Midterm Q3

Yufei Yin

Question 3: Project

Consider the Best New Albums section of the popular music site Pitchfork. Write an R function to retrieve the text of a review of an album, and also the rating that the reviewer gives to the album (on the scale 0.0 to 10.0). Provide a function named pitchfork which takes a single parameter: a URL specified by a string. When this function is passed the URL to a review of an album in the Best New Albums section of Pitchfork, the pitchfork function should return a list with two elements:

1) an element named text with value given by the body of the review as a character string (with all HTML removed),

2) an element named rating with value given by the decimal value of the review, as a numeric. You can find URLs for these reviews by navigating to https://pitchfork.com/reviews/best/albums/ and right clicking on an album and then selecting something like Copy Link.

Example usage follows. In this listing, the ellipsis indicates cropped R output.

Question 3: Part I

Provide a pdf containing all of the code that you wrote for this project. If your code spans multiple files, put all of the code into a single pdf and use comments to indicate the beginning of the files and the file names. You may use code from outside sources in your project provided that it is released under an open license and provided that you clearly indicate the code with comments.

```r
library(httr)
library(stringr)
library(rvest)

pitchfork = function(url){

  df = data.frame(text = NA,
                  rating = NA)

  # text
  text = read_html(url) %>%
    html_nodes(".dropcap p") %>%
    html_text()

  df$text = str_c(text, collapse = "\n\n")

  # rating
  s = GET(url)
  s = content(s, "text")
  # remove space
  s1 = str_replace_all(s, "[[:space:]]", "")
```

```r
    # match
    match = str_match_all(s1, "<spanclass=\"score\">(.*?)</span>")

    df$rating = as.numeric(match[[1]][1,2])

    return(df)
}
```

Question 3: Part II

Write a short report (in prose) arguing that the code you've provided works as required and provide a pdf of the report. Include output of R sessions in labeled figures. Provide some example calls of your implementation of the pitchfork function.

We are asked to apply web scraping by writing a function called **pitchfork()**. **pitchfork()** will take url as parameter, and return a data frame that contains "text" stands for album review text, and "rating" stands for the rate of the album.

We first use `data.frame()` function to build an empty data frame. The data frame has two columns: text, rating. Then, we use `read_html()`, pipe with `html_nodes()` and `html_text()` to extract the plain text. The output is a vector. Each element of this vector is one paragraph of the review article, so we use `str_c()` to concatenate them into one single string, and store it in the data frame. Moreover, we will extract album rating. we use `GET()` and `content()` to get the HTML source page, then use regular expression to get the rating, and store it in our data frame. Finally, our function should return the data frame.

After writing the **pichfork()** function, we need to test whether it works as required.

Test 1

rating

```
url = 'https://pitchfork.com/reviews/albums/jeff-parker-suite-for-max-brown/'
review = pitchfork(url)
review$rating
```

```
## [1] 8.4
```


Figure 1: Jeff Parker Rating

The rating from our function is the same as the website.

text

```
cat(review$text)
```

The output is too long, so we only show a few paragraphs from the beginning.

As we can see, the first paragraph from the website is the same as our function output.

```r
```{r, eval=FALSE}
cat(review$text)
```
```

Jeff Parker always writes parts that sound unassuming at first listen and unavoidable by the fifth. It's the X-factor that the guitarist and master collaborator has brought to every project on his long and still-growing list of projects, jazz or rock or otherwise: Tortoise, Isotope 217, the recently reunited Chicago Underground Quartet, his solo work as a bandleader, his work as a soloist, and his supporting contributions for countless others. Despite his ability to do backflips with a guitar, his best-known lick from Tortoise's 1998 song "TNT" is more like a heel-click—an easy, humble gesture, perfectly timed and placed.

It's a preternatural thing, of course, but it's also a skill that he's cultivated by changing up his scenery and embracing the unfamiliar. A few years ago, while splitting his time between his longtime home of Chicago and his new home of Los Angeles, Parker connected with the players at International Anthem, a new collective of jazz-raised, boundary-challenging musicians based in Chicago. Some of them were exploring the intersection of live improvisation and modern digital recording—loops, samples, beats—which was an area that had fascinated Parker ever since Madlib's jazz-ensemble project Yesterdays New Quintet blew his mind over a decade ago. Parker had been messing around with these elements for years during occasional DJ sets, thinking about how they could apply to his own playing, and had accumulated several hours of experiments. With this new label, he saw the right opportunity to formally release those results. He turned them into his 2016 album The New Breed, some of the most soulful and beat-driven music he'd ever put out under his own name and his first release for International Anthem.

