Data Job Postings Summarization

Group 4: Grace Xie, Sydney Li, Oliver Zhou, Xiyi Lin

Text Analytics

Problem Statement & Data Overview

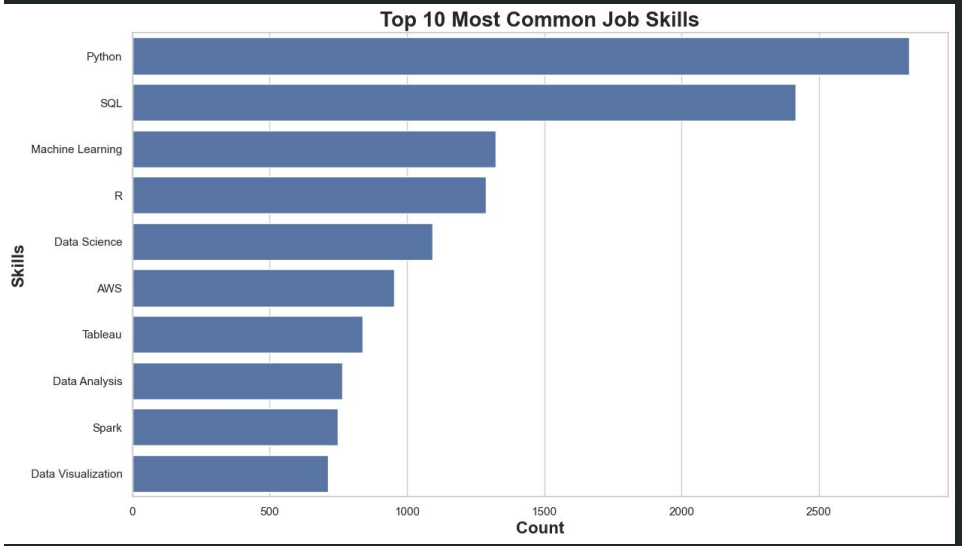
The demand for data science professionals is rapidly increasing, and LinkedIn serves as a major platform for connecting job seekers with opportunities in this domain. However, the vast amount of job postings makes it challenging to extract actionable insights.

This project aims to **analyze and summarize LinkedIn data science job postings** to uncover trends in required skills, software proficiencies, and job descriptions. The goal is to provide a concise and comprehensive understanding of employer expectations to help job seekers better align their profiles with industry demands.

Cr. https://www.kaggle.com/datasets/asaniczka/data-scientist-linkedin-job-postings

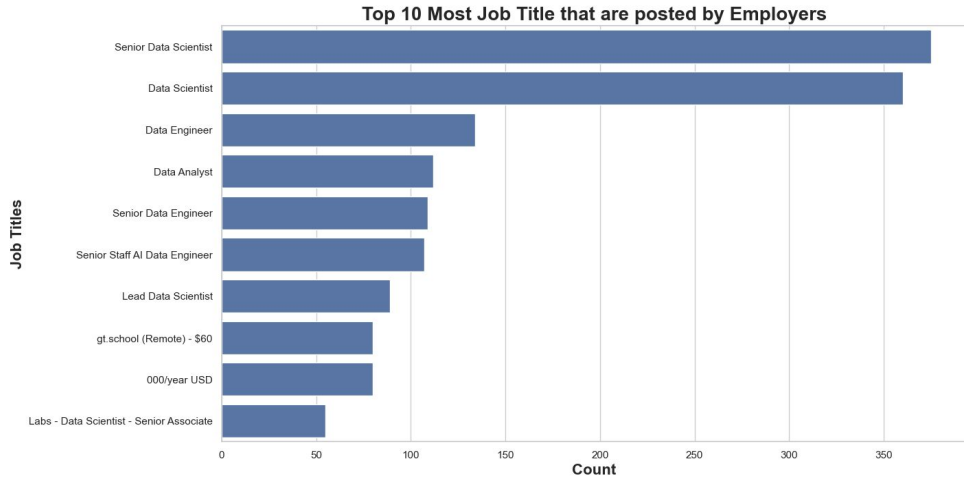
Top 10 Most Common Job Skills

Python, SQL, Machine Learning, R, Data Science, AWS, Tableau, Data Analysis, Spark, Data Visualization



