C750 - Wrangling OpenStreetMap Data

April 25, 2019

1 Wrangling Chattanooga's OpenStreetMap Data

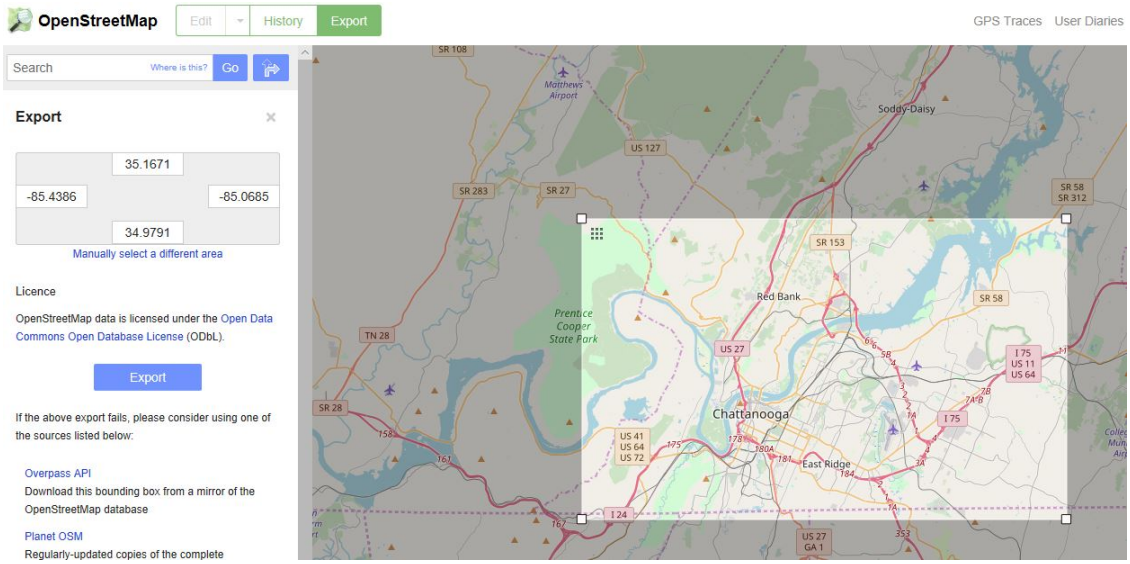
Using OpenStreetMap I exported data for Chattanooga TN and will be attempting to wrangle and normalize this data. I chose Chattanooga due to the fact that it's on my short list of possible cities to relocate to in the near future and it's one my wife and I's favorite cities to visit.

The dataset can be downloaded using this link for the Overpass API

Below is the map area used.

```
In [2]: from IPython.display import Image
        Image("pictures/chatmap.jpg")
```

Out[2]:



I'll be using python for the data manipulation coding, so the first thing to do will be to import the needed libraries. To start things off, I will define the OSM file as well that we will be using

```
In [1]: import os
        import xml.etree.cElementTree as ET
        import pprint
```

```
import re
from collections import defaultdict
import codecs
import json
import pymongo

osm_file = 'Chattanooga.osm'
```

1.1 Auditing the dataset

To start off working with the file, I will audit the data, in an effort to get a better understanding of the information it contains. The first step in doing this is to count the tag types found in the dataset.

```
In [15]: def count_tags(filename):
             tags = {}

             for event, elem in ET.iterparse(filename):

                 if elem.tag not in tags:

                     tags[elem.tag] = 1

                 else:

                     tags[elem.tag] += 1

             return tags

         def test():

             tags = count_tags(osm_file)
             print ('Types and counts elements in Chattanooga.osm:\n')
             pprint.pprint(tags)

         if __name__ == "__main__":
             test()
```

```
Types and counts elements in Chattanooga.osm:

{'bounds': 1,
 'member': 19043,
 'meta': 1,
 'nd': 546974,
 'node': 477107,
 'note': 1,
 'osm': 1,
 'relation': 287,
```

```
 'tag': 198577,
 'way': 56861}
```

Referencing the OpenStreetMap Wiki about the different elements types will help us understand the data we have here.

The wiki can be found here

Using the information from the above source, the elements hold the description of the specific element they are attached to via the 'key' and 'value' properties.

elements describe the meaning of the particular element to which they are attached by holding two free format text fields; a 'key' and a 'value'. I am also interested in getting more information about the child tags of and to gain more information about the entry.

