Using Python for Record Linkage: Entrepreneurship, Research and Development, and Lobbying in the Unmanned Aerial Vehicle Industry

Russell J. Funk
funk@umich.edu

May 20, 2013

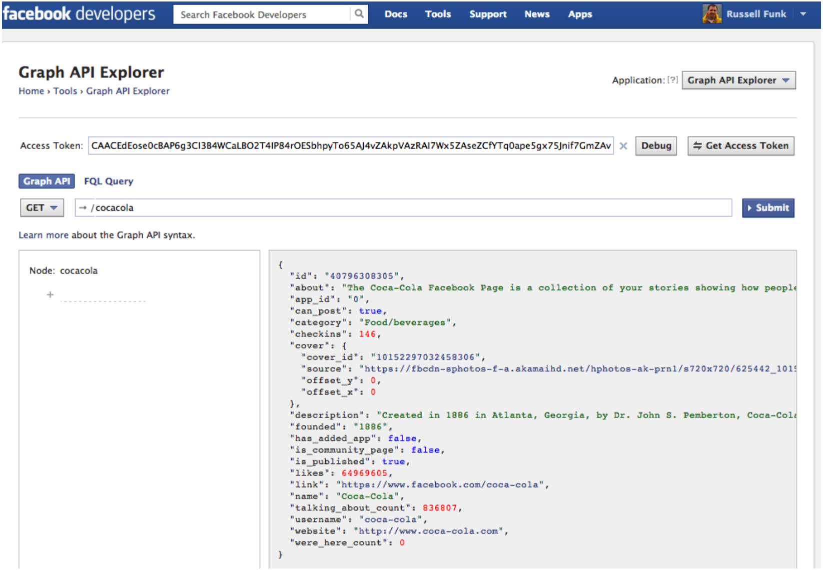
UNIVERSITY OF MICHIGAN

Session goals

We'll learn how to. . .

1. pull data from disparate online sources
2. link messy data with python using fuzzy matching
3. use python to build a data set ready for analysis

Motivation. . .



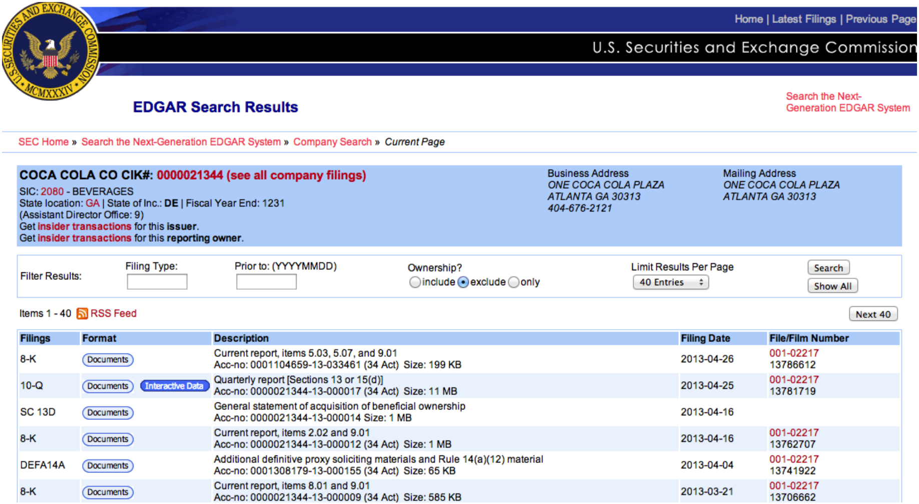
Access Token: CAACEdEose0cBAP6g3CI3B4WCaLBO2T4IP84rOESbhpyTo65AJ4vZAkpVAzRAI7Wx5ZAseZCfYTq0ape5gx75Jnif7GmZAv

