Information Extraction and Aggregation from Unstructured Web Data for Business Profiling

Student Team : Alexander Michels, Himanshu Ahuja

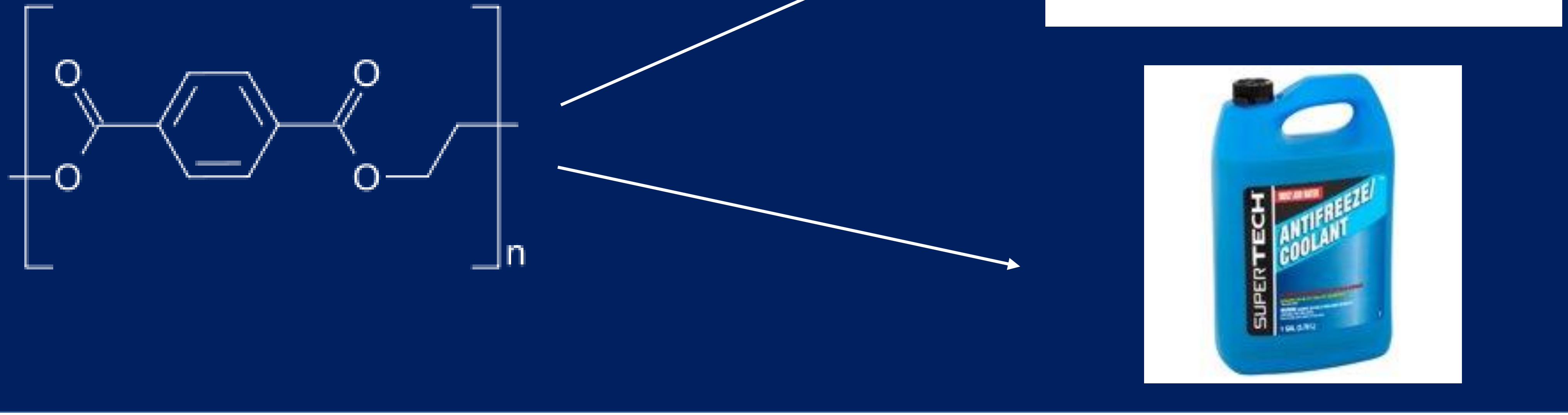
Industry Mentors : Dr. Stephen DeSalvo, Urjit Patel

Praedicat: An Insurance Tech Company

Determining Risk

Ethylene glycol

- moderately toxic
- generally used in anti-freeze

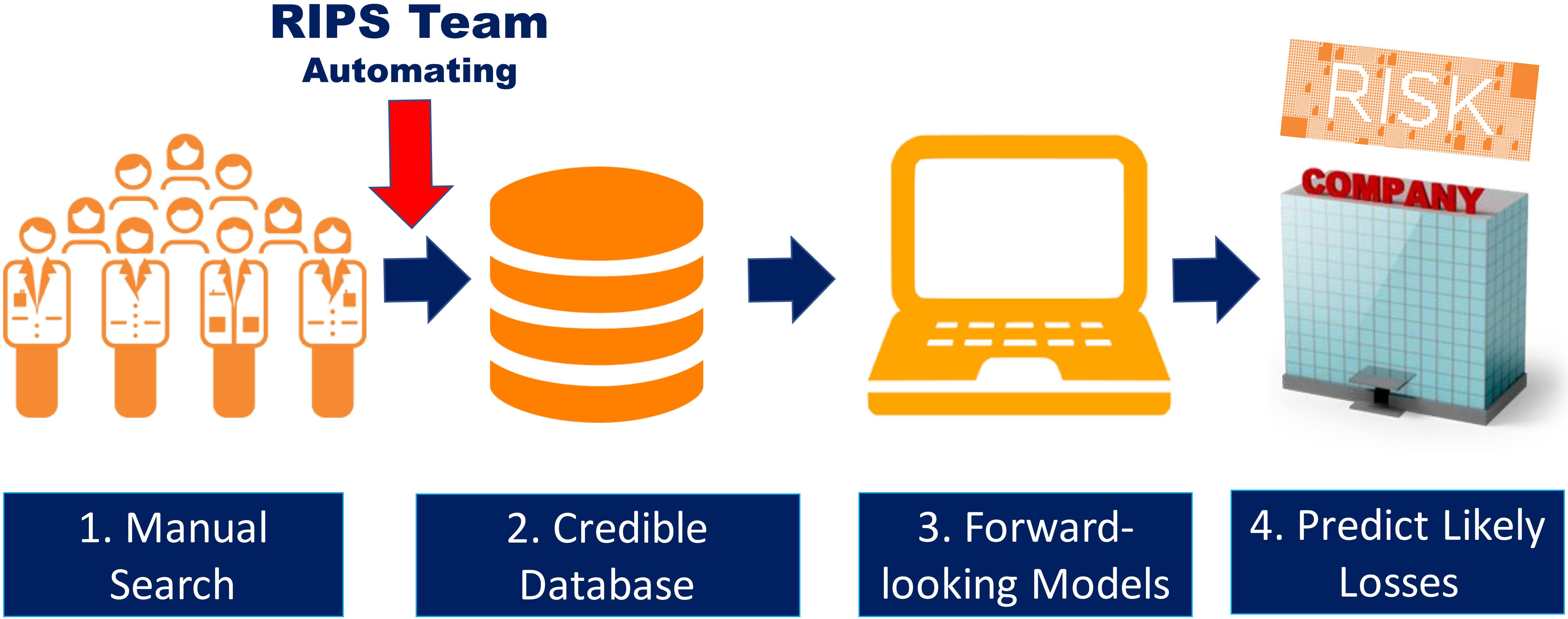


[1] https://iaspub.epa.gov/triexplorer/release_fac_profile?TRI=38063MRCNG1236A&TRILIB=TRIQ1&FLD=&FLD=RELLBY&FLD=TSFDSP&OFFDISPD=&OTHDISPD=&ONDISPD=&OTHOFFD=&YEAR=2016

Praedicat: An Insurance Tech Company

Predicting the likely amount of losses

Where Do We Fit in?



RIPS Team

Automating

1. Manual Search

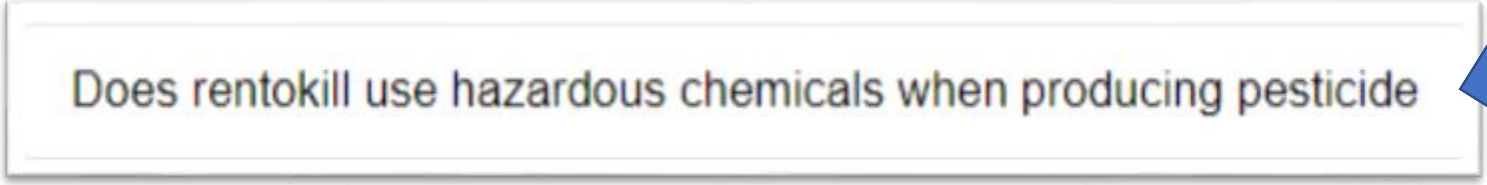
2. Credible Database

3. Forward-looking Models

4. Predict Likely Losses

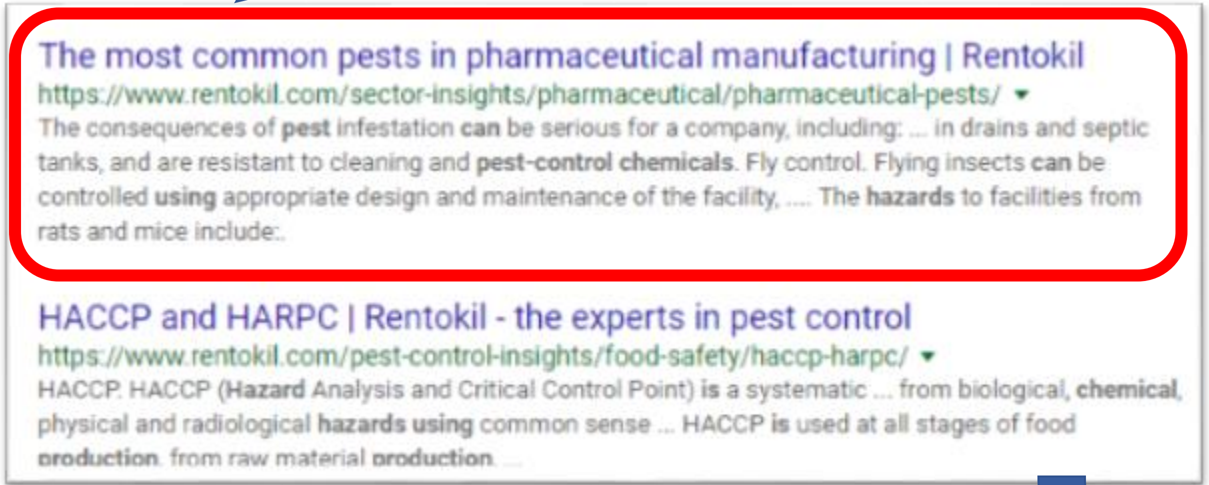
Difficulty of Searching Information

1. Search Engine



Does rentokill use hazardous chemicals when producing pesticide

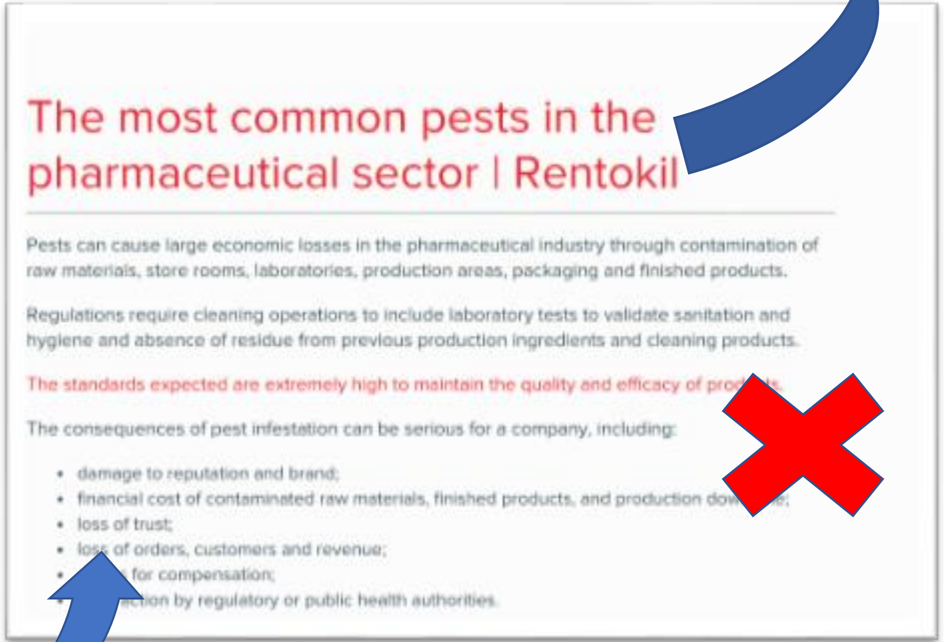
Change Keywords



The most common pests in pharmaceutical manufacturing | Rentokil
https://www.rentokil.com/sector-insights/pharmaceutical/pharmaceutical-pests/ ▾
The consequences of **pest** infestation **can** be serious for a company, including: ... in drains and septic tanks, and are resistant to cleaning and **pest-control chemicals**. Fly control. Flying insects **can** be controlled **using** appropriate design and maintenance of the facility, ... The **hazards** to facilities from rats and mice include:.

HACCP and HARPC | Rentokil - the experts in pest control
https://www.rentokil.com/pest-control-insights/food-safety/haccp-harpc/ ▾
HACCP. HACCP (**Hazard** Analysis and Critical Control Point) **is** a systematic ... from biological, **chemical**, physical and radiological **hazards using** common sense ... HACCP **is** used at all stages of food **production**, from raw material **production** ...

2. Site Search



The most common pests in the pharmaceutical sector | Rentokil

Pests can cause large economic losses in the pharmaceutical industry through contamination of raw materials, store rooms, laboratories, production areas, packaging and finished products.

Regulations require cleaning operations to include laboratory tests to validate sanitation and hygiene and absence of residue from previous production ingredients and cleaning products.

The standards expected are extremely high to maintain the quality and efficacy of products.

The consequences of pest infestation can be serious for a company, including:

- damage to reputation and brand;
- financial cost of contaminated raw materials, finished products, and production downtime;
- loss of trust;
- loss of orders, customers and revenue;
- claims for compensation;
- action by regulatory or public health authorities.

