

Semantic Entity Resolution

Jeremiah Greer

Outline

- CERCIS
- Entity Resolution
- Semantic Contribution
- Google Search
- Neo4j Graph
- Truth Set Survey
- Additional Features/Future Work

CERCIS

Cincinnati Entity Resolution,
Combination, and Information
System

Identifying Companies

Resolving into single entities

Discrepancies/Fraud Detection

CERCIS

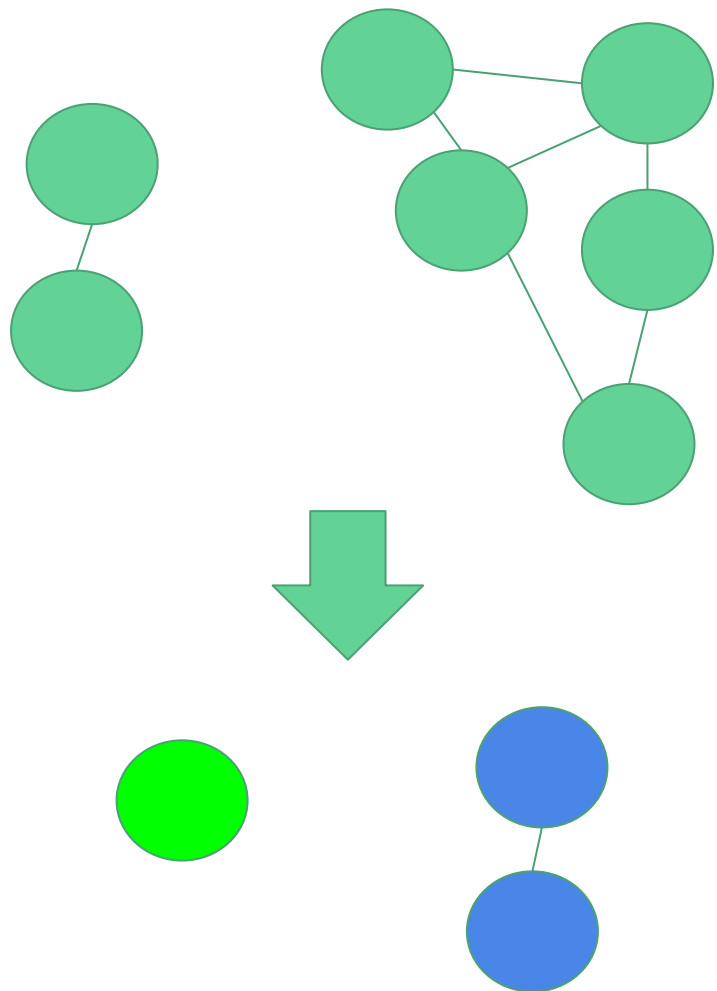
- World Bank “Major Contract Awards” Dataset
 - No address or distinct identifiers
 - Missing cells
 - Only consistent point is Supplier Name
- Company names are inconsistent
 - Input by individuals
 - Different representations of the same name
- Prevent giving loans to fraudulent companies
 - Purposeful misspelling of names
 - Name changes
 - Predict fraudulent behavior

```
53390 PRICEWATERHOUSECOOPER
53391 PRICEWATERHOUSE COOPER
53392 PRICEWATERHOUSE & COOPER
53393 PRICEWATERHOUSE COOPERES&PRONET INT.
53394 PRICEWATERHOUSE COOPER (LAO)
53395 PRICEWATERHOUSE COOPER LTD
53396 PRICEWATERHOUSECOOPERS
53397 PRICE WATERHOUSECOOPERS
53398 PRICE WATER HOUSE COOPERS
53399 PRICE WATERHOUSE COOPERS
53400 PRICE WATERHOUSE/COOPERS
53401 PRICEWATER HOUSE COOPERS
53402 PRICEWATERHOUSE COOPERS
53403 PRICEWATERHOUSE & COOPERS
53404 PRICEWATERHOUSECOOPERS AFRICA ASSOCIAT
53405 PRICEWATERHOUSECOOPERS AFRIQUE AND CECAF
53406 PRICEWATERHOUSE COOPERS AND BCEOM INDIA (P) LTD. [JV]
53407 PRICE WATERHOUSE COOPERS AND HOWARD HUMPHREYS
53408 PRICEWATERHOUSECOOPERS (ANGOLA), LDA
53409 PRICE WATERHOUSE COOPERS ANTIGUA
53410 PRICEWATERHOUSECOOPERS ASESORES
53411 PRICE WATERHOUSECOOPERS ASESORES GERENCIALES LTDA.
53412 PRICEWATERHOUSECOOPERS-ASSESSORIA DE G
53413 PRICEWATERHOUSECOOPERS ASSOCIATES AFRICA LIMITED
53414 PRICE WATERHOUSE COOPERS ASSOCIATES AFRICA LTD.
53415 PRICEWATERHOUSECOOPERS ASSOCIATES AFRICA LTD
53416 PRICEWATERHOUSECOOPERS ASSOCIATES AFRICA LTD.
53417 PRICEWATERHOUSECOOPERS AUDIT SRL
53418 PRICEWATERHOUSECOOPERS (AUSTRALIA)
53419 PRICEWATERHOUSECOOPERS BULGARIA FOOD
53420 PRICEWATERHOUSECOOPERS BUSINESS RECOVER
53421 PRICEWATERHOUSE COOPERS/CABRAM
```

