

# Web Scraping with rvest

Anisha BharathSingh

R-Ladies NYC Lightning Talks  
September 1<sup>st</sup>, 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



# Scraping Text with rvest

- `read_html()` - reads HTML from a URL or local file
- `html_nodes()` - selects nodes from the HTML document
- `html_text()` - extracts the text content of the selected nodes
- `html_attr()` - extracts the attribute values of the selected nodes
- `html_table()` - extracts tables from the HTML document
- `read_csv()` - reads CSV files into a data frame
- `write_csv()` - writes data frames to CSV files
- `read_json()` - reads JSON files into a list
- `write_json()` - writes lists to JSON files
- `read_xml()` - reads XML files into a list
- `write_xml()` - writes lists to XML files
- `read_rss()` - reads RSS feeds into a list
- `write_rss()` - writes lists to RSS files
- `read_sitemap()` - reads sitemap files into a list
- `write_sitemap()` - writes lists to sitemap files
- `read_xlsx()` - reads Excel files into a data frame
- `write_xlsx()` - writes data frames to Excel files
- `read_yaml()` - reads YAML files into a list
- `write_yaml()` - writes lists to YAML files
- `read_tsv()` - reads TSV files into a data frame
- `write_tsv()` - writes data frames to TSV files
- `read_dbf()` - reads DBF files into a data frame
- `write_dbf()` - writes data frames to DBF files
- `read_dxf()` - reads DXF files into a list
- `write_dxf()` - writes lists to DXF files
- `read_ods()` - reads OpenDocument Spreadsheet files into a data frame
- `write_ods()` - writes data frames to OpenDocument Spreadsheet files
- `read_pptx()` - reads PowerPoint files into a list
- `write_pptx()` - writes lists to PowerPoint files
- `read_docx()` - reads Word documents into a list
- `write_docx()` - writes lists to Word documents
- `read_md()` - reads Markdown files into a list
- `write_md()` - writes lists to Markdown files
- `read_html_github()` - reads HTML from a GitHub repository
- `html_github_nodes()` - selects nodes from the HTML document of a GitHub repository
- `html_github_text()` - extracts the text content of the selected nodes from the HTML document of a GitHub repository
- `html_github_attr()` - extracts the attribute values of the selected nodes from the HTML document of a GitHub repository
- `html_github_table()` - extracts tables from the HTML document of a GitHub repository
- `html_github_csv()` - reads CSV files from a GitHub repository
- `html_github_json()` - reads JSON files from a GitHub repository
- `html_github_xml()` - reads XML files from a GitHub repository
- `html_github_rss()` - reads RSS feeds from a GitHub repository
- `html_github_sitemap()` - reads sitemap files from a GitHub repository
- `html_github_xlsx()` - reads Excel files from a GitHub repository
- `html_github_yaml()` - reads YAML files from a GitHub repository
- `html_github_tsv()` - reads TSV files from a GitHub repository
- `html_github_dbf()` - reads DBF files from a GitHub repository
- `html_github_dxf()` - reads DXF files from a GitHub repository
- `html_github_ods()` - reads OpenDocument Spreadsheet files from a GitHub repository
- `html_github_pptx()` - reads PowerPoint files from a GitHub repository
- `html_github_docx()` - reads Word documents from a GitHub repository
- `html_github_md()` - reads Markdown files from a GitHub repository

# Scraping Text with rvest

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

1. Install & Load the rvest package

# Scraping Text with rvest

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

1. Install & Load the rvest package

```
library(rvest)
```

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)
```

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text\_n\_809993'
```

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

1. Install & Load the rvest package

2. Assign URL to a variable

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

# Scraping Text with rvest

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

```
library(rvest)
```

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text\_n\_809993'
```

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

1. Install & Load the rvest package

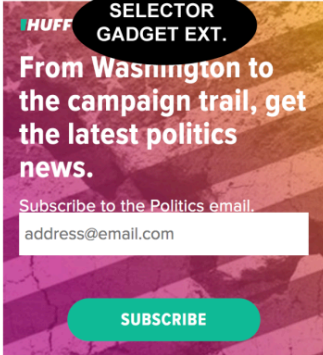
2. Assign URL to a variable

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

4. Use html\_nodes() to identify & return desired element



## 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)
```

```
URL <- 'https://www.huffpost.com/entry/i-have-a-dream-speech-text\_n\_809993'
```