3. Evaluate Contents

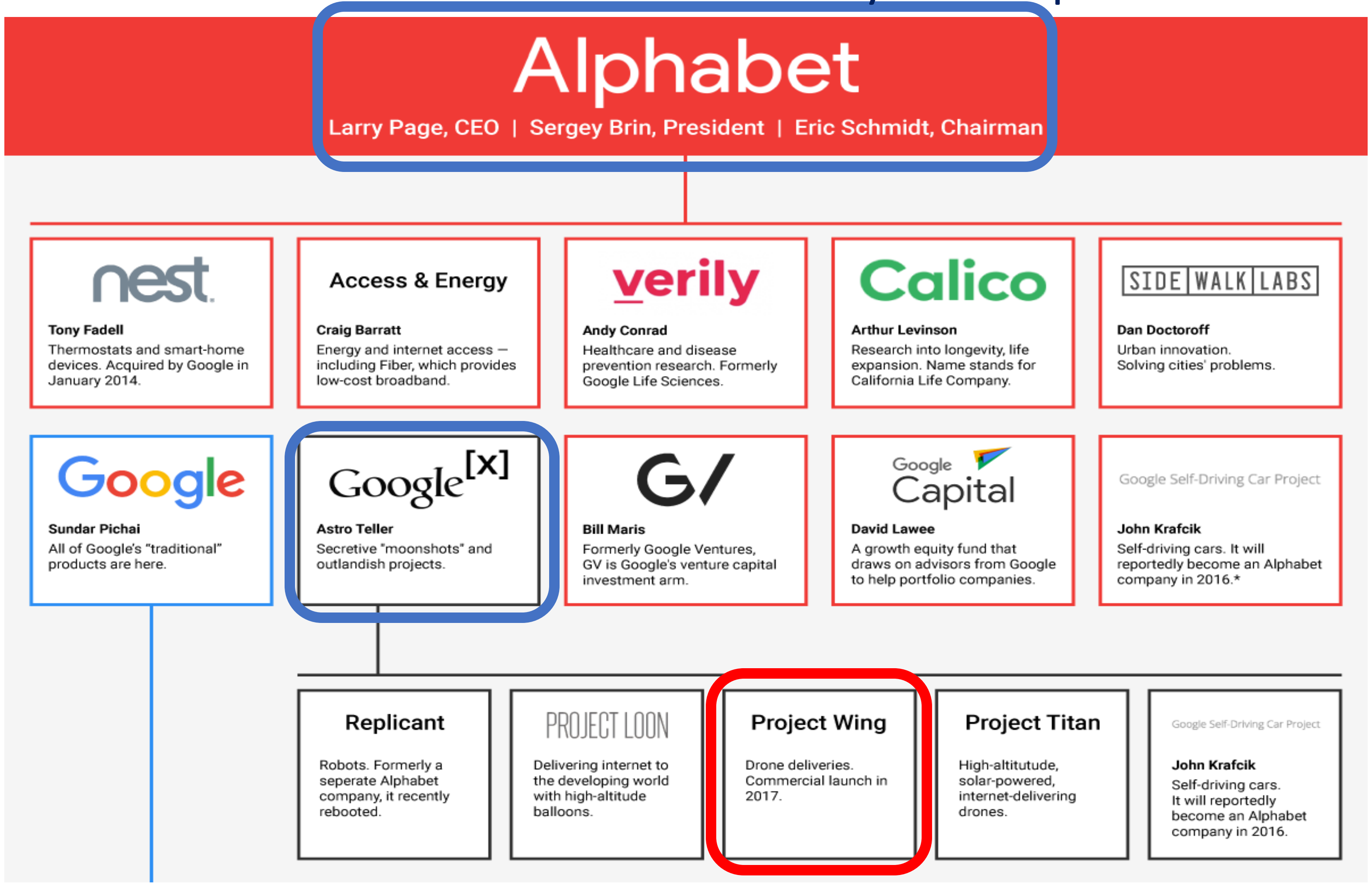
Difficulty of Allocating Risk

Example: Allocating Risk for Parent and Subsidiary Companies

Parent Company

Parent Company

Subsidiary Company



Alphabet

Larry Page, CEO | Sergey Brin, President | Eric Schmidt, Chairman

nest
Tony Fadell
Thermostats and smart-home devices. Acquired by Google in January 2014.

Access & Energy
Craig Barratt
Energy and internet access — including Fiber, which provides low-cost broadband.

verily
Andy Conrad
Healthcare and disease prevention research. Formerly Google Life Sciences.

Calico
Arthur Levinson
Research into longevity, life expansion. Name stands for California Life Company.

SIDEWALK LABS
Dan Doctoroff
Urban innovation. Solving cities' problems.

Google
Sundar Pichai
All of Google's "traditional" products are here.

Google[x]
Astro Teller
Secretive "moonshots" and outlandish projects.

GV
Bill Maris
Formerly Google Ventures, GV is Google's venture capital investment arm.

Google Capital
David Lawee
A growth equity fund that draws on advisors from Google to help portfolio companies.

Google Self-Driving Car Project
John Krafcik
Self-driving cars. It will reportedly become an Alphabet company in 2016.*

Replicant
Robots. Formerly a seperate Alphabet company, it recently rebooted.

PROJECT LOON
Delivering internet to the developing world with high-altitude balloons.

Project Wing
Drone deliveries. Commercial launch in 2017.

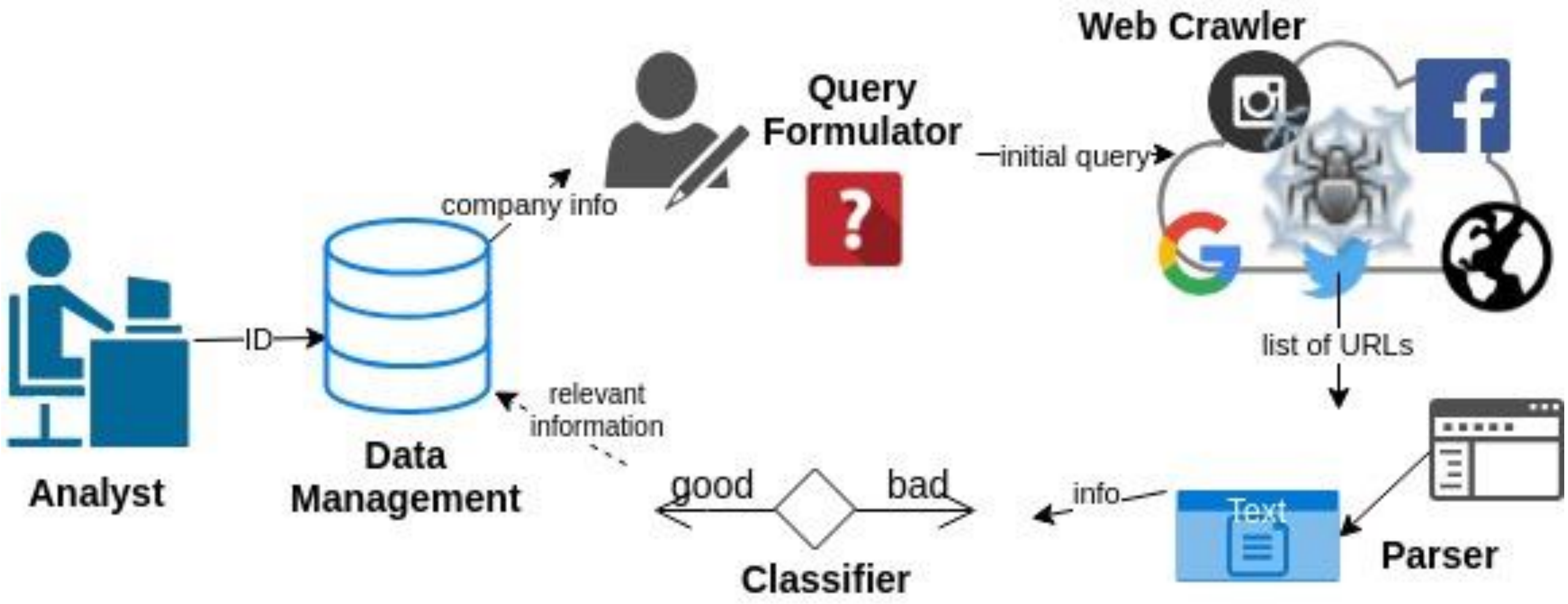
Project Titan
High-altitutude, solar-powered, internet-delivering drones.

Google Self-Driving Car Project
John Krafcik
Self-driving cars. It will reportedly become an Alphabet company in 2016.

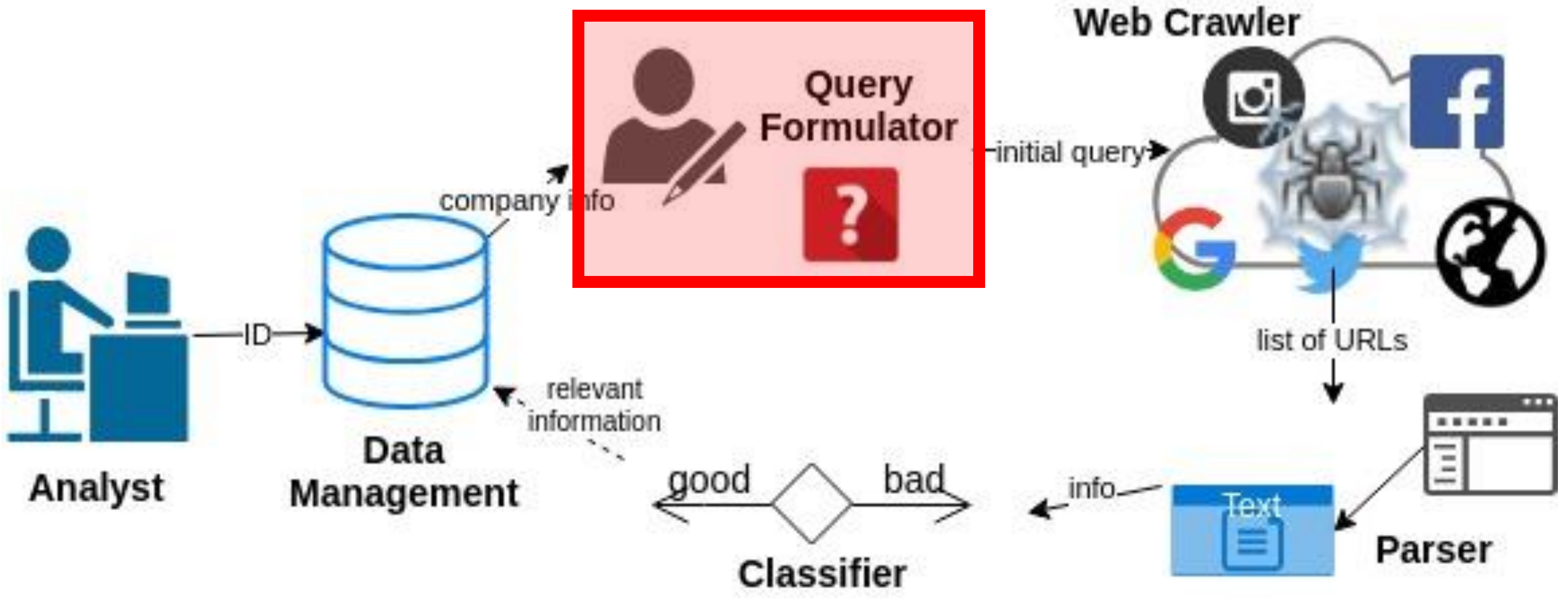
Problem Statement

How to automate information extraction and aggregation from unstructured data on the Internet for business profiling

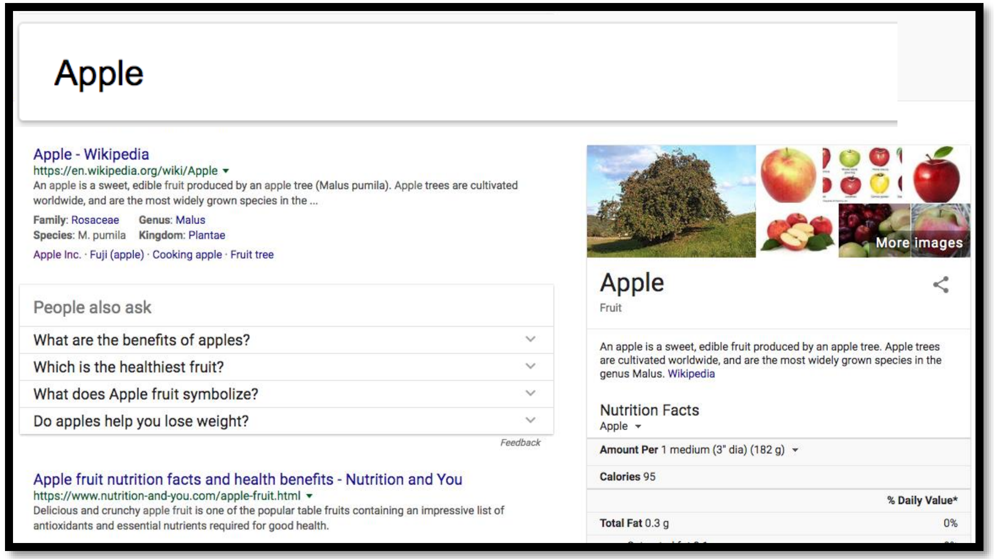
Solution Overview



Solution Overview



Query Formulator: Asking about the right things!



Apple

Apple - Wikipedia
https://en.wikipedia.org/wiki/Apple ▾
An apple is a sweet, edible fruit produced by an apple tree (Malus pumila). Apple trees are cultivated worldwide, and are the most widely grown species in the ...
Family: Rosaceae Genus: Malus
Species: M. pumila Kingdom: Plantae
Apple Inc. · Fuji (apple) · Cooking apple · Fruit tree

People also ask
What are the benefits of apples?
Which is the healthiest fruit?
What does Apple fruit symbolize?
Do apples help you lose weight?
Feedback

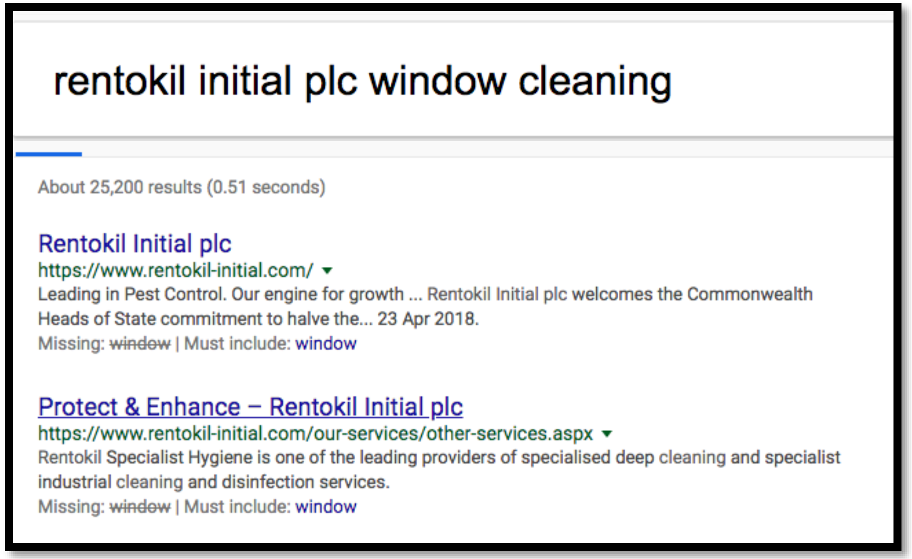
Apple fruit nutrition facts and health benefits - Nutrition and You
https://www.nutrition-and-you.com/apple-fruit.html ▾
Delicious and crunchy apple fruit is one of the popular table fruits containing an impressive list of antioxidants and essential nutrients required for good health.

