Named Entity Extraction and Geo-location extraction

Hong Wang

2015/06/24 16:23

Table of Contents

Named Entity Extraction

First, a list of entities is extracted from the text or the user profile location of the corresponding documents using Stanford CoreNlp library (http://nlp.stanford.edu/software/corenlp.shtml). There are 10 types of entities: PERSON, LOCATION, ORGANIZATION, MISC, DATE, TIME, DURATION, MONEY, PERCENT, and NUMBER. Only the location entities are used for geocoding service.

Geocoding

The geocoding is done using Geoname library (http://www.geonames.org/). It converts the location entity with an official name, and its corresponding coordinates, along with other information. An example of the geoname field looks like this:

```
geoname " : {
                  " location " : " Texas ",
                  " countryName " : " United States ",
                  " continentCode " : " NA ",
                  " population " : 22875689,
                  " coord " : {
                        " lat " : 31.25044,
                        " lng " : -99.25061
                  },
                  " boundingBox " : {
                        " north " : 36.5007,
                        " west " : -106.64565,
                        " east " : -93.50804,
                        " south " : 25.83716
                  },
                  " featureCode " : " ADM1 ",
                  " featureClass " : " country, state, region, ..."
}
```

Location:contains the official name of the location for the entity.

countryName (Nullable): is the name of the country in which the location is located.

continentCode (Nullable): is the continent code of the location.

Population (Nullable): is the population for the location.

Coord: is the center coordinate of the location.

boundingBox: is the bounding box of the location. The fields within the object are the distance from the center of the location to the edge of the bounding box, along 4 different directions.

featureCode: represents the detailed type of the location, which can be looked up here

featureClass: represents the type of the location, such as “A country, state, region,...” and “H stream, lake, ...”. Each feature class contains a set of feature codes. Both feature code and feature class can be looked up here (http://www.geonames.org/export/codes.html).

Deal With Usage Limits:

The Geoname service limits hourly and daily usage. For each account the maximum hourly usage is 2000 request. Therefore 10 accounts are created in order to get around the usage limit. Also, each unique result from the request is stored inside a geoname cache in the mongodb database.

Combined Field for NER and Geoname:

The results from both NER and Geoname serviced are stored in a single array, here is an example:

Document:

Text: RT @goodnewsnetwork: Ebola is spreading in Guinea, but in one town (pop.300K) doctors stopped the outbreak &amp; city is now Ebola-free #Ebola …

user.location : London, UK

```
[{
                "entity" : "Guinea",
                "entityType" : "LOCATION",
                "from" : "text",
                "geoname" : {
                        "location" : "Guinea",
                        "countryName" : "Guinea",
                        "continentCode" : "AF",
                        "population" : 10324025,
                        "coord" : {
                                "lat" : 10.83333,
                                "lng" : -10.66667
                        },
                        "boundingBox" : {
                                "north" : 12.67622,
                                "west" : -15.08626,
                                "east" : -7.64107,
                                "south" : 7.19355
                        },
                        "featureCode" : "PCLI",
                        "featureClass" : "country, state, region,..."
                }
        }, {
                "entity" : "one",
                "entityType" : "NUMBER",
                "from" : "text"
```

```
}, {
    "entity" : ".300",
    "entityType" : "NUMBER",
    "from" : "text"
}, {
    "entity" : "now",
    "entityType" : "DATE",
    "from" : "text"
}, {
    "entity" : "Ebola-free Ebola",
    "entityType" : "MISC",
    "from" : "text"
}, {
    "entity" : "London",
    "entityType" : "LOCATION",
    "from" : "user.location",
    "geoname" : {
        "location" : "London",
        "countryName" : "United Kingdom",
        "continentCode" : "EU",
        "population" : 7556900,
        "coord" : {
            "lat" : 51.50853,
            "lng" : -0.12574
        },
        "boundingBox" : {
            "north" : 51.86537,
            "west" : -0.70361,
            "east" : 0.45212,
            "south" : 51.15169
        },
        "featureCode" : "PPLC",
        "featureClass" : "city, village,..."
    }
}, {
    "entity" : "UK",
    "entityType" : "LOCATION",
    "from" : "user.location",
    "geoname" : {
            "location" : "London",
            "countryName" : "United Kingdom",
            "continentCode" : "EU",
            "population" : 7556900,
            "coord" : {
                "lat" : 51.50853,
                "lng" : -0.12574
            },
            "boundingBox" : {
                "north" : 51.86537,
                "west" : -0.70361,
                "east" : 0.45212,
                "south" : 51.15169
            },
            "featureCode" : "PPLC",
            "featureClass" : "city, village,..."
        }
    }
]
```

Combined location field:

To enable quick look up of locations in the database, a combined location field is inserted to each document, which combines the location coordinates from NER, the coordinate field, the location field, and the place field. The field is stored as a GeometryCollection in the GeoJSON format, which is compatible with mongodb geo-spatial index, here is an example:

```
"locationCollection" : {
        "type" : "GeometryCollection",
        "geometries" : [
            {
                "type" : "Point",
                "coordinates" : [
                    28.374493,
                    -16.793856
                ]
            },
            {
                "type" : "Point",
                "coordinates" : [
                    40.4165,
                    -3.70256
            ]
        }
    ]
}
```