Data Science Skills

Team Tidy

Alec McCabe

Chinedu Onyeka

Cliff Lee

Preston Peck

Santiago Torres

Agenda

Section	Team Member	Time
Approach ● Tools ● Assumptions	Santiago	1 minute
Data Collection ● Web Scraping ● Persistent Storage	Alec & Cliff	3 minutes
Data Transformation ● EMSI	Preston	1 minute
Data Analysis	Chinedu & Preston	2 minutes
Conclusion	Santiago	1 minute

Approach

Approach

For DATA 607 Project 3, all teams must use data to answer the question, “Which are the most valued data science skills?” Consider your work as an exploration; there is not necessarily a “right answer.”

Gather job postings and survey what the job market finds valuable

LinkedIn

stack overflow

indeed

Collaboration Tools

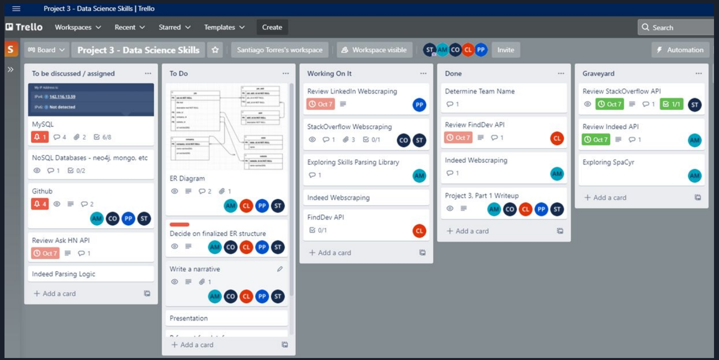
Communication
- Slack
- Zoom

Project Documentation
- Google Drive

Project Management
- Trello

Code Sharing
- GitHub

https://github.com/cliftonleesps/team_tidy



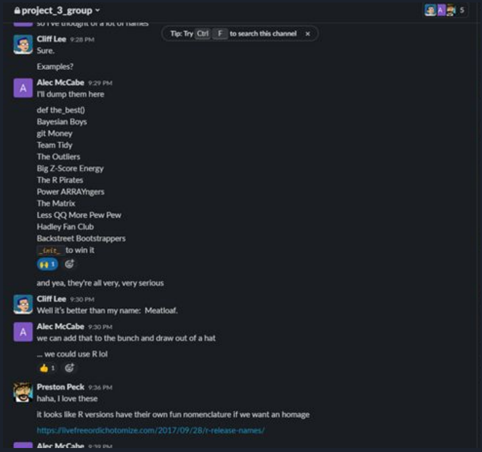
To be discussed / assigned
- MySQL — 1, 4, 2, 6/8
- NoSQL Databases - neo4j, mongo, etc — 1, 0/2
- Github — 4, 2 — AM CO PP ST
- Review Ask HN API — Oct 7, 1
- Indeed Parsing Logic

To Do
- ER Diagram — 2, 1 — AM CL PP ST
- Decide on finalized ER structure — AM CO CL PP ST
- Write a narrative — 1 — AM CO CL PP ST
- Presentation

Working On It
- Review LinkedIn Webscraping — Oct 7 — PP
- StackOverflow Webscraping — 1, 3, 0/1 — CO ST
- Exploring Skills Parsing Library — 1 — AM
- Indeed Webscraping
- FindDev API — 0/1 — CL

Done
- Determine Team Name — 1
- Review FindDev API — Oct 7, 1 — CL
- Indeed Webscraping — 1 — AM
- Project 3. Part 1 Writeup — AM CO CL PP ST

Graveyard
- Review StackOverflow API — Oct 7, 1, 1/1 — ST
- Review Indeed API — Oct 7, 1 — AM
- Exploring SpaCyr — AM



so I've thought of a lot of names

Cliff Lee 9:28 PM
Sure.

Examples?

Alec McCabe 9:29 PM
I'll dump them here

def the_best()
Bayesian Boys
git Money
Team Tidy
The Outliers
Big Z-Score Energy
The R Pirates
Power ARRAYngers
The Matrix
Less QQ More Pew Pew
Hadley Fan Club
Backstreet Bootstrappers
`_init_` to win it
🙌 1

and yea, they're all very, very serious

Cliff Lee 9:30 PM
Well it's better than my name: Meatloaf.

Alec McCabe 9:30 PM
we can add that to the bunch and draw out of a hat

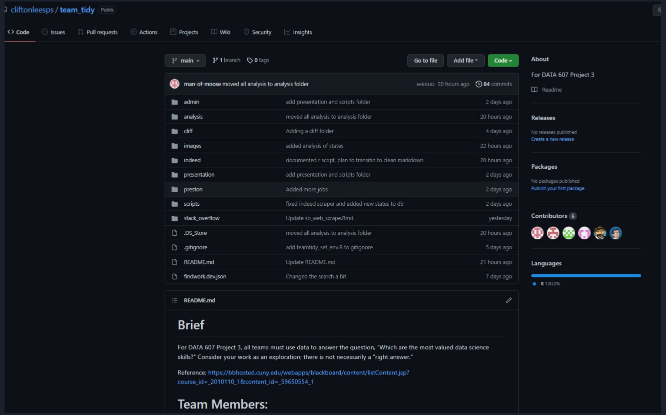
... we could use R lol
👍 1

Preston Peck 9:36 PM
haha, I love these

it looks like R versions have their own fun nomenclature if we want an homage

https://livefreeordichotomize.com/2017/09/28/r-release-names/

Alec McCabe 9:39 PM



man-of-moose moved all analysis to analysis folder — eeb5162 20 hours ago — 84 commits