My next step in going deeper into the data is to list out and count the attributes it contains. I'm going to do this by building a dictionary of those and their count.

```
In [20]: # This loop will create out dictionary and count
         def getattributes(filename):
             attrs = {}
             for event, elem in ET.iterparse(filename, events=('start', 'end')):
                 if event == 'end':
                     for attr in elem.attrib:
                         if attr not in attrs:
                             attrs[attr] = 1
                         else:
                             attrs[attr] += 1
             return attrs


         attrs = getattributes(osm_file)

         print ('Attributes and counts:\n')
         pprint.pprint(attrs)
```

```
Attributes and counts:

{'changeset': 534255,
 'generator': 1,
 'id': 534255,
 'k': 198577,
 'lat': 477107,
 'lon': 477107,
 'maxlat': 1,
 'maxlon': 1,
 'minlat': 1,
 'minlon': 1,
 'osm_base': 1,
 'ref': 566017,
 'role': 19043,
```

```
 'timestamp': 534255,
 'type': 19043,
 'uid': 534255,
 'user': 534255,
 'v': 198577,
 'version': 534256}
```

As I mentioned earlier the two attributes that contains the most relevant data are k and v, key and value respectively. The next step is to further look into those values. Using a similar loop, I will build a dictionary of values for k.

```
In [21]: def getkeys(filename):
             keys = {}
             for event, elem in ET.iterparse(filename, events=('start', 'end')):
                 if event == 'end':
                     key = elem.attrib.get('k')
                     if key:
                         if key not in keys:
                             keys[key] = 1
                         else:
                             keys[key] += 1
             return keys


         keys = getkeys(osm_file)

         print ('Keys and counts:\n')
         pprint.pprint(keys)
```

```
Keys and counts:

{'CHA:bicycle': 245,
 'FIXME': 43,
 'HFCS': 19,
 'ISO3166-1': 1,
 'ISO3166-1:alpha2': 1,
 'ISO3166-1:alpha3': 1,
 'ISO3166-1:numeric': 1,
 'ISO3166-2': 2,
 'NHD:ComID': 2160,
 'NHD:Elevation': 338,
 'NHD:FCode': 2120,
 'NHD:FDate': 340,
 'NHD:FType': 2120,
 'NHD:GNIS_ID': 10,
 'NHD:GNIS_Name': 10,
 'NHD:OBJECTID': 340,
```

```
'NHD:RESOLUTION': 1780,
'NHD:ReachCode': 2108,
'NHD:Resolution': 340,
'NHD:way_id': 1780,
'NHS': 45,
'OBJECTID': 1,
'Rec_Area': 1,
'Road': 1,
'Trail': 1,
'USGS-LULC:CLASS': 3,
'USGS-LULC:CNTYNAME': 3,
'USGS-LULC:LEVEL_I': 3,
'USGS-LULC:LEVEL_II': 3,
'USGS-LULC:STATECTY': 3,
'access': 415,
'addr:city': 4210,
'addr:country': 52,
'addr:full': 8,
'addr:housename': 119,
'addr:housenumber': 2328,
'addr:inclusion': 7,
'addr:postcode': 3483,
'addr:state': 3586,
'addr:street': 5278,
'addr:unit': 24,
'addr:use': 61,
'admin_level': 136,
'aeroway': 76,
'alcohol': 1,
'alt_name': 157,
'alt_name:vi': 1,
'amenity': 2920,
'area': 274,
'artist_name': 3,
'artwork_type': 4,
'atm': 5,
'attraction': 1,
'attribution': 1829,
'barrier': 84,
'basin': 2,
'beds': 2,
'bench': 4,
'bicycle': 1222,
'bicycle_parking': 1,
'board_type': 1,
'boat': 5,
'border_type': 19,
'boundary': 151,
```

```
'brand': 81,
'brand:wikidata': 73,
'brand:wikipedia': 67,
'brewery': 1,
'bridge': 389,
'bridge:name': 2,
'building': 29415,
'building:levels': 1585,
'building:part': 14,
'bunker_type': 4,
'bus': 2,
'cables': 4,
'capacity': 51,
'capacity:disabled': 12,
'capacity:parent': 4,
'capacity:women': 5,
'category': 1,
'census:population': 7,
'change:lanes:backward': 3,
'change:lanes:forward': 3,
'club': 1,
'construction': 7,
'content': 2,
