Scraping tripadvisor.ca for Airline Reviews-Case Air Canada

Mustafa Koroglu

NYC Data Science Academy
New York, NY

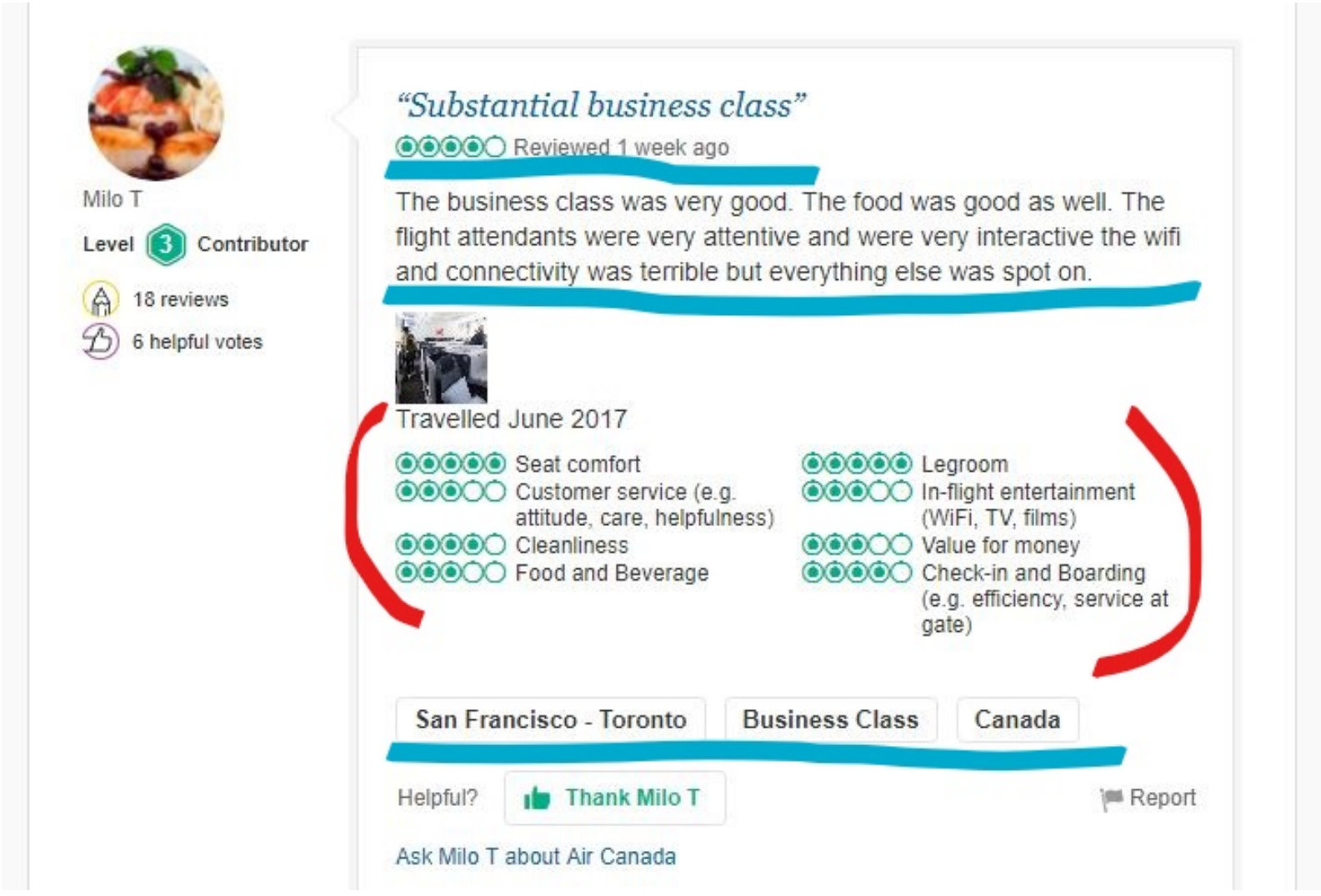
August 2, 2017

Outline

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



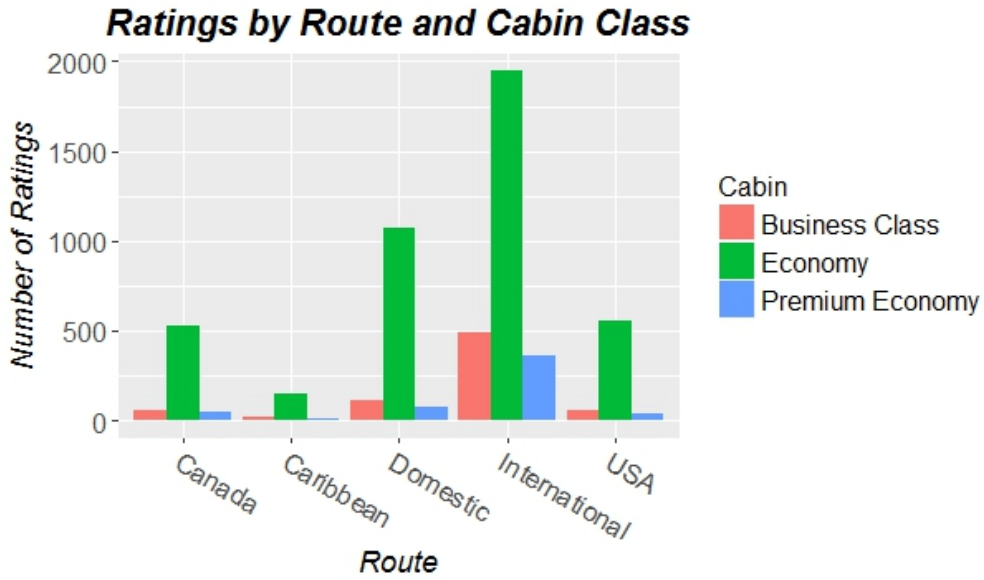
Spider Code

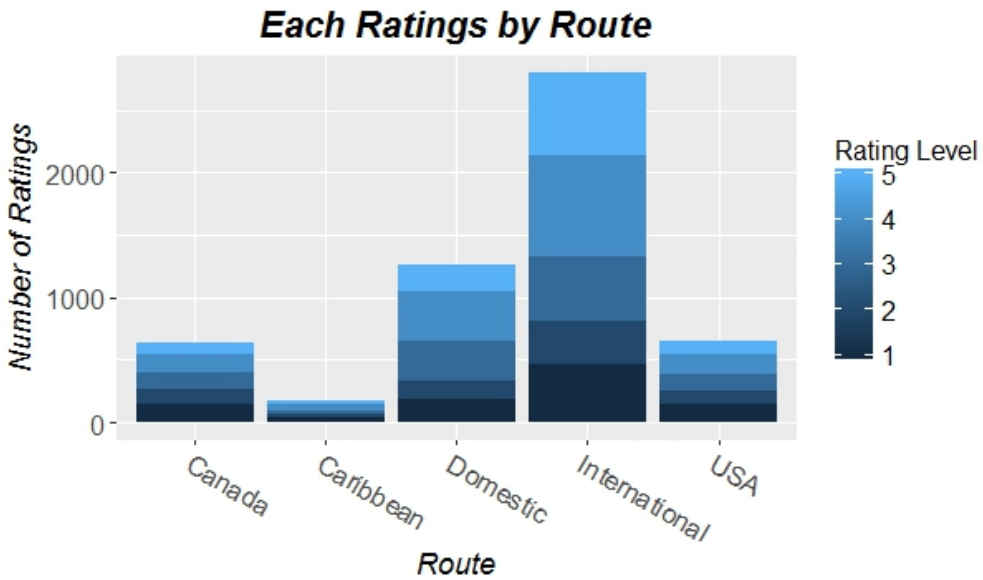
```
6     name='trip_advisor'
7     allowed_urls=['https://www.tripadvisor.ca']
8     start_urls=['https://www.tripadvisor.ca/Airline_Review-d8728998-Reviews-Cheap-Flights-Air-Canada']
9
10    def verify(self, page_list):
11        if isinstance(page_list, list):
12            if len(page_list) == 1:
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
  file.
15                # The rule of thumb is to encode it with ascii using the following command.
16                # return content.encode('ascii','ignore')
17            else:
18                return page_list[1]
19
20    def parse(self, response):
21
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.xpath('.//div[@class="rating reviewItemInline"]/span/text()').extract()
29            if len(date_)==1:
30                if re.search('(ago)$', date_[0]) != None:
31                    date=review.xpath('.//div[@class="rating reviewItemInline"]/span/@title').extract_first()
32                else:
33                    date=date_[0]
34            else:
35                date=review.xpath('.//div[@class="rating reviewItemInline"]/span/@title').extract_first()
36
37
38            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.