Apple
Fruit

An apple is a sweet, edible fruit produced by an apple tree. Apple trees are cultivated worldwide, and are the most widely grown species in the genus Malus. Wikipedia

Nutrition Facts
Apple ▾
Amount Per 1 medium (3" dia) (182 g) ▾
Calories 95
% Daily Value*
Total Fat 0.3 g 0%

Zero useful results

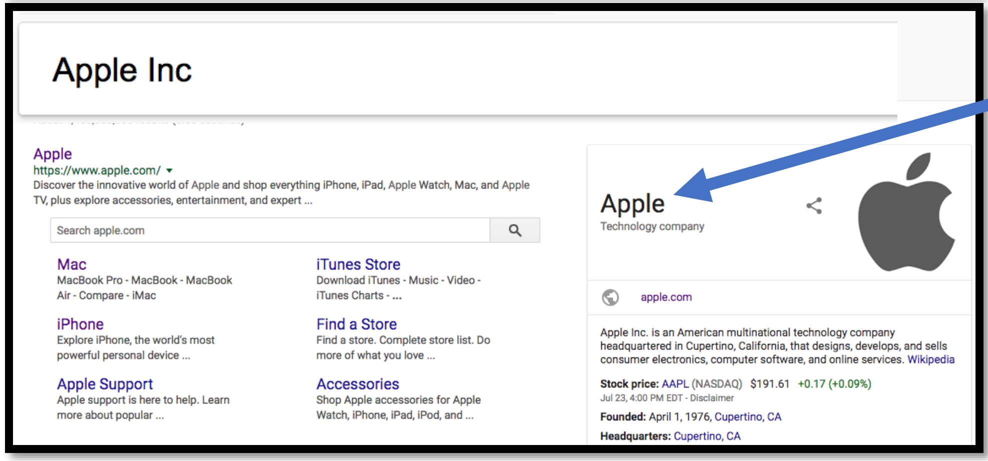


rentokil initial plc window cleaning

About 25,200 results (0.51 seconds)

Rentokil Initial plc
https://www.rentokil-initial.com/ ▾
Leading in Pest Control. Our engine for growth ... Rentokil Initial plc welcomes the Commonwealth Heads of State commitment to halve the... 23 Apr 2018.
Missing: ~~window~~ | Must include: window

Protect & Enhance – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/other-services.aspx ▾
Rentokil Specialist Hygiene is one of the leading providers of specialised deep cleaning and specialist industrial cleaning and disinfection services.
Missing: ~~window~~ | Must include: window



Apple Inc

Apple
https://www.apple.com/ ▾
Discover the innovative world of Apple and shop everything iPhone, iPad, Apple Watch, Mac, and Apple TV, plus explore accessories, entertainment, and expert ...

Search apple.com

Mac
MacBook Pro - MacBook - MacBook Air - Compare - iMac

iTunes Store
Download iTunes - Music - Video - iTunes Charts - ...

iPhone
Explore iPhone, the world's most powerful personal device ...

Find a Store
Find a store. Complete store list. Do more of what you love ...

Apple Support
Apple support is here to help. Learn more about popular ...

Accessories
Shop Apple accessories for Apple Watch, iPhone, iPad, iPod, and ...

Apple
Technology company

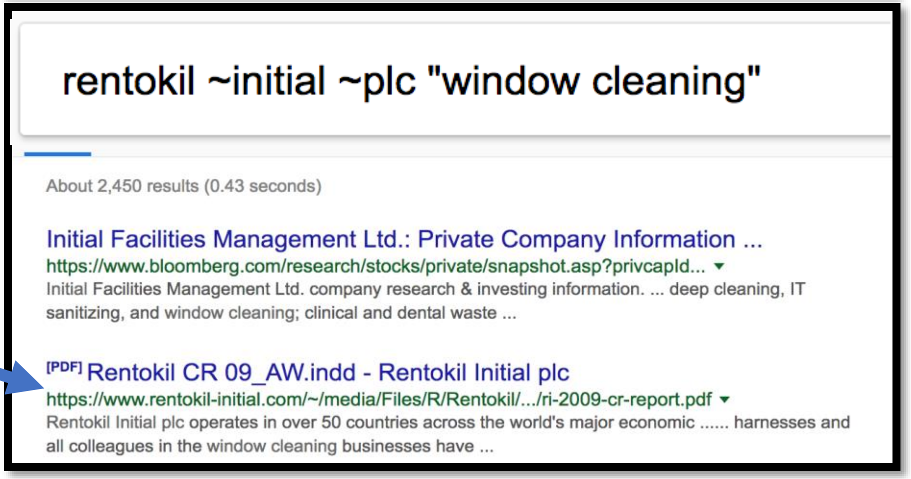
apple.com

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. Wikipedia

Stock price: AAPL (NASDAQ) $191.61 +0.17 (+0.09%)
Jul 23, 4:00 PM EDT - Disclaimer
Founded: April 1, 1976, Cupertino, CA
Headquarters: Cupertino, CA

'Apple Inc.' returns the right results.

PDF result mentions Rentokil Initial PLC involvement in window cleaning.



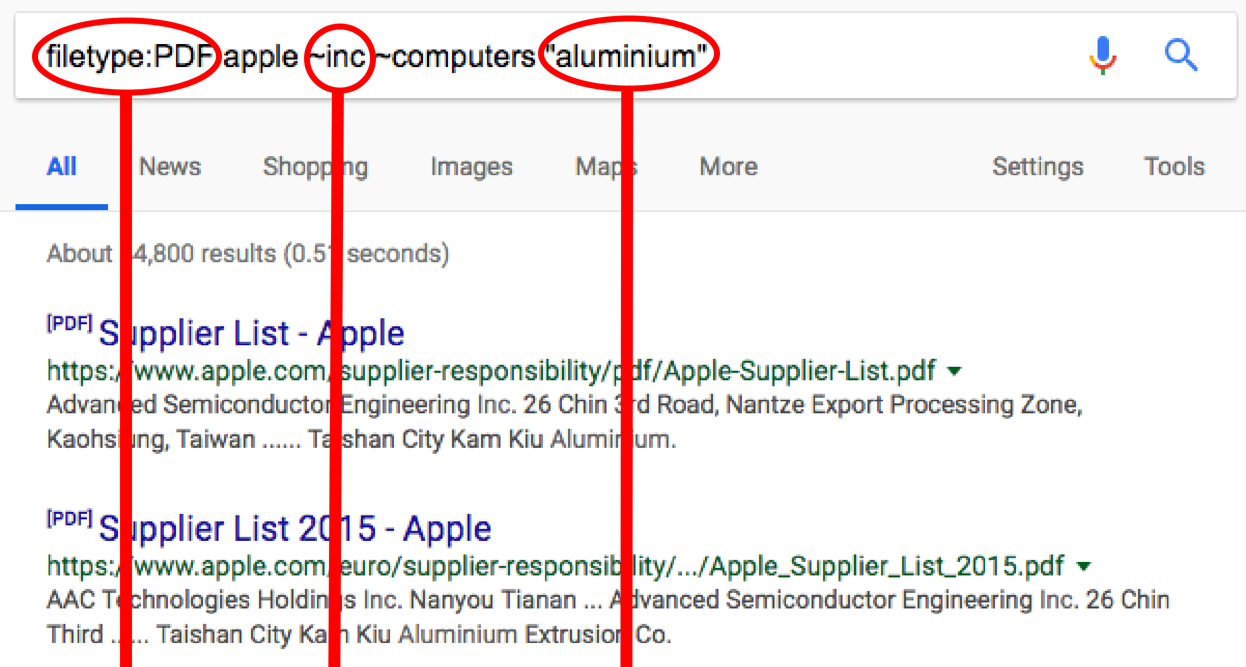
rentokil ~initial ~plc "window cleaning"

About 2,450 results (0.43 seconds)

Initial Facilities Management Ltd.: Private Company Information ...
https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId... ▾
Initial Facilities Management Ltd. company research & investing information. ... deep cleaning, IT sanitizing, and window cleaning; clinical and dental waste ...

[PDF] Rentokil CR 09_AW.indd - Rentokil Initial plc
https://www.rentokil-initial.com/~/media/Files/R/Rentokil/.../ri-2009-cr-report.pdf ▾
Rentokil Initial plc operates in over 50 countries across the world's major economic ...... harnesses and all colleagues in the window cleaning businesses have ...

Query Formulator: How did we ask the right things?



filetype:PDF apple ~inc ~computers "aluminium"

About 4,800 results (0.51 seconds)

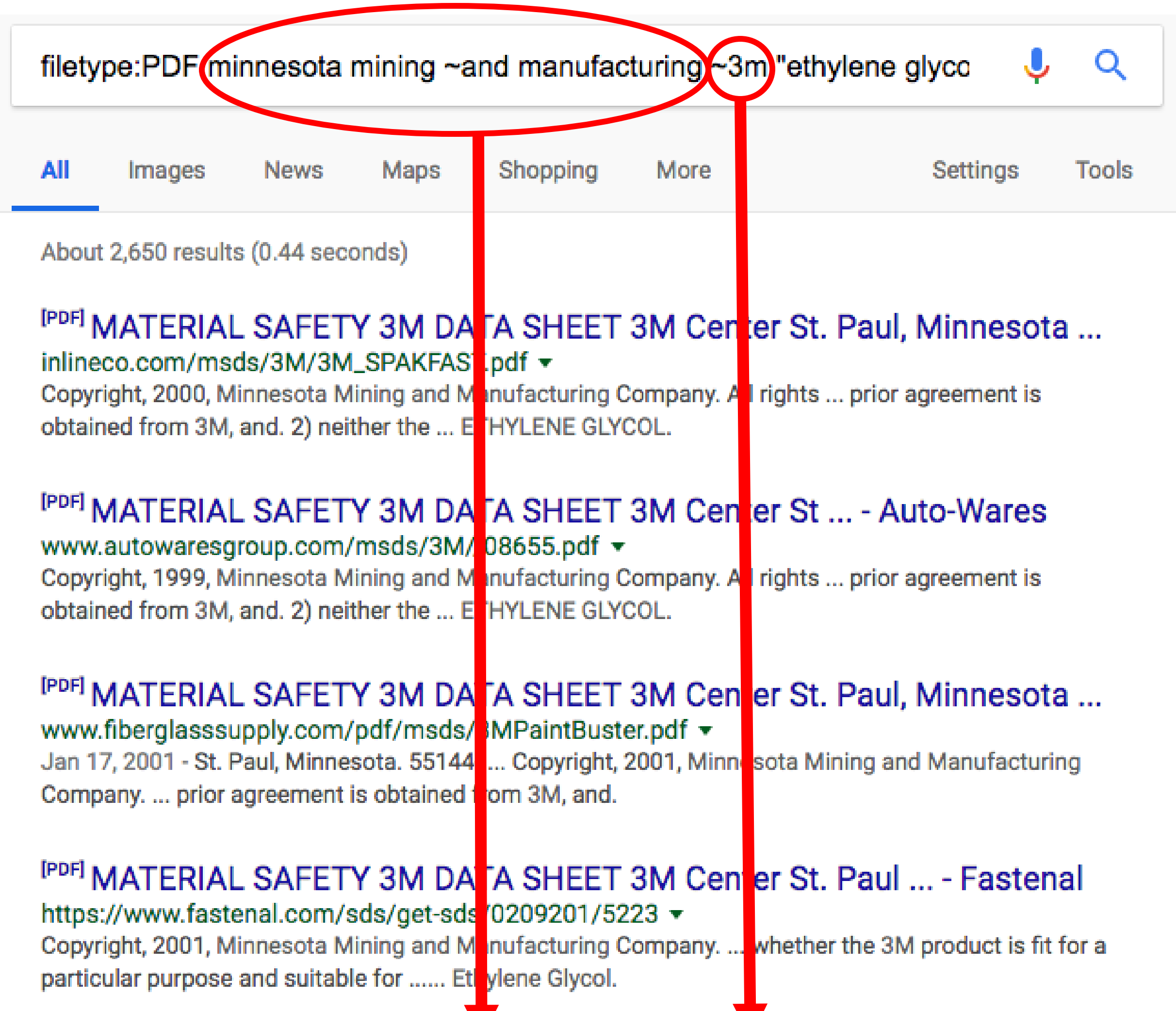
[PDF] Supplier List - Apple
https://www.apple.com/supplier-responsibility/pdf/Apple-Supplier-List.pdf
Advanced Semiconductor Engineering Inc. 26 Chin 3rd Road, Nantze Export Processing Zone, Kaohsiung, Taiwan ...... Taishan City Kam Kiu Aluminium.

[PDF] Supplier List 2015 - Apple
https://www.apple.com/euro/supplier-responsibility/.../Apple_Supplier_List_2015.pdf
AAC Technologies Holdings Inc. Nanyou Tianan ... Advanced Semiconductor Engineering Inc. 26 Chin Third ...... Taishan City Kam Kiu Aluminium Extrusion Co.

Making some words optional

Mention the file-type

Making keywords mandatory



filetype:PDF minnesota mining ~and manufacturing ~3m "ethylene glyco

About 2,650 results (0.44 seconds)

[PDF] MATERIAL SAFETY 3M DATA SHEET 3M Center St. Paul, Minnesota ...
inlineco.com/msds/3M/3M_SPAKFAST.pdf
Copyright, 2000, Minnesota Mining and Manufacturing Company. All rights ... prior agreement is obtained from 3M, and. 2) neither the ... ETHYLENE GLYCOL.

[PDF] MATERIAL SAFETY 3M DATA SHEET 3M Center St ... - Auto-Wares
www.autowaresgroup.com/msds/3M/08655.pdf
Copyright, 1999, Minnesota Mining and Manufacturing Company. All rights ... prior agreement is obtained from 3M, and. 2) neither the ... ETHYLENE GLYCOL.

