Exploring Data-Related Jobs Around the World: Visualization, Topics Modeling, and Predictive Salary Modeling

Team 14 CSE6242 Final Project

Benoit Bailly Marc-Henri Bleu-Laine Alexander Gurung Hymee Huang Haley Xue

Motivation

Jobs in the data industry are often grouped together under the vague term "Data Scientist," despite having drastically different job requirements and expectations. In addition, traditional job posting boards do little to help job seekers compare roles between locations, especially in terms of salary and respective costs of living.

These failings make it difficult for job seekers to identify which postings are a good fit for them and make it challenging for employers to tailor their postings to specific types of candidates.

Data

We leveraged existing datasets of Indeed and Glassdoor postings on Kaggle and extended them with our own job postings scraped from Indeed. We used the entirety of the dataset for our global visualizations, but sub-selected for U.S. postings for our more in-depth analysis to ensure language and cultural standardization. The salary model uses the dataset sourced from Indeed which was further filtered for postings with salary information (U.S. jobs only).

Dataset	Size	Rows	Application
Kaggle Glassdoor	870MB	~ 165K	Tableau Visualizations, Topics Model
Indeed	4.4MB	~ 1K	Salary Model

Approach

We created three tools to provide insight into data-related job postings

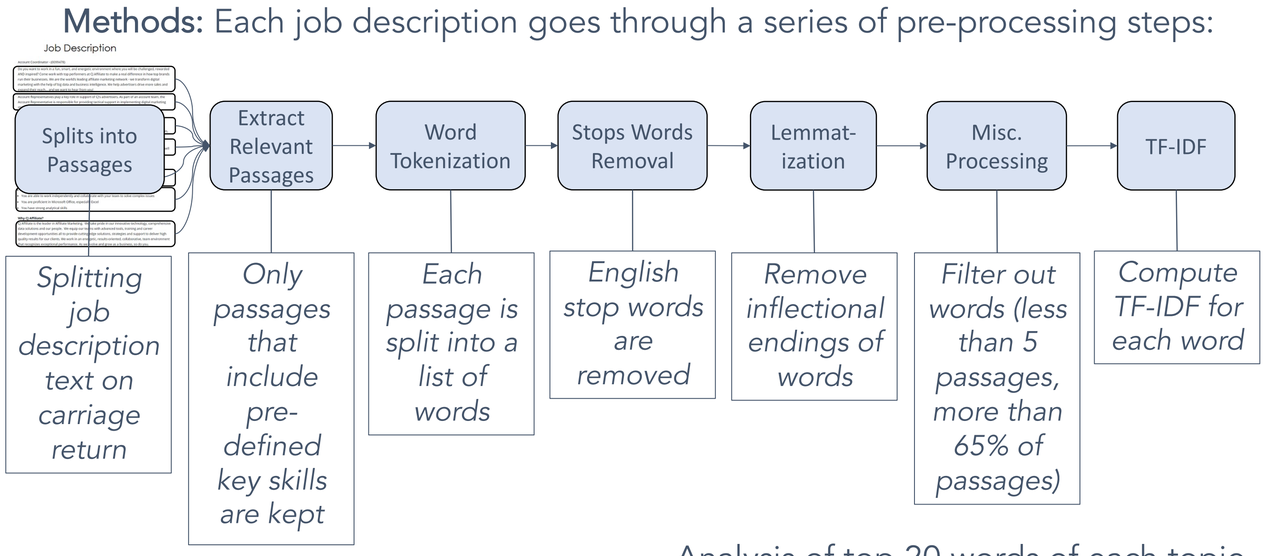
- Interactive visualizations with global and U.S. dashboard views
- Topic modeling of data-related job descriptions to identify sub-categories
- Machine learning model normalized to cost of living to predict salaries

Additional details regarding our analytical approaches can be found in the *Experiments & Results* section.

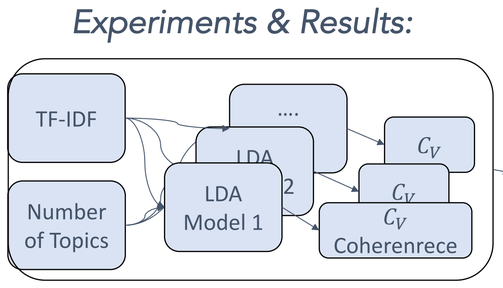
Experiments & Results

Topics Modeling

Methods: Each job description goes through a series of pre-processing steps:



Experiments & Results:



Hyperparameter Search: Select number of topics that yields highest C_V

Ideal Number of Topics → LDA

Analysis of top 20 words of each topic yields:

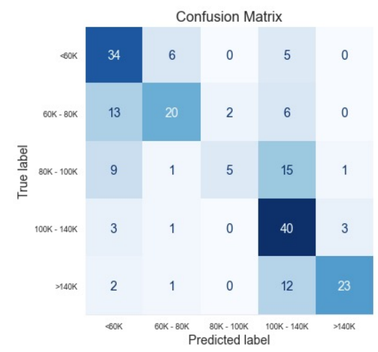
- **Topic 1:** Communication and Teamwork
- **Topic 2:** Project Management/Product Development and Testing
- **Topic 3:** Machine Learning, Statistical Analysis and Algorithm
- **Topic 4:** Educational Credidentials
- **Topic 5:** MS Office, Cloud Tech, Deployment
- **Topic 6:** Software Engineering w/ Statistics and Cloud Knowledge
- **Topic 7:** Research Focus
- **Topic 8:** Project Management (emphasis on control/support)
- **Topic 9:** Software Engineering Paradigm

Predictive Salary Model

Methods: The tool uses a random forest classifier to predict the salary for a job posting. This is helpful to users as few job postings have salaries listed, yet compensation is probably one of the most important pieces of information for job applicants. TF-IDF vectorization was used to extract features from the job descriptions. Numerical (cost of living) and categorical features (job type, level, and company) are also used as model inputs. Given a set of these inputs, the model classifies a job into one of five salary buckets.

