

Exploring enigma.io

enigma.io is a company that is aggregating open data from accross the web into one platform and providing powerful ways to search and filter the data to discover insight. In this example we shall look at employment trends in software companies over the last number of years. By keeping the data within the enigma.io platform allows the use of powerful filters wihtout having to download, store and filter all of the data locally.

Step 1

Browse to enigma.io and search for **google**, you will be prompted to register an account or login.

Step 2

The result screen will display all the datasets that contain our search term. For this exercise we are going to look at the **Office of Foreign Labor Certification** datasets.

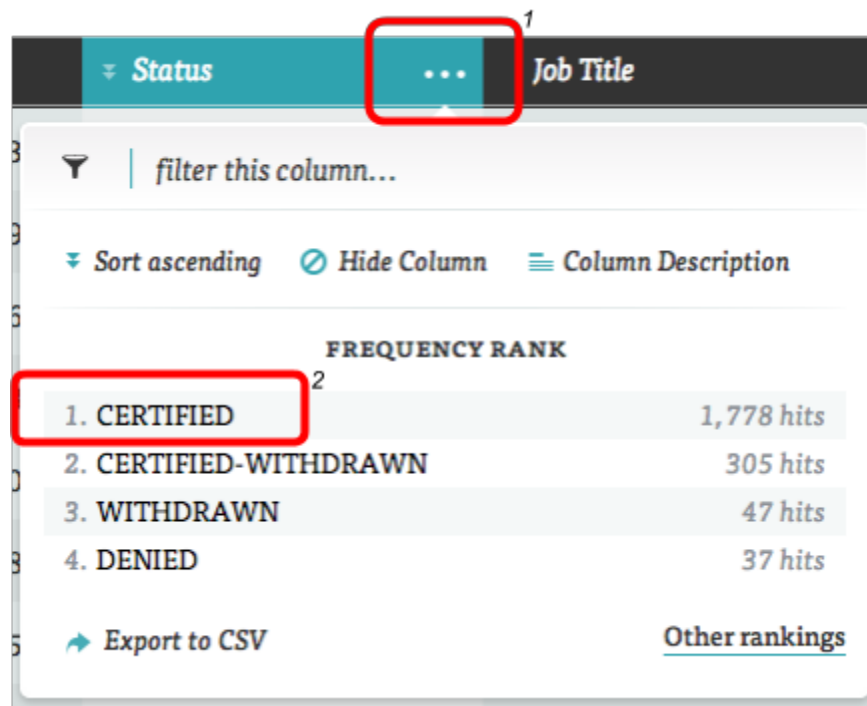
SOURCES	TOPICS	Sort
ENIGMA		
UNITED STATES	372	
COMPANIES	25	
GOVERNMENTS	7	
CURATED COLLECTIONS	3	
ORGANIZATIONS	1	

Office of Foreign Labor Certification		
H-1B Visa Applications 2013	Employer Name	Employer Address
The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations.	GOOGLE INC.	1600 AMPHITHEAT
2.2k hits 442.3k rows 36 columns	GOOGLE INC.	1600 AMPHITHEAT
	GOOGLE INC.	1600 AMPHITHEAT
H-1B Visa Applications 2012	Employer Name	Employer Address
The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations.	GOOGLE INC.	1600 AMPHITHEAT
1.7k hits 415.8k rows 36 columns	GOOGLE INC.	1600 AMPHITHEAT
	GOOGLE INC.	1600 AMPHITHEAT
H-1B Visa Applications 2011	Employer Name	Employer Address
The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations.	GOOGLE INC.	1600 AMPHITHEAT
1.3k hits 358.9k rows 28 columns	GOOGLE INC.	1600 AMPHITHEAT
	GOOGLE INC.	1600 AMPHITHEAT

This dataset lists the non-immigrant visa applications in the US allowing companies to temporarily employ foreign workers. Start by exploring the latest dataset to see what sort of information it contains.