Entity Resolution

Resolving Entities

- Company names are the consistent minimum
- Syntactic and Semantic Resolution
- Build relationships between entities
- Resolve entities to their single identities
- Results in unique, non-duplicate entities for reference



Semantic Contribution

Syntactic concerns the physical structure of the name

Semantic attempts to find what the name means/represents intuitively

Semantic Contribution

- Syntactic attempts to find similarity between names, but sometimes that can fail
 - Complete name changes: Valujet, Airtran => Southwest
 - Alternate representations/abbreviations: SMEC, Snowy Mountain Engineering Corporation
 - Individual people's names
- Semantic finds missing connections by meaning of terms
 - IBM and International Business Machines
- Variety of methods/features available

Semantic Contribution

International Business Machines

IBM

IBM - United States

www.ibm.com/ - IBM -
The IBM corporate home page, entry point to information about IBM products and services.
Join Us - IBM Support Portal - Jobs at IBM - Select a country/region

IBM

www.ibm.com
Google+ page

IBM

www.ibm.com
Google+ page

IBM

www.ibm.com
Google+ page

Map results for international business machines

- A Ellipse Lobby Shop, 4350 Fairfax Drive, Arlington, VA (571) 312-8140
- B 3050 Chain Bridge Road, Fairfax, VA (703) 393-2400
- C 1430 Spring Hill Road, McLean, VA (703) 921-0195



Map for international business machines

IBM - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/IBM - Wikipedia -
The International Business Machines Corporation (IBM) is an American multinational technology and consulting corporation, with headquarters in Armonk, New York ...
Ginni Rometty - History of IBM - List of IBM products - Thomas J. Watson

In the news



International Business Machines Corp. (IBM) To Layoff Over 100000 Employees

Bidness ETC - 9 hours ago
It has been reported that International Business Machines Corp. (NYSE:IBM) will be laying ...

Don't Panic Over International Business Machines Corp.'s Earnings (IBM)
The Motley Fool - 3 days ago

IBM

Computer hardware company

The International Business Machines Corporation is an American multinational technology and consulting corporation, with headquarters in Armonk, New York, United States. Wikipedia

Stock price: IBM (NYSE) \$155.87 +0.48 (+0.31%)
Jan 23, 4:03 PM EST - Disclaimer

CEO: Ginni Rometty

Customer service: 1 (877) 426-6006

Sales: 1 (888) 746-7426

Headquarters: Armonk, NY

Founded: June 16, 1911, Endicott, NY

Founders: Charles Ranlett Flint, Thomas J. Watson



About 1,130,000,000 results (0.51 seconds)

In the news



IBM job cuts could come soon

Poughkeepsie Journal - 1 day ago
IBM's executives have confirmed that more restructuring is planned this year. Now a ...

International Business Machines Corp. (IBM) To Layoff Over 100000 Employees

Bidness ETC - 9 hours ago

Here's Why IBM Is Still Building Mainframes

Motley Fool - 1 day ago

More news for ibm

IBM - United States

www.ibm.com/ - IBM -
The IBM corporate home page, entry point to information about IBM products and services.

