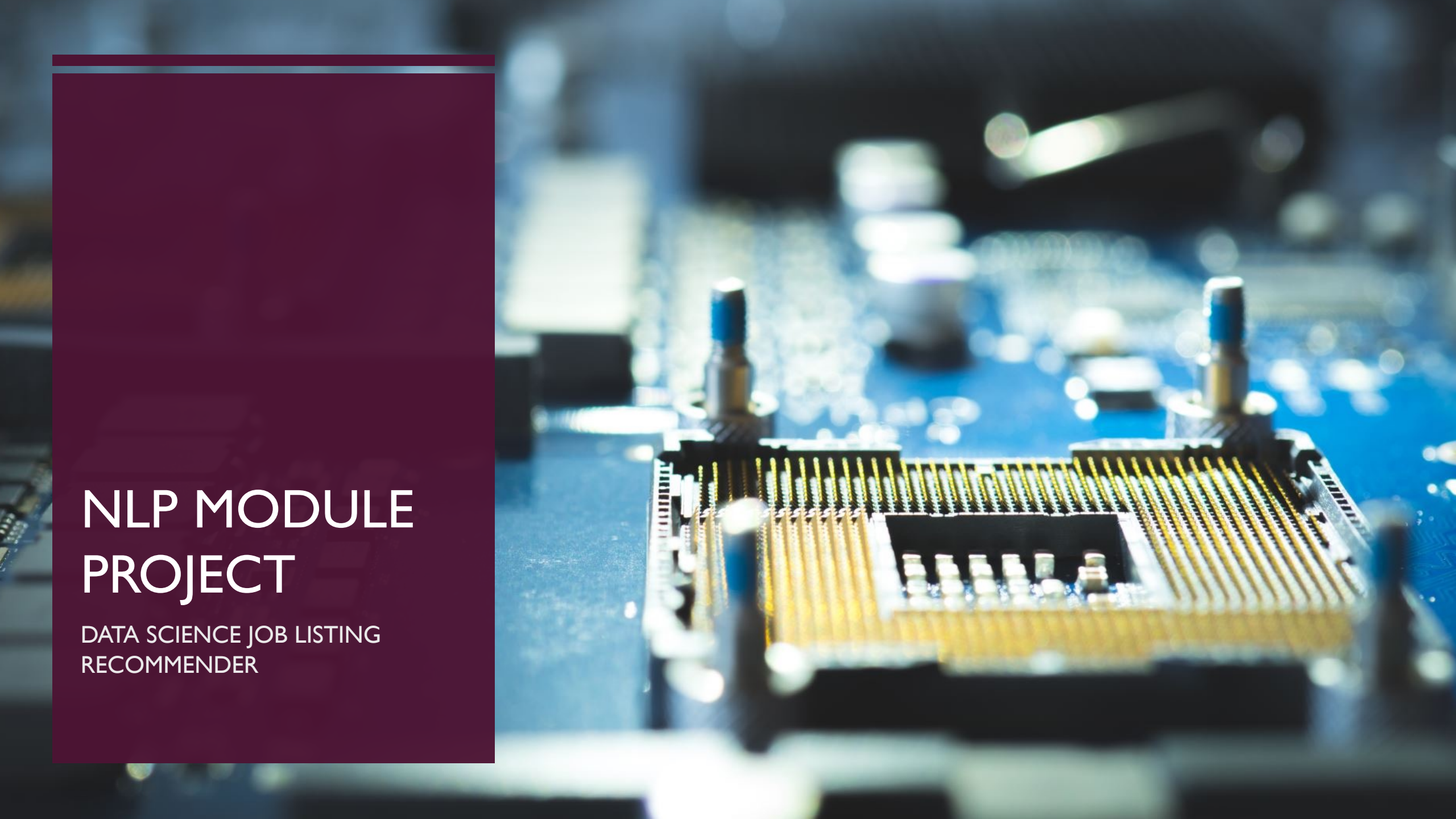


NLP MODULE PROJECT

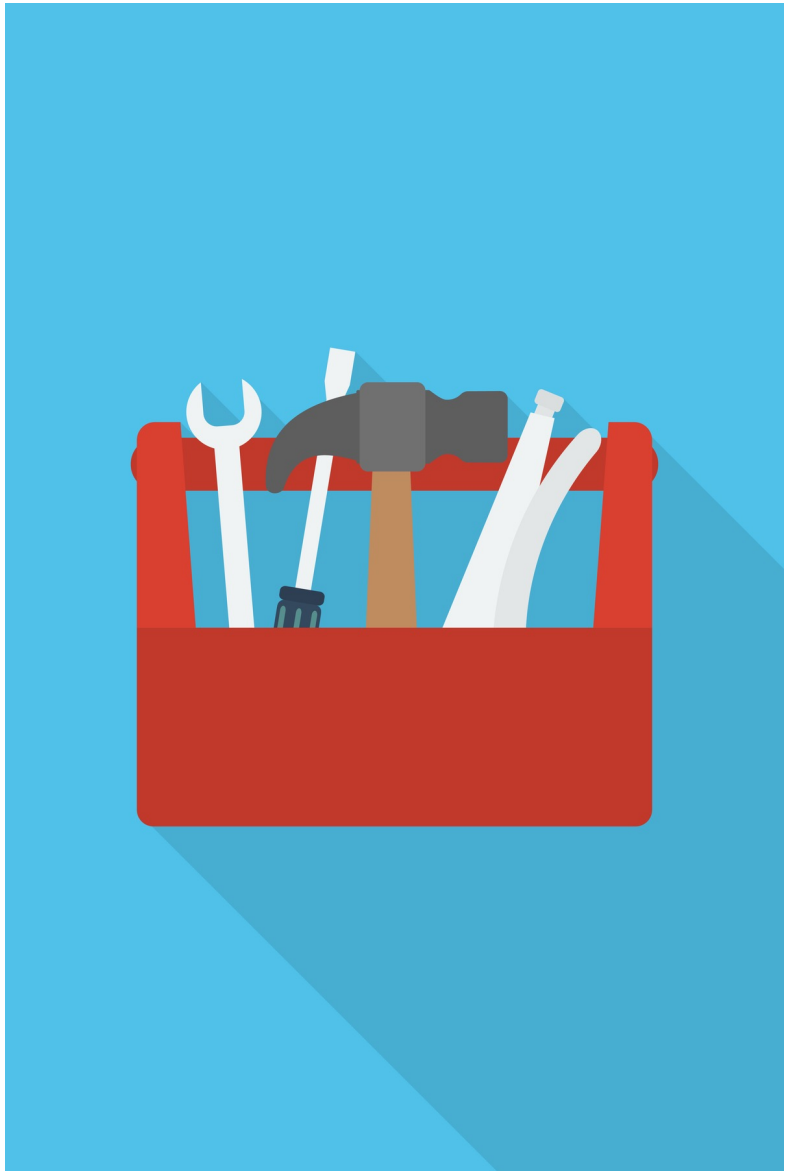
DATA SCIENCE JOB LISTING RECOMMENDER

INTRODUCTION

- Data science jobs vary significantly
- The language used to describe similar jobs can vary
- Data science job vacancies are very numerous
- Could an NLP-powered job listing recommender help?

