Assignment #5 – Take-home Final

Due 20:00pm, Tuesday, Dec 19, 2017.

Problem 1.

There are three 512x512 pixel image files stored as floating point numbers, in csv format. They are images of Paris, in different wavelengths, from the near infrared (nir.csv), to red (red.csv) and green (grn.csv). The task is to use a PCA expansion of these images and apply various thresholds to the PCA components to identify

(a) Green areas corresponding to parks
(b) The river Seine

Hint: first create a 263144 long vector of each image, then build a 263144x3 matrix, then to the PCA. After this convert back into individual images in 512x512 dimensions, and apply various cuts, and visually inspect the images until satisfied. An example of the desired result is shown below:

Problem 2.

The file hough.csv contains 1560 points with x,y coordinates. Some of the points are on straight lines. Use the Hough transform to find the equations of these three lines. Overplot the lines on the scatter plot of the points.

Problem 3.

In the CasJobs tool (https://skyserver.sdss.org/CasJobs) select ‘IPDSDB’ as the database context. The database contains data from the famous Panama Papers and Paradise Papers websites. Additional hints can be found in the appendix of this PDF file.
Using SQL,

a) Find Rex Tillerson in the database. Information about him was obtained from the Paradise papers, hence use the tables with names starting with icij.PD_. Note, there is no guarantee that his name shows up exactly as 'Rex Tillerson'. Use SQL's LIKE clause to find approximate matches.

b) Find the company ("entity") of which Rex Tillerson is a shareholder of. Note, the Paradise papers database we downloaded does not yet give a specific rel_type for shareholder. So any relationship is ok.

c) Find all other officers of this company. You will note that there are more than in the graph below. Make therefore a further sub-selection of those whose name indicates a relation to Yemen.

In all cases, create queries that take the name of Rex Tillerson as input, or some version of that. I.e. for the 2nd and 3rd queries do not use results of the previous queries as input!
Hints to P3

Use a relational database with data obtained from the site https://offshoreleaks.icij.org/. This site publishes information obtained from various sets of leaked documents, showing relationships between more or less well-known people and companies located in various tax-sheltered localities. In this test we ONLY want to use this database as a target for testing your skills in writing SQL queries and not make any value judgements. We completely accept the disclaimer form the ICJ site itself\(^1\).

We have loaded the data underlying the graph database used on that site in the relational database for this class, and would like you to reproduce some of the results from the ICJ site by querying this database using SQL. We have loaded data from two sets of data, the “Panama Papers” and the more recently leaked “Paradise Papers”.

The database tables can be queried in the CasJobs tool (https://skyserver.sdss.org/CasJobs/). On the Query tab, select ‘IPDSDB’ as the context. To find out the tables you can query, on the schema browser tab, select the same database in the left-most drop-down, select ‘Tables’ in the middle drop down. The right drop down shows a couple of names starting with ‘icij.’ Those can be used in the queries. Select them to see the structure of these tables. The tables with name starting with PN_ are from the Panama papers data set, the data from the Paradise papers are stored in tables with name starting with PD_.

To get you going, we first investigate Sigmundur David Gunnlaugsson, the prime minister of Iceland. The ICJ web site, at https://offshoreleaks.icij.org/stories/sigmundur-david-gunnlaugsson shows links between him and several entities displayed as a graph, which is reproduced below. The data was obtained from the “Panama Papers”.

---

\(^1\) Disclaimer There are legitimate uses for offshore companies and trusts. We do not intend to suggest or imply that any people, companies or other entities included in the ICJ Offshore Leaks Database have broken the law or otherwise acted improperly. Many people and entities have the same or similar names. We suggest you confirm the identities of any individuals or entities located in the database based on addresses or other identifiable information. If you find an error in the database please get in touch with us.
So how would we first find the prime minister? From the graph we see he is represented as an “officer”. In our database these are stored in the table icij.PN_officers. To find the details about we could run the following query:

```sql
SELECT * FROM icij.PN_officers
WHERE name = 'Sigmundur David Gunnlaugsson'
```

Remember, to be able to run SQL queries you must import the SciServer.CasJobs library. The query can be run as follows, where the result, a pandas.DataFrame is stored in a variable. Note that the target database, ‘IPDSDB’ must be specified.

Now we would like to follow the link to the “entity” of which he is/was a shareholder. To do so we must follow a link, which in the panama papers part of our database is stored in the table icij.PN_edges. Since the link seems to go from the Officer to the Entity, we can try the following query:

```sql
SELECT ed,rel_type, en.*
FROM icij.PN_Officers o
    INNER JOIN icij.PN_Edges ed ON ed.node_1=o.node_id
    INNER JOIN icij.PN_Entities en ON en.node_id=ed.node_2
WHERE o.name = 'Sigmundur David Gunnlaugsson'
```

Indeed this query shows that the prime minister of Iceland is/was a shareholder (the value of the rel_type column) in the entity named ‘Wintris Inc.’, as shows in the graph.