E6895 Advanced Big Data Analytics:

Analysis of Tourism Destination Prediction

Yihan Dai / yd2349
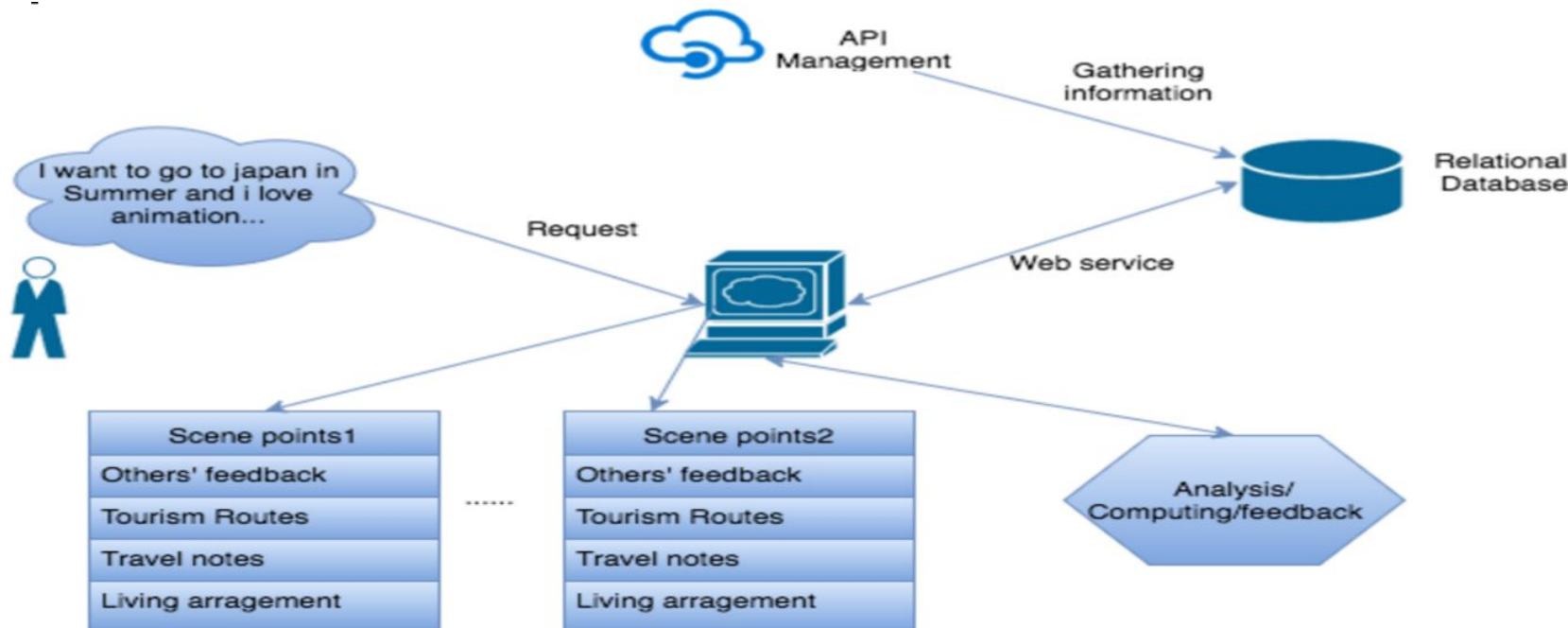Zikai Lin / zl2442

May 12, 2016

E6895 Advanced Big Data Analytics – Final Project

A **searching platform**, where **tourists in road trip** can receive the **recommendations** of interest of places based on their **input information**.
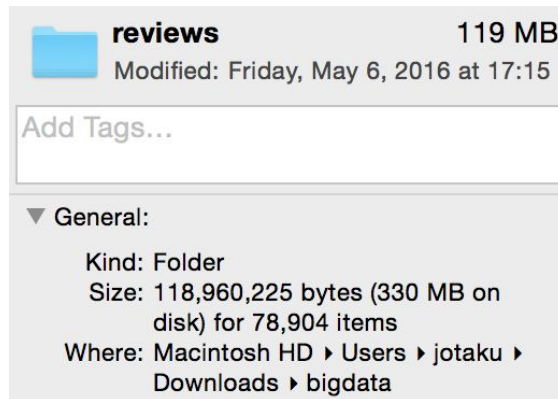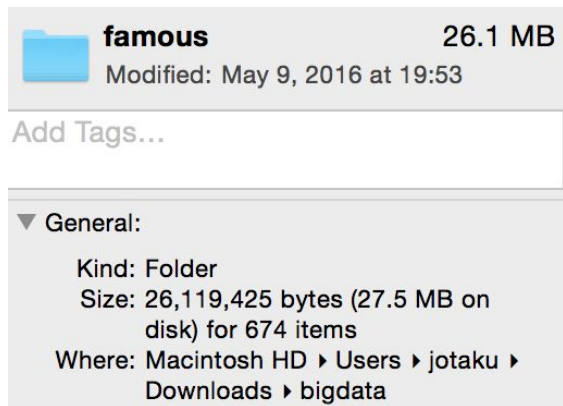
Data Set Source