METHODOLOGY

- Web-scraping: Selenium and Beautiful soup
- Data manipulation: Pandas
- Data storage: Pickle
- Text preprocessing: Python re and spaCy
- Vectoriser: TF-IDF
- Topic Modelling: NMF
- App: Streamlit

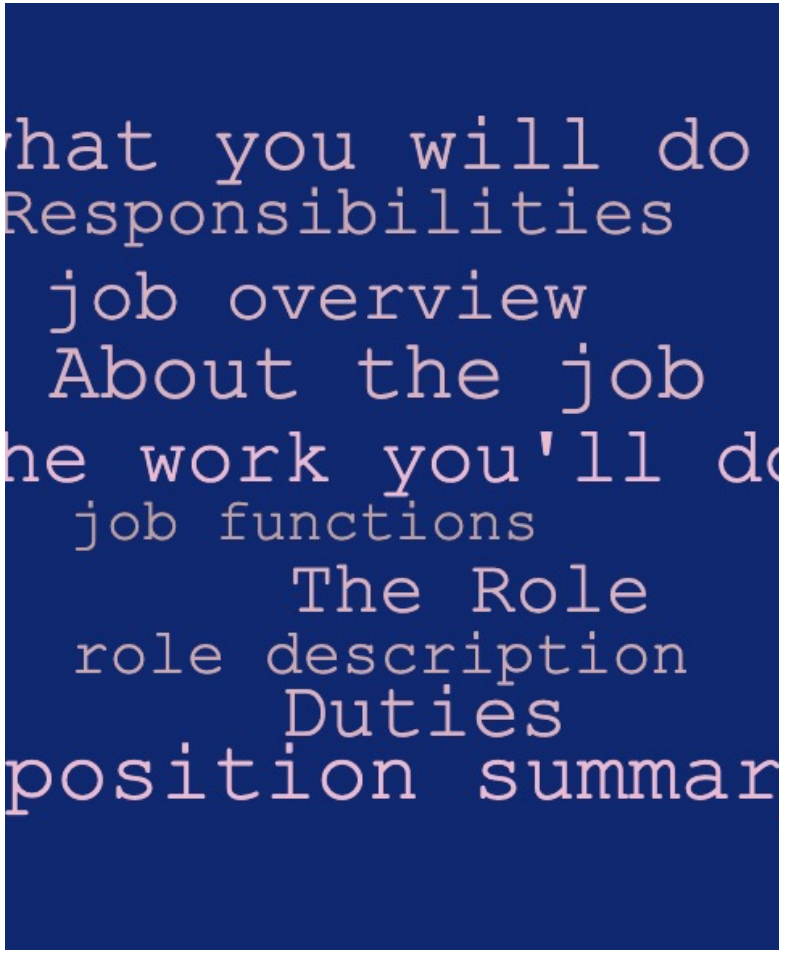


WEB SCRAPING AND DATA

- Linkedin search on "data scientist" in about 50 locations
- Job listings with "data scientist" or "machine learning" in title
- Nearly 13,000 listings
- Roughly one third duplicates