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

1. Install & Load the rvest package

2. Assign URL to a variable

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

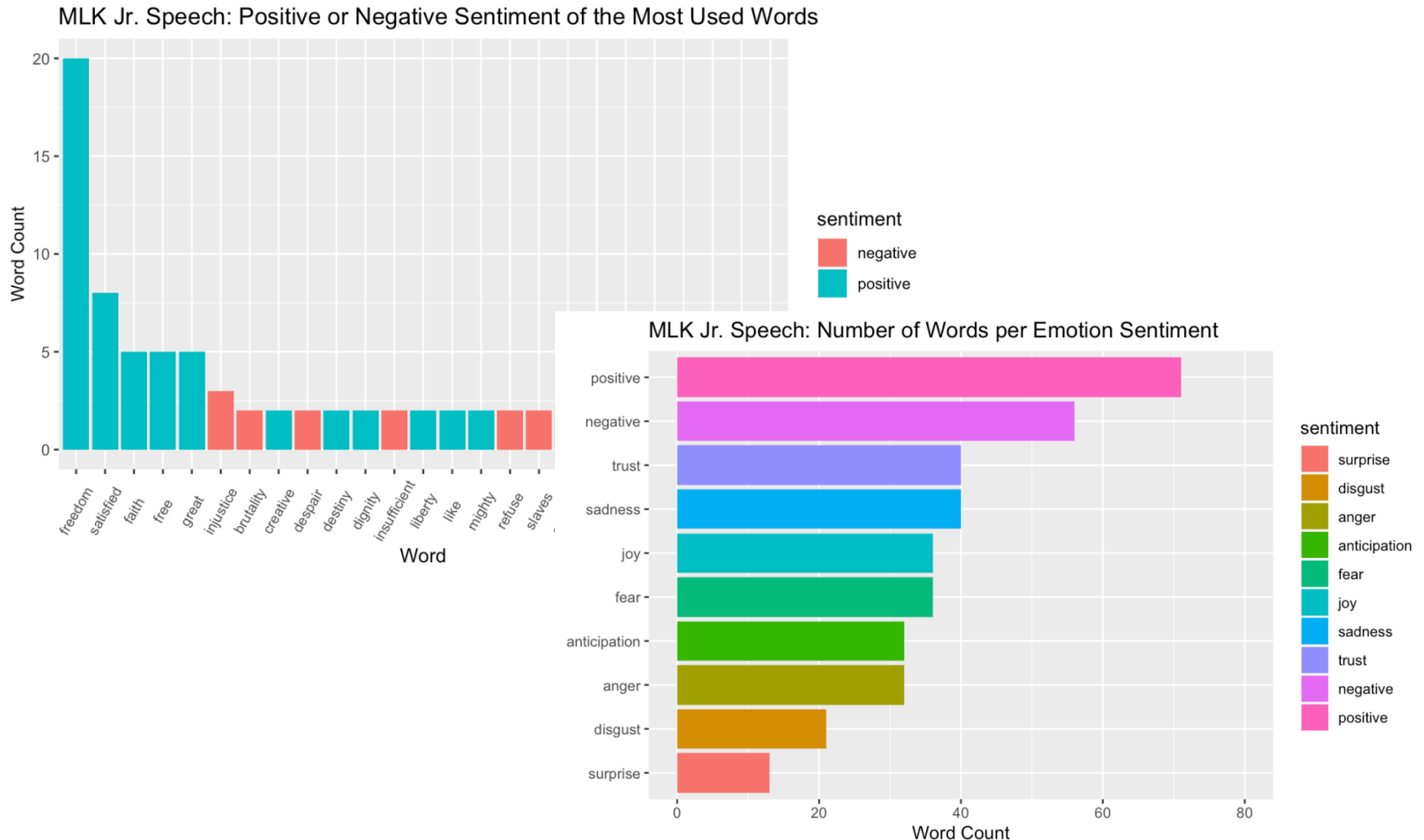
4. Use html\_nodes() to identify & return desired element

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 Scrapping Applications



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

Thank You!