Google+ page - Be the first to review

6710 Rockledge Drive, Bethesda, MD 20817
(800) 426-4968

Join Us - IBM Support Portal - Jobs at IBM

IBM - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/IBM - Wikipedia -
The International Business Machines Corporation (IBM) is an American multinational technology and consulting corporation, with headquarters in Armonk, New York ...
Ginni Rometty - History of IBM - List of IBM products - Thomas J. Watson

IBM - Yahoo Finance

finance.yahoo.com/q?s=IBM - Yahoo! Finance -
2 days ago - View the basic IBM stock chart on Yahoo! Finance. Change the date range, chart type and compare International Business Machines against ...



IBM

Software Company

Address: 6710 Rockledge Drive, Bethesda, MD 20817

Phone: (800) 426-4968

Reviews

Be the first to review

Are you the business owner?

Feedback

See results about

IBM (Computer hardware company)
Stock price: IBM (NYSE) \$155.87 +0.48 (+0.31%)
Jan 23, 4:03 PM EST



Semantic Contribution

- Valujet, AirTran, Southwest
- No Google Search Connection
- No Syntactic Connection
- Wikipedia Connection

AirTran Airways

From Wikipedia, the free encyclopedia

Not to be confused with Air Transat.

AirTran Airways was an American low-cost airline headquartered originally in Orlando, Florida then in Dallas, Texas after its acquisition by Southwest Airlines, with which it was integrated into. AirTran operated nearly 700 daily flights, primarily in the eastern and midwestern United States, with its principal hub at Hartsfield-Jackson Atlanta International Airport where it operated nearly 200 daily departures. AirTran's fleet consisted of Boeing 717 aircraft, of which it was the largest operator, and Boeing 737-700 aircraft. It was fully integrated into Southwest Airlines on December 28, 2014.

Contents [hide]
1 History
1.1 Foundation & early years
1.2 2000s
1.2.1 Failed acquisitions
1.3 2010s
1.3.1 Buyout and wind-down
2 Corporate affairs
2.1 Employee relations
3 Destinations
3.1 Top served cities
3.2 Codeshare agreements
4 Fleet
4.1 Retired
5 Cabin
6 Livery
7 Incidents and accidents
8 References
9 External links

History [edit]

Foundation & early years [edit]

See also: *Valujet Airlines*

The original *AirTran Airways*, a Boeing 737 operator with service to/from Orlando, was founded by AirTran Corporation, the holding company of Mesaba Airlines of Minneapolis, Minnesota, operating as a northwest Arctic carrier with hubs in Minneapolis and Detroit. In 1994, AirTran Holdings purchased a start-up 737 operator named **Conquest Sun** and renamed the airline AirTran Airways. Conquest Sun similar to Valujet, was an airline started by former *Eastern Air Lines* employees. The original AirTran Airways moved its headquarters to Orlando, Florida, and grew to 11 Boeing 737 aircraft serving 24 cities

ValuJet Airlines

From Wikipedia, the free encyclopedia



This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unourced material may be challenged and removed. (November 2007)

ValuJet Airlines was an American low-cost carrier, headquartered in unincorporated Clayton County, Georgia, that operated regulary scheduled domestic and international flights in the Eastern United States and Canada during the 1990s. The company was founded in 1992 and was known for its cost-cutting measures. All of the airline's planes were purchased used from other airlines, very little training was provided to workers, and contractors were used for maintenance and other services. The company quickly developed a reputation for its lax safety in 1996; the military refused ValuJet's bid to fly military personnel over safety worries, and officials at the FAA wanted the airline to be grounded.

The May 1996 crash of Flight 592, caused by illegally stored hazardous materials on board, spelled doom for the airline. ValuJet was grounded the next month and not allowed to fly again until September, with a greatly reduced fleet. The airline's major customers never returned and the company suffered major losses. In 1997, the company merged with the much smaller regional airline AirWays Corp., holding company for AirTran Airways. Although ValuJet was the nominal survivor, the merged airline operated as AirTran until its merger with Southwest Airlines in 2011. AirTran never acknowledged its past involvement with ValuJet and kept the airline's memorabilia locked in a warehouse in Atlanta.