Top 10 On-Demand Job Titles

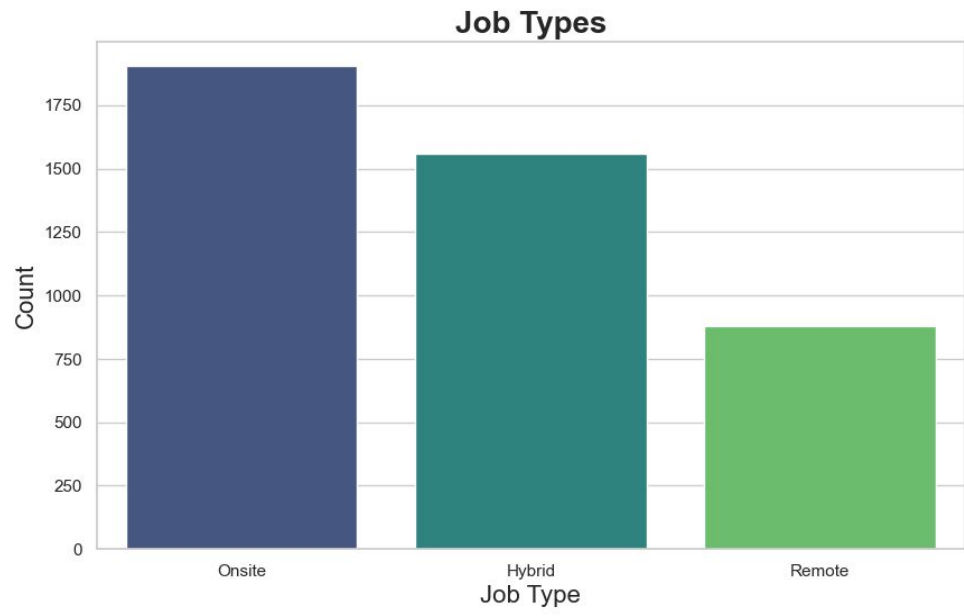
Senior Data Scientist, Data Scientist, Data Engineer, Data Analyst, Senior Data Engineer, ...