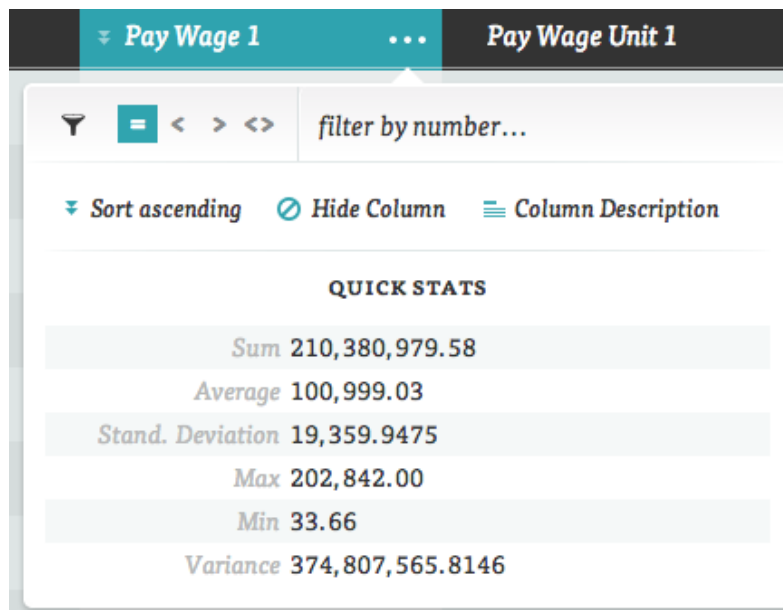
Step 3

For this exercise we would like to discover employment trends within some of the major software companies and chart the rise and fall of certain job titles. Starting with the latest dataset we first need to filter out those applications that were not certified. The data relating to this is contained in the **status** column. To view a summary of which types of status exist for Google, click the three ... that appear when you hover over the status column.



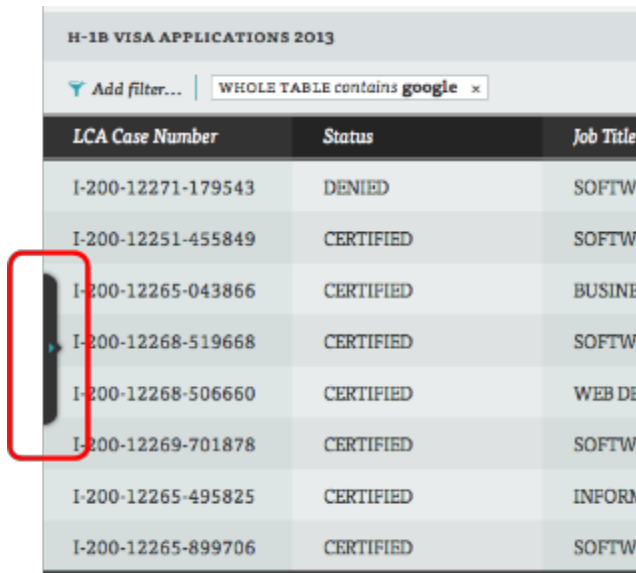
We want to filter this column for only those applications that were certified. From the drop down simply click the **certified** option in the frequency rank list.

With the data filtered it is now possible to look at how much Google spent on employing temporary non-immigrant workers within the US. To view the stats, take a look at the drop down stats from the **Pay Wage 1** column.



Step 4

In order to make sense of the data we are going to group it by **Job Title**, to do this first pull in the **tools sidebar**.