Contents [hide]
1 History
1.1 Inception
1.2 Safety problems
1.3 Fallout from the crash of Flight 592
2 Fleet
3 Destinations
4 Incidents and accidents
4.1 Flight 597
4.2 Flight 592
5 References
6 External links

History [edit]

Inception [edit]

ValuJet was founded in 1992 and began operations on October 26, 1993. It originally offered service from Atlanta to Orlando, Jacksonville and Tampa with a single Douglas DC-9 that previously belonged to Delta Air Lines. The first flight, Flight 901, flew from Atlanta to Tampa. The carrier was headed by a group of industry veterans including co-founder and chairman Robert Priddy, who had started a string of successful airlines including Atlantic Southeast Airlines (ASA), Air Midwest Airlines, and Florida Gulf Airlines. Board members Maury Gallagher and Tim Flynn, the other co-founders, developed and ran

AirTran Airways



IATA FL	KCAO TNS	Callign CITRUS
Founded	1992 as ValuJet Airlines Georgia, USA ^[1]	
Commenced operations	October 26, 1993 as ValuJet Airlines ^[1] November 17, 1997 as AirTran Airways	
Ceased operations	December 28, 2014 (integrated into Southwest Airlines)	
Hubs	<div> <ul style="list-style-type: none">Baltimore–Washington International Thurgood Marshall AirportGeneral Mitchell International Airport (Iowa)Hartsfield-Jackson Atlanta International AirportOrlando International Airport </div>	
Frequent-flyer program	A+ Rewards	
Fleet size	138	
Destinations	69	
Company slogan	Go. There's nothing stopping you.	
Parent company	Southwest Airlines Co. NYSE: LUV ^[6]	
Headquarters	Dallas, Texas, U.S.	
Key people	<div> <ul style="list-style-type: none">Gary C. Kelly (Chairman & CEO)Bob Jordan (President)Bob Fomoro (Former Chairman, President and CEO) </div>	
Website	airtran.com ^[6]	

ValuJet Airlines



IATA ZJ	KCAO VAN	Callign CITRER
Founded	1992 Georgia, USA ^[1]	
Commenced operations	October 26, 1993 ^[1]	
Ceased operations	November 17, 1997 ^[1] (AirTran Airways)	
Hubs	<div> <ul style="list-style-type: none">Hartsfield-Jackson Atlanta International AirportLogan International Airport (Iowa)Miami International AirportOrlando International AirportPhiladelphia International AirportWashington Dulles International Airport </div>	
Fleet size	56	
Destinations	29	
Headquarters	Clayton County, Georgia, USA	
Key people	<div> <ul style="list-style-type: none">Robert Priddy (Chairman)Maurice J. Gallagher, Jr. (CEO & President) </div>	
Website	valujet.com ^[6]	

Google Search

Basis

- Multiple representations (names) of same entity
- Google provides results “most relevant” to user
- Page rank, search selection, etc.
- Searching IBM and International Business Machines returns largely same results
- First results are considered most relevant