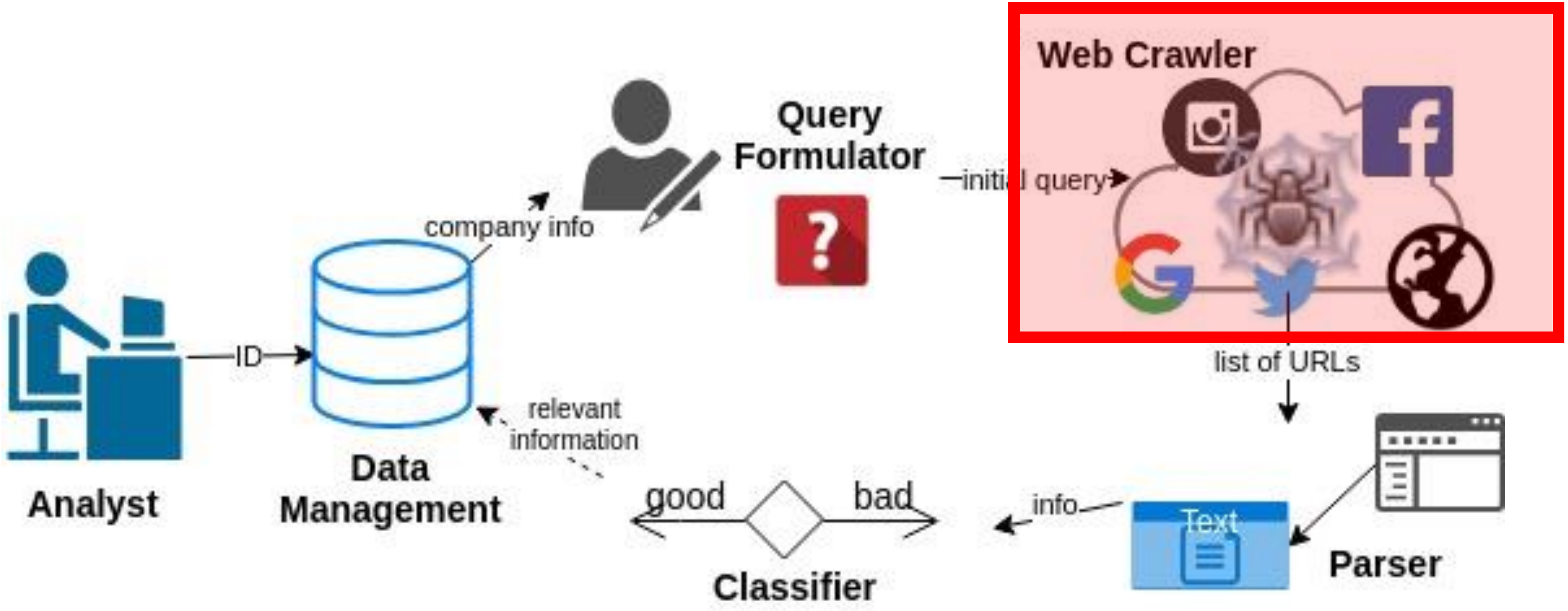
[PDF] MATERIAL SAFETY 3M DATA SHEET 3M Center St. Paul, Minnesota ...
www.fiberglasssupply.com/pdf/msds/3MPaintBuster.pdf
Jan 17, 2001 - St. Paul, Minnesota. 55144 ... Copyright, 2001, Minnesota Mining and Manufacturing Company. ... prior agreement is obtained from 3M, and.

[PDF] MATERIAL SAFETY 3M DATA SHEET 3M Center St. Paul ... - Fastenal
https://www.fastenal.com/sds/get-sds/0209201/5223
Copyright, 2001, Minnesota Mining and Manufacturing Company. ... whether the 3M product is fit for a particular purpose and suitable for ...... Ethylene Glycol.

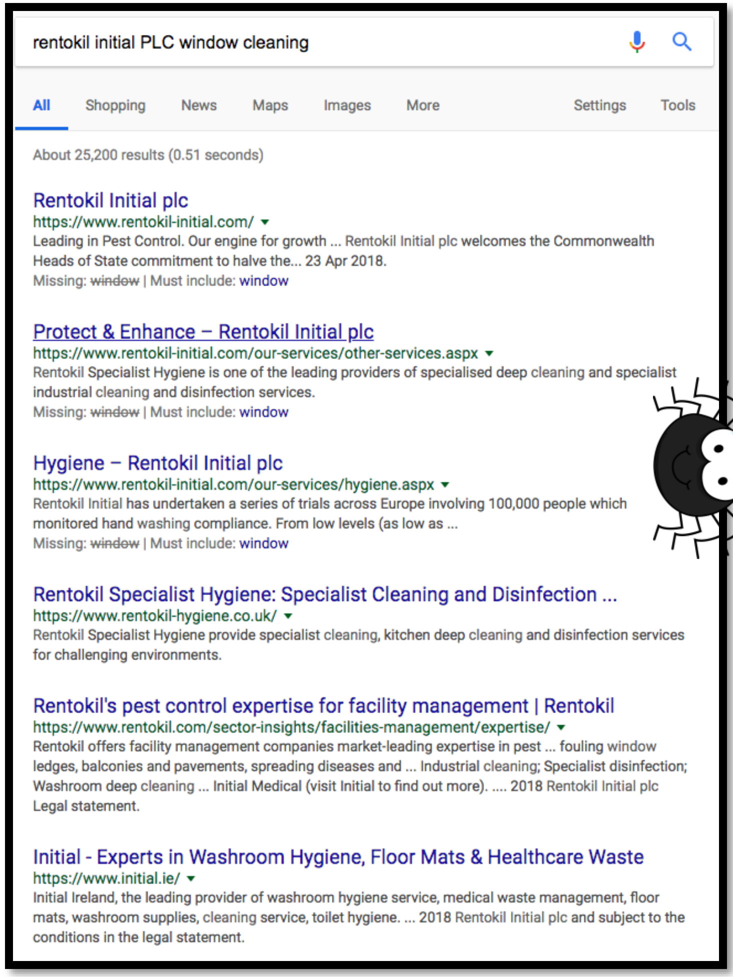
Name of the company

Optional alias

Solution Overview



Web Crawling: What is web Crawling?



rentokil initial PLC window cleaning

About 25,200 results (0.51 seconds)

Rentokil Initial plc
https://www.rentokil-initial.com/ ▾
Leading in Pest Control. Our engine for growth ... Rentokil Initial plc welcomes the Commonwealth Heads of State commitment to halve the... 23 Apr 2018.
Missing: ~~window~~ | Must include: window

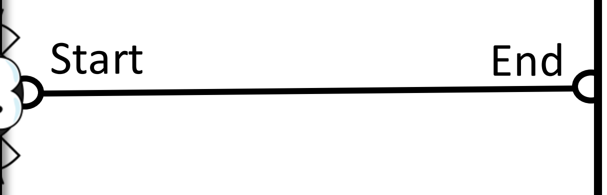
Protect & Enhance – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/other-services.aspx ▾
Rentokil Specialist Hygiene is one of the leading providers of specialised deep cleaning and specialist industrial cleaning and disinfection services.
Missing: ~~window~~ | Must include: window

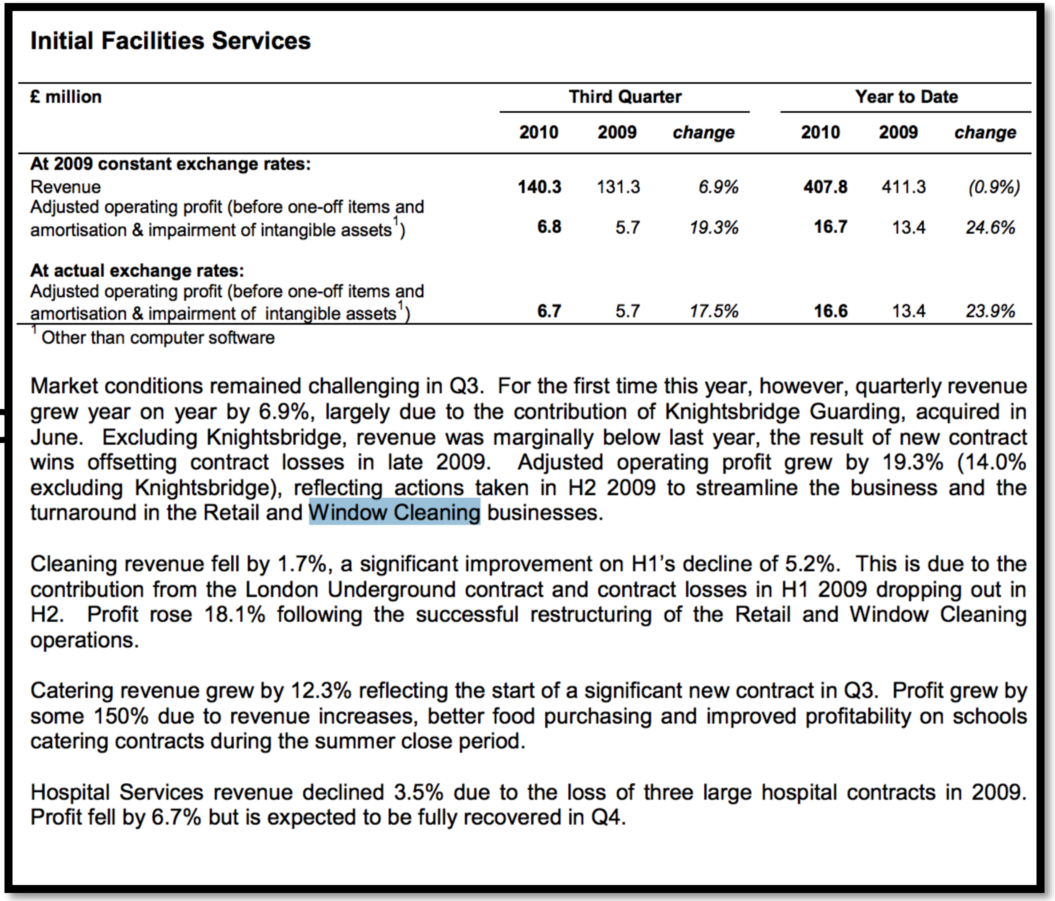
Hygiene – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/hygiene.aspx ▾
Rentokil Initial has undertaken a series of trials across Europe involving 100,000 people which monitored hand washing compliance. From low levels (as low as ...
Missing: ~~window~~ | Must include: window

Rentokil Specialist Hygiene: Specialist Cleaning and Disinfection ...
https://www.rentokil-hygiene.co.uk/ ▾
Rentokil Specialist Hygiene provide specialist cleaning, kitchen deep cleaning and disinfection services for challenging environments.

Rentokil's pest control expertise for facility management | Rentokil
https://www.rentokil.com/sector-insights/facilities-management/expertise/ ▾
Rentokil offers facility management companies market-leading expertise in pest ... fouling window ledges, balconies and pavements, spreading diseases and ... Industrial cleaning; Specialist disinfection; Washroom deep cleaning ... Initial Medical (visit Initial to find out more). .... 2018 Rentokil Initial plc Legal statement.

Initial - Experts in Washroom Hygiene, Floor Mats & Healthcare Waste
https://www.initial.ie/ ▾
Initial Ireland, the leading provider of washroom hygiene service, medical waste management, floor mats, washroom supplies, cleaning service, toilet hygiene. ... 2018 Rentokil Initial plc and subject to the conditions in the legal statement.





Initial Facilities Services

£ million	Third Quarter 2010	2009	*change*	Year to Date 2010	2009	*change*
At 2009 constant exchange rates:						
Revenue	**140.3**	131.3	*6.9%*	**407.8**	411.3	*(0.9%)*
Adjusted operating profit (before one-off items and amortisation & impairment of intangible assets[1])	**6.8**	5.7	*19.3%*	**16.7**	13.4	*24.6%*
At actual exchange rates:						
Adjusted operating profit (before one-off items and amortisation & impairment of intangible assets[1])	**6.7**	5.7	*17.5%*	**16.6**	13.4	*23.9%*

[1] Other than computer software

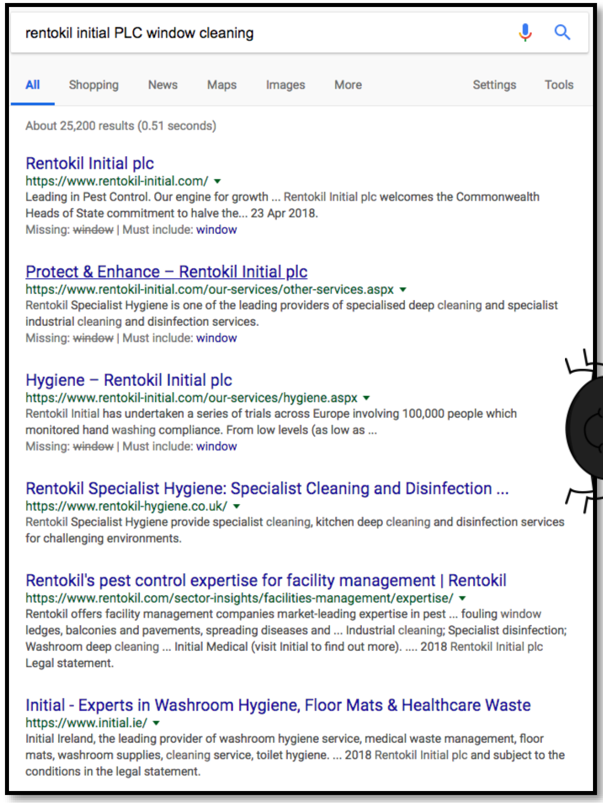
Market conditions remained challenging in Q3. For the first time this year, however, quarterly revenue grew year on year by 6.9%, largely due to the contribution of Knightsbridge Guarding, acquired in June. Excluding Knightsbridge, revenue was marginally below last year, the result of new contract wins offsetting contract losses in late 2009. Adjusted operating profit grew by 19.3% (14.0% excluding Knightsbridge), reflecting actions taken in H2 2009 to streamline the business and the turnaround in the Retail and Window Cleaning businesses.

Cleaning revenue fell by 1.7%, a significant improvement on H1's decline of 5.2%. This is due to the contribution from the London Underground contract and contract losses in H1 2009 dropping out in H2. Profit rose 18.1% following the successful restructuring of the Retail and Window Cleaning operations.

Catering revenue grew by 12.3% reflecting the start of a significant new contract in Q3. Profit grew by some 150% due to revenue increases, better food purchasing and improved profitability on schools catering contracts during the summer close period.

Hospital Services revenue declined 3.5% due to the loss of three large hospital contracts in 2009. Profit fell by 6.7% but is expected to be fully recovered in Q4.