→ /cocacola

Node: cocacola

```
{
  "id": "40796308305",
  "about": "The Coca-Cola Facebook Page is a collection of your stories showing how people
  "app_id": "0",
  "can_post": true,
  "category": "Food/beverages",
  "checkins": 146,
  "cover": {
    "cover_id": "10152297032458306",
    "source": "https://fbcdn-sphotos-f-a.akamaihd.net/hphotos-ak-prn1/s720x720/625442_1015
    "offset_y": 0,
    "offset_x": 0
  },
  "description": "Created in 1886 in Atlanta, Georgia, by Dr. John S. Pemberton, Coca-Col
  "founded": "1886",
  "has_added_app": false,
  "is_community_page": false,
  "is_published": true,
  "likes": 64969605,
  "link": "https://www.facebook.com/coca-cola",
  "name": "Coca-Cola",
  "talking_about_count": 836807,
  "username": "coca-cola",
  "website": "http://www.coca-cola.com",
  "were_here_count": 0
}
```

Motivation. . .


Home | Latest Filings | Previous Page
U.S. Securities and Exchange Commission
Search the Next-Generation EDGAR System
Search
Show All
Next 40

EDGAR Search Results

SEC Home » Search the Next-Generation EDGAR System » Company Search » *Current Page*

COCA COLA CO CIK#: 0000021344 (see all company filings)
SIC: 2080 - BEVERAGES
State location: GA | State of Inc.: **DE** | Fiscal Year End: 1231
(Assistant Director Office: 9)
Get **insider transactions** for this **issuer**.
Get **insider transactions** for this **reporting owner**.

Business Address
ONE COCA COLA PLAZA
ATLANTA GA 30313
404-676-2121

Mailing Address
ONE COCA COLA PLAZA
ATLANTA GA 30313

Filter Results: Filing Type: ______ Prior to: (YYYYMMDD) ______ Ownership? ○ include ● exclude ○ only Limit Results Per Page: 40 Entries

Items 1 - 40 RSS Feed

Filings	Format	Description	Filing Date	File/Film Number
8-K	Documents	Current report, items 5.03, 5.07, and 9.01 Acc-no: 0001104659-13-033461 (34 Act) Size: 199 KB	2013-04-26	001-02217 13786612
10-Q	Documents Interactive Data	Quarterly report [Sections 13 or 15(d)] Acc-no: 0000021344-13-000017 (34 Act) Size: 11 MB	2013-04-25	001-02217 13781719
SC 13D	Documents	General statement of acquisition of beneficial ownership Acc-no: 0000021344-13-000014 Size: 1 MB	2013-04-16	
8-K	Documents	Current report, items 2.02 and 9.01 Acc-no: 0000021344-13-000012 (34 Act) Size: 1 MB	2013-04-16	001-02217 13762707
DEFA14A	Documents	Additional definitive proxy soliciting materials and Rule 14(a)(12) material Acc-no: 0001308179-13-000155 (34 Act) Size: 65 KB	2013-04-04	001-02217 13741922
8-K	Documents	Current report, items 8.01 and 9.01 Acc-no: 0000021344-13-000009 (34 Act) Size: 585 KB	2013-03-21	001-02217 13706662

Why study the unmanned aerial vehicle industry?

A research question...

What is the correlation between lobbying expenditures and research and development contracts for small businesses from Department of Defense?

Quick background...

- **Small Business Innovation Research Program (SBIR)**—requires that all federal agencies with extramural research budgets in excess of $100 million reserve 2.5% for contracts or grants to small businesses
- **Small Business Technology Transfer Program (STTR)**—similar to SBIR, but smaller, and emphasizes funding partnerships between small businesses and nonprofit organizations

How can we find data?

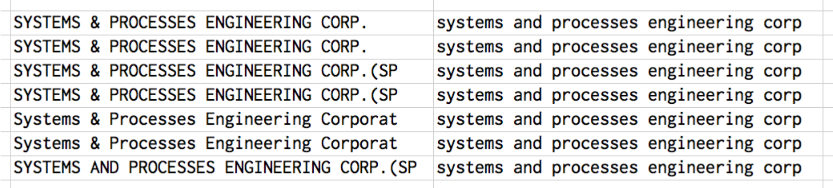
OpenSecrets.org
Center for Responsive Politics

DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA

Check **README.md** in the **/data** folder for instructions.

