This report describes the various locations of San Diego, detailing shops, roads, etc. I chose San Diego because it is the town I feel most familiar to. The purpose of this report is to expose myself to Python code involving parsing and manipulating large XML files.

DIFFICULTIES:

There were some mapping difficulties encountered, mostly due to abbreviated street and shop names. For example, not all 7-Eleven convenience stores were recorded by the official name. I have addressed the most obvious issues by standardizing all abbreviations to a single common element. This helped greatly in translating abbreviated street names to their full names. A regex was used to arbitrate between skipped street names. Street names that were only numbers or blank spaces were not passed.

Every node was examined using loops and regexes to extract relevant address info and data. I did not anticipate that nowhere in the original OSM file is the mention of a "user" field in any node. This can be attributed to the fact that OpenStreetMap can be modified anonymously.

Every node element was checked for redundant information, such as country/city attributes, as well as repeated information such as zip codes. This was done by compiling regexes to filter unwanted elements containing colons. The subsequent file assumes that all ways and nodes are in the San Diego area.

Another difficulty in properly recording street and shop names was, capitalizing them with correct grammar. A filter was implemented that applied this only to element keys ending with "name".

Some database posts were updated via. Pymongo API, such as amenity fields for certain chain cafes and restaurants.

###CODE

Interesting Observations:

Top Five Cafes in San Diego:

```
{u'count': 53, u'_id': u'Starbucks'}
{u'count': 3, u'_id': u'Fairbanks'}
{u'count': 3, u'_id': u'Jamba Juice'}
{u'count': 3, u'_id': u"Peet's Coffee & Tea"}
{u'count': 3, u'_id': u'The Coffee Bean'}
```

Pipeline:

```
pipeline = [
    {"$match": {"name":{"$exists":1},"amenity":{"$in":[re.compile("cafe", re.I),
re.compile("coffee", re.I)]}}},
    {"$group": {"_id": "$name", "count": {"$sum": 1}}},
    {"$sort": {"count":-1}},
```

```
    {"$limit":100}
]
```

Top Five “Restaurants” in San Diego:

```
{u'count': 74, u'_id': u"Mcdonald's"}
{u'count': 71, u'_id': u'Jack In The Box'}
{u'count': 62, u'_id': u'Subway Sandwiches Sandwiches'}
{u'count': 37, u'_id': u'Taco Bell'}
{u'count': 32, u'_id': u'Carls Jr.'}
```

```
pipeline = [
    {"$match": {"name":{"$exists":1},"amenity":{"$in":[re.compile("cafe", re.I),
re.compile("coffee", re.I)]}}},
    {"$group": {"_id": "$name", "count": {"$sum": 1}}},
    {"$sort": {"count":-1}},
    {"$limit":100}
]
```

```
Total # Nodes: 1371754
db.small_posts.find({"type":"node"}).count()

Total # Ways: 98438
db.small_posts.find({"type":"way"}).count()

Total # Unique Users: 0 (???)
db.small_posts.distinct("created.user")

Total # Documents: 1470192
db.small_posts.find().count()

# Unique Cuisines: 92
len(db.small_posts.distinct("cuisine"))

Top 5 Ethnic Cuisines:
...
{u'count': 169, u'_id': u'mexican'}
...
{u'count': 42, u'_id': u'american'}
...
{u'count': 32, u'_id': u'chinese'}
{u'count': 30, u'_id': u'italian'}
...
{u'count': 23, u'_id': u'sushi'}
{u'count': 21, u'_id': u'thai'}
{u'count': 20, u'_id': u'japanese'}
```

pipeline = [

```
    {"$match": {"cuisine":{"$exists":1}}},
```

```
    {"$group": {"_id": "$cuisine", "count": {"$sum": 1}}},
    {"$sort": {"count":-1}},
]
```

Having lived in this area for some time, I sense that certain restaurants are incorrectly labeled or neglected. I know that there are more Korean restaurants than Japanese, most of them barbecue joints or cafes.

Convoy Street:

Convoy Street is a lively and popular hang-out area, and is the site of many popular asian restaurants and markets. I will compare data from this portion of the map to the rest of San Diego.

Coordinates of interest:
32.8383
32.8134
-117.1693
-117.1353

Code:

```
pipeline = [
    {"$match": {"amenity":{"$exists":1}, "$and":[{"pos.lat":{"$gt":32.8134,
"$lt":32.8383}}, {"pos.lon":{"$gt":-117.1693, "$lt":-117.1353}}]}},
    {"$group": {"_id": "$amenity", "count": {"$sum": 1}}},
    {"$sort": {"count":-1}},
    {"$limit":20}
]
```

Top 10 Amenities:
```
{u'count': 18, u'_id': u'restaurant'}
{u'count': 17, u'_id': u'fast food'}
{u'count': 11, u'_id': u'place of worship'}
{u'count': 9, u'_id': u'cafe'}
{u'count': 7, u'_id': u'fuel'}
{u'count': 6, u'_id': u'bar'}
{u'count': 6, u'_id': u'dentist'}
{u'count': 5, u'_id': u'shelter'}
{u'count': 3, u'_id': u'bank'}
{u'count': 2, u'_id': u'school'}
```

Top 10 Cuisines:

```
{u'count': 8, u'_id': u'sandwich'}
{u'count': 5, u'_id': u'burger'}
{u'count': 4, u'_id': u'mexican'}
{u'count': 4, u'_id': u'chinese'}
{u'count': 3, u'_id': u'korean'}
{u'count': 2, u'_id': u'pizza'}
{u'count': 2, u'_id': u'japanese'}
{u'count': 2, u'_id': u'asian'}
{u'count': 1, u'_id': u'vietnamese'}
```

{u'count': 1, u'_id': u'chinese;japanese'}

Because “sandwich” does not specifically describe ethnicity (all burger joints are generic fast-food), I have included the names of the shops.

{u'count': 3, u'_id': u'Subway'}
{u'count': 1, u'_id': u'Cali Baguette'}
{u'count': 1, u'_id': u"Mama's Grill"}
{u'count': 1, u'_id': u"Jersey Mike's"}
{u'count': 1, u'_id': u'Subway Sandwiches Sandwiches'}
{u'count': 1, u'_id': u"Arby's"}

It is evident that many restaurants are neglected in the OSM file, and warrants an update. But a preliminary study shows an abundance of South-East Asian restaurants. I was surprised to not see any mention of Mitsuwa or Marukai, which are my favorite Japanese supermarkets.