openstreetmap数据分析报告

Author: mengyu

Date: 2017/11/05

---

1. 问题

下载的北京的数据，遇到问题如下： 1. 信息中英文混杂； 2. 地址信息(英文)类型包含缩写情况； 3. 餐馆数据中的菜系信息，有些相同意义但不同的词。

信息中英文混杂

数据中同时包含中文，英文和拼音，此次分析中不打算非常细节的分析，所以只需要保证相对较多的数据统一就不会影响到分析结果。修复数据的方式是先统计出异常数据，然后手动生成字典，入库的时候将异常数据转化。查询代码如下（为了方便统计，先忽略异常，将数据导入mongodb）

```
    # 各种设施数量top10
    for item in db.openstreetmap.aggregate([
        {"$match": {"address.street":{"$exists": 1}}},
        {"$group": {"_id":"$address.street", "count": {"$sum": 1}}},
        {"$sort": {"count": -1}},
        {"$limit": 100}
    ]):
        print("{:30} {}".format(item["_id"], item["count"]))
```

根据结果生成字典

```
{
    "光明zhong jie": "光明中街",
    "St. Shuangqiao": "双桥路",
    "Santilun West 6th Street": "三里屯西六街",
    "Xizhaosi Street": "夕照寺街"
}
```

重新清洗数据时加入处理代码。

地址信息(英文)类型包含缩写

需要找到所有缩写，清洗数据的时候需要将缩写替换成全称。 使用脚本分析数据

```
python main.py show_incorrect_street
```

分析结果，生成dict

```
{
    "St": "Street",
    "St.": "Street",
    "Ave": "Avenue",
    "Rd.": "Road",
    "road": "Road",
    "Str": "Street",
    "ave.": "Avenue"
}
```

在清洗数据的代码中加入相关逻辑。

餐馆数据中的菜系信息同义词

菜系信息中存在“chinese”、“中国”、“中餐”、“chinese_food”等数据，这些都可以归属为中餐。数据中绝大多数都是英文，最多的是“chinese”,所以将上述几种数据都转为“chinese”。 查询所有菜系信息数据：

```
for item in db.openstreetmap.aggregate([
    {"$match": {"cuisine":{"$exists": 1}}},
    {"$group": {"_id":"$cuisine", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}},
    {"$limit": 100}
]):
    print("{:30} {}".format(item["_id"], item["count"]))
```

2. 数据汇总

文件大小：

beijing_china.osm

192.10MB

总数据量

```
# 总数据量
db.openstreetmap.find().count()
```

```
1042475
```

节点数据量

```
# 节点数据量
db.openstreetmap.find({"type": "node"}).count()
```

```
907927
```

途径数据量

```
# 途径数据量
db.openstreetmap.find({"type": "way"}).count()
```

```
134534
```

节点类型数量top10

```
# 节点类型数量top10
for item in db.openstreetmap.aggregate([
    {"$match": {"amenity":{"$exists": 1}}},
    {"$group": {"_id":"$amenity", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}},
    {"$limit": 10}
]):
    print("{:30} {}".format(item["_id"], item["count"]))
```

```
restaurant                     1507
parking                        865
school                         549
bank                           476
toilets                        439
fast_food                      351
cafe                           308
fuel                           305
hospital                       197
bar                            168
```

3. 其他数据分析

建筑宗教分布，佛教寺庙远高于其他宗教

```
# 建筑宗教分布
print("{:30} {}".format("所属宗教","建筑数量"))
for i, item in enumerate(db.openstreetmap.aggregate([
    {"$match":{"amenity":"place_of_worship", "name":{"$exists": 1}, "religion": {"$exists":
1}}},
    {"$group": {"_id": "$religion", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}}
])):
    print("{:30} {}".format(item["_id"], item["count"]))
```

```
所属宗教                           建筑数量
buddhist                       43
christian                      12
muslim                         6
taoist                         6
confucian                      1
hindu                          1
```