'cost:coffee': 1,
'covered': 19,
'craft': 23,
'created_by': 59,
'crossing': 73,
'cuisine': 143,
'culvert': 3,
'cutting': 2,
'cycle_network': 8,
'cycleway': 187,
'cycleway:left': 2,
'cycleway:right': 4,
'deep_draft': 4,
'default_language': 1,
'delivery': 5,
'denomination': 178,
'description': 24,
'description:en': 1,
'designation': 4,
'destination': 60,
'destination:lanes': 3,
'destination:ref': 47,
'destination:ref:to': 6,
'destination:street': 50,
'diet:vegetarian': 1,
```

```
'direction': 4,
'dispensing': 5,
'distance': 56,
'distillery': 1,
'drinking_water': 1,
'drive_through': 8,
'ele': 739,
'electrified': 164,
'email': 8,
'emergency': 37,
'entrance': 4,
'faa': 1,
'fax': 12,
'fee': 46,
'fence_type': 3,
'fire_hydrant:type': 13,
'fireplace': 1,
'fixme': 6,
'flag': 1,
'foot': 833,
'footway': 172,
'ford': 1,
'frequency': 7,
'fuel': 1,
'fuel:cng': 7,
'fuel:diesel': 2,
'fuel:octane_98': 1,
'gauge': 161,
'gdot:grip': 2,
'generator:method': 12,
'generator:output:electricity': 7,
'generator:source': 11,
'generator:type': 11,
'gnis:Class': 222,
'gnis:County': 221,
'gnis:County_num': 221,
'gnis:ST_alpha': 221,
'gnis:ST_num': 221,
'gnis:county_id': 471,
'gnis:county_name': 40,
'gnis:created': 492,
'gnis:edited': 5,
'gnis:feature_id': 1006,
'gnis:feature_type': 21,
'gnis:id': 220,
'gnis:import_uuid': 19,
'gnis:reviewed': 19,
'gnis:state_id': 471,
```

```
'golf': 56,
'healthcare': 5,
'healthcare:speciality': 3,
'height': 4,
'heritage': 4,
'heritage:operator': 4,
'hgv': 278,
'hgv:national_network': 72,
'highway': 21161,
'historic': 50,
'historic:amenity': 11,
'history': 146,
'horse': 1737,
'hours': 1,
'iata': 1,
'icao': 1,
'imminate': 1,
'import_uuid': 220,
'incline': 1,
'industrial': 1,
'information': 6,
'inscription': 1,
'int_name': 1,
'intermittent': 4,
'internet_access': 9,
'internet_access:fee': 4,
'is_in': 220,
'is_in:continent': 2,
'is_in:country': 17,
'is_in:country_code': 17,
'is_in:iso_3166_2': 15,
'is_in:state': 44,
'is_in:state_code': 15,
'junction': 25,
'junction:ref': 25,
'landuse': 1018,
'lanes': 836,
'lanes:backward': 9,
'lanes:forward': 9,
'layer': 431,
'leaf_cycle': 2,
'leaf_type': 4,
'leisure': 1047,
'leisure_1': 1,
'length': 31,
'length_unit': 13,
'level': 1,
'line': 3,
```

```
'lit': 20,
'location': 69,
'lock': 1,
'man_made': 172,
'material': 4,
'maxheight': 1,
'maxspeed': 269,
'maxspeed:advisory': 29,
'maxspeed:hgv': 11,
'memorial': 1,
'microbrewery': 2,
'military': 70,
'minspeed': 8,
'motor_vehicle': 500,
'motorcar': 5,
'motorcycle': 5,
'mtb:scale': 7,
'mtb:scale:imba': 22,
'name': 11343,
'name:ab': 1,
'name:ace': 1,
'name:af': 1,
'name:als': 1,
'name:am': 1,
'name:an': 2,
'name:ang': 1,
'name:ar': 4,
'name:arc': 1,
'name:arz': 1,
'name:as': 1,
'name:ast': 1,
'name:av': 1,
'name:ay': 1,
'name:az': 3,
'name:ba': 1,
'name:bar': 1,
'name:bat-smg': 1,
'name:bcl': 1,
'name:be': 2,
'name:be-tarask': 1,
'name:bg': 3,
'name:bi': 1,
'name:bm': 1,
'name:bn': 3,
'name:bo': 1,
'name:bpy': 1,
'name:br': 3,
'name:bs': 1,
```

```
'name:bxr': 1,
'name:ca': 2,
'name:cbk-zam': 1,
'name:cdo': 1,
'name:ce': 1,
'name:ceb': 1,
'name:chr': 1,
'name:chy': 1,
'name:ckb': 1,
'name:co': 1,
'name:crh': 1,
'name:cs': 2,
'name:csb': 1,
'name:cu': 1,
'name:cv': 2,
'name:cy': 1,
'name:da': 1,
'name:de': 1,
'name:diq': 1,
'name:dsb': 1,
'name:dv': 1,
'name:dz': 1,
'name:ee': 1,
'name:el': 3,
'name:eml': 1,
'name:en': 7,
'name:eo': 3,
'name:es': 1,
'name:et': 1,
'name:etymology:wikidata': 11,
'name:eu': 1,
'name:ext': 1,
'name:fa': 3,
'name:ff': 1,
'name:fi': 1,
'name:fiu-vro': 1,
'name:fo': 1,
'name:fr': 2,