Job Work Mode

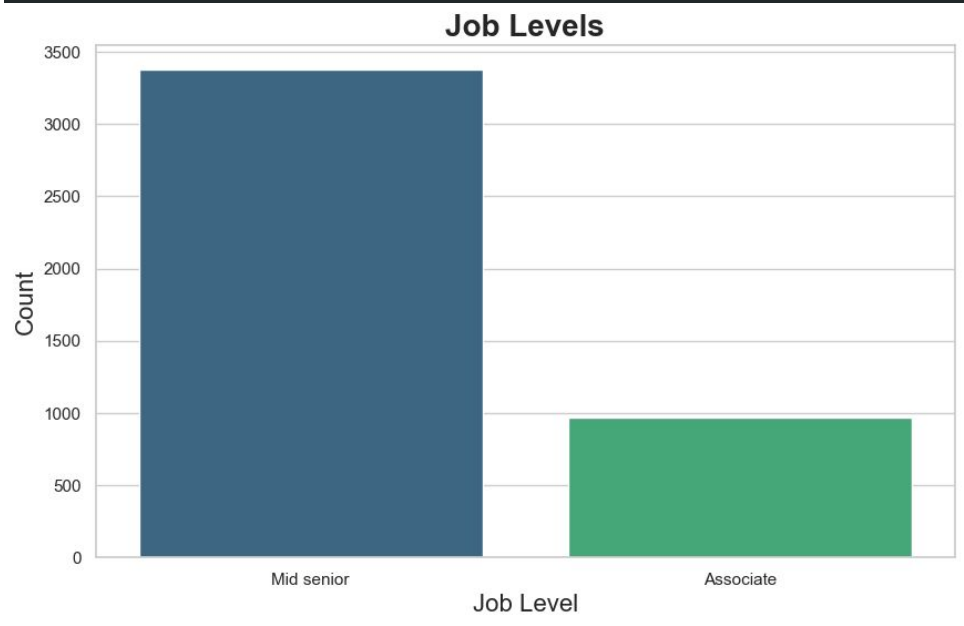
💻🏠💼

job level	job_type	
Associate	Hybrid	327
	Onsite	435
	Remote	204
Mid senior	Hybrid	1231
	Onsite	1469
	Remote	676