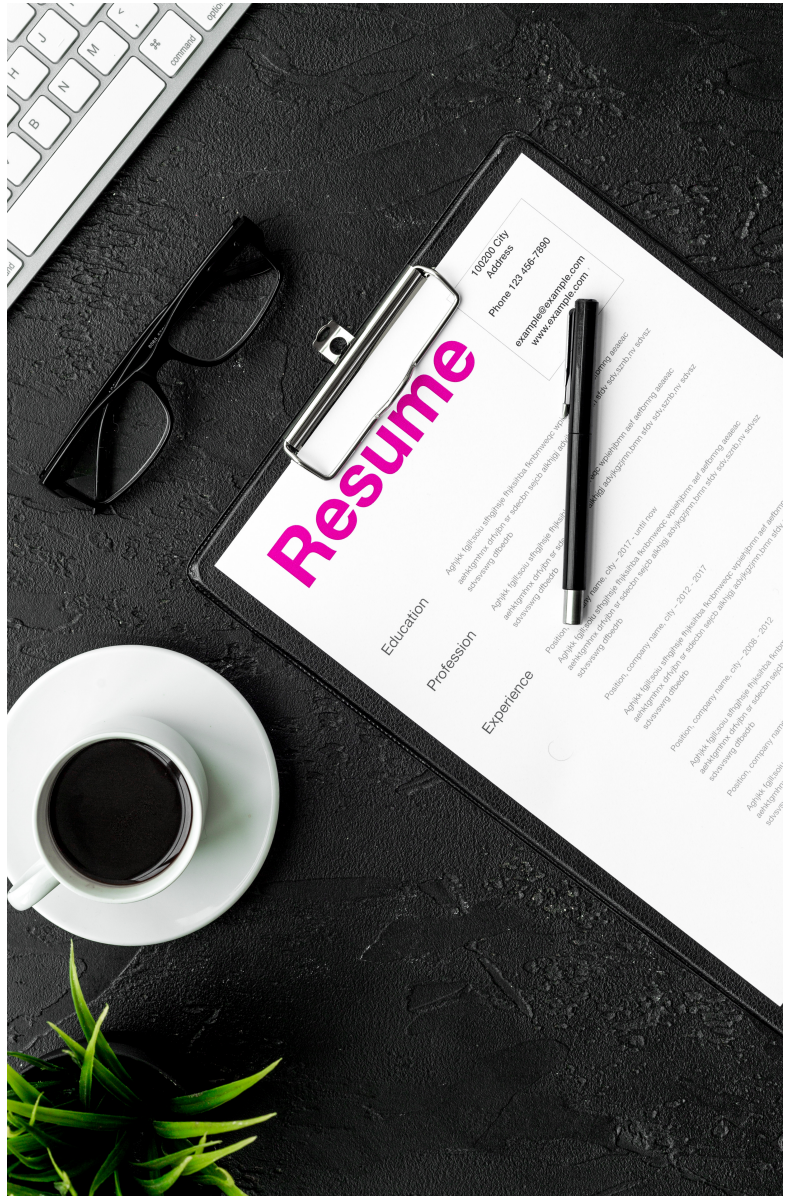
WHAT IS AN OBSERVATION?

- Extracting only sections describing contents of job was difficult
- Many listings lengthy and contain all sorts of information
- No consistency of formatting or language
- I tackled this task with regex