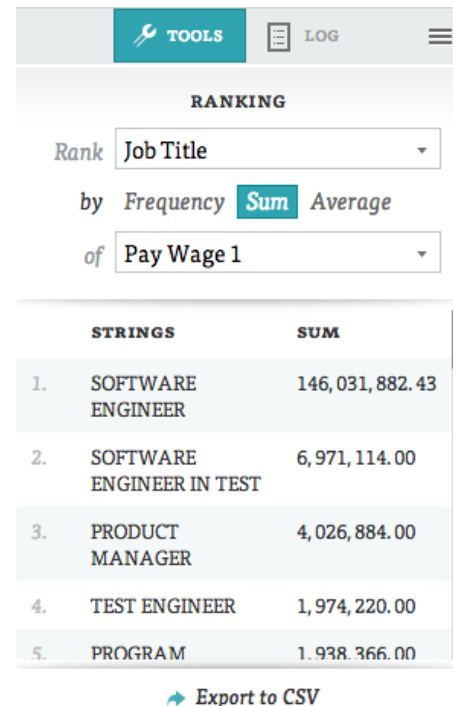


H-1B VISA APPLICATIONS 2013

Add filter... | WHOLE TABLE contains google x

LCA Case Number	Status	Job Title
I-200-12271-179543	DENIED	SOFTWARE ENGINEER
I-200-12251-455849	CERTIFIED	SOFTWARE ENGINEER
I-200-12265-043866	CERTIFIED	BUSINESS DEVELOPER
I-200-12268-519668	CERTIFIED	SOFTWARE ENGINEER
I-200-12268-506660	CERTIFIED	WEB DEVELOPER
I-200-12269-701878	CERTIFIED	SOFTWARE ENGINEER
I-200-12265-495825	CERTIFIED	INFORMATION SYSTEMS MANAGER
I-200-12265-899706	CERTIFIED	SOFTWARE ENGINEER

We can then group the data by **Job Title** summed by the numerical values in the **Pay Wage 1** column.



TOOLS LOG

RANKING

Rank Job Title

by Frequency Sum Average

of Pay Wage 1

	STRINGS	SUM
1.	SOFTWARE ENGINEER	146,031,882.43
2.	SOFTWARE ENGINEER IN TEST	6,971,114.00
3.	PRODUCT MANAGER	4,026,884.00
4.	TEST ENGINEER	1,974,220.00
5.	PROGRAM	1.938.366.00

Export to CSV

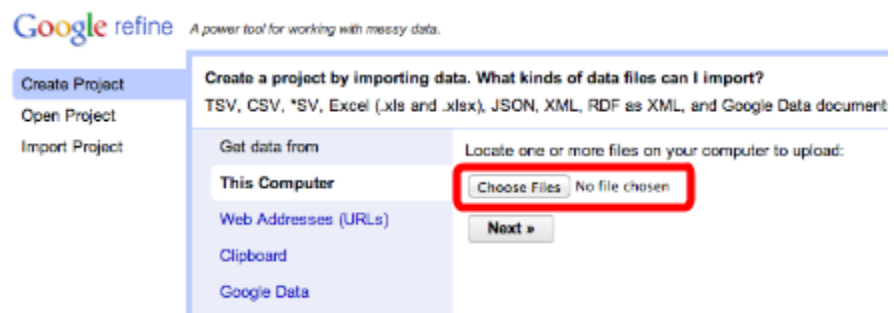
Finally we can export this data to csv in order to build a dataset for visual analysis.

Step 5

Now repeat all the steps for **facebook** in order to end up with two CSV files.

Step 6

The next step involves cleaning the data. All data is dirty, so we need a copy of Google/Open Refine. Upload both datasets into refine, creating a separate project for each one.



Google refine A power tool for working with messy data.

Create Project
Open Project
Import Project

Create a project by importing data. What kinds of data files can I import?
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data document

Get data from
This Computer
Web Addresses (URLs)
Clipboard
Google Data

Locate one or more files on your computer to upload:
Choose Files No file chosen

Next »

« Start Over Configure Parsing Options

Project name: | | lca_case_job_title | sum |
| --- | --- | --- |
| 1. | SOFTWARE ENGINEER | 38527255 |
| 2. | PRODUCTION ENGINEER | 3099114 |
| 3. | RESEARCH SCIENTIST | 2788694 |
| 4. | PRODUCT DESIGNER | 1496789 |
| 5. | DATA SCIENTIST | 1364951 |
| 6. | ENGINEERING MANAGER | 1052265 |
| 7. | DATA ENGINEER | 933180 |

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

Character encoding:

Columns are separated by: ☒ commas (CSV) ☐ tabs (TSV) ☐ custom:

Escape special characters with:

Update Preview

☐ Ignore first 0 line(s) at beginning of file

☒ Parse next 1 line(s) as column headers

☐ Discard 0 row(s) at end of file

With the data loaded, apply a **text facet** to the job title column. From the facet box, click the **cluster** button.

lca_case_job_title

92 choices Sort by: name count

Cluster

ACCOUNT MANAGER 1

ANALYST, MONETIZATION ANALYTICS 1

ANALYST, VERTICAL MEASUREMENT - CPG 1

ANALYTICS ENGINEER 1

APPLICATION ENGINEER 1

ASSOCIATE SMR 1

lca_case_job_title sum

Facet Text facet

Text filter Numeric facet

Edit cells Timeline facet

Edit column Scatterplot facet

Transpose Custom text facet...