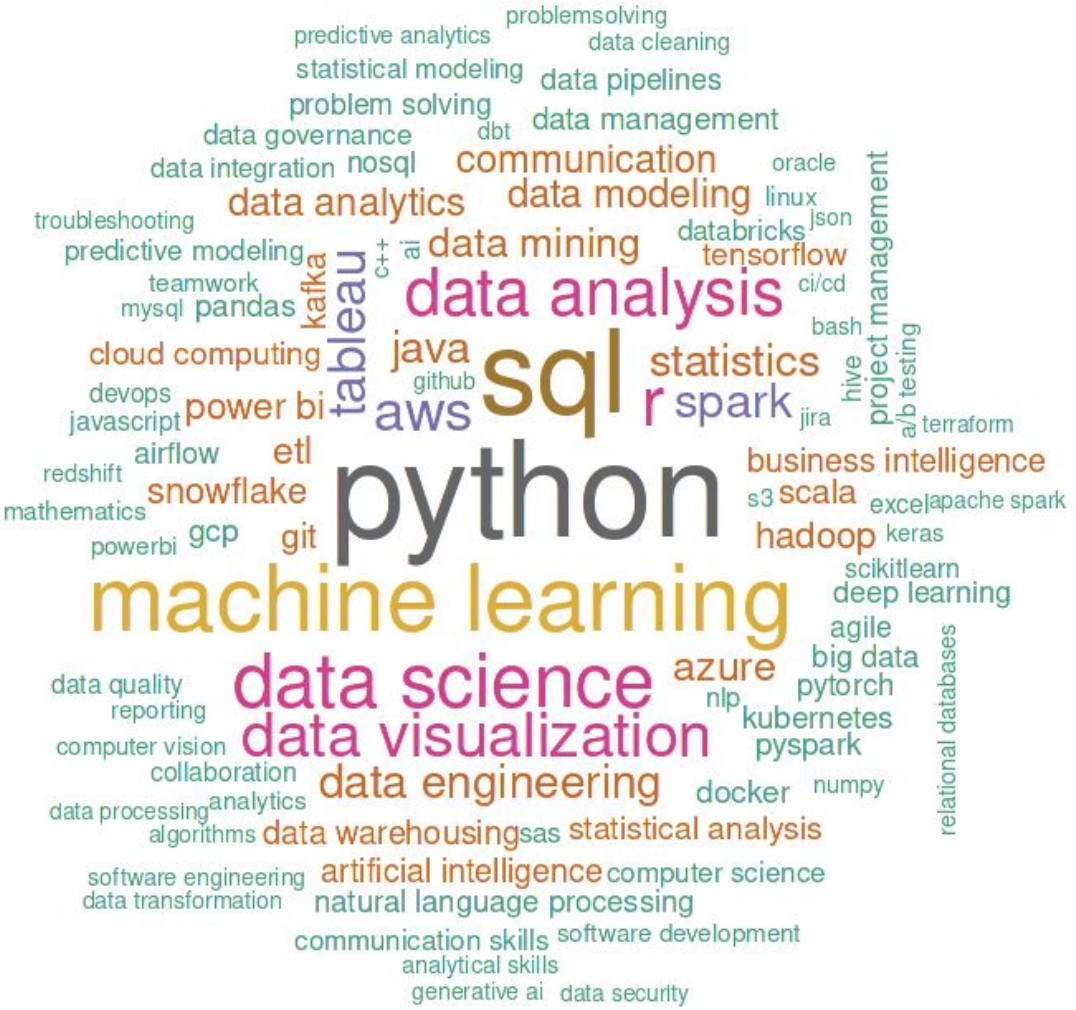


Job Levels

Mid-senior >> Associate



Key Skill Sets

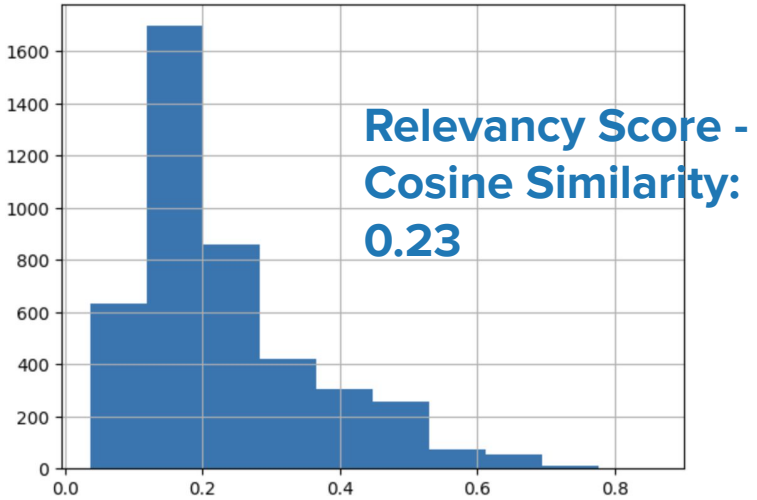


Text Processing

- Remove special characters & digits
- Lowercase
- Tokenization
- Retain data-related keywords (skills)
- Remove stop words

Baseline model

1. Input: cleaned job summary text
2. Model: ***facebook/bart-large-cnn***
3. Output: a shortened concise summary
4. Evaluation: **ROUGE scores** - measure the summary's quality



	ROUGE-1	*ROUGE-2*	*ROUGE-L*	*ROUGE-L sum*
Base line	**14.4**	**13.3**	**14.0**	**14.1**

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), is a set of metrics and a software package specifically designed for evaluating automatic summarization, but that can be also used for machine translation. The metrics compare an automatically produced summary or translation against reference (high-quality and human-produced) summaries or translations.

Cont.

For baseline models or simple approaches, these scores may be reasonable, as early-stage or simpler models might achieve ROUGE scores in the range of 10-20%.

Further Improvement:

- Achieve higher ROUGE Scores (depends on context and use)
- May use LLM or Prompt Engineering to prettify the summary we extracted (more smooth and human-sound words)

Chatgpt Generated Baseline model 2

- Reduce redundancy word
- Use **`gpt-3.5-turbo`**
- Prompt:

"You are a helpful assistant. The following data are the job description in the LinkedIn. Summarize the following job description with fluent sentences with the company name, job title, how many years of experience, skills needed."

- Limits the length of the generated summary to 100
- **`top_p=1`**: Ensures diversity in output
- **`temperature=0.7`**: Controls the creativity of the output