admin	add presentation and scripts folder	2 days ago
analysis	moved all analysis to analysis folder	20 hours ago
cliff	Adding a cliff folder	4 days ago
images	added analysis of states	22 hours ago
indeed	documented r script, plan to transitin to clean markdown	20 hours ago
presentation	add presentation and scripts folder	2 days ago
preston	Added more jobs	2 days ago
scripts	fixed indeed scraper and added new states to db	2 days ago
stack_overflow	Update so_web_scrape.Rmd	yesterday
.DS_Store	moved all analysis to analysis folder	20 hours ago
.gitignore	add teamtidy_set_env.R to gitignore	5 days ago
README.md	Update README.md	21 hours ago
findwork.dev.json	Changed the search a bit	7 days ago

About

For DATA 607 Project 3

Readme

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 6

Languages

R 100.0%

README.md

Brief

For DATA 607 Project 3, all teams must use data to answer the question, "Which are the most valued data science skills?" Consider your work as an exploration; there is not necessarily a "right answer."

Reference: https://bbhosted.cuny.edu/webapps/blackboard/content/listContent.jsp?course_id=_2010110_1&content_id=_59650554_1

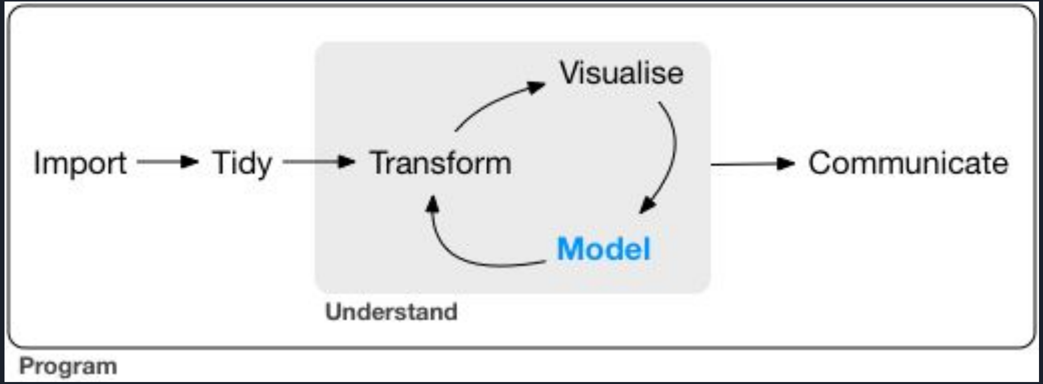
Team Members:

Assumptions

We'd see skills that fall along equally on each part of the model:

- Import / Tidy / Transform skills
- Visualization skills
- Modeling skills
- Communication skills