'name:frp': 2,
'name:frr': 1,
'name:fur': 1,
'name:fy': 1,
'name:ga': 1,
'name:gag': 1,
'name:gan': 1,
'name:gd': 2,
'name:gl': 2,
'name:glk': 1,
```

```
'name:gn': 1,
'name:gu': 1,
'name:gv': 2,
'name:ha': 1,
'name:hak': 3,
'name:haw': 3,
'name:he': 3,
'name:hi': 3,
'name:hif': 1,
'name:hr': 1,
'name:hsb': 1,
'name:ht': 3,
'name:hu': 1,
'name:hy': 3,
'name:ia': 1,
'name:id': 1,
'name:ie': 1,
'name:ig': 1,
'name:ik': 1,
'name:ilo': 1,
'name:io': 1,
'name:is': 2,
'name:it': 1,
'name:iu': 1,
'name:ja': 3,
'name:jbo': 1,
'name:jv': 1,
'name:ka': 3,
'name:kaa': 1,
'name:kab': 1,
'name:kbd': 1,
'name:ki': 1,
'name:kk': 1,
'name:kl': 1,
'name:km': 1,
'name:kn': 2,
'name:ko': 3,
'name:koi': 1,
'name:krc': 1,
'name:ks': 1,
'name:ksh': 1,
'name:ku': 1,
'name:kv': 1,
'name:kw': 2,
'name:ky': 1,
'name:la': 1,
'name:lad': 1,
'name:lb': 1,
```

```
'name:lbe': 1,
'name:lez': 1,
'name:lg': 1,
'name:li': 1,
'name:lij': 1,
'name:lmo': 1,
'name:ln': 1,
'name:lo': 1,
'name:lt': 5,
'name:ltg': 1,
'name:lv': 3,
'name:lzh': 1,
'name:map-bms': 1,
'name:mdf': 1,
'name:mg': 1,
'name:mhr': 1,
'name:mi': 1,
'name:min': 1,
'name:mk': 2,
'name:ml': 3,
'name:mn': 2,
'name:mo': 1,
'name:mr': 3,
'name:mrj': 1,
'name:ms': 1,
'name:mt': 1,
'name:mwl': 1,
'name:my': 1,
'name:myv': 1,
'name:mzn': 1,
'name:na': 1,
'name:nah': 1,
'name:nan': 1,
'name:nap': 1,
'name:nds': 1,
'name:nds-nl': 1,
'name:ne': 1,
'name:new': 1,
'name:nl': 1,
'name:nn': 1,
'name:no': 1,
'name:nov': 1,
'name:nrm': 1,
'name:nso': 1,
'name:nv': 3,
'name:oc': 1,
'name:om': 1,
'name:or': 1,
```

```
'name:os': 2,
'name:pa': 1,
'name:pag': 1,
'name:pam': 1,
'name:pap': 1,
'name:pcd': 1,
'name:pdc': 1,
'name:pfl': 1,
'name:pih': 1,
'name:pl': 3,
'name:pms': 1,
'name:pnb': 2,
'name:ps': 1,
'name:pt': 2,
'name:qu': 3,
'name:rm': 1,
'name:rn': 1,
'name:ro': 1,
'name:roa-tara': 1,
'name:ru': 6,
'name:rue': 1,
'name:rw': 1,
'name:sa': 1,
'name:sah': 1,
'name:sc': 1,
'name:scn': 1,
'name:sco': 1,
'name:sd': 1,
'name:se': 1,
'name:sg': 1,
'name:sh': 1,
'name:si': 1,
'name:sk': 1,
'name:sl': 1,
'name:sm': 1,
'name:sn': 1,
'name:so': 1,
'name:sq': 1,
'name:sr': 3,
'name:srn': 1,
'name:ss': 1,
'name:stq': 1,
'name:su': 1,
'name:sv': 1,
'name:sw': 1,
'name:szl': 1,
'name:ta': 3,
'name:te': 2,
```

```
'name:tet': 1,
'name:tg': 1,
'name:th': 3,
'name:tk': 1,
'name:tl': 1,
'name:tn': 1,
'name:to': 1,
'name:tpi': 1,
'name:tr': 1,
'name:ts': 1,
'name:tt': 1,
'name:tw': 1,
'name:ty': 1,
'name:tzl': 1,
'name:udm': 1,
'name:ug': 1,
'name:uk': 4,
'name:ur': 2,
'name:uz': 3,
'name:vec': 1,
'name:vep': 1,
'name:vi': 1,
'name:vls': 1,
'name:vo': 1,
'name:wa': 1,
'name:war': 1,
'name:wo': 1,
'name:wuu': 1,
'name:xal': 1,
'name:xh': 1,
'name:xmf': 1,
'name:yi': 3,
'name:yo': 1,
'name:yue': 1,
'name:za': 1,
'name:zea': 1,
'name:zh': 3,
'name:zh-Hans': 1,
'name:zh-Hant': 1,
'name:zu': 1,
'name_1': 104,
'name_2': 4,
'natural': 3665,
'natural_1': 1,
'network': 146,
'nist:fips_code': 8,
'nist:state_fips': 8,
'noexit': 1,
```

```
'noref': 18,
'note': 42,
'note:lanes': 8,
'note:old_railway_operator': 54,
'office': 21,
'official_name': 3,
'official_name:cs': 1,
'official_name:en': 1,
'official_name:eo': 1,
'official_name:nl': 1,
'official_name:pl': 1,
'official_name:vi': 1,
'old_name': 12,
'old_name:ru': 1,
'old_name:vi': 1,
'old_railway_operator': 114,
'old_short_name:ru': 1,
'oneway': 1458,
'opening_hours': 65,
'operator': 352,
'operator:wikidata': 2,
'operator:wikipedia': 2,
'outdoor_seating': 3,
'park_ride': 20,
'parking': 173,
'parking:condition:both': 1,
'parking:condition:left': 1,