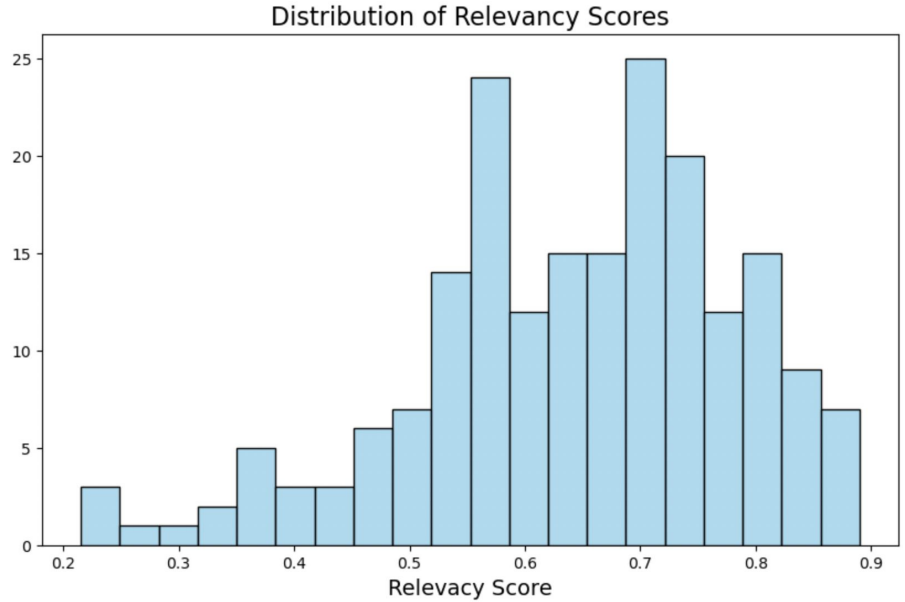
Model Result

- Randomly subset 200 samples

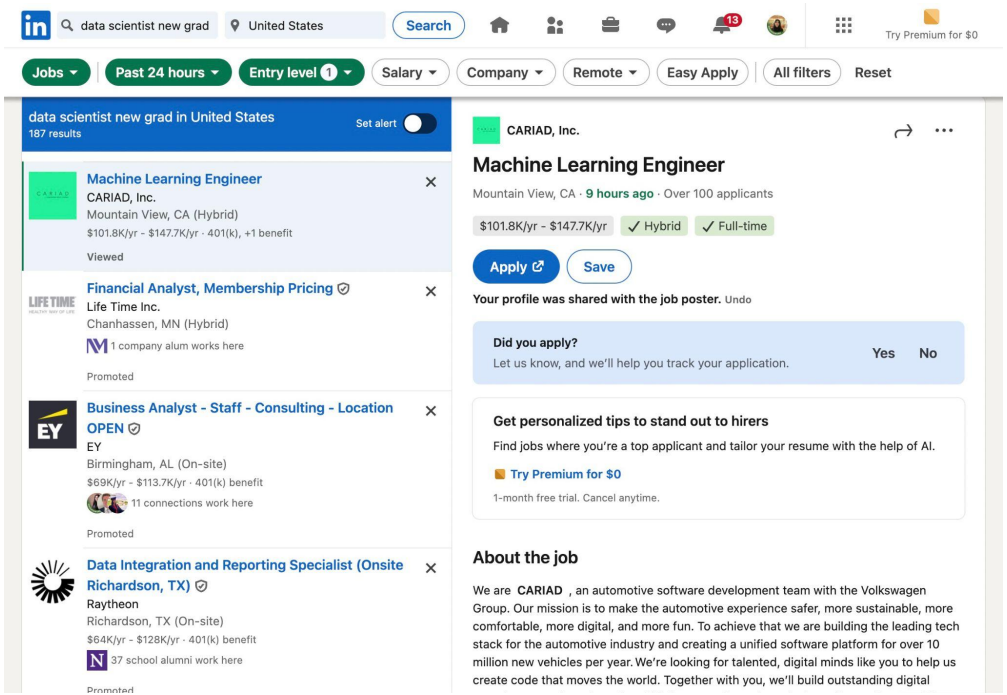
Format of the Result

The summarized output in the `job_summary_summary` column will include:

1. **Company Name**: The name of the company.
2. **Job Title**: The title of the position (e.g., Data Scientist).
3. **Years of Experience**: The required years of experience for the role.
4. **Skills Needed**: Key skills and tools required for the position.