Experiments & Results: Precision, recall, and F1 scores were used to compare different methods. Our final TF-IDF + Random Forest model had a F1 accuracy of ~0.6, which outperformed our DistilBERT + Random Forest and BERT classifiers. Grid search and hyperparameter methods were used to improve model performance. While our BERT models were not deployed, incorporating NLP models into a salary classification problem is a relatively novel approach.

	precision	recall	f1-score	support
100K - 140K	0.51	0.85	0.64	47
60K - 80K	0.69	0.49	0.57	41
80K - 100K	0.71	0.16	0.26	31
<60K	0.56	0.76	0.64	45
>140K	0.85	0.61	0.71	38
accuracy			0.60	202
macro avg	0.67	0.57	0.56	202
weighted avg	0.65	0.60	0.58	202

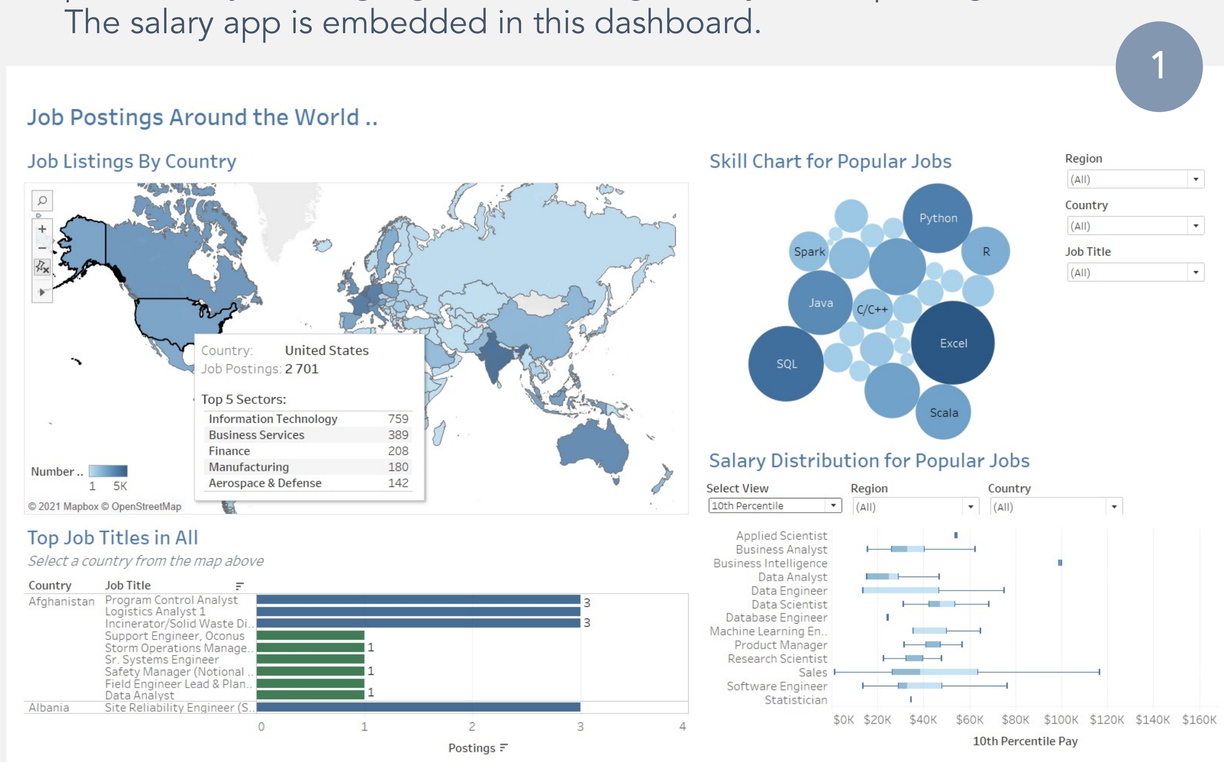


Dashboards

Our Tableau visualizations are divided into four views (two are shown below):

1. **World View** – describes job postings around the world with an emphasis on the required skills, the most common job titles per country, and the salary distribution for common job titles.
2. **USA View** – describes job postings in the U.S. with a breakdown per state/per industry and highlights the average salary for all postings in our dataset. The salary app is embedded in this dashboard.
3. **Topics Model** – shows a full listing of jobs in the United States and the identified topics per posting. Users can filter by various columns (title, topic, etc.). Users can also search for the full job description for a job ID.
4. **LDA Visualization** – provides an in-depth visualization of the LDA analysis. This view is an embedded Jupyter Notebook; users can view the intertopic distance and view most salient terms by topic.

1



2