Data Scrawler for yelp

March 8, 2015

1 Data Structure

1.1 Business

Fields:

1. business_id
2. full_address
3. hours, dictionary, key=M/T,..., value=close/hr, open/hr
4. open, true/false
5. categories, list
6. city
7. review_count
8. name
9. neighborhoods,list
10. longitude,latitude
11. stars
12. state
13. attributes, dic
14. type

Algorithm 1 ScrapyBusiness

```
Input: url-keyword-search_result
Output: users, businesses, reviews
Initialize
Business:
    i. Parse businesses via $xpath('//ul/[@class = ''ylistylist - borderedsearch - results'']')$, about 10 on each page
    ii. Find each business local url by $xpath('.//span[@class =' indexed-biz-name']/a[@class =' biz-name']/@href').extract()$
    iii. fecth www.yelp.com/url
    iv. Adjust $s(u_i)$ by equation (5);
    v. Update $s(r_j)$ by equation (4);
    Until converge or achieve maximum iteration
Output the scores
```

1.2 User

fields

1. yelping since, (date)
2. votes, dic{'funny',count,'useful','cool'}
3. review_count
4. name
5. user_id
6. friends, list of user_ids
7. fans
8. averge_stars
9. type
10. compliments,dic
11. elite,list

1.3 Review

fields

1. votes, dic{'funny','useful','cool',count}
2. user_id
3. review_id

4. stars
5. date, 2013-04-19
6. text
7. type, review
8. business_id
9. check_in
10. not_recommend