	ROUGE-1	*ROUGE-2*	*ROUGE-L*	*ROUGE-L sum*
GPT	**24.5**	**16**	**19.1**	**19.1**




```
job_1 = "We are CARIAD, an automotive software development team with the Volkswagen Group. Our mission
test1 = summarize_job(job_1)
print(test1)
```

```
Company: CARIAD, automotive software development team within the Volkswagen Group.

Job Title: Digital Software Engineer.

Experience: Not specified in the job description.

Skills needed: The ideal candidate should possess talent and digital skills to create innovative code
for the automotive industry. They should have a strong background in software development and be pass
ionate about creating outstanding digital experiences and products. Additionally, the candidate shoul
d have the ability to work collaboratively in a team to shape the future of mobility.
```

Role Responsibilities:

AI Application Development

- Research and develop AI application to improve engineering efficiency and productivity
- Design and create RAG and RL application pipelines
- Define and create vector datastores for different application use
- Design and develop application frontend and backend services

Prompt Engineering

- Develop application specific prompt tuning pipeline
- Create and design prompt engineering architecture for different applications
- Develop scripts to process data and generate ground truth dataset for fine-tuning

Model Fine-tuning

- Fine tune base models with different quantization techniques for application specific usage
- Optimize models for performance and size
- Evaluate and improve the efficacy of the fine-tuned model
- Define and create multi-model and multi-adapter inferencing workflow

Example:

https://www.linkedin.com/jobs/view/4079518767

DistilBART

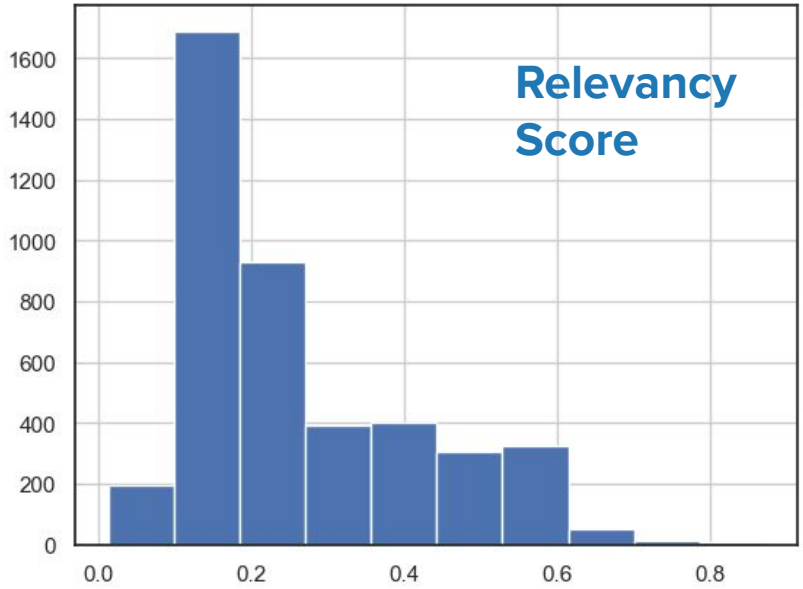
- Model: sshleifer/distilbart-cnn-12-6
- Reasons to choose this model:

  Lightweight (distilled version of BART)

  Fine-tuned for summarization

  Handles diverse inputs
- Rouge scores:

```
{'rouge1': np.float64(21.7),
 'rouge2': np.float64(20.5),
 'rougeL': np.float64(21.1),
 'rougeLsum': np.float64(21.1)}
```



Average relevancy score: 0.26

TF-IDF + T5

- Model: t5-small
- Reasons to use it:
  Extractive-abstractive hybrid approach
  Improves fluency, making summaries human-like
- Rouge scores:

```
'rouge1': np.float64(15.9), 'rouge2': np.float64(9.0), 'rougeL': np.float64(15.8), 'rougeLsum': np.float64(15.8)
```

Other Models

LDA - not suitable for summarization tasks; focused on topics rather than creating coherent, context aware summary

BERTSUM - used for extractive summarization and requires clearly separable sentences as input

LongT5 - extremely time-consuming, not efficient; not fine-tuned for summarization

Open Pretrained Transformer: designed for text generation, not sequence-to-sequence modeling; the decoder-only structure does not provide functionalities for coherent summarization

FLAN-T5 Example:

- Model: *google/flan-t5-base*
- *Reasons to choose this model:*
  - *High semantic relevance*
  - *Flexibility*
  - *Scalability*