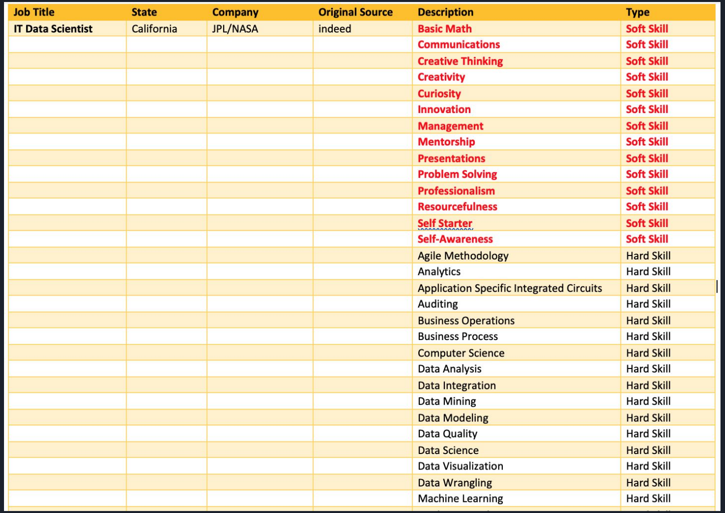
Hadley's Data Science Model



Data Collection

Persistent Storage

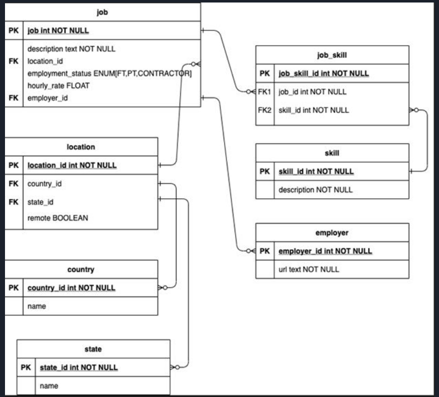
MySQL™



Job Title	State	Company	Original Source	Description	Type
IT Data Scientist	California	JPL/NASA	indeed	**Basic Math**	**Soft Skill**
				Communications	**Soft Skill**
				Creative Thinking	**Soft Skill**
				Creativity	**Soft Skill**
				Curiosity	**Soft Skill**
				Innovation	**Soft Skill**
				Management	**Soft Skill**
				Mentorship	**Soft Skill**
				Presentations	**Soft Skill**
				Problem Solving	**Soft Skill**
				Professionalism	**Soft Skill**
				Resourcefulness	**Soft Skill**
				Self Starter	**Soft Skill**
				Self-Awareness	**Soft Skill**
				Agile Methodology	Hard Skill
				Analytics	Hard Skill
				Application Specific Integrated Circuits	Hard Skill
				Auditing	Hard Skill
				Business Operations	Hard Skill
				Business Process	Hard Skill
				Computer Science	Hard Skill
				Data Analysis	Hard Skill
				Data Integration	Hard Skill
				Data Mining	Hard Skill
				Data Modeling	Hard Skill
				Data Quality	Hard Skill
				Data Science	Hard Skill
				Data Visualization	Hard Skill
				Data Wrangling	Hard Skill
				Machine Learning	Hard Skill

Sample Data
(Soft Skills in Red)

ER Diagram



job	
PK	job int NOT NULL
	description text NOT NULL
FK	location_id
	employment_status ENUM[FT,PT,CONTRACTOR]
	hourly_rate FLOAT
FK	employer_id

job_skill	
PK	job_skill_id int NOT NULL
FK1	job_id int NOT NULL
FK2	skill_id int NOT NULL

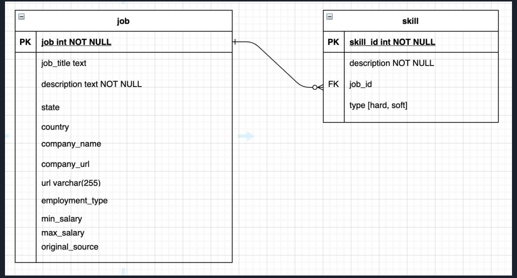
location	
PK	location_id int NOT NULL
FK	country_id
FK	state_id
	remote BOOLEAN

skill	
PK	skill_id int NOT NULL
	description NOT NULL

employer	
PK	employer_id int NOT NULL
	url text NOT NULL

country	
PK	country_id int NOT NULL
	name

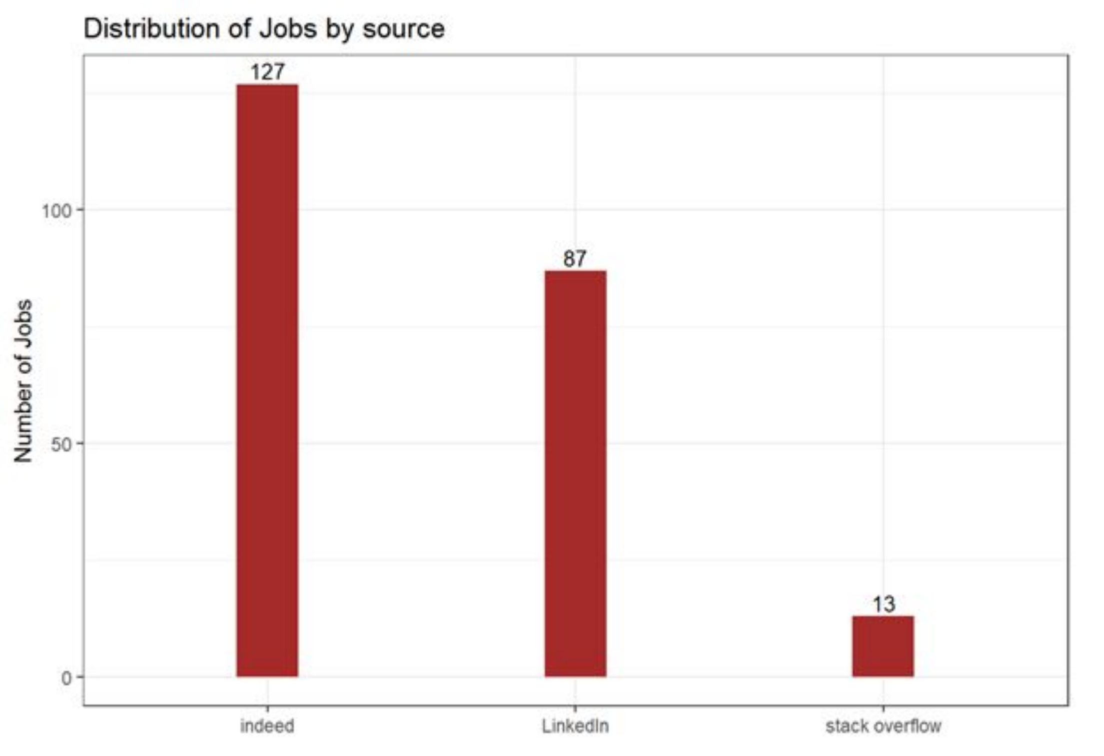
state	
PK	state_id int NOT NULL
	name



job	
PK	job int NOT NULL
	job_title text
	description text NOT NULL
	state
	country
	company_name
	company_url
	url varchar(255)
	employment_type
	min_salary
	max_salary
	original_source

skill	
PK	skill_id int NOT NULL
	description NOT NULL
FK	job_id
	type [hard, soft]