What Parker tapped into on The New Breed, he blows wide open on Suite for Max Brown, a mesmerizing follow-up and informal companion piece. While his electric guitar remains a highlight, Parker builds out a fast-slashing range of ideas using dozens of other sounds and instruments, most of which he plays himself. They're disparate in color and texture, pronounced and often short, each one elbowing or sliding its way in front of the one before it, impatient to steal the show. As a player and composer, Parker shines throughout. As an arranger, he catches fire.

Suite for Max Brown is a place where a 26-second, Dilla-indebted loop of an Otis Redding sample and 10 minutes of a jazz quintet weaving around what sounds like someone stacking plastic cups can share a tracklist; each is equally meaningful. He's less of a specialist and more open to breadth than ever before. "Del Rio" opens with vintage-Korg cloud cover, which parts to reveal an easy-stepping melody that Parker plays on the African mbira—which in turn tees up a warm gust of wind in the form of a big, swooping guitar lick. That last move is one that Parker makes a few times on Suite: entering on guitar at just the right, unexpected moment of a tune. It's as smooth and elegant as calligraphy; it sounds like he's literally signing these new ideas with his signature instrument.

Figure 2: Jeff Parker text

Jeff Parker always writes parts that sound unassuming at first listen and unavoidable by the fifth. It's the X-factor that the guitarist and master collaborator has brought to every project on his long and still-growing list of projects, jazz or rock or otherwise: Tortoise, Isotope 217, the recently reunited Chicago Underground Quartet, his solo work as a bandleader, his work as a soloist, and his supporting contributions for countless others. Despite his ability to do backflips with a guitar, his best-known lick from Tortoise's 1998 song "TNT" is more like a heel-click—an easy, humble gesture, perfectly timed and placed.

Figure 3: Jeff Parker first paragraph

We will try a different url to further test our function.

Test 2

rating

```
url = 'https://pitchfork.com/reviews/albums/20390-to-pimp-a-butterfly/'
review = pitchfork(url)
review$rating
```
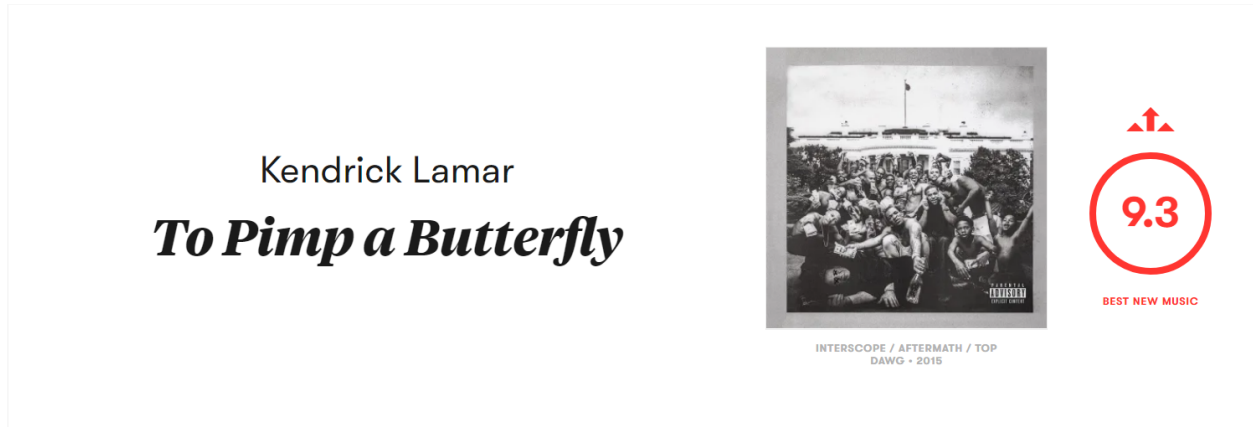
```
## [1] 9.3
```


Figure 4: Kendrick Lamar Rating

text

```
cat(review$text)
```

Again, we obtain the same data from the website.

Figure 5: Kendrick Lamar text

Figure 6: Kendrick Lamar first paragraph

Final test

Finally, we create a vector contains 4 url of Kanye west's albums, and use for loop to store all the data frame from **pitchfork()** function to a list called `kanye_west`, and test if the data are the same as the website.