Original Job Description:

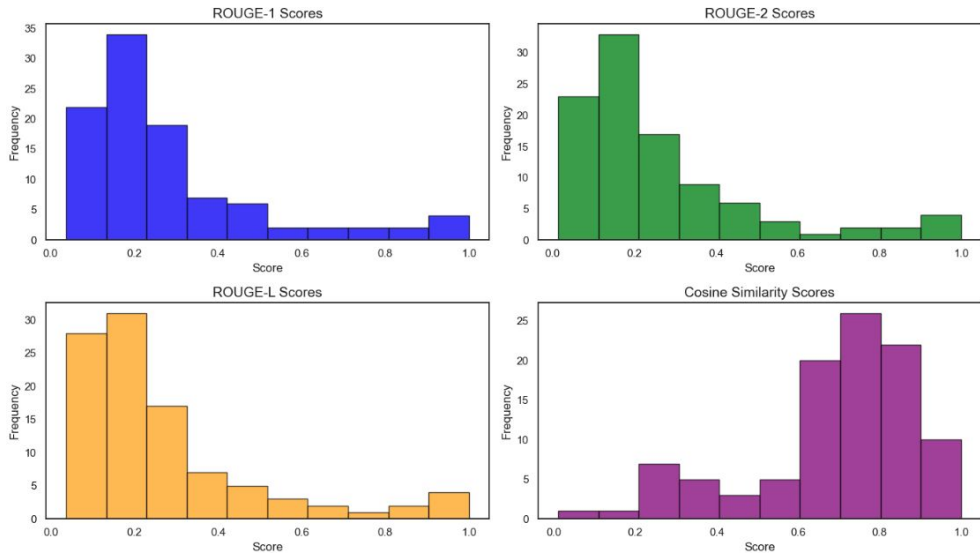
Why Choose Jefferson Health Plans?
We are an award-winning, not-for-profit health maintenance organization. We are committed to creating a community where everyone belongs, acknowledges, and celebrates diversity and has opportunities to grow to their fullest potential.
While this job currently provides a flexible remote option, due to in-office meetings, training as required, or other business needs, our employees are to be residents of PA or the nearby states of DE or NJ.
Perks of JHP and why you will love it here:
Competitive Compensation Packages including 401(k) Savings Plan with Company Match and Profit Sharing
Flextime and Work-at-Home Options
Benefits & Wellness Program including generous Time Off
Impact on the communities we service
We are seeking a talented and enthusiastic Technical Data Analyst to join our team!
The Data Operations' Team primary function is to provide access to various clinical and financial data to management to support the business decision process. Databases are maintained by this team for analyst use. Business users can access analyst data and analysis through the iQ portal which is also maintained by the Data Operations Team. This particular role will ensure consistent, accurate data for reporting. The technical data analyst will serve as the business lead for the corporate enterprise data warehouses project and software conversion efforts. They will work with IS, our consultants and other business users to verify the data model design, test data builds, verify algorithms and integrate the data using our new business intelligence reporting tool. The person will enable other Healthcare Economics analysts to use current and future data sources effectively.
As the Technical Data Analyst, your daily duties may include:
Use KNIME, QlikView, SQL, MS Access, MS Excel and other reporting tools as necessary to maintain Data Analytics Division databases.
Review logical and physical data model design. Make recommendations for improvement.
Review and publish application and analysis to user portal.
Extract, transform and load data from internal and external databases into Data Analytics Division databases.
Coordinate and conduct testing of data as it is loaded into databases by consultants, vendors and IS.
Be a key contributor to the Data Operations team. Model and verify data from business perspective.
Conduct trainings when appropriate for Data Analytics division staff.
Use appropriate business intelligence tools to obtain, validate and analyze relevant data.
Verify and analyze report data and prepare documentation of processes needed for development and improvement of applications.
Help determine optimal analysis and presentation of medical data to improve service quality.
Effectively communicate results of analysis.
Research issues to determine source of discrepancies.
Identify, communicate and monitor areas for improvement.
Share knowledge with other Data Analytics Division associates to improve processes, quality and data integrity.
Qualifications
B.S. or higher degree in computer science, information systems, statistics or other analytical field of study.
3 or more years of data reporting and analysis experience including report and dashboard development, database development, data connectivity and technical problem resolution.
Experience with delivery and reimbursement of medical services in a managed care environment.
Skills, We Value:
Strong problem-solving, quantitative and analytical skills

