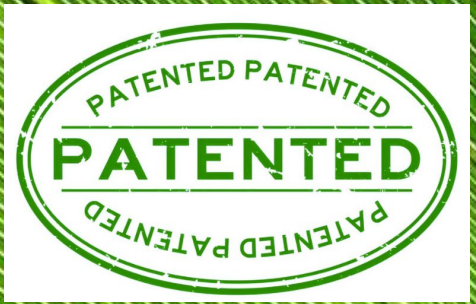
MAS DSE 203
Final Project



December 10, 2021

Josue Garcia, Eleanor Hope-Bell, Ifti Mirza

Questions we wanted to answer

1. **Do patents predict whether an AI company is more likely to receive capital funding?**

2. **What sub categories of AI are applying for patents?**

3. **What percentage of funding goes to women owned businesses?**

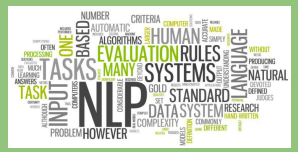
Planned Approach

1. Search / extract patent & funding data sources
2. Find ways to integrate those data sources
3. Build knowledge graph based on integrated dataset
4. Provide query access to the knowledge graph

Architecture

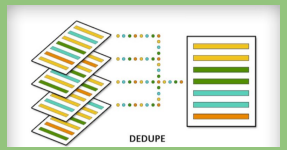
SBIR

Patents granted

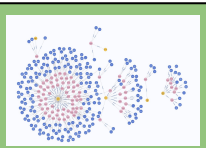


Venurebeat

Entity Matching



Knowledge Graph



Overview of Datasets

Structured

1. **SBIR** - Small Business Innovation Research - America's Seed Fund. Source: SBIR.gov, (189k rows x 36 columns).

2. **Location** - List of companies and their addresses. (~20 Mb)

3. **Assignee (**~25 Mb)

Semi- Structured

4. **Patents Granted** 6 Gb, excluded all withdrawn applications . Years 1976-2021 filtered for last 5 years.

Un-Structured

5. **Venturebeat -** Webscraped for articles containing "AI funding"

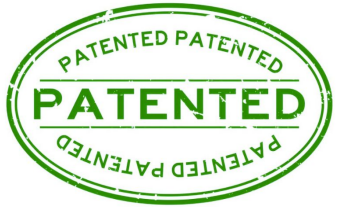
Dataset: Patent

Approach

1. US patents granted from the last 5 years
2. Filtered patents for AI classifications only
3. Topic extraction for each patent

Technology Used

1. Pytextrank for topic extraction
2. Jupyter Noteboook used for processing datasets
3. Dedupe using recordlinking



Dataset: SBIR

Approach

1. SBIR => The Small Business Innovation Research
2. “America’s Seed Fund” - Government funded research & development
3. Through 2019, over 179k awards have been made totaling more than $54.3 billion.

Technology Used

1. Jupyter Notebook
2. Dedupe using recordlinking

Dataset: Venturebeat

Approach

1. VentureBeat, a US **technology website**, publishes news

2. VentureBeat is the leader in coverage of **artificial intelligence and machine learning**, with two of our AI writers ranked as #1 and #3 respectively.