Webscraping



Web-scraped 220 jobs from:

- ○ Indeed
- ○ LinkedIn
- ○ StackOverfow

Technologies used:

- ○ Xml2
- ○ Rselenium

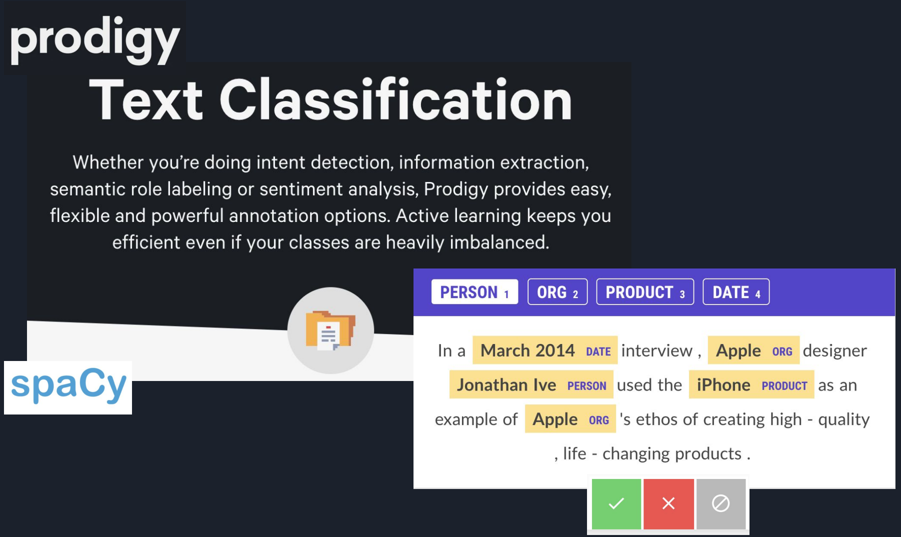
Similar structure between websites:

- ○ Jobcards collected from page
- ○ Iterate through pages

Data Transformation

Data Transformation
Extracting meaning

- With all our **job description text**, how do we extract the key **skill data** effectively and efficiently?
- We could parse manually or hope for predictable patterns to parse by, but this is tedious and inconsistent
- We explored options for Natural Language Processing (NLP) services with Named Entity Recognition (NER) models...



prodigy

Text Classification

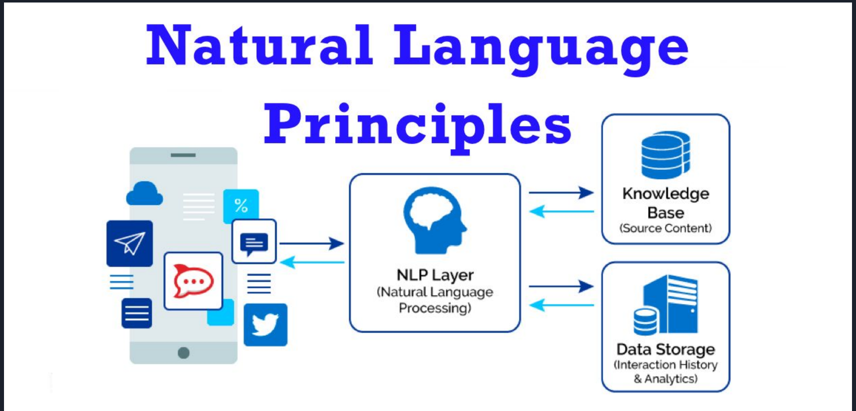
Whether you're doing intent detection, information extraction, semantic role labeling or sentiment analysis, Prodigy provides easy, flexible and powerful annotation options. Active learning keeps you efficient even if your classes are heavily imbalanced.

spaCy

PERSON 1 ORG 2 PRODUCT 3 DATE 4

In a March 2014 DATE interview , Apple ORG designer Jonathan Ive PERSON used the iPhone PRODUCT as an example of Apple ORG 's ethos of creating high - quality , life - changing products .