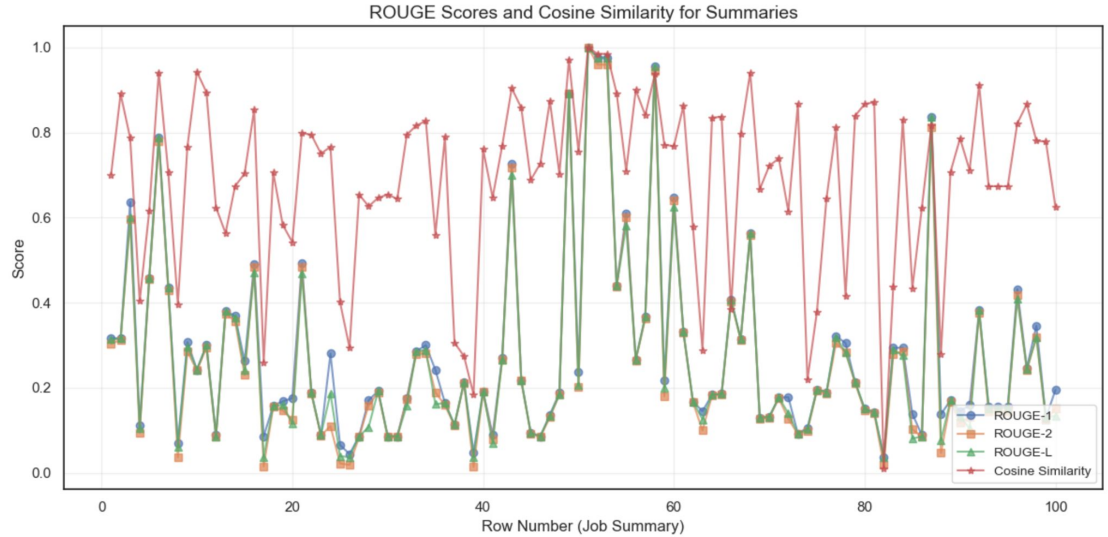
Flan-T5 Summary:

Data Operations' iQ portal project and software conversion efforts. This particular role will ensure consistent, accurate data for reporting. The technical data analyst will serve as the business lead for the corporate enterprise data warehouses project and software conversion efforts. They will work with IS, our consultants and other business users to verify the data model design, test data builds, verify algorithms and integrate the data using our new business intelligence reporting tool. The person will enable other Healthcare Economics analysts to use current and future data sources effectively. Be a key contributor to the Data Operations team. Model and verify data from business perspective. Conduct trainings when appropriate for Data Analytics division staff.

FLAN-T5

- Evaluation:
  - **ROUGE scores** - measure the summary's quality
  - **Relevancy scores** - Cosine Similarity





Average Cosine Similarity: 0.69
Average ROUGE-1: 0.28
Average ROUGE-2: 0.27
Average ROUGE-L: 0.27

Why FLAN-T5 is the Best Model so far?

- Instruction-Tuned for Better Understanding
- Handles Long Inputs
- Efficient and Versatile
- Minimized Repetition Issues
- Straightforward Integration

But still some challenges....

- Computational Requirements
- Fine-Tuning Needs

Conclusion

- Addresses the growing demand for data science professionals by tackling the challenge of extracting actionable insights
- Leverage advanced summarization models to provide a detailed understanding of trends in required skills, software proficiencies, and job descriptions
- Empower job seekers to align their profiles with industry expectations effectively, bridging the gap

Thank You For

You Listening