Web Crawling: Unsupervised machines cannot be trusted



About 25,200 results (0.51 seconds)

Rentokil Initial plc
https://www.rentokil-initial.com/ ▾
Leading in Pest Control. Our engine for growth ... Rentokil Initial plc welcomes the Commonwealth Heads of State commitment to halve the... 23 Apr 2018.
Missing: ~~window~~ | Must include: window

Protect & Enhance – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/other-services.aspx ▾
Rentokil Specialist Hygiene is one of the leading providers of specialised deep cleaning and specialist industrial cleaning and disinfection services.
Missing: ~~window~~ | Must include: window

Hygiene – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/hygiene.aspx ▾
Rentokil Initial has undertaken a series of trials across Europe involving 100,000 people which monitored hand washing compliance. From low levels (as low as ...
Missing: ~~window~~ | Must include: window

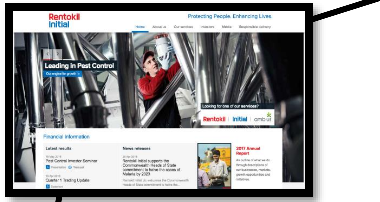
Rentokil Specialist Hygiene: Specialist Cleaning and Disinfection ...
https://www.rentokil-hygiene.co.uk/ ▾
Rentokil Specialist Hygiene provide specialist cleaning, kitchen deep cleaning and disinfection services for challenging environments.

Rentokil's pest control expertise for facility management | Rentokil
https://www.rentokil.com/sector-insights/facilities-management/expertise/ ▾
Rentokil offers facility management companies market-leading expertise in pest ... fouling window ledges, balconies and pavements, spreading diseases and ... Industrial cleaning; Specialist disinfection; Washroom deep cleaning ... Initial Medical (visit Initial to find out more). .... 2018 Rentokil Initial plc Legal statement.

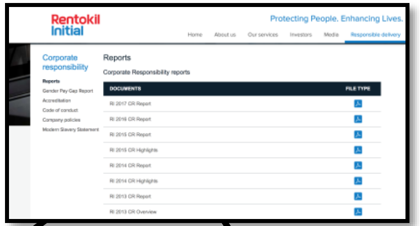
Initial - Experts in Washroom Hygiene, Floor Mats & Healthcare Waste
https://www.initial.ie/ ▾
Initial Ireland, the leading provider of washroom hygiene service, medical waste management, floor mats, washroom supplies, cleaning service, toilet hygiene. ... 2018 Rentokil Initial plc and subject to the conditions in the legal statement.

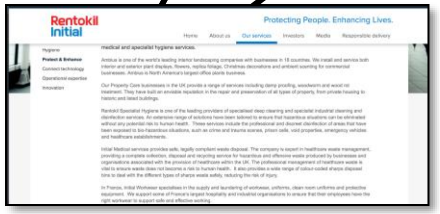
Start with a google search of the company and its business activity.

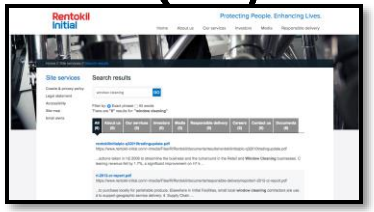
Start



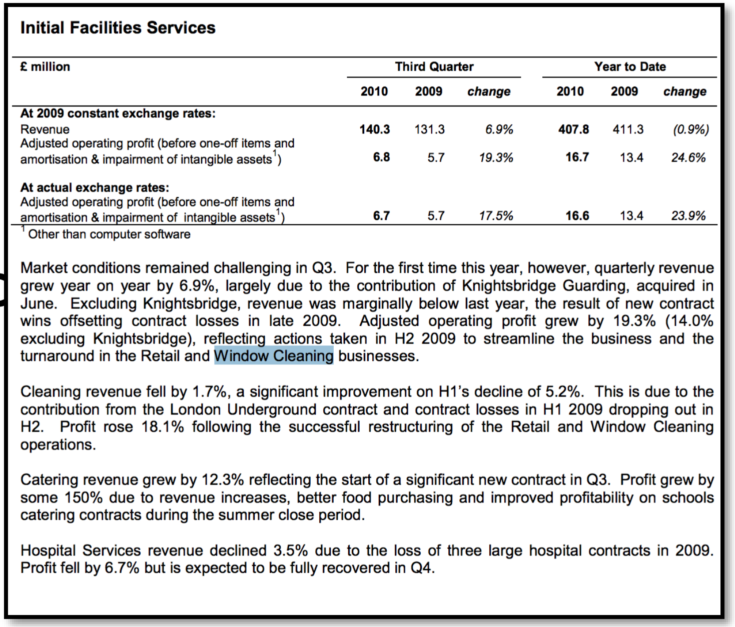
Rentokil Initial
Protecting People. Enhancing Lives.
Our services
Creating a more effective business


Rentokil Initial
Protecting People. Enhancing Lives.
Corporate responsibility
Reports
Corporate Responsibility reports
DOCUMENTS
FILE TYPE





End



Initial Facilities Services

£ million	**Third Quarter**			**Year to Date**		
	2010	**2009**	***change***	**2010**	**2009**	***change***
At 2009 constant exchange rates:						
Revenue	**140.3**	131.3	*6.9%*	**407.8**	411.3	*(0.9%)*
Adjusted operating profit (before one-off items and amortisation & impairment of intangible assets[1])	**6.8**	5.7	*19.3%*	**16.7**	13.4	*24.6%*
At actual exchange rates:						
Adjusted operating profit (before one-off items and amortisation & impairment of intangible assets[1])	**6.7**	5.7	*17.5%*	**16.6**	13.4	*23.9%*

[1] Other than computer software

Market conditions remained challenging in Q3. For the first time this year, however, quarterly revenue grew year on year by 6.9%, largely due to the contribution of Knightsbridge Guarding, acquired in June. Excluding Knightsbridge, revenue was marginally below last year, the result of new contract wins offsetting contract losses in late 2009. Adjusted operating profit grew by 19.3% (14.0% excluding Knightsbridge), reflecting actions taken in H2 2009 to streamline the business and the turnaround in the Retail and Window Cleaning businesses.

Cleaning revenue fell by 1.7%, a significant improvement on H1's decline of 5.2%. This is due to the contribution from the London Underground contract and contract losses in H1 2009 dropping out in H2. Profit rose 18.1% following the successful restructuring of the Retail and Window Cleaning operations.

Catering revenue grew by 12.3% reflecting the start of a significant new contract in Q3. Profit grew by some 150% due to revenue increases, better food purchasing and improved profitability on schools catering contracts during the summer close period.

Hospital Services revenue declined 3.5% due to the loss of three large hospital contracts in 2009. Profit fell by 6.7% but is expected to be fully recovered in Q4.

The business activity appears in the financial report that specifically appears on search services provided by the website.

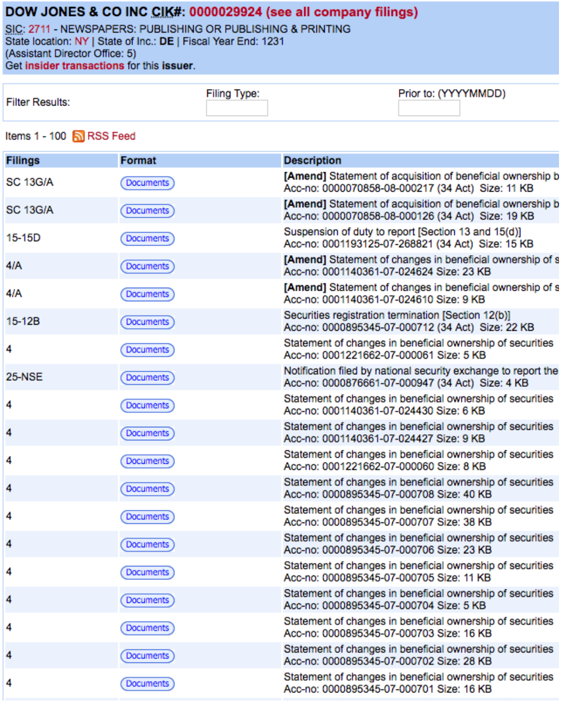
Web Crawling: Where and how far?

The problem:

We don’t know how far to dig, and where to dig?

We don’t know the credible sources and where the information lies on the credible sources.

Web Crawling: Credible data to the rescue

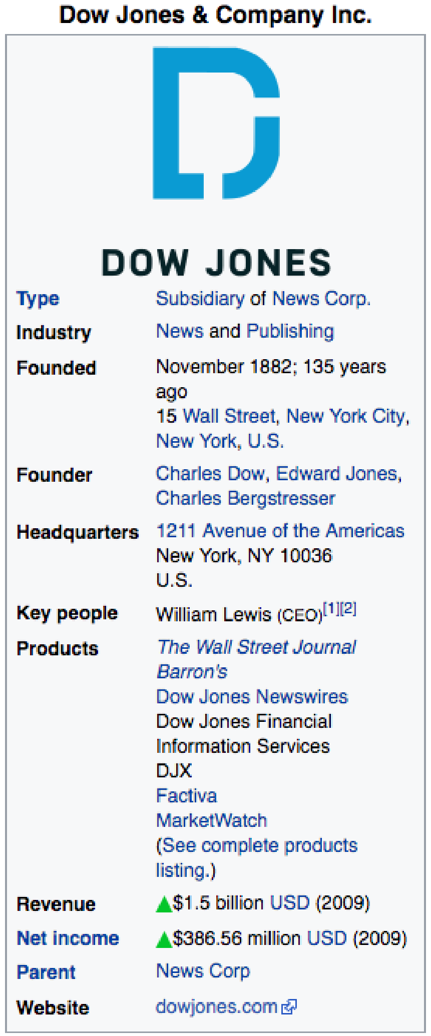


DOW JONES & CO INC CIK#: 0000029924 (see all company filings)
SIC: 2711 - NEWSPAPERS: PUBLISHING OR PUBLISHING & PRINTING
State location: NY | State of Inc.: **DE** | Fiscal Year End: 1231
(Assistant Director Office: 5)
Get **insider transactions** for this **issuer**.

Filter Results: Filing Type: Prior to: (YYYYMMDD)

Items 1 - 100 RSS Feed

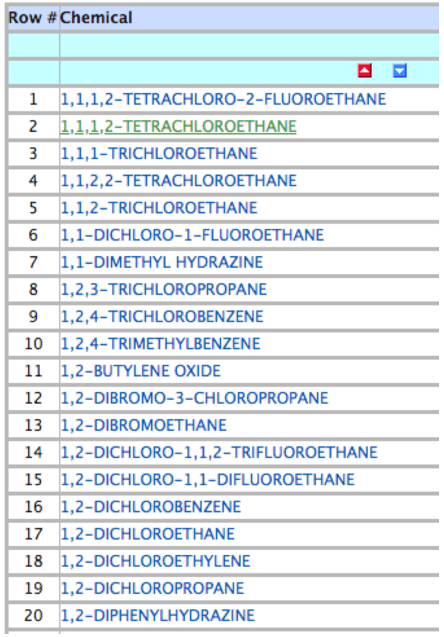
Filings	Format	Description
SC 13G/A	Documents	**[Amend]** Statement of acquisition of beneficial ownership b Acc-no: 0000070858-08-000217 (34 Act) Size: 11 KB
SC 13G/A	Documents	**[Amend]** Statement of acquisition of beneficial ownership b Acc-no: 0000070858-08-000126 (34 Act) Size: 19 KB
15-15D	Documents	Suspension of duty to report [Section 13 and 15(d)] Acc-no: 0001193125-07-268821 (34 Act) Size: 15 KB
4/A	Documents	**[Amend]** Statement of changes in beneficial ownership of s Acc-no: 0001140361-07-024624 Size: 23 KB
4/A	Documents	**[Amend]** Statement of changes in beneficial ownership of s Acc-no: 0001140361-07-024610 Size: 9 KB
15-12B	Documents	Securities registration termination [Section 12(b)] Acc-no: 0000895345-07-000712 (34 Act) Size: 22 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001221662-07-000061 Size: 5 KB
25-NSE	Documents	Notification filed by national security exchange to report the Acc-no: 0000876661-07-000947 (34 Act) Size: 4 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001140361-07-024430 Size: 6 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001140361-07-024427 Size: 9 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001221662-07-000060 Size: 8 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000708 Size: 40 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000707 Size: 38 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000706 Size: 23 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000705 Size: 11 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000704 Size: 5 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000703 Size: 16 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000702 Size: 28 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000701 Size: 16 KB



Dow Jones & Company Inc.