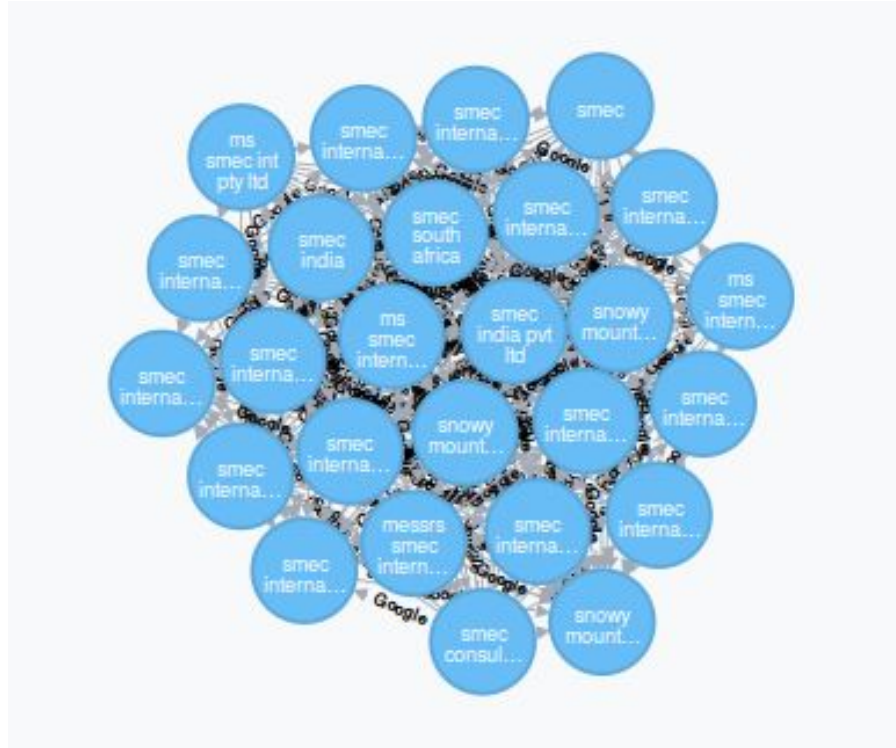
Implementation

1. Entity names are normalized and duplicates are removed
2. Search using normalized name on Google, scrape first 10 result links
 - a. Want to keep only the most pertinent links, few go beyond 1st page
3. Store result links, get number shared between 2 entities (intersection)
4. Add connection to Neo4j Graph
5. Query

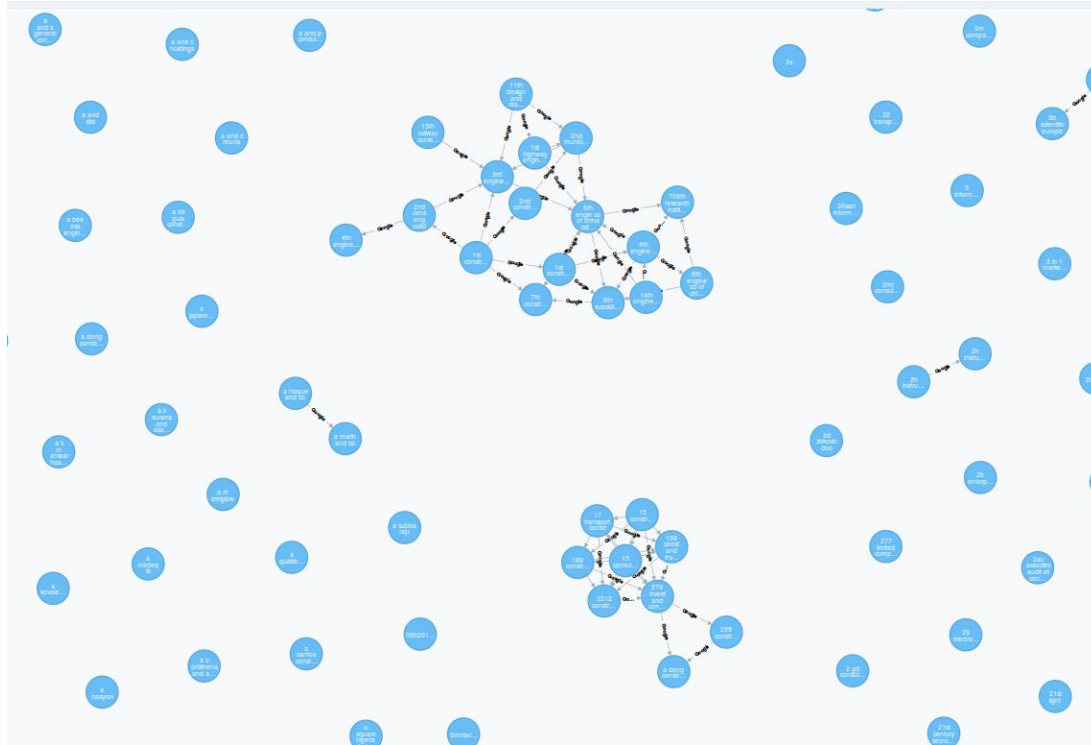
14th engineering bureau of china railway	5	china railway 16th bureau group 1st engineering coltd
14th engineering bureau of china railway	4	china railway 13th engineering bureau group co ltd
14th engineering bureau of china railway	4	china railway 20th bureau co ltd
14th engineering bureau of china railway	4	china railway 4th bureau group corporation
14th engineering bureau of china railway	4	china railway bureau group co ltd
14th engineering bureau of china railway	4	china railway bureau no 18 cor ltd group
14th engineering bureau of china railway	4	china railway construction 14 bureau
14th engineering bureau of china railway	4	china railway sixteen bureau group co
14th engineering bureau of china railway	4	the 17th engineering bureau of china railway the 1st engine
14th engineering bureau of china railway	3	china railway 10th bureau group coltd
14th engineering bureau of china railway	3	china railway 12th bureau group co ltd
14th engineering bureau of china railway	3	china railway 15 bureeau group coporation
14th engineering bureau of china railway	3	china railway 19th bureau group corp
14th engineering bureau of china railway	3	china railway 19th bureau group corp ltd
14th engineering bureau of china railway	3	china railway 20th bureau group co ltd
14th engineering bureau of china railway	3	china railway 25th engineering bureau group coltd
14th engineering bureau of china railway	3	china railway bureau for large bridge corp ltd group