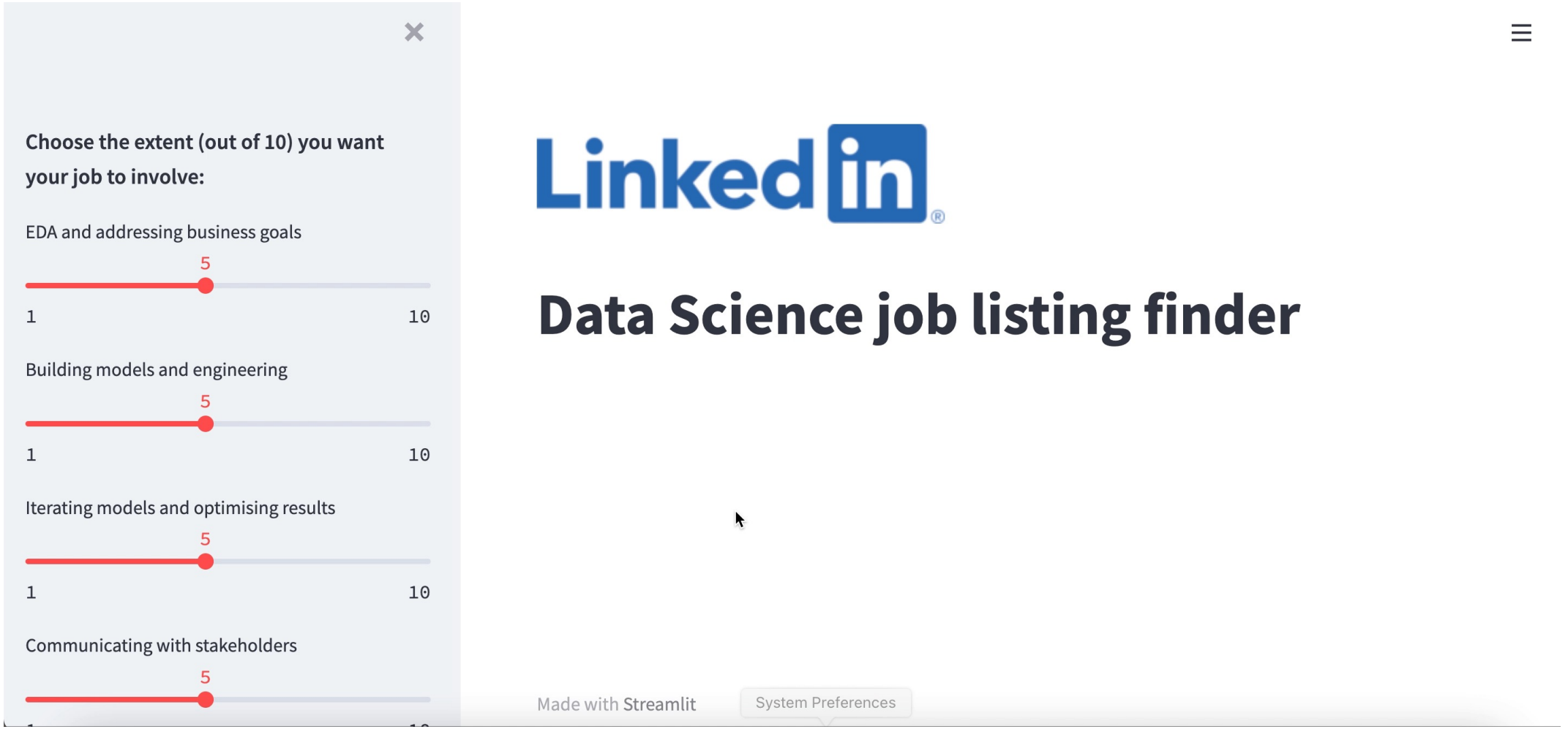


TOPIC MODELING

- Experimented with the following:
  - LSA
  - NMF (with different vectorizers)
  - LDA
  - CorEx