Figure 7: Kanye's 4 albums

```
url = c('https://pitchfork.com/reviews/albums/21542-the-life-of-pablo/',
        'https://pitchfork.com/reviews/albums/18172-kanye-west-yeezus/',
        'https://pitchfork.com/reviews/albums/15725-watch-the-throne/',
        'https://pitchfork.com/reviews/albums/14880-my-beautiful-dark-twisted-fantasy/')
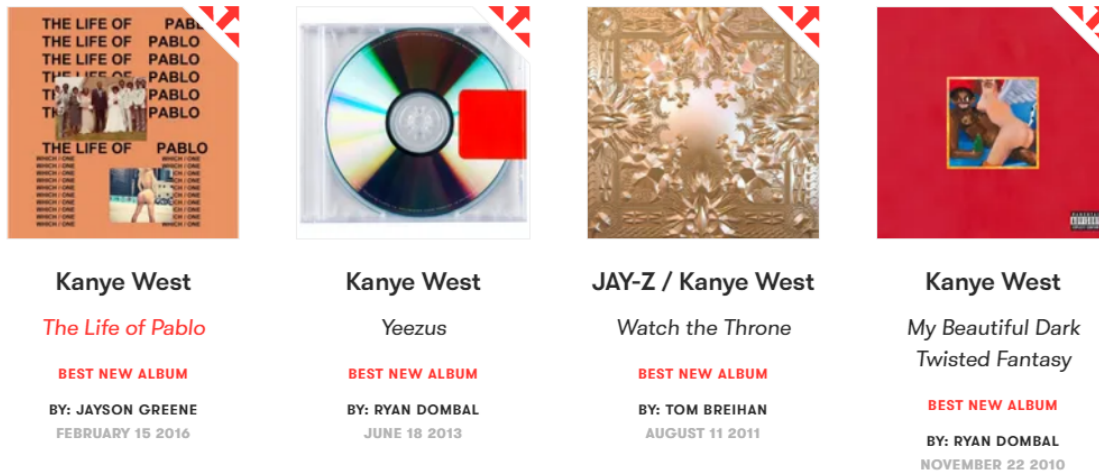
kanye_west = vector(mode = "list", length = length(url))
name = vector(mode = "character", length = length(url))

for (i in seq_along(url)){
  kanye_west[[i]] = pitchfork(url[i])

  # extract album name
  match = str_match_all(url, "-(.*)/")
  name[i] = match[[i]][1,2]
}
names(kanye_west) = name
```

The 4th album "My Beautiful Dark Twisted Fantasy" rating
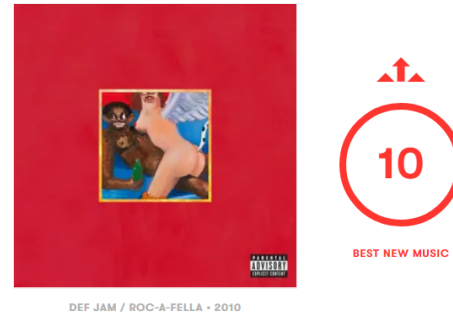
```r
kanye_west[[4]]$rating
```

```
## [1] 10
```


Figure 8: Kanye West MBDTF rating

The 4th album "My Beautiful Dark Twisted Fantasy" text

```r
cat(kanye_west[[4]]$text)
```


Figure 9: Kanye west MBDTF text

Kanye West's 35-minute super-video, *Runaway*, peaks with a parade. Fireworks flash while red hoods march through a field. At the center of the spectacle is a huge, pale, cartoonish rendering of Michael Jackson's head. *My Beautiful Dark Twisted Fantasy*'s gargantuan "All of the Lights" soundtracks the procession, with Kanye pleading, "Something wrong, I hold my head/ MJ gone, our nigga dead." The tribute marks another chapter in West's ongoing obsession with the King of Pop.

Figure 10: Kanye west MBDTF first paragraph

kanye_west	list [4]	List of length 4
the-life-of-pablo	list [1 x 2] (S3: data.frame)	A data.frame with 1 row and 2 columns
text	character [1]	'Pablo Picasso and Kanye West share many qualities—impatience with formal school ...
rating	double [1]	9
kanye-west-yeezus	list [1 x 2] (S3: data.frame)	A data.frame with 1 row and 2 columns
text	character [1]	'Look at the song title "I Am a God" and it seems like easy blasphemy-- Kanye We ...
rating	double [1]	9.5
watch-the-throne	list [1 x 2] (S3: data.frame)	A data.frame with 1 row and 2 columns
text	character [1]	'Watch the Throne features the following things: absurdly expensive samples, a p ...
rating	double [1]	8.5
my-beautiful-dark-twis...	list [1 x 2] (S3: data.frame)	A data.frame with 1 row and 2 columns
text	character [1]	'Kanye West\'s 35-minute super-video, Runaway, peaks with a parade. Fireworks fl ...
rating	double [1]	10

Figure 11: kanye west list

we only showed the output of the "My Beautiful Dark Twisted Fantasy" album here, but the other albums' outputs are also the same as the website.

According to all the test results above, I am confident that **pitchfork()** function satisfies the requirements which are extract plain text and rating from the website.