'parking:lane:both': 1,
'parking:lane:left': 1,
'payment:american_express': 1,
'payment:bitcoin': 3,
'payment:cash': 3,
'payment:cheque': 1,
'payment:coins': 1,
'payment:credit_cards': 3,
'payment:debit_cards': 1,
'payment:mastercard': 1,
'payment:visa': 1,
'payments': 1,
'phone': 155,
'place': 283,
'placement': 1,
'plant:output:electricity': 1,
'plant:source': 1,
'playground': 1,
'population': 15,
'power': 1182,
'project': 2,
```

```
'protect_class': 8,
'protection_title': 8,
'public_transport': 3,
'railway': 1316,
'railway:preserved': 12,
'railway:switch': 4,
'railway:traffic_mode': 87,
'rcn': 3,
'rcn_ref': 3,
'recycling_type': 1,
'ref': 768,
'ref:expressoil': 3,
'ref:fips': 2,
'ref:left': 3,
'ref:nrhp': 4,
'ref:right': 3,
'ref:walmart': 9,
'religion': 325,
'residential': 9,
'restriction': 9,
'roof:levels': 1,
'roof:shape': 8,
'rooftop': 1,
'rooms': 1,
'route': 83,
'sac_scale': 202,
'salt': 1,
'segregated': 62,
'service': 8159,
'service:bicycle:chain_tool': 1,
'service:bicycle:pump': 1,
'service:bicycle:tools': 1,
'service_times': 4,
'shelter': 8,
'shelter_type': 10,
'ship': 5,
'shop': 314,
'short_name': 3,
'short_name:cs': 1,
'short_name:en': 1,
'short_name:mo': 1,
'short_name:nl': 1,
'short_name:pl': 1,
'short_name:ru': 1,
'short_name:vi': 1,
'short_name:zh': 1,
'sidewalk': 210,
'sidewalk:both:surface': 1,
```

```
'sidewalk:right:surface': 2,
'smoking': 17,
'social_facility': 3,
'social_facility:for': 1,
'socket:chademo': 1,
'socket:type1_combo': 1,
'source': 3525,
'source:deep_draft': 4,
'source:hgv:national_network': 72,
'source:name': 17,
'source:name:br': 2,
'source:population': 3,
'source_ref': 2,
'sport': 491,
'stars': 1,
'start_date': 1,
'state': 4,
'state_id': 2,
'store': 1,
'substation': 2,
'supervised': 10,
'surface': 2215,
'surveillance': 2,
'swimming_pool': 1,
'switch': 1,
'symbol': 27,
'tactile_paving': 2,
'takeaway': 11,
'theatre:type': 1,
'tiger:CLASSFP': 15,
'tiger:CPI': 15,
'tiger:FUNCSTAT': 15,
'tiger:LSAD': 15,
'tiger:MTFCC': 15,
'tiger:NAME': 15,
'tiger:NAMELSAD': 15,
'tiger:PCICBSA': 15,
'tiger:PCINECTA': 15,
'tiger:PLACEFP': 15,
'tiger:PLACENS': 15,
'tiger:PLCIDFP': 15,
'tiger:STATEFP': 15,
'tiger:cfcc': 7173,
'tiger:county': 7210,
'tiger:mtfcc': 35,
'tiger:name_base': 6485,
'tiger:name_base_1': 415,
'tiger:name_base_2': 175,
```

```
'tiger:name_base_3': 150,
'tiger:name_base_4': 30,
'tiger:name_direction_prefix': 680,
'tiger:name_direction_prefix_1': 5,
'tiger:name_direction_prefix_3': 1,
'tiger:name_direction_suffix': 48,
'tiger:name_type': 5895,
'tiger:name_type_1': 68,
'tiger:name_type_2': 4,
'tiger:name_type_3': 2,
'tiger:name_type_4': 2,
'tiger:reviewed': 6846,
'tiger:separated': 662,
'tiger:source': 822,
'tiger:tlid': 818,
'tiger:upload_uuid': 816,
'tiger:zip_left': 5298,
'tiger:zip_left_1': 132,
'tiger:zip_left_2': 17,
'tiger:zip_left_3': 3,
'tiger:zip_right': 5106,
'tiger:zip_right_1': 40,
'tiger:zip_right_2': 6,
'timezone': 2,
'toilets:disposal': 1,
'toilets:position': 1,
'toilets:wheelchair': 3,
'tourism': 204,
'tower:type': 9,
'tracks': 3,
'traffic_calming': 1,
'traffic_signals': 91,
'traffic_signals:direction': 1,
'trail_visibility': 57,
'train': 2,
'tunnel': 122,
'turn': 2,
'turn:lanes': 4,
'turn:lanes:forward': 1,
'type': 350,
'unsigned_ref': 270,
'usage': 151,
'user_defined': 1,
'vending': 8,
'voltage': 25,
'voltage:primary': 1,
'was:building': 1,
'water': 59,
```

```
 'waterway': 1941,
 'website': 188,
 'wetland': 1,
 'wheelchair': 41,
 'wheelchair:description': 1,
 'width': 9,
 'wifi': 2,
 'wikidata': 93,
 'wikipedia': 83,
 'zip_left': 27,
 'zip_right': 27}
```