3. VentureBeat's unique audience of 6M monthly unique readers

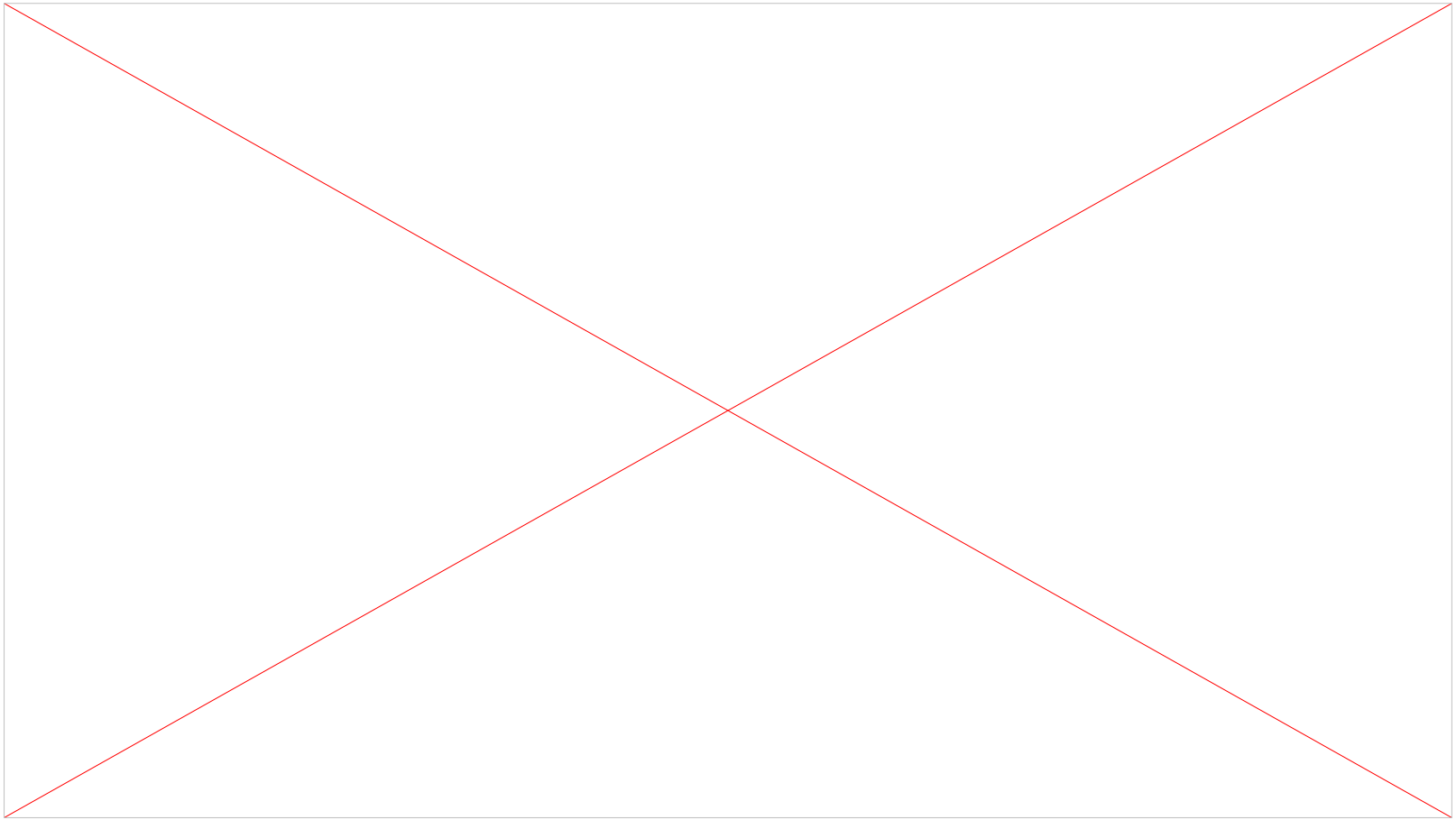
Technology Used

1. Jupyter Notebook

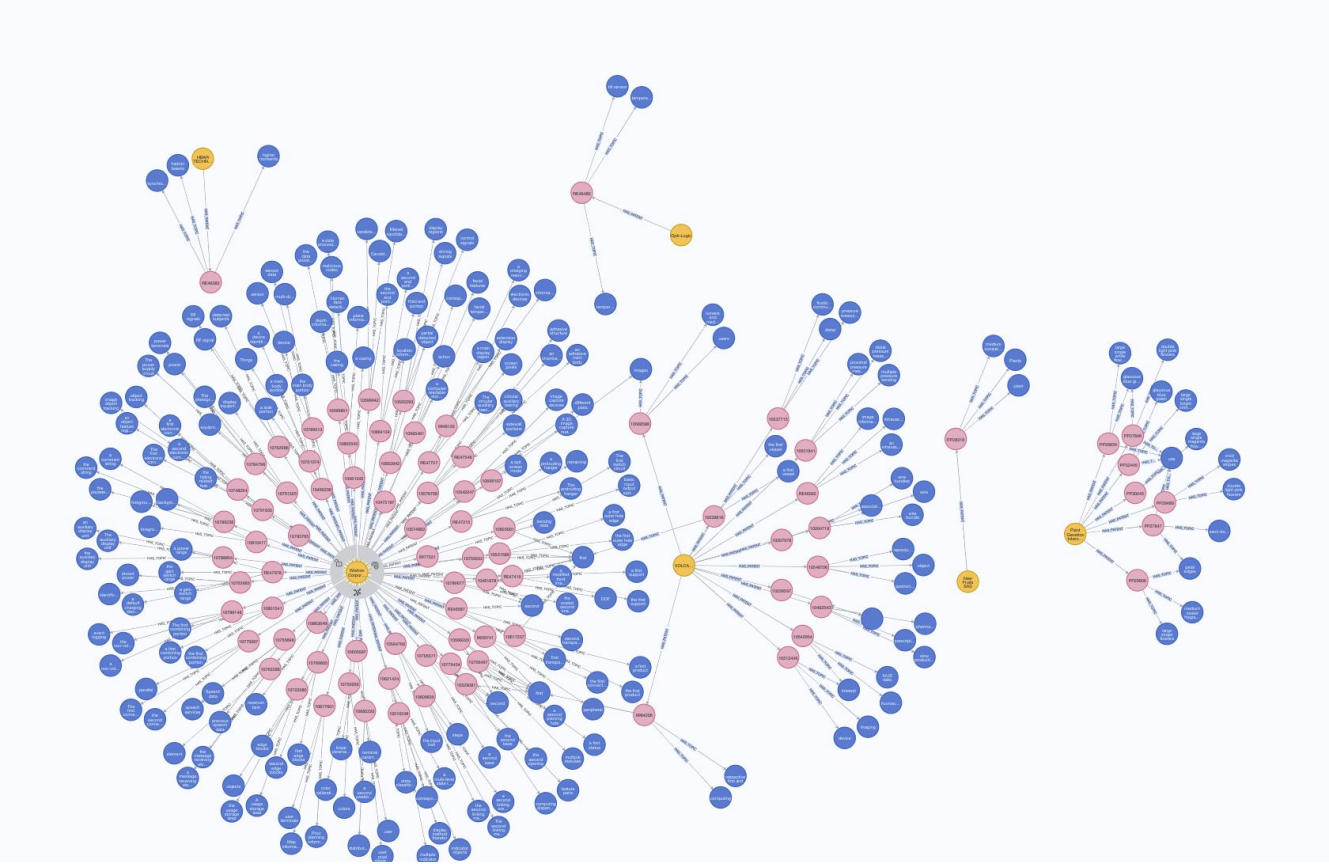
2. Selenium

3. Regex / user defined functions

Demo



Knowledgegraph Sample Result

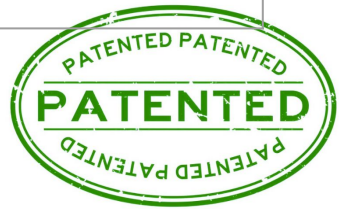


Lessons Learned & Next Steps

- Pytext extraction utilizes too much computing resources and time consuming
- Dedupe with large datasets requires more computing resources
- Neo4J crashed on large datasets
- Captcha issues when using Selenium
- Complicated web scraping process

Sources

Datasets:	
SBIR	https://www.sbir.gov
Angel List	https://angel.co
Angel List (processed)	https://github.com/rodrigosnader/angel-scraper/blob/master/data/startups.csv
Location and Assignee	https://patentsview.org/download/data-download-tables
Patents Granted	https://www.uspto.gov
Venturebeat	https://venturebeat.com



Questions?

Our GitHub repository:

https://github.com/Ifti007/dse-203-project

Thank you!