DOW JONES

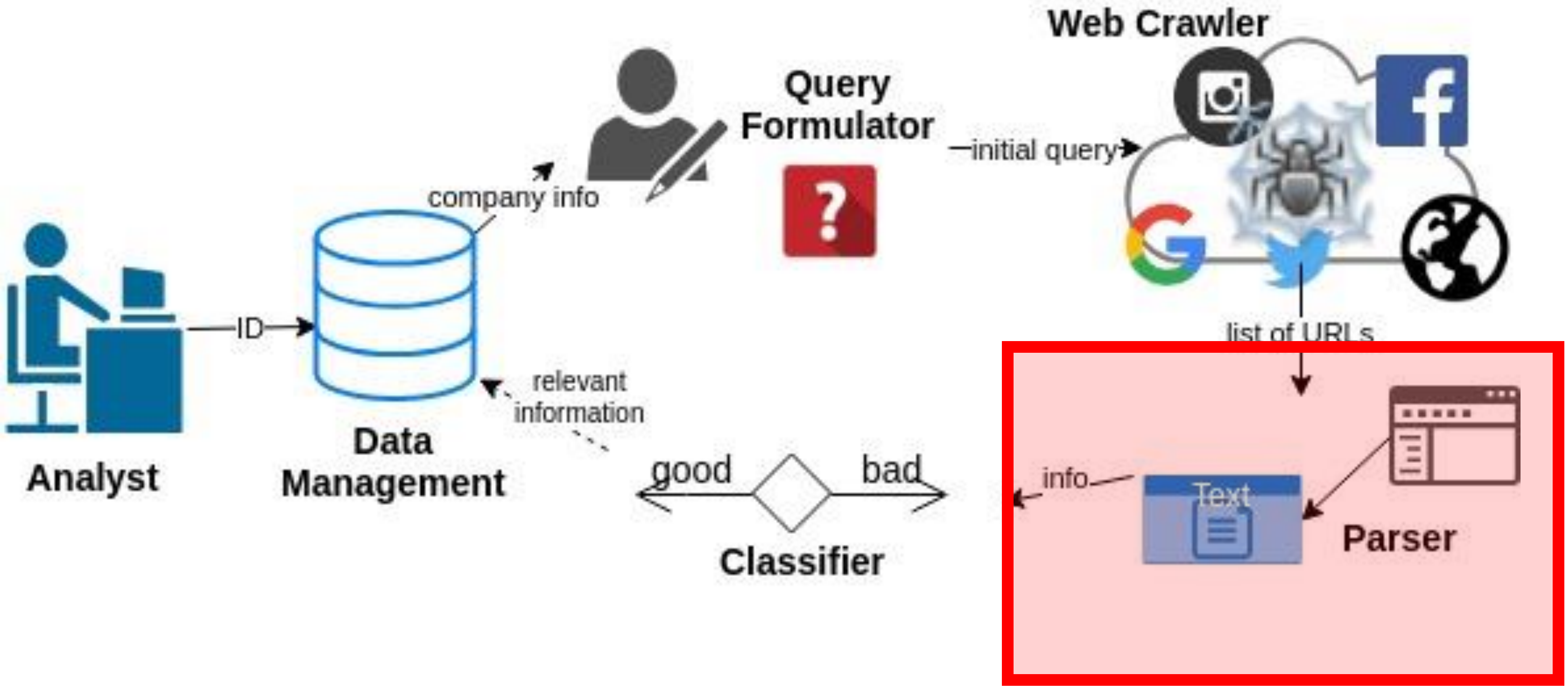
Type	Subsidiary of News Corp.
Industry	News and Publishing
Founded	November 1882; 135 years ago 15 Wall Street, New York City, New York, U.S.
Founder	Charles Dow, Edward Jones, Charles Bergstresser
Headquarters	1211 Avenue of the Americas New York, NY 10036 U.S.
Key people	William Lewis (CEO)[1][2]
Products	*The Wall Street Journal* *Barron's* Dow Jones Newswires Dow Jones Financial Information Services DJX Factiva MarketWatch (See complete products listing.)
Revenue	▲$1.5 billion USD (2009)
Net income	▲$386.56 million USD (2009)
Parent	News Corp
Website	dowjones.com



Row #	Chemical
1	1,1,1,2-TETRACHLORO-2-FLUOROETHANE
2	1,1,1,2-TETRACHLOROETHANE
3	1,1,1-TRICHLOROETHANE
4	1,1,2,2-TETRACHLOROETHANE
5	1,1,2-TRICHLOROETHANE
6	1,1-DICHLORO-1-FLUOROETHANE
7	1,1-DIMETHYL HYDRAZINE
8	1,2,3-TRICHLOROPROPANE
9	1,2,4-TRICHLOROBENZENE
10	1,2,4-TRIMETHYLBENZENE
11	1,2-BUTYLENE OXIDE
12	1,2-DIBROMO-3-CHLOROPROPANE
13	1,2-DIBROMOETHANE
14	1,2-DICHLORO-1,1,2-TRIFLUOROETHANE
15	1,2-DICHLORO-1,1-DIFLUOROETHANE
16	1,2-DICHLOROBENZENE
17	1,2-DICHLOROETHANE
18	1,2-DICHLOROETHYLENE
19	1,2-DICHLOROPROPANE
20	1,2-DIPHENYLHYDRAZINE

- Interestingly the structured data (available on Federal websites & Wikipedia) is also credible!
- Design of specific crawlers to get data from specific types of data.
- Create the baseline data.

Solution Overview

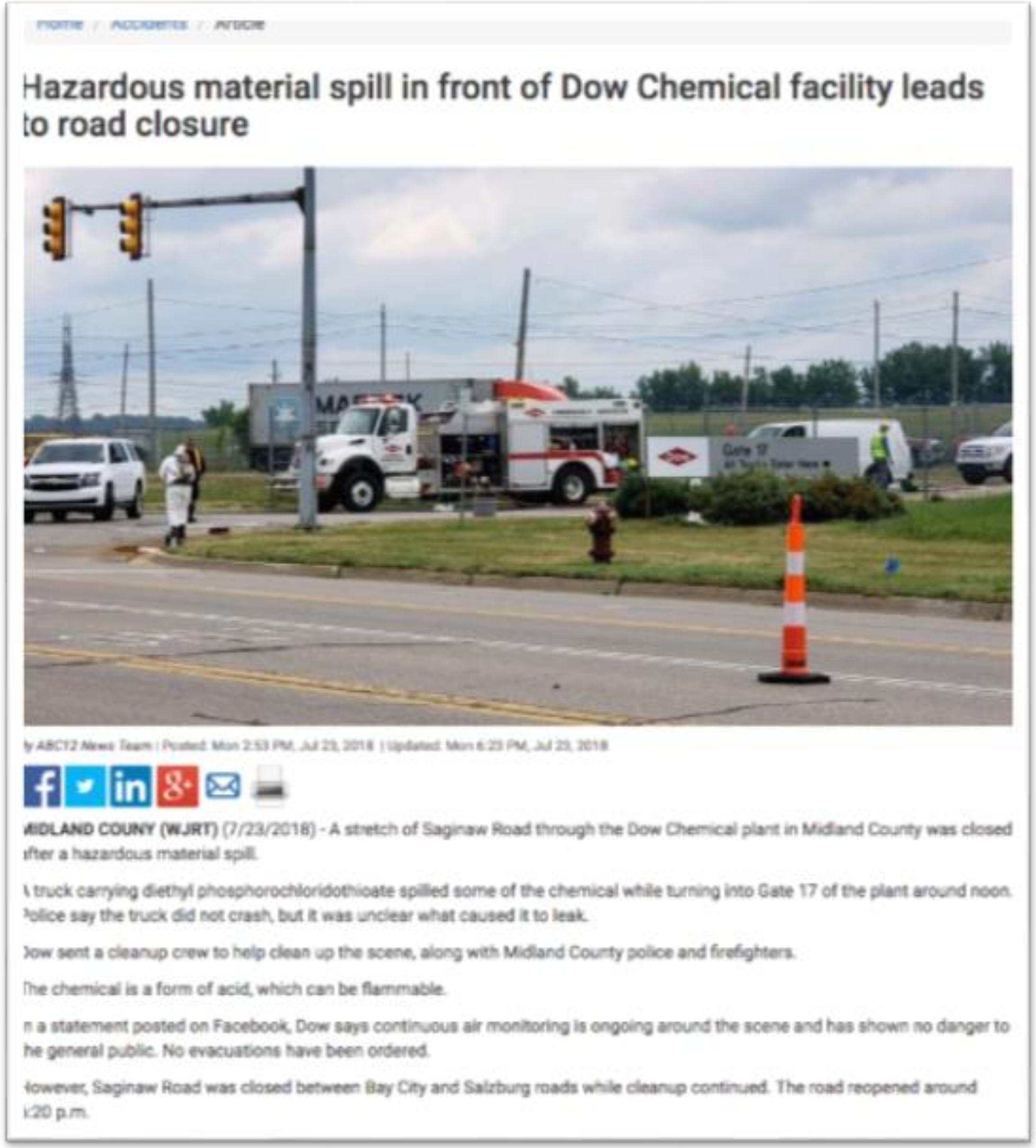


Parser:
Getting unstructured data

Use of text abundance to locate meaningful paragraphs.

Filtering out tags containing social media redirects.

Removing graphic contents, advertisements.



Home / Accidents / Article

Hazardous material spill in front of Dow Chemical facility leads to road closure

y ABC12 News Team | Posted: Mon 2:53 PM, Jul 23, 2018 | Updated: Mon 6:23 PM, Jul 23, 2018

IDLAND COUNY (WJRT) (7/23/2018) - A stretch of Saginaw Road through the Dow Chemical plant in Midland County was closed ıfter a hazardous material spill.

\ truck carrying diethyl phosphorochloridothioate spilled some of the chemical while turning into Gate 17 of the plant around noon. ⁾olice say the truck did not crash, but it was unclear what caused it to leak.

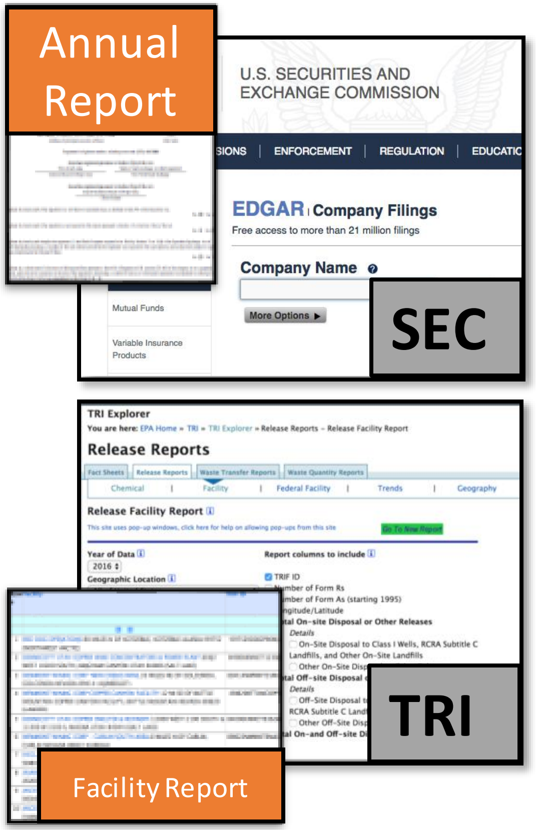
)ow sent a cleanup crew to help clean up the scene, along with Midland County police and firefighters.

'he chemical is a form of acid, which can be flammable.

n a statement posted on Facebook, Dow says continuous air monitoring is ongoing around the scene and has shown no danger to he general public. No evacuations have been ordered.

lowever, Saginaw Road was closed between Bay City and Salzburg roads while cleanup continued. The road reopened around ⁞:20 p.m.

Parser: Why go towards unstructured data?



Annual Report

U.S. SECURITIES AND EXCHANGE COMMISSION

SIONS | ENFORCEMENT | REGULATION | EDUCATIO

EDGAR | Company Filings

Free access to more than 21 million filings

Company Name

More Options ▸

Mutual Funds

Variable Insurance Products

SEC

TRI Explorer

You are here: EPA Home » TRI » TRI Explorer » Release Reports » Release Facility Report

Release Reports

Fact Sheets | Release Reports | Waste Transfer Reports | Waste Quantity Reports

Chemical | Facility | Federal Facility | Trends | Geography

Release Facility Report

Year of Data
2016

Geographic Location

Report columns to include

☑ TRIF ID

umber of Form Rs

umber of Form As (starting 1995)

ngitude/Latitude

tal On-site Disposal or Other Releases

Details

On-Site Disposal to Class I Wells, RCRA Subtitle C Landfills, and Other On-Site Landfills

Other On-Site Disp

tal Off-site Disposal

Details

Off-Site Disposal t

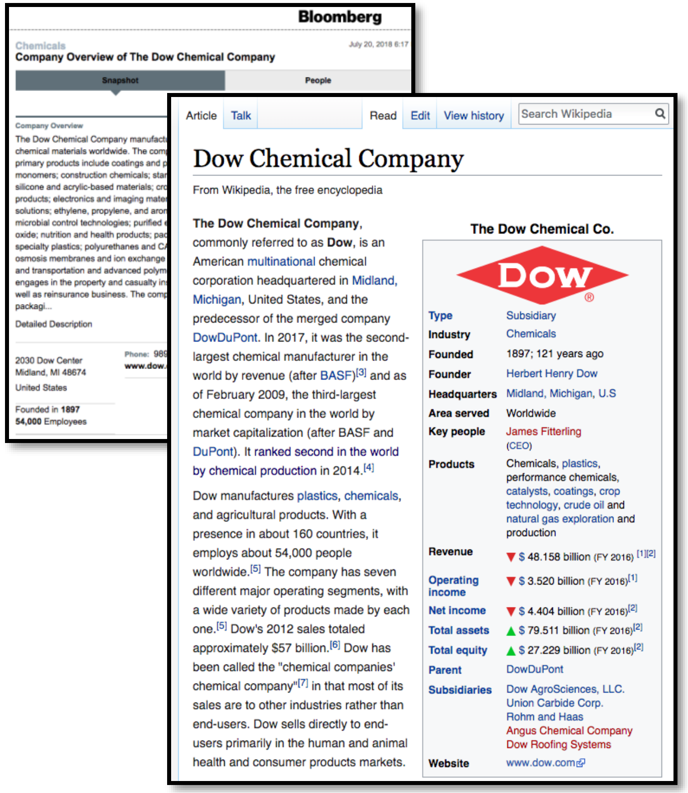
RCRA Subtitle C Land

Other Off-Site Disp

tal On-and Off-site Di

TRI

Facility Report



Bloomberg

Chemicals

Company Overview of The Dow Chemical Company

July 20, 2018 6:17

Snapshot | People

Company Overview

The Dow Chemical Company manufact... chemical materials worldwide. The com... primary products include coatings and p... monomers; construction chemicals; star... silicone and acrylic-based materials; cro... products; electronics and imaging mater... solutions; ethylene, propylene, and aron... microbial control technologies; purified... oxide; nutrition and health products; pac... specialty plastics; polyurethanes and C... osmosis membranes and ion exchange... and transportation and advanced polym... engages in the property and casualty in... well as reinsurance business. The comp... packagi...

Detailed Description

2030 Dow Center
Midland, MI 48674
United States

Phone: 989...
www.dow...