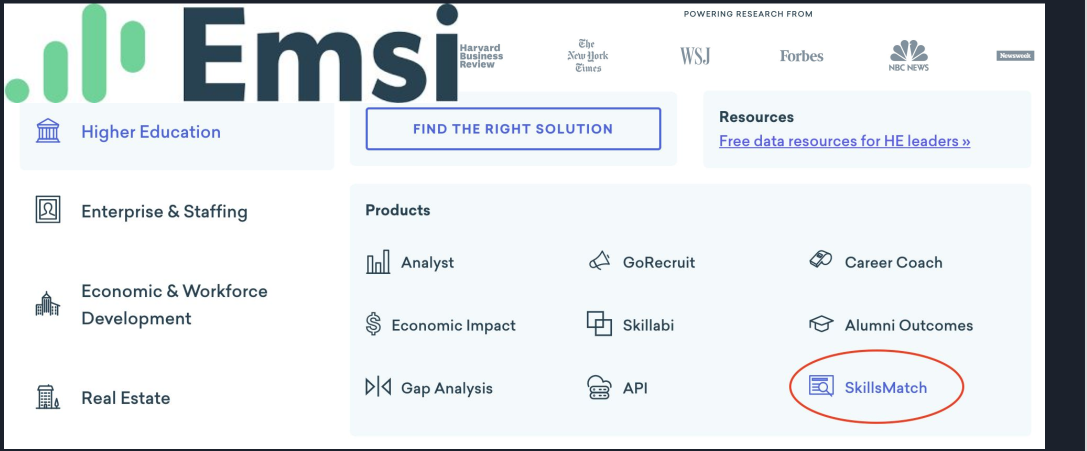
✓ ✕ ⊘



Natural Language Principles
NLP Layer
(Natural Language Processing)
Knowledge Base
(Source Content)
Data Storage
(Interaction History & Analytics)
%

Data Transformation

NLP using Labor Market Analytics & Economic Data API



Emsi

POWERING RESEARCH FROM

Higher Education

Enterprise & Staffing

Economic & Workforce Development

Real Estate

FIND THE RIGHT SOLUTION

Resources

Free data resources for HE leaders »

Products

Analyst

GoRecruit

Career Coach

Economic Impact

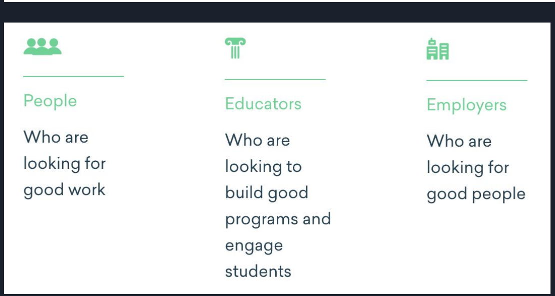
Skillabi

Alumni Outcomes

Gap Analysis

API

SkillsMatch



People

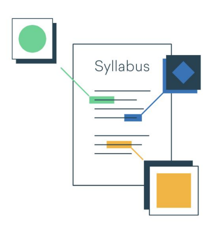
Who are looking for good work

Educators

Who are looking to build good programs and engage students

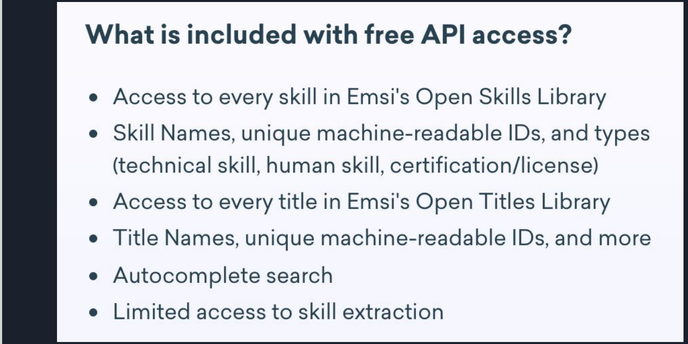
Employers

Who are looking for good people



OpenSkills API

"30,000+ skills that we've collected from hundreds of millions of job postings, resumes, and online profiles"

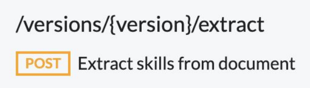


What is included with free API access?

- Access to every skill in Emsi's Open Skills Library
- Skill Names, unique machine-readable IDs, and types (technical skill, human skill, certification/license)
- Access to every title in Emsi's Open Titles Library
- Title Names, unique machine-readable IDs, and more
- Autocomplete search
- Limited access to skill extraction

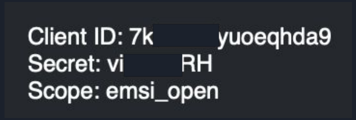
Data Transformation

Emsi Skills API

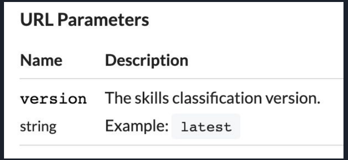


/versions/{version}/extract

POST Extract skills from document

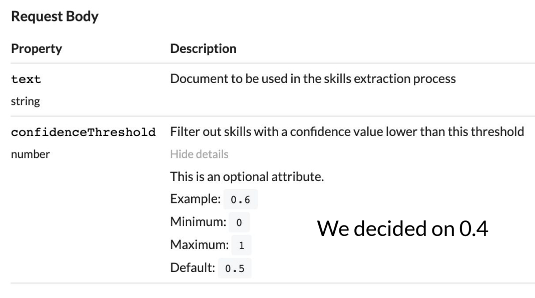


Client ID: 7k [redacted] yuoeqhda9
Secret: vi [redacted] RH
Scope: emsi_open



URL Parameters

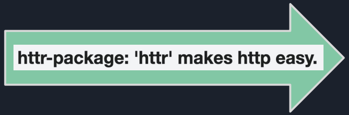
Name	Description
`version` string	The skills classification version. Example: `latest`



