OpenStreetMap Project
Data Wrangling with MongoDB

Map Area: Shanghai, China

CONTENTS

1. Problems Encountered in the Map

I downloaded a XML OSM dataset of Shanghai city area from OpenStreetMap. After inspecting city names and street names, I found several problems as following：

Problems in city name

- There are other city names("苏州","Hangzhou"), not only Shanghai city or his alias
- Some data have district information, but miss city name("松江区")
- One city may have both name in English and Chinese("上海 Shanghai")
- There is inconsistent writing style("Shanghai","shanghai","上海","上海市")

Methods:

- keep the other city names. looking at the picture on the right or click OpenStreetMap , we can find this dataset not only includes Shanghai city but also areas of surrounding cities like Hangzhou, Wuxi, which shouldn't be considered as incorrect data
- keep the district information and add exact city name if it missed
- considering of conciseness, get rid of its English name if it has both English and Chinese names
- unify inconsistent names
  - change English name to Chinese
  - add administrative level("区"，"市") after district and city name

Examples:

problematic city name	updated city name
"Hangzhou"	"杭州市"
"奉化"	"奉化市"
"静安区"	"上海市静安区"
"Pudong District, Shanghai"	"上海市浦东区"
"上海 Shanghai"	"上海市"
"Shanghai"	"上海市"
"上海"	"上海市"
"上海嘉定区"	"上海市嘉定区"
"Pudong"	"上海市浦东区"

Problems in street name

- Some street name is over-abbreviated("Haigang Ave.","Yongfu Rd.")
- There are inconsistent and incorrect writing styles ("Lu","lu","Rode","Raod")
- There are invalid information without exact street name("长白二村 Changbai Community #2", "S308","浦江镇")
- Some street names include other redundant information like house number, place name or meaningless characters("高新区商业街18号2-3楼","秀沿路1028弄2支弄")
- One street may have both Chinese and English names("世纪大道 Century Avenue"), which I think is redundant

Methods:

- if a street name is over-abbreviated or inconsistent or incorrect, change it to official writing
- keep these invalid information about street in original status
- if street name includes other redundant information like house number, place name or other meaningless characters, get rid of them
- if a street has both Chinese and English name("世纪大道 Century Avenue"), keep Chinese part("世纪大道")

Examples:

problematic street name	update street name
"Yongfu Rd."	"Yongfu Road"
"Haigang Ave."	"Haigang Avenue"
"XingHai street"	"XingHai Street"
"沪南路2729弄"	"沪南路"
"大连西路30弄 Lane 30 of West Dalian Road"	"大连西路"
"高新区商业街18号2-3楼"	"高新区商业街"
"CaoXi Bei lu 99"	"CaoXi Bei Road"
"世纪大道 Century Avenue"	"世纪大道"

2. Data Overview

This section contains basic statistics about the dataset and relative MongoDB queries.

File sizes

shanghai_china.osm———————626MB

```
# Number of documents saved in MongoDB
    > db.p3_project.find().count()
    3317919

# Number of nodes
     > db.p3_project.find({'type':'node'}).count()
     2958645

# Number of ways
     > db.p3_project.find({'type':'way'}).count()
     358965

# Number of unique users
     > db.p3_project.distinct("created.user").length
     1905

# Number of unique cities including district
     > db.p3_project.distinct("address_detail.city").length
     53
```

3. Additional Ideas

Suggestion for analyzing land-use data

According to Wiki, "landuse" tag is mainly used to describe the primary use of land by humans. As for this dataset, the top 5 appearing usages of land are "residential", "industrial", "grass", "forest" and "farmland".
As time goes by, the preferences of land-use could be a lot different. Comparing the "landuse" data of different periods, we might find out the changes of how people develop and use land resources.If the number of commercial place increased significantly, we could guess that people may have more entertainment consumption in past years. If the green area has increased, it's very possible that people have paid more attention on environmental protection.
One thing we should be aware is that some changes could be the result of data integrity improvement instead of human behavior, especially if all kinds of land-use have increased significantly during the same period.

Suggestion for improving dataset

When I queried for most popular cafe shops, I found six shops in top ten are "Starbucks" with different alias, like "Starbucks","星巴克咖啡 Starbucks" etc. Encountering this inconsistency problem, I think out two ways. One is making writing rules for specific name. And the other is providing a frequently used name list for contributors to select, which needs to be maintained by a dedicated worker. As convenience will make more people participate, increase data volume, improve data quality, I suggest the latter method, which does less for more.

Additional queries

```
# Top 5 appearing land-use
    > db.p3_project.aggregate([{"$match":{"landuse":{"$exists":1}}},
    {"$group":{"_id":"$landuse"know,"count":{"$sum":1}}}, {"$sort":
    {"count":-1}},{"$limit":5}])
    { "_id" : "residential", "count" : 4372 }
    { "_id" : "industrial", "count" : 2055 }
    { "_id" : "grass", "count" : 1780 }
    { "_id" : "forest", "count" : 656 }
    { "_id" : "farmland", "count" : 498 }

# Most popular cafe shops
    > db.p3_project.aggregate([{"$match":{"amenity":{"$exists":
    1},"amenity":"cafe","name":{"$exists":1}}},{"$group":
    {"_id":"$name","count":{"$sum":1}}},{"$sort":{"count":-1}},
    {"$limit":10}])
    { "_id" : "Starbucks", "count" : 42 }
    { "_id" : "星巴克", "count" : 15 }
    { "_id" : "星巴克咖啡", "count" : 11 }
    { "_id" : "两岸咖啡", "count" : 8 }
    { "_id" : "Costa Coffee", "count" : 5 }
    { "_id" : "Pacific Coffee", "count" : 5 }
    { "_id" : "Costa", "count" : 4 }
    { "_id" : "星巴克咖啡 Starbucks", "count" : 4 }
    { "_id" : "上岛咖啡", "count" : 4 }
    { "_id" : "星巴克Starbucks", "count" : 4 }

# Top 3 contributing users
    > db.p3_project.aggregate([{$match:{"created.user":{$ne:"null"}}},
    {$group:{_id:"$created.user",count:{$sum:1}}},{$sort:{count:-1}},
    {$limit:3}])
     { "_id" : "Chen Jia", "count" : 523160 }
     { "_id" : "aighes", "count" : 186219 }
     { "_id" : "katpatuka", "count" : 130723 }

# Most popular leisure places
    > db.p3_project.aggregate([{"$match":{"leisure":{"$exists":1}}},
    {"$group":{"_id":"$leisure","count":{"$sum":1}}}, {"$sort":
    {"count":-1}},{"$limit":3}])
    { "_id" : "pitch", "count" : 1882 }
    { "_id" : "park", "count" : 1782 }
    { "_id" : "common", "count" : 170 }

# Most popular convenience stores
    > db.p3_project.aggregate([{"$match":{"shop":{"$exists":
    1},"shop":"convenience","name":{"$exists":1}}},{"$group":
    {"_id":"$name","count":{"$sum":1}}},{"$sort":{"count":-1}},
    {"$limit":3}])
    { "_id" : "可的", "count" : 38 }
    { "_id" : "快客", "count" : 21 }
    { "_id" : "喜士多", "count" : 18 }
```