餐馆是否注明烹饪风格，结果可以发现有72%的餐馆没有标明烹饪风格

```
# 未注明烹饪风格的餐馆总数，注明烹饪风格的餐馆总数
no_cuisine_cnt = db.openstreetmap.find({"amenity":"restaurant", "name": {"$exists": 1}, "cui
sine":{"$exists": 0}}).count()
has_cuisine_cnt = db.openstreetmap.find({"amenity":"restaurant", "name": {"$exists": 1}, "cu
isine":{"$exists": 1}}).count()
no_cuisine_cnt, has_cuisine_cnt
```

```
(826, 317)
```

菜系风格top10，结果可以发现基本上都是中餐，中餐以外，意大利和日本略微多一些

```
# 菜系风格top10
for item in db.openstreetmap.aggregate([
    {"$match": {"amenity":"restaurant", "name": {"$exists": 1}, "cuisine":{"$exists": 1}}},
    {"$unwind": "$cuisine"},
    {"$group": {"_id":"$cuisine", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}},
    {"$limit": 10}
]):
    print("{:50} {}".format(item["_id"], item["count"]))
```

```
chinese                                         162
italian                                         17
japanese                                        17
international                                   12
pizza                                           12
american                                        9
asian                                           9
regional                                        7
german                                          6
korean                                          5
```

数据贡献者贡献数量以及比例, top5的用户贡献了55%左右的数据

```
all_count = db.openstreetmap.find().count()
for x in db.openstreetmap.aggregate([
    {"$group": {"_id": "$created.uid", "count":{"$sum":1}}},
    {"$sort": {"count": -1}},
    {"$limit": 10}
]):
    print("uid:{_id:10}   count:{count:10}    {perc:.2f}%".format(perc=x["count"]*100.0/all_
count,**x))
```

```
uid:    288524   count:    250334    24.01%
uid:    376715   count:    140477    13.48%
uid:    499500   count:     70806    6.79%
uid:   4814295   count:     65893    6.32%
uid:    139957   count:     51860    4.97%
uid:     17497   count:     23407    2.25%
uid:    486052   count:     22329    2.14%
uid:     83557   count:     18075    1.73%
uid:     75424   count:     15725    1.51%
uid:   2639622   count:     13635    1.31%
```

每年提交数据数量，结果可以看到从2012年开始数据量才开始多起来，并且2017年有较高的增长

```
# 创建时间分布
for obj in db.openstreetmap.aggregate([
    {
        "$group": {
            "_id": {"year": {"$year": "$created.timestamp"}},
             "count": {"$sum":1}
        }
    },
    {"$sort": {"_id.year": 1}}
]):
    print(obj["_id"]["year"], obj["count"])
```

```
2007 81
2008 7856
2009 24758
2010 37414
2011 38223
2012 168198
2013 142529
2014 101093
2015 138463
2016 165670
2017 218190
```

4. 建议

尽量减少直接输入，例如餐馆的菜系数据，可以把常见的菜系提供给用户选择的列表里。另外也需要提供给用户一个自己填写的入口，预防有一些新的未知选项，记录到数据库中的时候自动加上标注，定期分析用户填写的内容，逐渐完善选项。

好处：

1. 优化用户体验，选择题比填空题好做；
2. 收集到的数据尽量可控，能极大减少同义词的出现，以及各种随意的描述

预期的问题：

1. 用户填写的内容仍然是无法预期的，仍需要人工进行分析汇总。
2. 语言不通问题，用户恰好不懂选项用的语言。可以结合一些自动翻译的东西做些提示。

5. 总结

经过对北京地区的数据数据进行处理和分析，得到了需要的汇总数据。 过程当中也发现了一些问题，会影响一些分析的准确程度，例如多语言混合的问题，如果完全修复的话需要的代价会很大。

另外数据量也是一个问题，从每年提交数据数量的分析结果中可知，最近几年数据量都很多，而且今年有明显增多，相信数据量会随时间逐渐完善。

参考

中文正则 http://www.cnblogs.com/yitian/archive/2008/11/14/1333569.html

mongo语法 https://docs.mongodb.com/manual/reference/operator/aggregation/group/

openstreetmap数据说明 http://wiki.openstreetmap.org/wiki/Tag:amenity=restaurant?uselang=zh-CN