little bit About rvest

- rvest is a R package that helps you scrape information from the web
- Works well with magrittr to make common web scraping tasks easy
- Inspired by Python's beautiful soup library for parsing HTML files

More info: https://rdocumentation.org/packages/rvest/versions/0.3.6

Scraping Text with rvest

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package
2. Assign URL to a variable

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text_n_809993'
```

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package

2. Assign URL to a variable

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text_n_809993'
```

```
speech_text <- read_html(URL)
```

3. Use read_html() to read in content from the URL

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package

2. Assign URL to a variable

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text_n_809993'
```

```
speech_text <- read_html(URL) %>%
  html_nodes(              )
```

3. Use read_html() to read in content from the URL
4. Use html_nodes() to identify & return desired element

SelectorGadget Chrome Extension

Locate HTML Nodes with “SelectorGadget”

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package

2. Assign URL to a variable

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text_n_809993'
```

```
speech_text <- read_html(URL) %>%
  html_nodes('blockquote')
```

3. Use read_html() to read in content from the URL
4. Use html_nodes() to identify & return desired element

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package

2. Assign URL to a variable

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text_n_809993'
```

```
speech_text <- read_html(URL) %>%
  html_nodes('blockquote') %>%
  html_text()
```

3. Use read_html() to read in content from the URL
4. Use html_nodes() to identify & return desired element
5. Use html_text() to convert node(s) into text

Scraping Text with rvest

```
install.packages('rvest')
library(rvest)
```

1. Install & Load the rvest package

2. Assign URL to a variable

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text_n_809993'
```

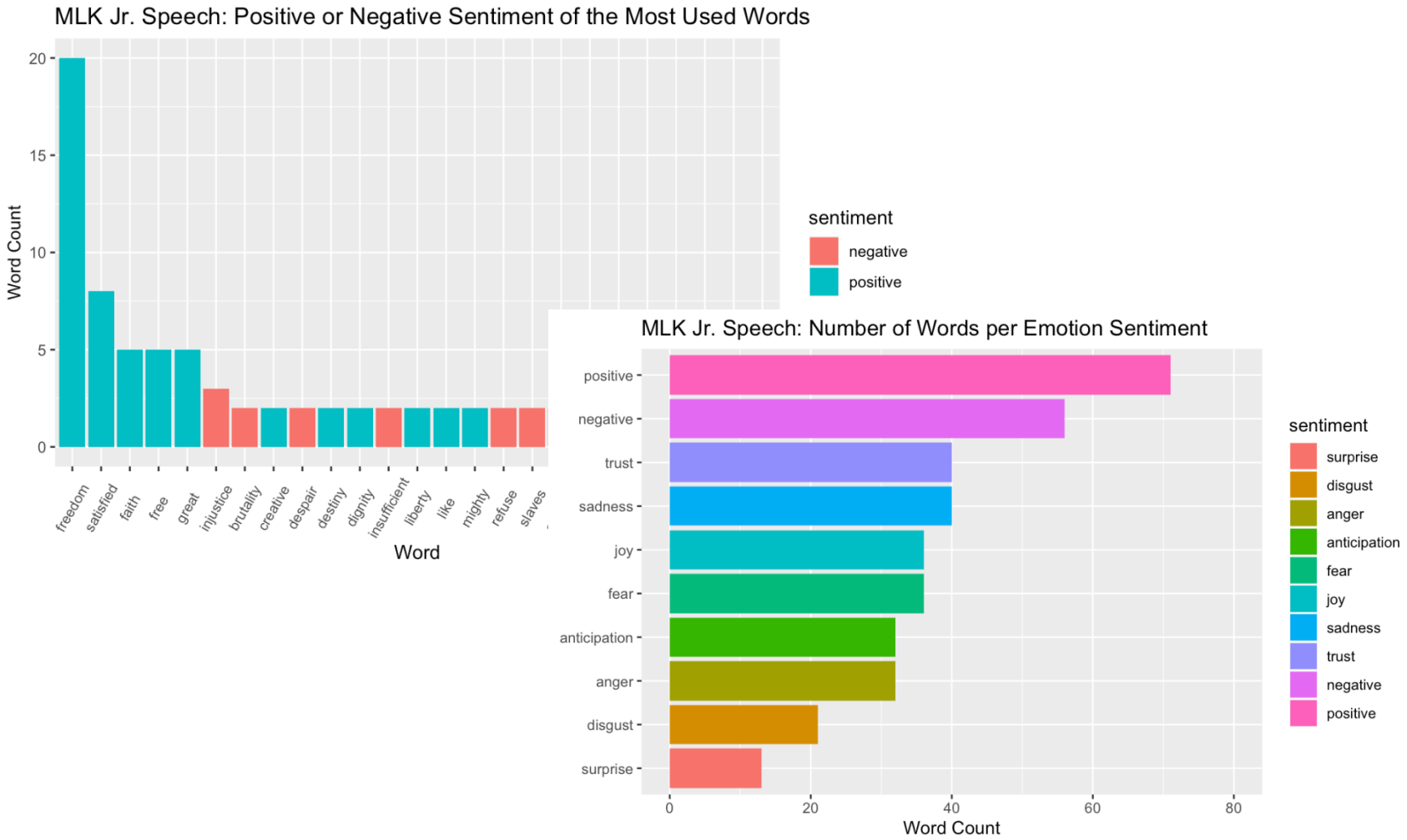
```
speech_text <- read_html(URL) %>%
  html_nodes('blockquote') %>%
  html_text()
```

3. Use read_html() to read in content from the URL
4. Use html_nodes() to identify & return desired element
5. Use html_text() to convert node(s) into text

```
speech_text
```

```
[1] "\nI am happy to join with you today in what will go down in history as the greatest demonstration
for freedom in the history of our nation.\nFive score years ago, a great American, in whose symbolic
shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon
light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It
came as a joyous daybreak to end the long night of their captivity.\nBut one hundred years later, the
Negro still is not free. One hundred years later, the life of the Negro is still sadly crippled by the
manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a
lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later,
the Negro is still languishing in the corners of American society and finds himself an exile in his own
land. So we have come here today to dramatize a shameful condition.\nIn a sense we have come to our
nation's capital to cash a check. When the architects of our republic wrote the magnificent words of the
Constitution and the Declaration of Independence, they were signing a promissory note to which every
American was to fall heir. This note was a promise that all men, yes, black men as well as white men,
would be guaranteed the unalienable rights of life, liberty, and the pursuit of happiness.\nIt is obvious
today that America has defaulted on this promissory note insofar as her citizens of color are concerned.
```

Use For Text Sentiment Analysis



Link to post: https://journeytodatascientist.blog/2020/01/20/martin-luther-king-jr-i-have-a-dream-speech/

Web Scraping Applications

Image Source: https://blog.apify.com/what-is-web-scraping-1b548f8d6ac1

Thank You!