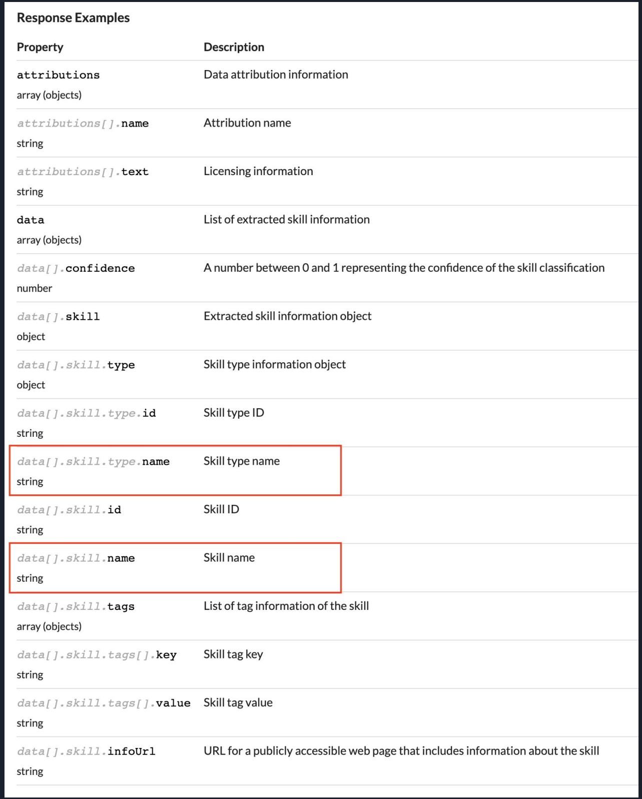
Request Body

Property	Description
`text` string	Document to be used in the skills extraction process
`confidenceThreshold` number	Filter out skills with a confidence value lower than this threshold Hide details This is an optional attribute. Example: `0.6` Minimum: `0` Maximum: `1` Default: `0.5`

We decided on 0.4



httr-package: 'httr' makes http easy.



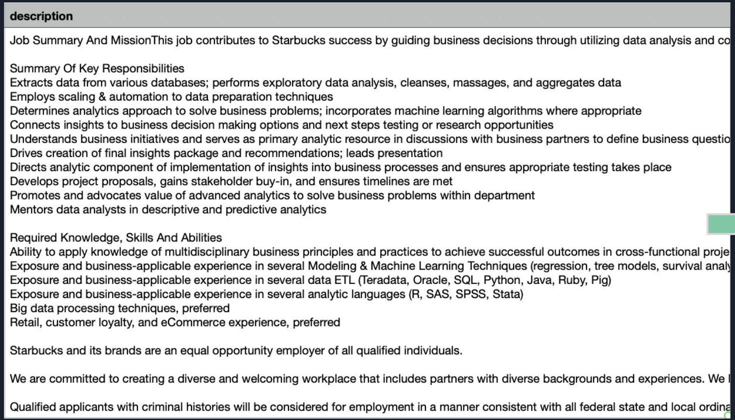
Response Examples

Property	Description
`attributions` array (objects)	Data attribution information
`attributions[].name` string	Attribution name
`attributions[].text` string	Licensing information
`data` array (objects)	List of extracted skill information
`data[].confidence` number	A number between 0 and 1 representing the confidence of the skill classification
`data[].skill` object	Extracted skill information object
`data[].skill.type` object	Skill type information object
`data[].skill.type.id` string	Skill type ID
`data[].skill.type.name` string	Skill type name
`data[].skill.id` string	Skill ID
`data[].skill.name` string	Skill name
`data[].skill.tags` array (objects)	List of tag information of the skill
`data[].skill.tags[].key` string	Skill tag key
`data[].skill.tags[].value` string	Skill tag value
`data[].skill.infoUrl` string	URL for a publicly accessible web page that includes information about the skill