Sort... Custom numeric facet...

View Customized facets

Reconcile 579392

573712

The clustering function allows you to fix errors in the dataset by grouping groups of cells. Select a number of different methods and keying functions in order to group together rows where the jobs are very similar in nature.

Method: Keying Function:

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
6	6	<ul style="list-style-type: none"> SOFTWARE ENGINEER (1 rows) SOFTWARE ENGINEER, INFRASTRUCTURE (1 rows) SOFTWARE ENGINEER, MOBILE IOS (1 rows) SOFTWARE ENGINEER, PRODUCTS (1 rows) SOFTWARE ENGINEER, SYSTEMS (1 rows) SOFTWARE ENGINEERING MANAGER (1 rows) 	<input type="checkbox"/>	SOFTWARE ENGINEER
2	2	<ul style="list-style-type: none"> PRODUCT MANAGER (1 rows) PRODUCT MANAGER (FINANCIAL ANALYSIS) (1 rows) 	<input type="checkbox"/>	PRODUCT MANAGER
2	2	<ul style="list-style-type: none"> INTERNET MARKETING ANALYST (MOBILE DATA) (1 rows) INTERNET MARKETING ANALYST (MOBILE ENGINEERING) (1 rows) 	<input type="checkbox"/>	INTERNET MARKETING AN/

Step 7

With both datasets cleaned it is time to align them into one single dataset. To do this upload them both to Google sheets at docs.google.com. Once done open the one relating to google and create a new column to the right of the google salary data (you might also want to rename the columns for clarity). Title this column **facebook salary**. Finally scroll to the bottom of this dataset.

The next stage is to copy **just** the job title data from the facebook dataset and paste it under the job titles in the google spreadsheet. Now do the same with the salary data **remembering to put it in column 3 and not column 2!**

At this point we should have a several hundred row spreadsheet with 3 columns of data. Now we can group the data again using a pivot table.

To create a pivot table in google sheets select the **pivot table report** option from the **data** menu.

The screenshot shows a Google Sheets spreadsheet with a pivot table. The pivot table has three columns: 'SUM of Google', 'SUM of Facebook', and an empty column. The rows list various job titles. To the right, the 'Report Editor' panel is open, showing the configuration for the pivot table. The 'Rows' section is set to 'Job Title'. The 'Columns' section is set to 'Google' and 'Facebook'. The 'Values' section is set to 'SUM' for both 'Google' and 'Facebook'. The 'Filter' section is empty. A red circle highlights the 'Rows', 'Columns', and 'Values' sections of the Report Editor.

	SUM of Google	SUM of Facebook
ACCOUNT ANAL	200783	0
ACCOUNT EXE	160436	0
ACCOUNT MAN	68931	138757
ACCOUNT STR	55411	0
ACCOUNTANT	85592	0
ACCOUNTING M	106246	0
AD SERVING S	324126	0
AGENCY LEAD	138674	0
ANALYST - TEC	450217	0
ANALYST (HR)	83990	0
ANALYST, MON	0	81264
ANALYST, VER	0	77358
ANALYTICAL LI	270442	0
ANALYTICAL LI	137321	0
ANALYTICS EN	0	100409
APPLICATION I	105500	0
APPLICATION I	0	153356
ASSOC ENTER	55536	0
ASSOCIATE MA	103400	0
ASSOCIATE PR	79976	0
ASSOCIATE PR	110822	0
ASSOCIATE PR	110781	0
ASSOCIATE PR	751734	0

By selecting the correct **rows** and **values** from the report editor (see above) we can group the data together, making our final report of facebook vs google job title and spending.

Extension exercises

Can you think of a way to visualise the data?

What are the problems with the data and how might you go about resolving them?

Can you add data from other companies and years to tell a more interesting story?