Python for Bioinformatics

adventures in bioinformatics

Thursday, June 25, 2020

Shapefiles

I've been working with maps using Python, primarily maps for the United States. The standard format for much geographic data is GeoJSON.

But there is another format that is even more official, and is maintained by the US Census Bureau. That is a collection of Shapefiles.

To recap, here is a site where GeoJSON files for the Us are available in multiple sizes (small, medium, large), as well as their .kml and Shapefile .shp equivalents. The sizes are 500k, 5m, 20m, from largest to smallest (the labels are the scales, 500k being the most detailed).

The files were obtained from links on this page. It includes data for the US, for the states, and for counties. And it contains Congressional districts, which would be useful to remember.

A Shapefile is binary-encoded geographic data in a particular format. A good discussion is here.

The specification was developed at least partly by ESRI, which develops geographic information software.. The encoding was undoubtedly designed to save space, back when space (storage, transmission bandwidth) was much more expensive. Now, the opaque data is a liability.

Shapefiles

There is actually not just one file, but always a minimum of three including .shp, .shx and .dbf inside a zip container. The Shapefiles from the US Census also have .prj and .xml.

I can't tell much by looking at one with hexdump, except that most of it is aligned (in sections), on 16-byte boundaries. The format is described at this link, but I haven't worked with that.

One way to open and read a Shapefile is to use geopandas. I grab that with `pip3 install geopandas`

Jackson's Mill WV

Search This Blog

Search

Labels

- 16S rRNA (10)
- alignments (7)
- bayes (17)
- binary (5)
- bindings (1)
- Bioconductor (7)
- bioinformatics (77)
- BLAST (8)
- book (1)
- C (8)
- calculus (14)
- command line (15)
- cool stuff (1)
- COVID-19 (17)
- crypto (10)
- ctypes (5)
- Cython (3)
- dental project (6)
- distributions (20)
- DNA binding sites (6)

The example is

```
>>> d = 'gz_2010_us_040_00_20m'
>>> fn = d + '/gz_2010_us_040_00_20m.shp'
>>> df = gpd.read_file(fn)
```

The columns are:

- GEO_ID
- STATE
- NAME
- LSAD
- CENSUSAREA
- geometry

```
>>> df.NAME.head()
0        Arizona
1       Arkansas
2     California
3       Colorado
4    Connecticut
Name: NAME, dtype: object
```

For some reason the order is several joined partial lists of states, each one alphabetized.

We need to extract the coordinates for a particular state:

```
import geopandas as gpd

fn = 'ex.shp'
df = gpd.read_file(fn)

sel = df['NAME'] == 'Maine'
g = me = df.loc[sel].geometry

from shapely.geometry import mapping
D = mapping(g)

for f in D['features']:
    print(f['id'])
    L = f['geometry']['coordinates']

    for m in L:
        print(len(m[0]))
        print(m[0][0])
```


- duly quoted (31)
- EMBOSS (3)
- fun (16)
- Geometry (26)
- go (4)
- HMM (6)
- homework (5)
- Illumina (12)
- Instant Cocoa (74)
- linear algebra (12)
- links (1)
- Linux (8)
- maps (5)
- matplotlib (38)
- matrix (7)
- maximum likelihood (5)
- meta (21)
- motif (11)
- Note to self (1)
- numpy (18)
- OS X (46)
- phy trees (32)
- phylogenetics (64)
- Pretty code (7)
- probability (8)
- puzzles (2)
- PyCogent (34)
- PyObjC (59)
- Python (2)
- Qiime (9)
- Quick Objective-C (15)
- Quick Python (4)
- Quick Unix (3)
- R (29)
- RPy2 (14)
- sequence models (11)
- simple math (68)
- simple Python (115)
- simulation (43)
- software installs (41)
- ssh (8)

```
        print(m[0][1])
        print()


> python3 script.py
39
11
(-69.307908, 43.773767)
(-69.30675099999999, 43.775095)
11
(-69.42792, 43.928798)
(-69.423323, 43.922871)
...
```

Posted by telliott99 at 6/25/2020 03:51:00 PM

Labels: maps

Newer Post Home Older Post

- stats (39)
- Sudoku (1)
- Swift (14)
- Ubuntu (8)
- Unifrac (8)
- Unix (7)
- What we're eating (2)
- what we're listening to (5)
- what we're reading (30)
- what we're saying (1)
- What we're thinking (2)
- Xcode (9)
- Xgrid (12)
- XML (9)

Unique visitors since Feb 21, 2011

Blog Archive

- May (19)
- April (4)
- March (1)

- 2018 (1)
- 2017 (7)
- 2016 (1)
- 2015 (16)
- 2014 (3)
- 2013 (2)
- 2012 (77)
- 2011 (174)
- 2010 (224)
- 2009 (265)
- 2008 (76)

About Me

telliott99

I'm retired, but used to teach and do research in Microbiology. This blog started as a record of my adventures learning bioinformatics and using Python. It has expanded to include Cocoa, R, simple math and assorted topics. As bbum says, it's so "google can organize my head." The programs here are developed on OS X using R and Python plus other software as noted. YMMV. I've had to turn comments off for the blog. Nothing but spam anymore. The intrepid reader will be able to find me. Hint: +"9" and I use gmail.

View my complete profile