Neo4j Graph

SMEC, r.score \geq 2, limit 100



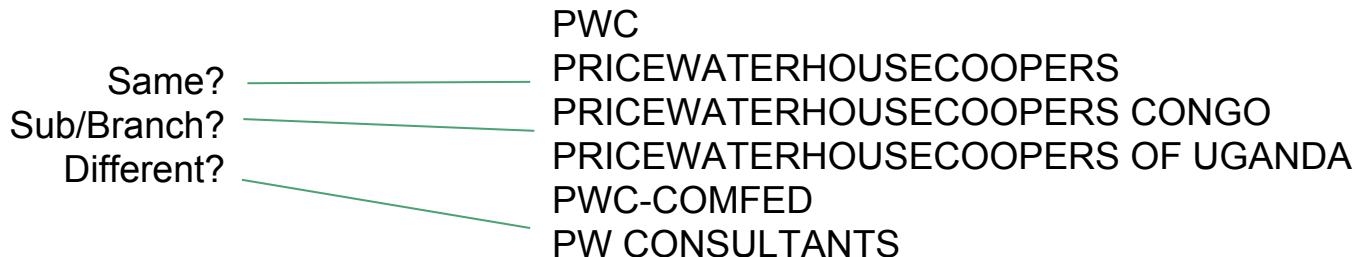
General Graph, limit 200



Truth Set/Survey

Problem Areas

- Currently have no knowledge of “correct” clusterings
- Companies believed to have three possible relationships:
 - Same
 - Subsidiary/Branch
 - Different
- Need a “Truth Set” to determine accuracy



Survey

- Relationship between two entities
- Confidence level
- Mixture of random and chosen
- Currently using SurveyMonkey
- Dynamic survey development

World Bank Company Survey

+ Add Page Title

Below are pairs of company names, separated by "||". Please decide which relationship these two companies share. The options are:

1. Same Company: If the two company names refer to the same entity.
2. Subsidiary/Branch: If one of the companies is a subsidiary or branch of the other.
3. Different Companies: If the companies share no relation in identity.

In addition, please select how confident you are of your answer.

* 1. **A 25.0%** 1-5 CORPORATE COMPANY || 1-5 JOINT STOCK COMPANY
B 25.0% AA KOMERC || AA KOMERC, D.O.O. PALE
C 25.0% CLIFFORD CHANCE PUNDER CIS LTD || CLIFFORD CHANCE SELAFA
D 25.0% OLYMPIA || OLYMPIA DE MEXICO

Same Company
 Subsidiary/Branch
 Different Companies

* 2. How confident are you in your answer?