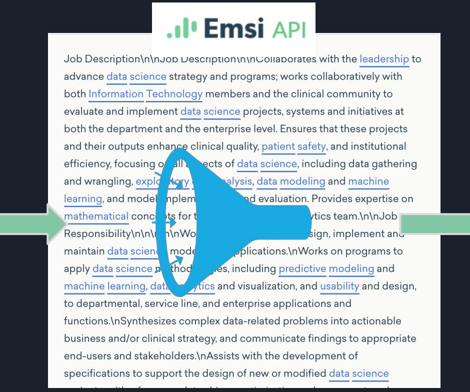
Data Transformation
Summary

Input: ***Job Description***

Output: ***Skills***

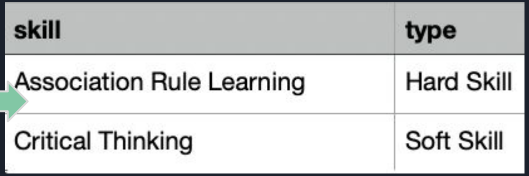


description
Job Summary And MissionThis job contributes to Starbucks success by guiding business decisions through utilizing data analysis and co Summary Of Key Responsibilities Extracts data from various databases; performs exploratory data analysis, cleanses, massages, and aggregates data Employs scaling & automation to data preparation techniques Determines analytics approach to solve business problems; incorporates machine learning algorithms where appropriate Connects insights to business decision making options and next steps testing or research opportunities Understands business initiatives and serves as primary analytic resource in discussions with business partners to define business questio Drives creation of final insights package and recommendations; leads presentation Directs analytic component of implementation of insights into business processes and ensures appropriate testing takes place Develops project proposals, gains stakeholder buy-in, and ensures timelines are met Promotes and advocates value of advanced analytics to solve business problems within department Mentors data analysts in descriptive and predictive analytics Required Knowledge, Skills And Abilities Ability to apply knowledge of multidisciplinary business principles and practices to achieve successful outcomes in cross-functional proje Exposure and business-applicable experience in several Modeling & Machine Learning Techniques (regression, tree models, survival analy Exposure and business-applicable experience in several data ETL (Teradata, Oracle, SQL, Python, Java, Ruby, Pig) Exposure and business-applicable experience in several analytic languages (R, SAS, SPSS, Stata) Big data processing techniques, preferred Retail, customer loyalty, and eCommerce experience, preferred Starbucks and its brands are an equal opportunity employer of all qualified individuals. We are committed to creating a diverse and welcoming workplace that includes partners with diverse backgrounds and experiences. We l Qualified applicants with criminal histories will be considered for employment in a manner consistent with all federal state and local ordina



Emsi API

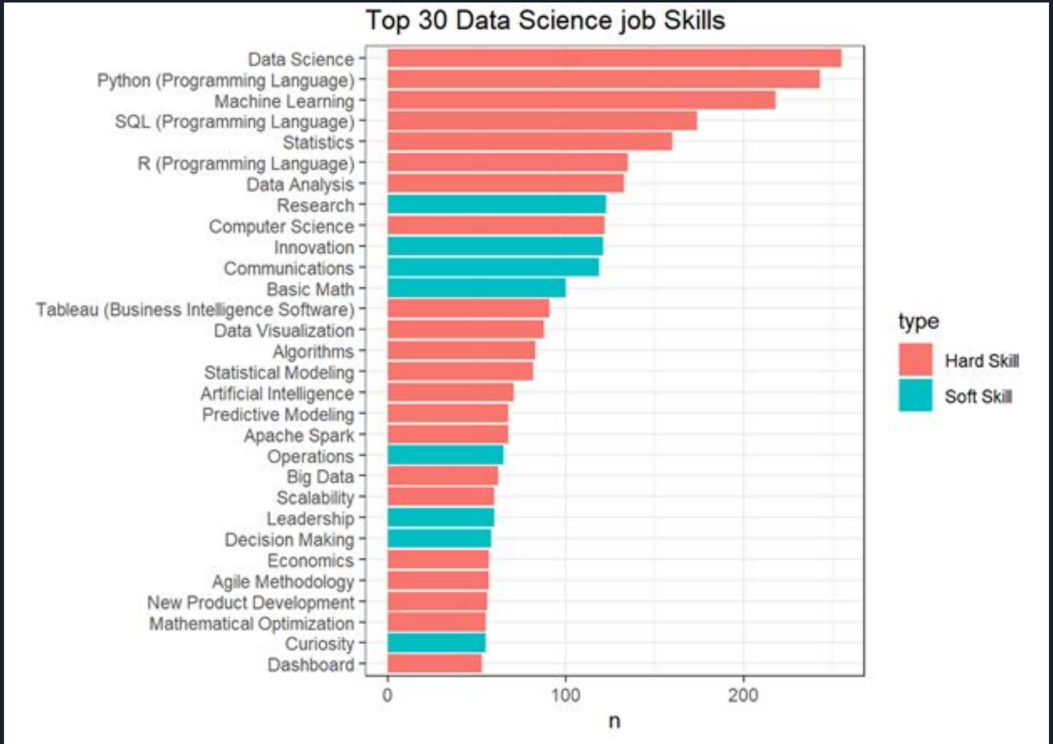
Job Description\n\nJob Description\n\nCollaborates with the leadership to advance data science strategy and programs; works collaboratively with both Information Technology members and the clinical community to evaluate and implement data science projects, systems and initiatives at both the department and the enterprise level. Ensures that these projects and their outputs enhance clinical quality, patient safety, and institutional efficiency, focusing o[illegible] all a[illegible]ects of data science, including data gathering and wrangling, explo[illegible]tory [illegible] [illegible]alysis, data modeling and machine learning, and model [illegible]mplem[illegible] [illegible]nd evaluation. Provides expertise on mathematical conc[illegible]pts for t[illegible] [illegible]ytics team.\n\nJob Responsibility\n\n\[illegible]n\nWo[illegible] [illegible]sign, implement and maintain data scien[illegible] mode[illegible] [illegible]pplications.\nWorks on programs to apply data science [illegible]ethod[illegible]es, including predictive modeling and machine learning, data [illegible]ytics and visualization, and usability and design, to departmental, service line, and enterprise applications and functions.\nSynthesizes complex data-related problems into actionable business and/or clinical strategy, and communicate findings to appropriate end-users and stakeholders.\nAssists with the development of specifications to support the design of new or modified data science