What to consider in your trip:

Weather: from Yahoo API.

Reviews and Tags: from roadtrippers.com, where provides ratings, reviews and tags(using Beautifulsoup).

In total: 78904 tourist attractions in 674 places.

Data Set Sample

first line: attraction name

second line: rating
third line: rating number

following: tags

rest: reviews

all in html format

```html
<h1 itemprop="name" title="Universal Studios Orlando">Universal Studios Orlando</h1>
<span itemprop="ratingValue">4.4</span>
<span itemprop="ratingCount">28 </span>
<div class="category-list">
<ul>
<li><span>Holiday Lights</span></li>
<li><span>Movie Filming Location</span></li>
<li><span>Children's Attraction</span></li>
<li><span>Theme Park</span></li>
<li><span>Tourist Attraction</span></li>
<li><span>Amusement Park</span></li>
</ul>
</div>
<div class="review-text" itemprop="reviewBody">
<p>Go behind the scenes, beyond the screen, and jump right into the action of your
favorite movies at Universal Studios®, the world's premier movie and TV based theme
park. </p>
<p>It's everybody's fantasy… to leap through the screen and find yourself in the middle of
your favorite motion pictures. Here at Universal Studios®, YOU become the star of some of
the greatest movies and TV shows ever made.
</p>
</div>
<div class="review-text" itemprop="reviewBody">
<p>An earlier review informed me that the Jaws ride is no longer operating. That's totally
bogus. Out of all the rides I remember from a former visit, the Jaws and the Back to the
Future ride were incredible.  </p>
</div>
<p class="review-text">
        They are nostalgic experiences for sure, but they were really starting to look
dated. I'll miss them, but I'm always looking forward to new attractions.
        </p>
<div class="review-text" itemprop="reviewBody">
<p>There's wayyyy too much stuff to do at Universal Studios to only spend a day there. So
I recommend spending at least a couple days at the park. Lots of people skip the park in
favor of Disney World, but honestly Universal is fantastic. Just as clean as Disney, but
with more thrill rides. Warning, it's pretty busy so expect long waits in lines. For
instance, the Harry Potter ride had a wait of 480 minutes this summer. Definitely go for
the Express Pass. Basically you go to each of the rides you wanna go on and grab a pass
that gives you a time when you should come back to ride it.  </p>
```
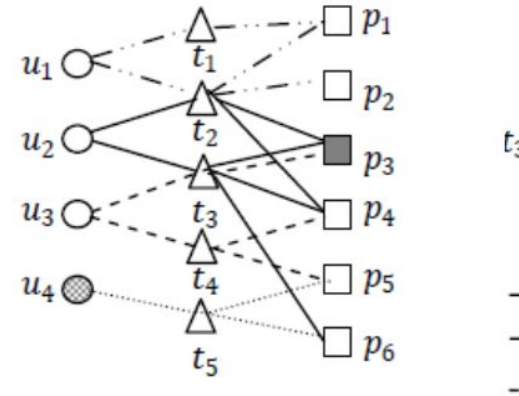
Recommendation System with Tags

Algorithm: Item-Based Collaborative filtering (IBCF)

⊡ User-Tag-Item relationship: records each user's personal tagging relationships

- ■ (User×Tag)-Item mapping
- ■ Item-(User×Tag) mapping

Tag Statistics

divide the tags into 10 scopes:

1. Park
2. nightlife
3. children's attraction
4. nature reserve
5. sports
6. history site
7. art & science
8. aquatic events
9. shows
10. adventure


```
allTags.txt

Other Sport Stadium : 1295
State Park : 1968
Architectural Site : 5750
Bar : 2998
Government Building : 1624
Museum : 13256
Theater : 5254
Science Place : 2664
Native Culture : 1214
Art Gallery : 5297
Music Venue : 6314
Drive-in Movie Theater : 1032
Natural Feature : 5711
Offbeat Attraction : 5512
Other Historical : 5123
Scenic Point : 6339
Concert Hall : 2863
Zoo : 4836
Other Amusement : 2059
Nature Reserve : 9072
Cinema : 1019
Hiking Area : 11447
Forest : 4091
Fishing Spot : 5164
Water Sports : 3279
Beach : 3087
Comedy Club : 2831
Community Park : 3938
Other Adventure Sport : 1862
Show : 2934
Roadside Attraction : 1457
Cycling Area : 1595
Science Museum : 2642
Children's Attraction : 13959
Stadium : 2216
Folk Art : 1306
Amusement Park : 2098
```

Tag Analysis

The relevance of a tag to an item
    The number of users used the tag for the item / the total number of users tagged the
item

$$\mathcal{P}(t_x \mid p_k) = \frac{\mathcal{P}(p_k, \ t_x)}{\mathcal{P}(p_k)} = \frac{|U_{p_k, t_x}|}{|U_{p_k}|}$$

The relevance of two tags in terms of each item
    Each item is equally important
    The average relevance of a tag to the collected items

$$r_{u_i, t_x}(t_y) = \sum_{p_k \in P_{u_i, t_x}} \frac{\mathcal{P}(t_y \mid p_k)}{|P_{u_i, t_x}|}$$

Data Set

Table 1

place	number	rate	park	nightlife	children	nature	sports	historic	art_science	aquatic	shows	adventure
abilene-tx	191	4.04	21.28	2	14.98	70.97	16.37	40.39	23.51	0	12.75	2
akron-oh	1189	4.37	36.83	15.81	45.44	31.43	62.3	124.8	92.59	0	24.16	26.67
alafaya-fl	3356	4.31	61.9	35.22	208.64	97.4	75.29	140.55	89.08	55.11	78.59	63.52
alameda-ca	3177	4.42	50.51	53.49	119.77	91.95	41.5	270.48	253.61	24.29	111.44	67.02
albany-ga	80	4.03	9.58	0	12.34	39.01	9.58	27.49	7.75	2.58	0	4.17
albany-ny	999	4.27	33.11	1	16.05	57.61	49.57	95.42	59.94	13.91	75.64	13.01
albany-or	375	4.17	36.45	0	31.05	37.7	16.66	32.51	17.17	18.14	9	13.5
alexandria-va	5008	4.59	68.6	91.94	46.76	106.76	61.88	497.25	234.59	22.36	137.84	31.53
allen-tx	3346	4.35	90.03	80.3	102.3	99.13	34.09	114.46	124.35	9.72	101.59	54.32
allentown-pa	963	4.09	80.58	2	76.67	58.24	23.56	132.94	57.18	22.4	30.06	46.09
alpharetta-ga	2959	4.38	85.85	52.62	94.76	189.02	82.89	197.52	167.69	19.53	243.5	39.68
amarillo-tx	500	4.11	20.94	0	24.69	49.53	23.36	38.31	48.57	12.17	17.09	18.47
ames-ia	638	4.25	23.55	9.4	43.59	24.43	19.04	72.54	59.75	0	18.09	9.3
anaheim-ca	4441	4.45	89.38	17.01	121.61	96.69	82.6	205.74	145.15	48.71	129.64	55.73
anchorage-ak	107	3.87	1	0	18	16.34	7.17	24.38	22.72	5.58	8.78	4.58
anderson-in	1779	4.32	72.95	43.37	41.15	61.24	126.77	161.95	119.62	30.11	89.72	12.49
ann-arbor-mi	2032	4.34	71.33	31.99	76.48	62.79	105.45	149.34	149.46	12.99	90.64	12.61

Testing

SEARCHING PLATFORM

boston-ma		Place3:
orlando-fl		Place4:

Tourism Start   05/17/2016

Tourism End   05/17/2016

Comfort

○ < $1000   ○ $1000 ~ $3000   ⊘ > $3000

Tag1   1 ▾   Tag3   5 ▾

Tag2   2 ▾

1. Park Attraction   6. Historic sites
2. Nightlife Attraction   7. Art_science
3. Children Center   8. Aquatic Events
4. Natural Points   9. Shows&Opera
5. Sports   10. Adventure experience

Submit

Thank you!