

Scraping tripadvisor.ca for Airline Reviews-Case Air Canada

Mustafa Koroglu

NYC Data Science Academy
New York, NY

August 2, 2017

Outline

Motivation

Sample Review and Spider Code

Data

Visualization-not completed yet

Future Work

Why Do Customer Reviews Matter for Airline Companies?

- ▶ Can we expose any pattern from customer reviews based on cabin class, route category and destination?
- ▶ Is there any seasonal pattern for the reviews and ratings? If so, can we use it for marketing strategy for Air Canada?
- ▶ Detail ranking categories will help us to evaluate the last question.

Sample Review



Milo T

Level **3** Contributor



18 reviews



6 helpful votes

"Substantial business class"

★★★★☆ Reviewed 1 week ago

The business class was very good. The food was good as well. The flight attendants were very attentive and were very interactive the wifi and connectivity was terrible but everything else was spot on.



Travelled June 2017

- ★★★★★ Seat comfort
- ★★★★☆ Customer service (e.g. attitude, care, helpfulness)
- ★★★★★ Cleanliness
- ★★★★☆ Food and Beverage

- ★★★★★ Legroom
- ★★★★☆ In-flight entertainment (WiFi, TV, films)
- ★★★★★ Value for money
- ★★★★☆ Check-in and Boarding (e.g. efficiency, service at gate)

San Francisco - Toronto

Business Class

Canada

Helpful?

Thank Milo T

Report

Ask Milo T about Air Canada

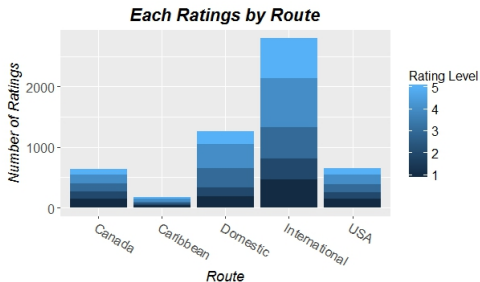
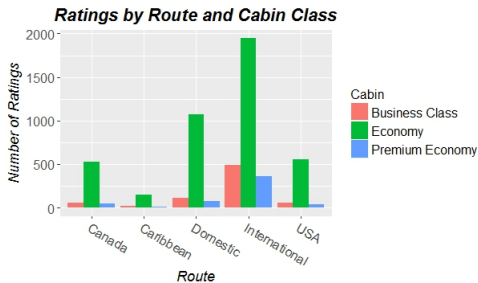
Spider Code

```
6 name='trip_advisor'
7 allowed_urls=['https://www.tripadvisor.ca']
8 start_urls=['https://www.tripadvisor.ca/Airline_Review-d8728998-Reviews-Cheap-Flights-Air-Canada']
9
10 def verify(self, page_list):
11     if isinstance(page_list, list):
12         if len(page_list) == 1:
13             return page_list[0]
14         # In Python 2, everything you scraped is in unicode, which might cause some trouble when you save it to local
15         # file.
16         # The rule of thumb is to encode it with ascii using the following command.
17         # return content.encode('ascii','ignore')
18     else:
19         return page_list[1]
20
21 def parse(self, response):
22     reviews=response.xpath('//div[@class="wrap"]')
23
24     for review in reviews:
25         title=review.xpath('//div/a/span/text()').extract_first()
26         rating=review.xpath('//div[@class="rating reviewItemInline"]/span/@class').extract_first()
27         content=review.xpath('//p[@class="partial_entry"]/text()').extract_first()
28         date_review=review.xpath('//div[@class="rating reviewItemInline"]/span/text()').extract()
29         if len(date_review)==1:
30             if re.search('ago$', date_review[0]) != None:
31                 date=review.xpath('//div[@class="rating reviewItemInline"]/span/@title').extract_first()
32             else:
33                 date=date_review[0]
34         else:
35             date=review.xpath('//div[@class="rating reviewItemInline"]/span/@title').extract_first()
36
37     categories=review.xpath('//div[@class="allLabels"]')
```

Data

- ▶ 5618 observations with 7 variables.
- ▶ Detail ranking categories will be scraped.
- ▶ Reviews for all airlines will be scraped.

Visualization



Work In Progress

- ▶ Scraping the detail ranking category.
- ▶ Working on word cloud for negative and positive reviews.
- ▶ Comparison of the main analysis across different major airlines.