skill	type
Association Rule Learning	Hard Skill
Critical Thinking	Soft Skill

Data Analysis

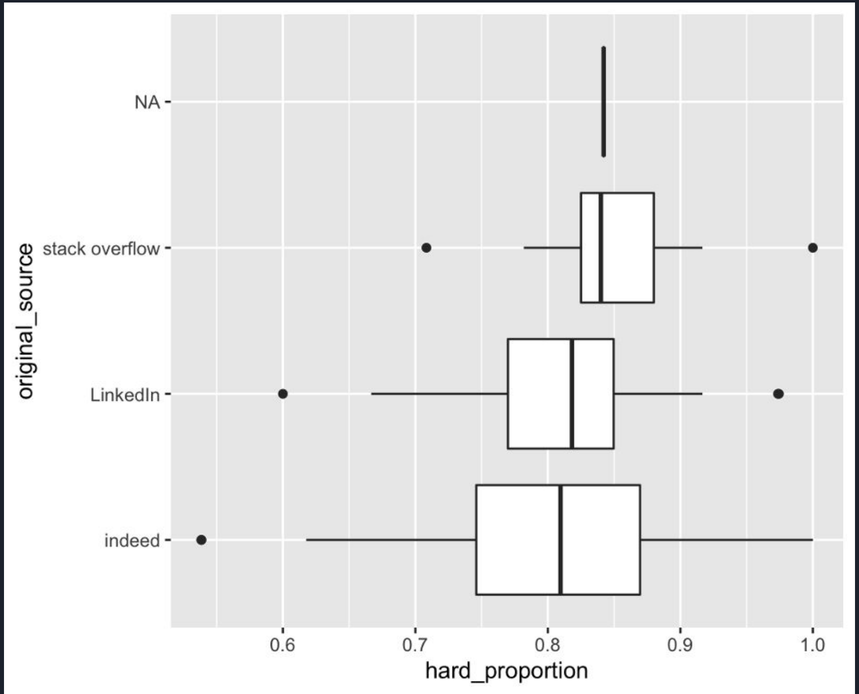
Top 30 Data Science Job Skills



- >900 Unique Skills
- Hard vs Soft Skills
    - ⅓ of top 12 are soft skills compared to 20% in overall dataset
- Programming skills highly ranked in top 30 skills

Statistical Modeling

“Does Stack Overflow’s distribution of hard_proportion in jobs have a statistically significant difference than the jobs of other sources”



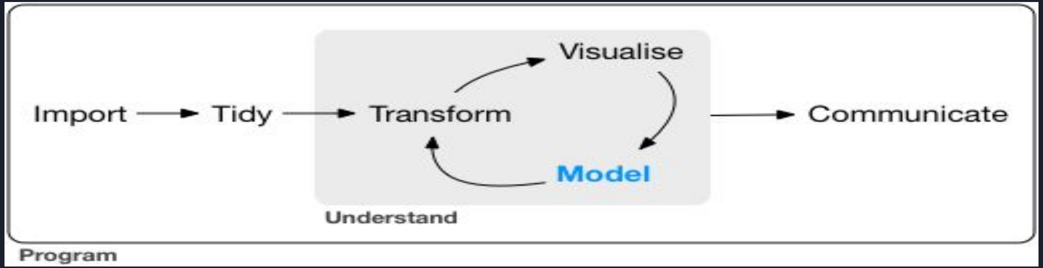
1. Visualize proportional differences
2. Stack Overflow displays a larger proportion of hard skills with a shorter spread
3. Is there a statistical difference?
4. Two Sample T-Test -> p-value = 0.1778
5. Fail to reject the null hypothesis

- Two Sample t-test
- data: so_props and other_props
- t = 1.3536, df = 161, p-value = 0.1778
- 95 percent confidence interval:
    a. -0.0168604 0.0903353
- sample estimates:
    a. mean of x mean of y
    b. 0.8475232 0.8107858

Conclusion

Most Valuable Data Science Skills

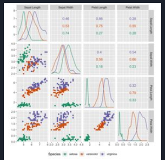
Hadley's Data Science Model



Programming

Import
Tidy
Transform
Visualise

Statistics & Machine Learning



Model

Communication

Communicate

Next Steps

1. Expand to other jobs / job sources
   a. POC for Data Science - what do other technology jobs look like?
   b. Increase sample sizes
2. Analyzing job titles
   a. We believe we may see different hard to soft skill proportion based on job title (Senior vs Staff Data Scientist)
3. Remove duplicates from sources
   a. Currently no checks in place to enforce uniqueness
4. Build our own model for skills categorization
   a. We depend heavily on EMSI to produce results