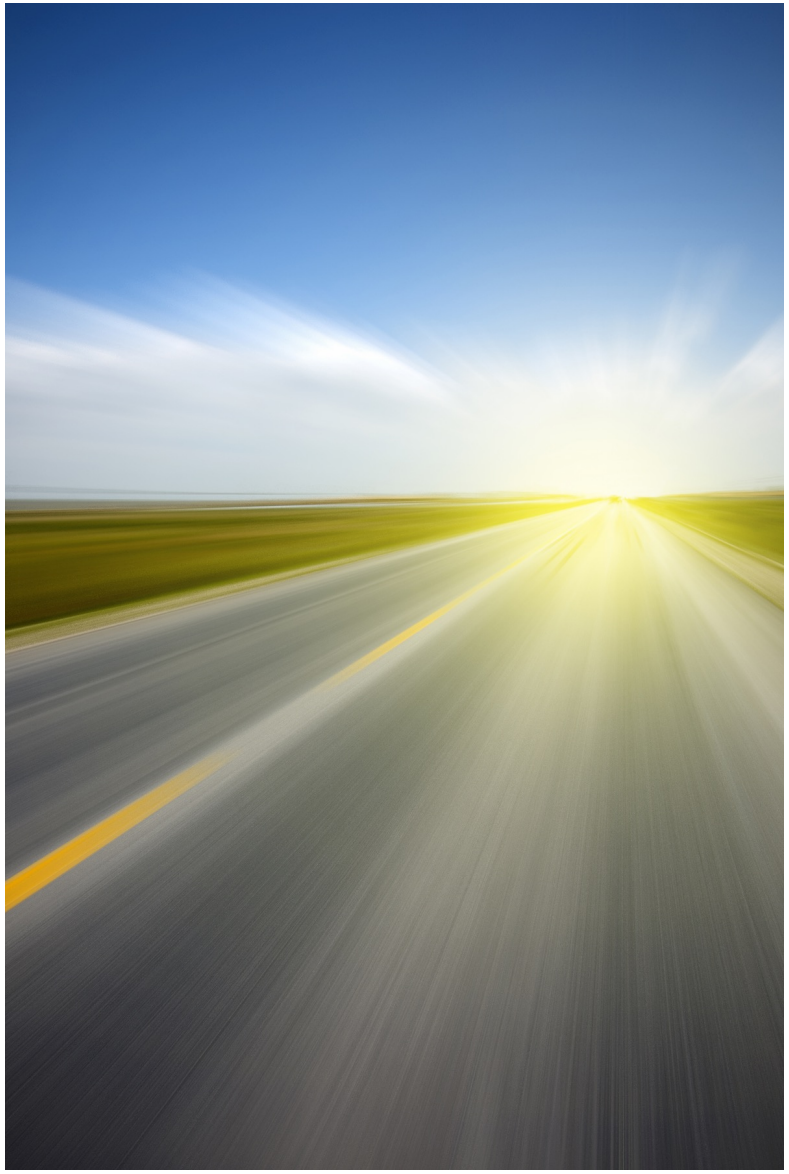


APP DEMO


Choose the extent (out of 10) you want your job to involve:
EDA and addressing business goals
5
1
10
Building models and engineering
5
1
10
Iterating models and optimising results
5
1
10
Communicating with stakeholders
5
Data Science job listing finder
Made with Streamlit
System Preferences

CONCLUSION AND FURTHER WORK

- A challenging project!
- It would be interesting to do:
  - Further iterations on ‘extracting observations’ piece
  - Location filtering for app
  - More investigation of CoRex



APPENDIX I

Regex used for extracting 'job contents' from listing:

```
(?:[Rr]esponsibilities|[Ww]hat
[Yy]ou['']ll|[Ww]hat [Yy]ou
[Ww]ill|[Dd]uties|[Tt]he
[Rr]ole.{0,10}\||[Oo]verview|[Ww]ork.{0,10}\|)(.*?
)(?:Requirements|[Qq]ualifications|Skills.{0,10}\|
|[Ll]ooking [Ff]or.{0,5}\||[Yy]ou [Hh]ave:)
```

APPENDIX 11

Top topic terms from final TF-IDF/NMF model with 10 topics:	Name given to topics for app:
['datum', 'analysis', 'data', 'statistical', 'business', 'model', 'develop', 'analytic', 'process']	EDA and addressing business goals
['learning', 'machine', 'machine learning', 'model', 'ml', 'engineer', 'learning model', 'ai', 'build']	Building models and engineering
['life', 'balance', 'culture', 'culture inclusion', 'career growth', 'employee', 'inclusion', 'career', 'team']	Good work environment and work-life balance
['business', 'model', 'analytical model', 'datum', 'support', 'problem', 'example', 'testing', 'business problem']	
['accuracy', 'outcome develop', 'model', 'data', 'monitor', 'outcome', 'tool monitor', 'effectiveness accuracy', 'gathering technique']	Iterating models and optimising results
['product', 'drive', 'team', 'metric', 'decision', 'insight', 'business', 'build', 'strategy']	Communicating with stakeholders
['data', 'science', 'data science', 'business', 'team', 'solution', 'project', 'data scientist', 'scientist']	
['experience', 'ability', 'skill', 'computer', 'degree', 'strong', 'work', 'field', 'language']	
['marketing', 'connect', 'term', 'modeling', 'mix', 'brand', 'partner', 'optimization', 'build enhance']	
['client', 'consulting', 'service', 'help client', 'professional', 'technology', 'consultant', 'market', 'mission']	

APPENDIX III

URL of app:
https://share.streamlit.io/billbell73/nlp_project/main/streamlit_app/my_app.py

URL of project Github repo:
https://github.com/billbell73/nlp_project