Founded in 1897
54,000 Employees

Article | Talk | Read | Edit | View history | Search Wikipedia

Dow Chemical Company

From Wikipedia, the free encyclopedia

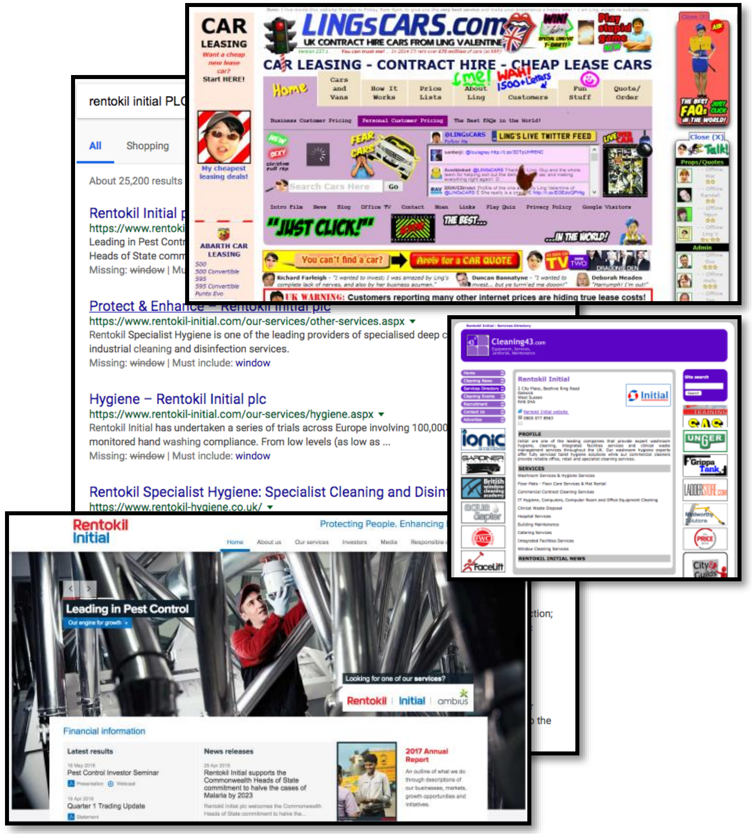
The Dow Chemical Company, commonly referred to as **Dow**, is an American multinational chemical corporation headquartered in Midland, Michigan, United States, and the predecessor of the merged company DowDuPont. In 2017, it was the second-largest chemical manufacturer in the world by revenue (after BASF)[3] and as of February 2009, the third-largest chemical company in the world by market capitalization (after BASF and DuPont). It ranked second in the world by chemical production in 2014.[4]

Dow manufactures plastics, chemicals, and agricultural products. With a presence in about 160 countries, it employs about 54,000 people worldwide.[5] The company has seven different major operating segments, with a wide variety of products made by each one.[5] Dow's 2012 sales totaled approximately $57 billion.[6] Dow has been called the "chemical companies' chemical company"[7] in that most of its sales are to other industries rather than end-users. Dow sells directly to end-users primarily in the human and animal health and consumer products markets.

The Dow Chemical Co.

DOW ®

Type	Subsidiary
Industry	Chemicals
Founded	1897; 121 years ago
Founder	Herbert Henry Dow
Headquarters	Midland, Michigan, U.S
Area served	Worldwide
Key people	James Fitterling (CEO)
Products	Chemicals, plastics, performance chemicals, catalysts, coatings, crop technology, crude oil and natural gas exploration and production
Revenue	▼ $ 48.158 billion (FY 2016) [1][2]
Operating income	▼ $ 3.520 billion (FY 2016)[1]
Net income	▼ $ 4.404 billion (FY 2016)[2]
Total assets	▲ $ 79.511 billion (FY 2016)[2]
Total equity	▲ $ 27.229 billion (FY 2016)[2]
Parent	DowDuPont
Subsidiaries	Dow AgroSciences, LLC. Union Carbide Corp. Rohm and Haas Angus Chemical Company Dow Roofing Systems
Website	www.dow.com



rentokil initial PLC

All | Shopping

About 25,200 results

Rentokil Initial p...
https://www.rentokil...
Leading in Pest Contr...
Heads of State comm...
Missing: ~~window~~ | Mu...

Protect & Enhance – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/other-services.aspx ▾
Rentokil Specialist Hygiene is one of the leading providers of specialised deep c... industrial cleaning and disinfection services.
Missing: ~~window~~ | Must include: window

Hygiene – Rentokil Initial plc
https://www.rentokil-initial.com/our-services/hygiene.aspx ▾
Rentokil Initial has undertaken a series of trials across Europe involving 100,000... monitored hand washing compliance. From low levels (as low as ...
Missing: ~~window~~ | Must include: window

Rentokil Specialist Hygiene: Specialist Cleaning and Disin...
https://www.rentokil-hygiene.co.uk/ ▾

More credible information

Structured Data

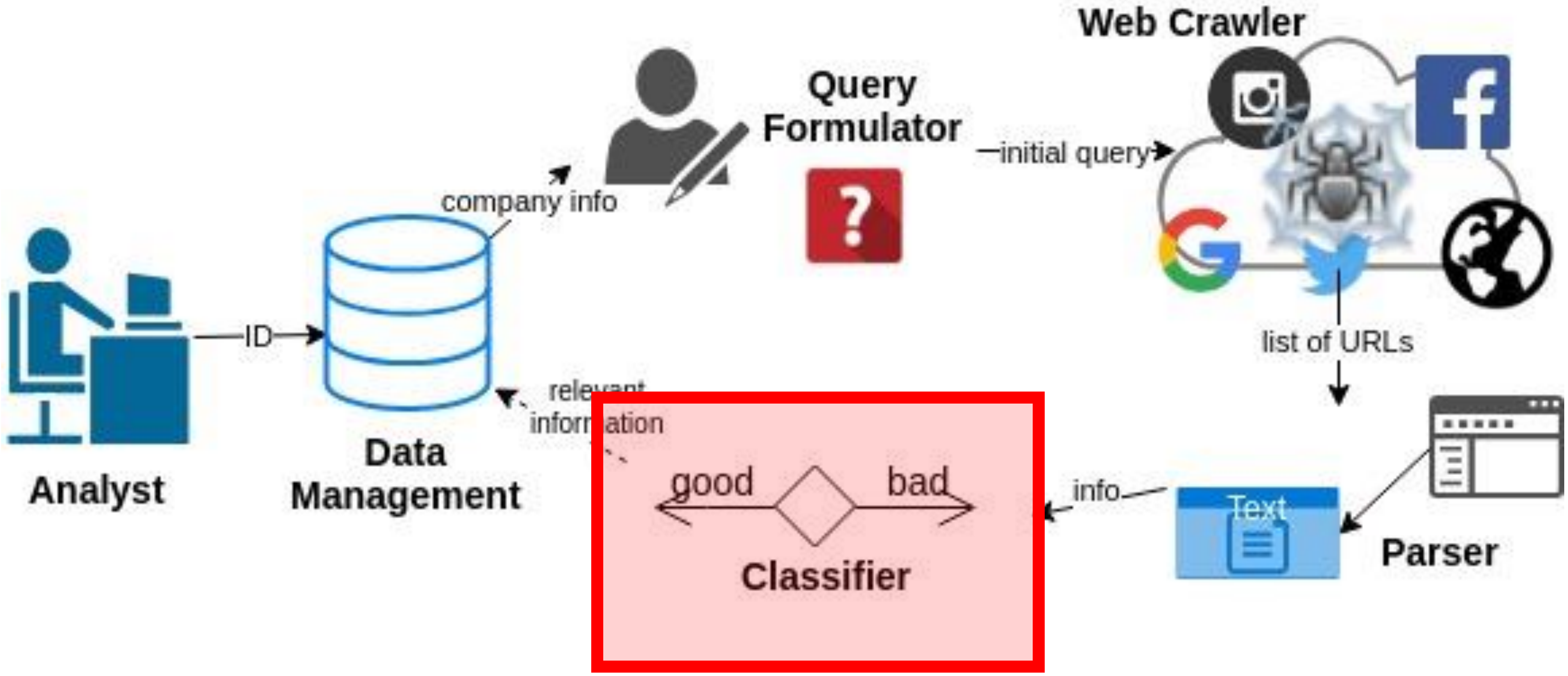
Lower volume of data

More indicative information

Unstructured data

Higher volume of data

Solution Overview



Classifiers

- Our web crawlers and parsers can give us a huge quantity of data, but not all sources are created the same.
- Our web crawler needs to distinguish between high quality data and YouTube comments.

Term Frequency-Inverse Document Frequency (TF-IDF)

- Each term (word, *i*) in a document (*j*) has its own *term frequency (tf)*

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

- Each word has an *inverse-document frequency (idf), which*

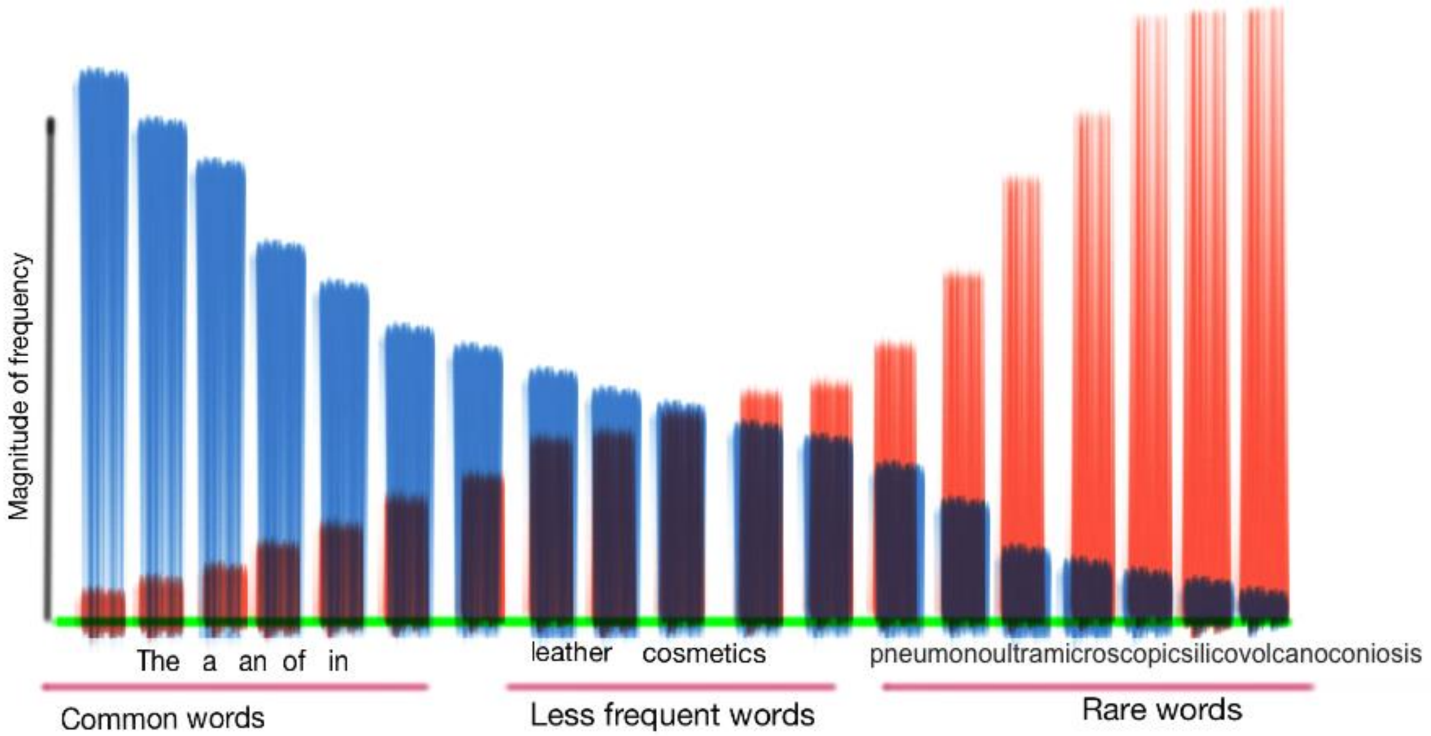
$$idf(w) = log(\frac{N}{df_t})$$

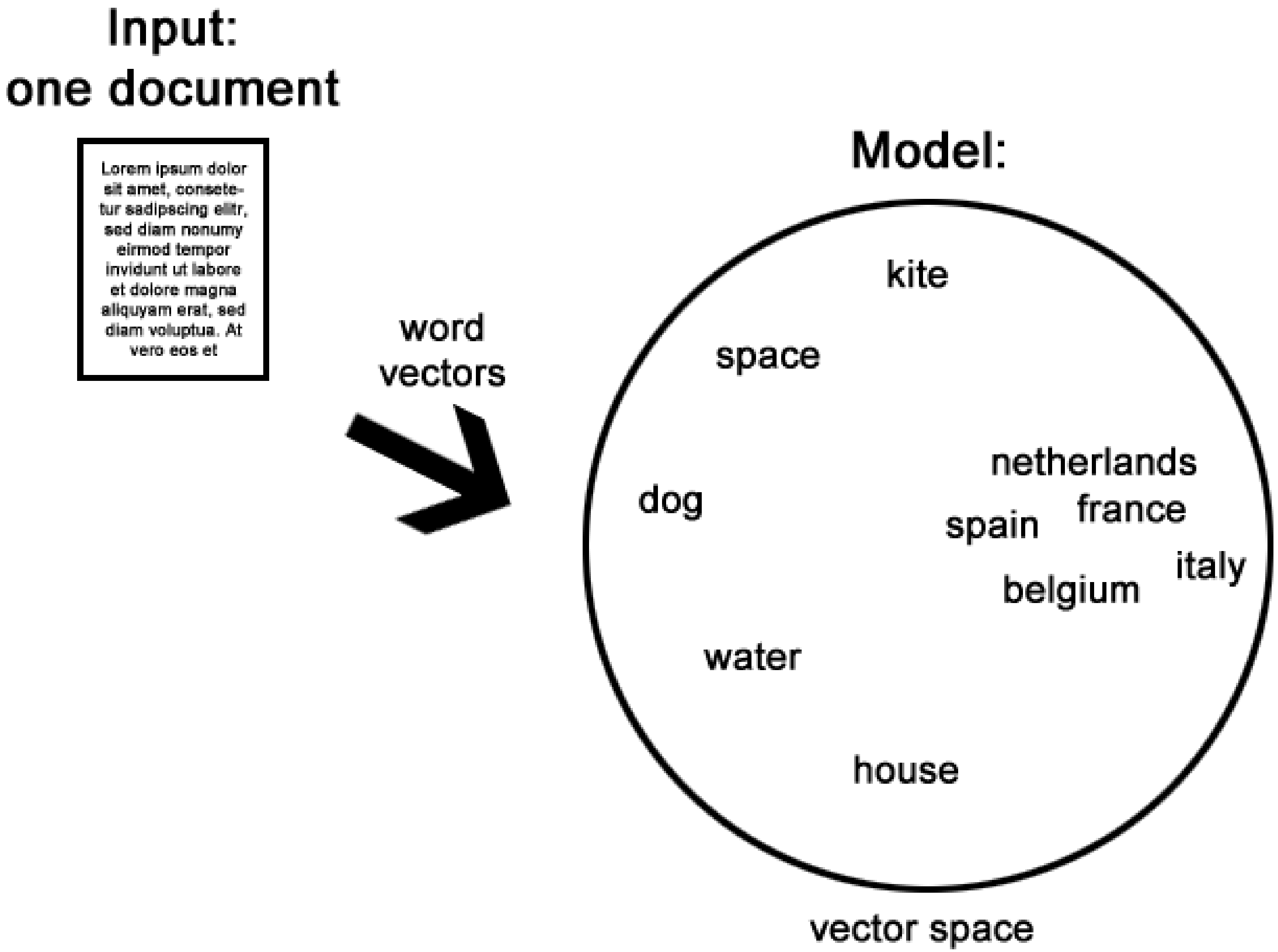
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