The list is quite large but I will look at the values for the keys a bit further along in the investigation to help find issues in the dataset.

So the list of keys is very very large, but looking through the dictionary it seems the tags can be lumped into a few general types. Using some expressions, I'll loop through and lump similar tags together.

```
In [29]: #The lower expression is going to count tags that only contain lowercase leters but a
         lower = re.compile(r'^([a-z]|_)*$')
         #The lower_colon expression checks for tags that should be valid but have a colon
         lower_colon = re.compile(r'^([a-z]|_)*:([a-z]|_)*$')
         #The probelm character check below is going to return of count of tags with character
         problemchars = re.compile(r'[=\+/&<>;\'"\?%#$@\,\. \t\r\n]')



         def key_type(element, keys):
             if element.tag == "tag":

                 if lower.search(element.attrib['k']):
                     keys['lower'] += 1
                 elif lower_colon.search(element.attrib['k']):
                     keys['lower_colon'] += 1
                 elif problemchars.search(element.attrib['k']):
                     keys['problemchras'] += 1
                 else:
                     keys['other'] += 1

             return keys


         def process_map(filename):
             keys = {"lower": 0, "lower_colon": 0, "problemchars": 0, "other": 0}
             for _, element in ET.iterparse(filename):
                 keys = key_type(element, keys)
```