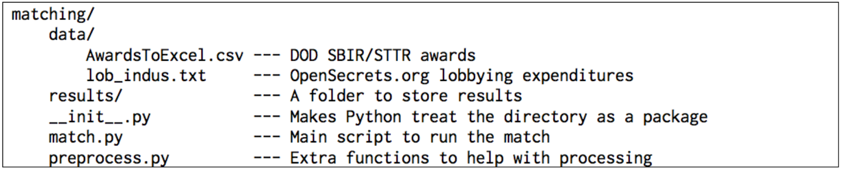
The challenge. . .

How can we link data records across sources without common unique identifiers?



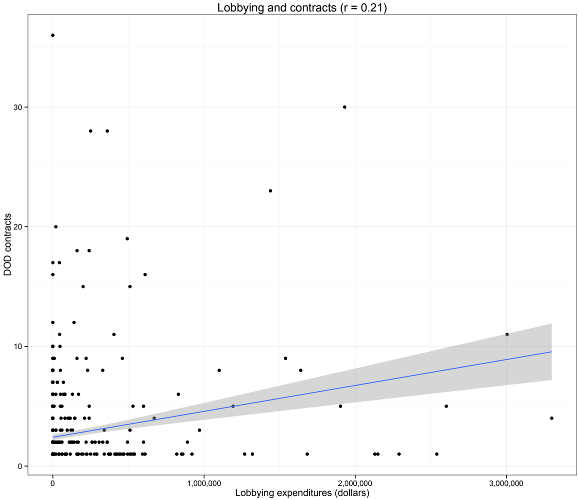
SYSTEMS & PROCESSES ENGINEERING CORP.	systems and processes engineering corp
SYSTEMS & PROCESSES ENGINEERING CORP.	systems and processes engineering corp
SYSTEMS & PROCESSES ENGINEERING CORP.(SP	systems and processes engineering corp
SYSTEMS & PROCESSES ENGINEERING CORP.(SP	systems and processes engineering corp
Systems & Processes Engineering Corporat	systems and processes engineering corp
Systems & Processes Engineering Corporat	systems and processes engineering corp
SYSTEMS AND PROCESSES ENGINEERING CORP.(SP	systems and processes engineering corp

Overview of the project directory. . .



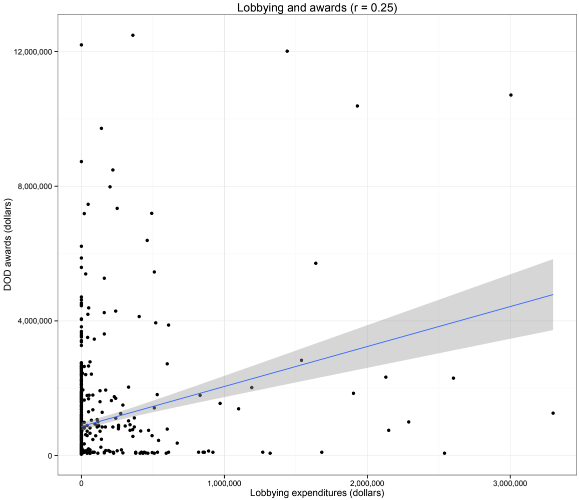
```
matching/
    data/
        AwardsToExcel.csv --- DOD SBIR/STTR awards
        lob_indus.txt     --- OpenSecrets.org lobbying expenditures
    results/              --- A folder to store results
    __init__.py           --- Makes Python treat the directory as a package
    match.py              --- Main script to run the match
    preprocess.py         --- Extra functions to help with processing
```

Finding (A)...but what does it mean?



UNIVERSITY OF MICHIGAN

Finding (B)…but what does it mean?



UNIVERSITY OF MICHIGAN