🟦 $tf_{i,j}$ = number of occurrences of i in j

🟥 df_i = number of documents containing i

N = total number of documents



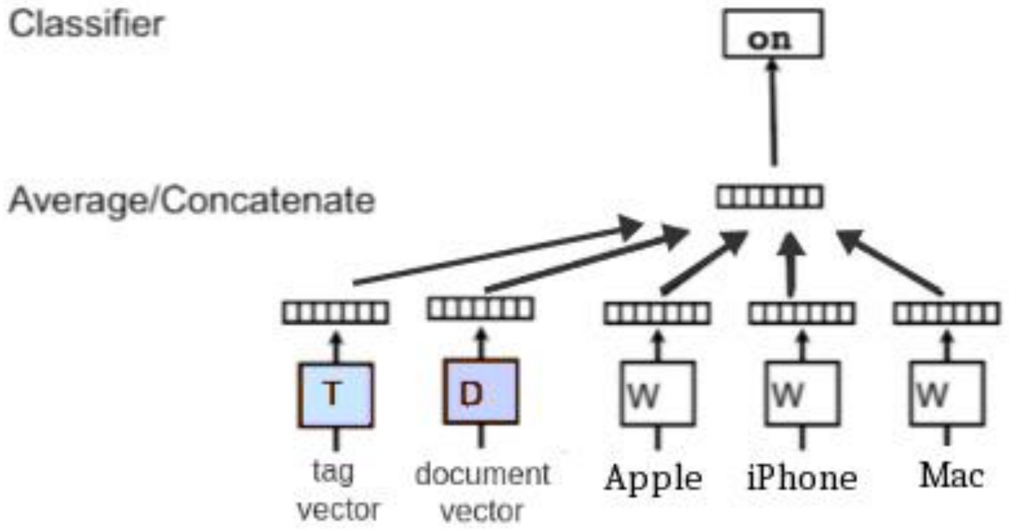

Input:
one document
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et
word vectors
Model:
kite
space
netherlands
dog
spain
france
italy
belgium
water
house
vector space

most_similar('france'):

```
spain          0.678515
belgium        0.665923
netherlands    0.652428
italy          0.633130
```

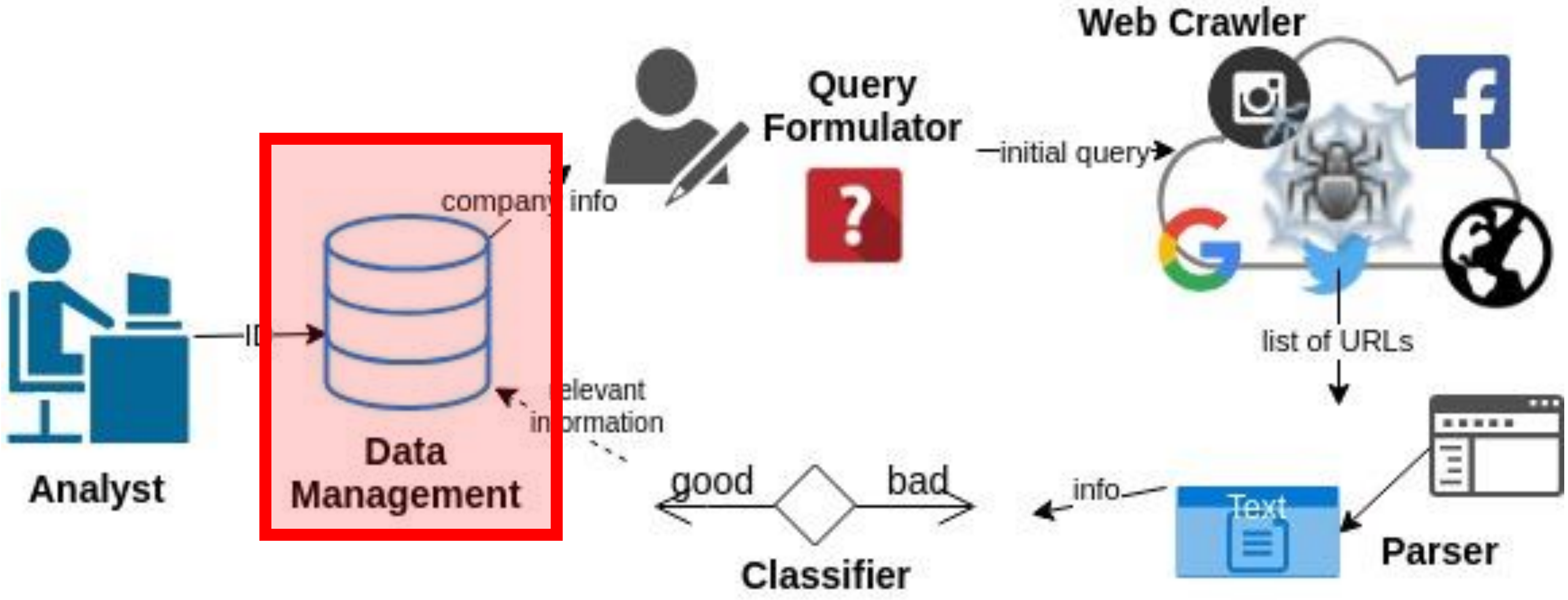
Keyword Generation with word2vec

doc2vec



- Doc2vec uses a ***Distributed Memory version of Paragraph Vector*** (PV-DM)
- You can "tag" documents with topics.
- We can attempt to cluster or classify documents using tags.

Solution Overview



Data Management

- Many companies can't make use of their data
- Our data needs to be:
  - Traceable back to the source
  - Queryable
- Our solution: ProfileManager and WebResourceManager.

The Story So Far...

- Our web crawling and parsing capabilities are bottlenecked by our ability to classify information.
- We currently have a database of high-credibility information about 52,629 companies and corporate entities including:
  - 10Ks
  - 8Ks
  - EX21s
  - Wikipedia

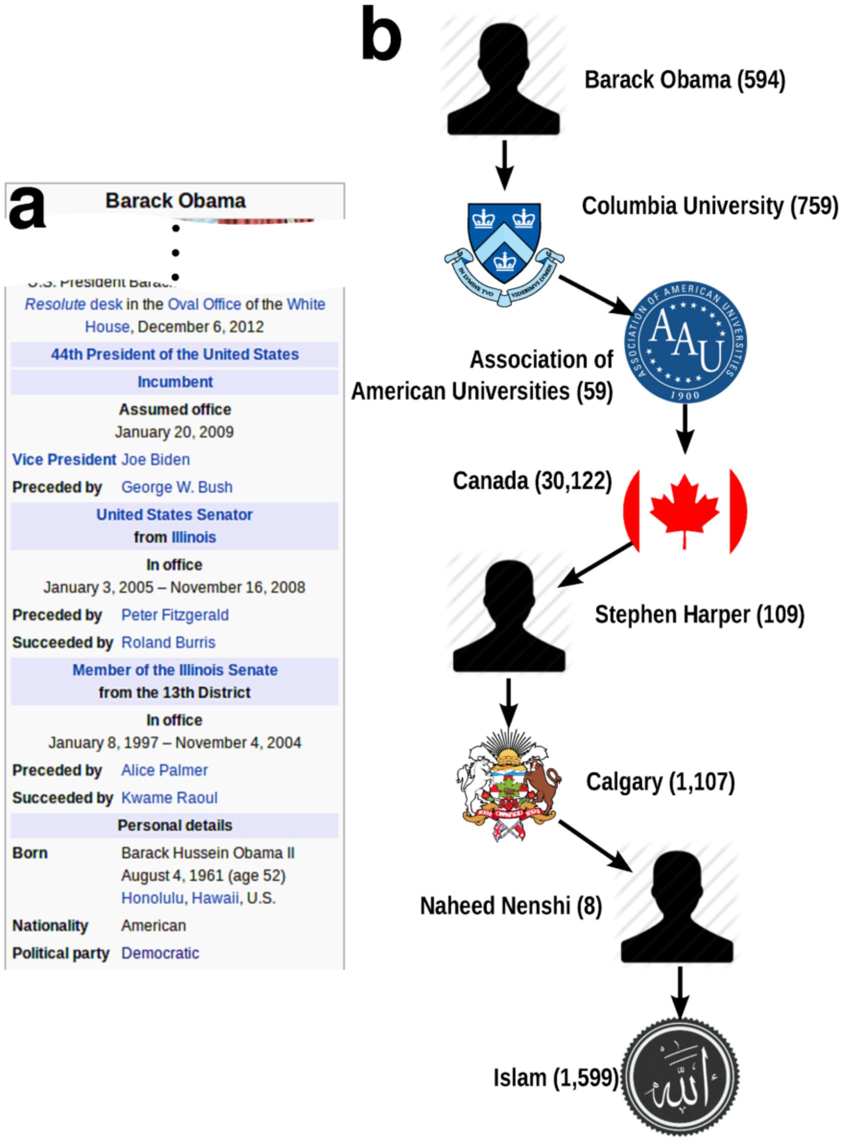
Going Forward

Knowledge Graph

- Knowledge graphs represent **statements of facts** as subject-predicate-object triples

  For example ("Barack Obama", "is a", "Muslim")

- G=(V,E) where V is a set of concept nodes and E is a set of predicate edges
- To determine the degree of truth of a statement of fact, the shortest path from subject to object is traversed and paths that pass through high-degree nodes are assigned a low truth value



Barack Obama

U.S. President Barac

Resolute desk in the Oval Office of the White House, December 6, 2012

44th President of the United States

Incumbent

Assumed office
January 20, 2009

Vice President	Joe Biden
Preceded by	George W. Bush

United States Senator from Illinois

In office
January 3, 2005 – November 16, 2008

Preceded by	Peter Fitzgerald
Succeeded by	Roland Burris

Member of the Illinois Senate from the 13th District

In office
January 8, 1997 – November 4, 2004

Preceded by	Alice Palmer
Succeeded by	Kwame Raoul

Personal details

Born	Barack Hussein Obama II August 4, 1961 (age 52) Honolulu, Hawaii, U.S.
Nationality	American
Political party	Democratic

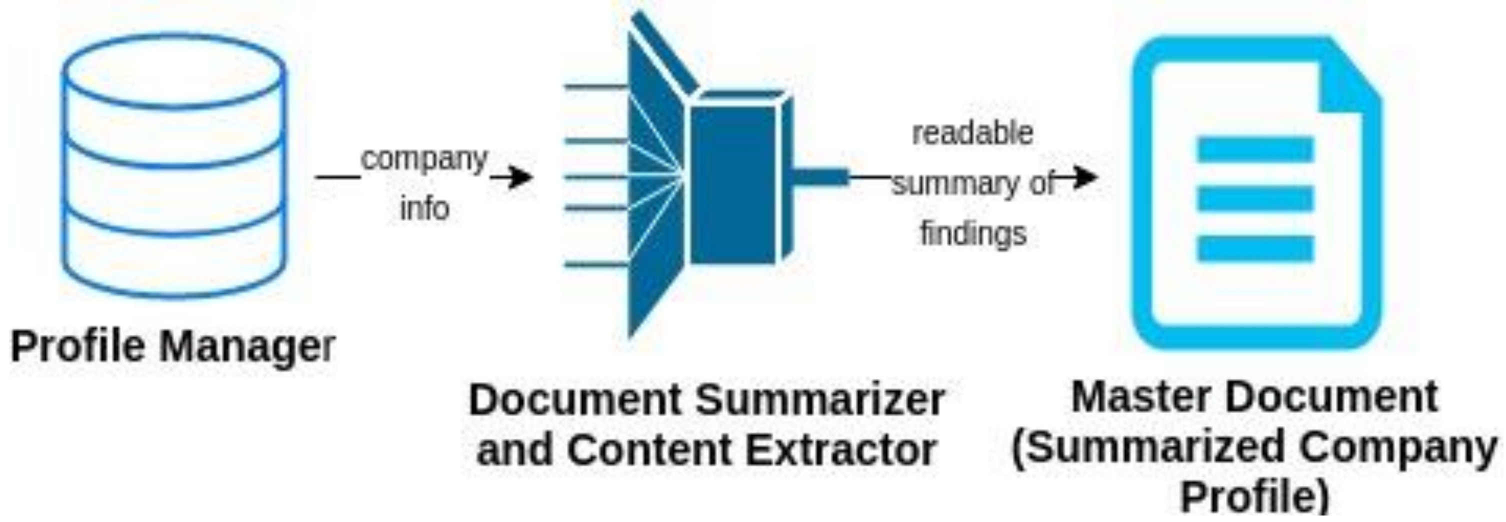
How we are hoping to use the knowledge graph

Credibility Checking and Classification

- A document that contains a statement of fact **P** that is contradictory to the set of possible-worlds as described by our knowledge graph is rejected.
- We can also use competitive learning to prioritize trustworthy sources.

Aggregates Information in a Queryable Format

- This makes the information much more accessible to the analysts at Praedicat, Inc. and makes the job of summarization easier


Profile Manager
company info
Document Summarizer and Content Extractor
readable summary of findings
Master Document (Summarized Company Profile)

Milestone 2

Questions?