Not Confident Somewhat Confident Confident

Additional Features

Top-Level Domains

- Each of the 10 urls saved per entity has a TLD
- Potential additional identifier (ibm.com, pwc.com, etc.)
- Problem:
 - Important (ibm.com)
 - Non-Important (wikipedia.org)

```
4015 alibaba.com 1212
4016 ftaarea.com 1428
4017 justdial.com 1455
4018 bancomundial.org 1460
4019 wordpress.com 1467
4020 banquemondiale.org 1721
4021 yelp.com 1928
4022 kompass.com 1971
4023 manta.com 2164
4024 hoovers.com 2309
4025 viadeo.com 2448
4026 developmentaid.org 2537
4027 blogspot.com 2629
4028 twitter.com 5538
4029 devex.com 5869
4030 youtube.com 8605
4031 bloomberg.com 10167
4032 wikipedia.org 19513
4033 worldbank.org 34782
4034 linkedin.com 36702
4035 facebook.com 39588

findthecompany.com.mx 54.7945205479 146
ftaarea.com 58.7096774194 155
justdial.com 63.4146341463 164
bancomundial.org 67.8362573099 171
wordpress.com 68.6390532544 169
yelp.com 72.192513369 187
banquemondiale.org 74.1573033708 178
manta.com 75.6613756614 189
kompass.com 78.1420765027 183
viadeo.com 79.3650793651 189
hoovers.com 83.2432432432 185
blogspot.com 86.8421052632 190
developmentaid.org 88.1443298969 194
devex.com 99.0 200
twitter.com 99.4974874372 199
facebook.com 100.0 200
youtube.com 100.0 200
bloomberg.com 100.0 200
linkedin.com 100.0 200
worldbank.org 100.0 200
wikipedia.org 100.0 200
```

Wikipedia Scrape

- Find Wikipedia page
- Scrape links and terms
- Build connections
- Problems:
 - Wikipedia Search
 - Wikipedia Urls

```
ningbo municipal facility construction and dev https://en.wikipedia.org/wiki/Ningbo
ov[]
posco a and c co ltd https://en.wikipedia.org/wiki/POSCO
[u'POSCO', u'Hyundai Steel', u'Samsung', u'Hyundai Glovis', u'S&T Motors', u'Hyundai Engineering & Con
struction', u'Steel', u'KCC Corporation', u'Hyundai Hysco', u'Kumho Tire']
etablissement kabys https://fr.wikipedia.org/wiki/K-bis
[]
programa de las naciones unidas para el desarrollo pnud https://es.wikipedia.org/wiki/Programa_de_Las_
Naciones_Unidas_para_el_Desarrollo
[u'Northwestern Mexico', u'Santa B\xe1rbara (canton)', u'Nicaragua']
kirloskar co ltd https://en.wikipedia.org/wiki/Kirloskar_Group#References
[u'Pune', u'Dharwad district', u'Kirloskar Group', u'Sichuan FAW Toyota Motor', u'The Times Group', u'
Daihatsu', u'List of Toyota manufacturing facilities', u'Future Group', u'SEMT Pielstick', u'Sangli di
strict']
bosiljka vucic https://sr.wikipedia.org/sr-el/%D0%92%D1%83%D1%87%D0%B8%D1%9B-%D0%9F%D0%B5%D1%80%D0%B8
%D1%88%D0%B8%D1%9B%D0%B8
[]
beijing space machinery and electricity engineering and techn https://en.wikipedia.org/wiki/North_Ch
ina_Electric_Power_University
[]
webber david https://en.wikipedia.org/wiki/Like_a_Mighty_Army
[u'Red Bull RB3', u'Phil Brewer', u'Turner Sports', u'How Do You Solve a Problem like Maria?', u'Steve
Hardy', u'2006 ARCA Re/Max Series season', u'The Woman in White (musical)', u'Justus Ward', u'Jessie
Brewer', u'Mary Mae Ward']
cyprus https://en.wikipedia.org/wiki/History_of_Cyprus
[u'Cyprus', u'Districts of Cyprus', u'Foreign relations of Cyprus', u'Cyprus\u2013United Kingdom relat
ions', u'Districts of Northern Cyprus', u'Cyprus\u2013United States relations', u'Coat of arms of Cypr
us', u'President of Cyprus', u'Outline of Cyprus', u'List of airports in Northern Cyprus']
ms cooper motors corporation u ltd https://en.wikipedia.org/wiki/Cooper_Motor_Corporation
[u'Kyocera', u'Mazda', u'GS Yuasa', u'Massachusetts Institute of Technology', u'Wright brothers', u'Mo
rris Minor', u'List of companies named after people', u'1973 Birthday Honours', u'1997 New Year Honour
s', u'1951 Birthday Honours']
jeremiah@jeremiah-desktop:~/CERCIS/src/semantic/working$
```