PANCHAMI DINESH

Search Behavior–Driven Skill Demand Analysis Using Analytics, NLP, and RAG

Abstract

With the rapid evolution of technology, students often struggle to identify which technical skills will remain relevant in the future. This project proposes a data-driven decision support system that analyzes online search behavior to anticipate future skill demand. Google Trends data is used as a proxy for public learning interest, which is combined with job demand indicators, regional analysis, natural language processing (NLP), and a Retrieval-Augmented Generation (RAG) framework. The system performs time-series analysis, volatility measurement, semantic skill clustering, and interactive question answering through a Gradio interface. The results demonstrate that search behavior, when analyzed systematically and validated with external indicators, can provide meaningful insights into future-relevant skills for students.

---

1. Introduction

Choosing the right technical skills is a critical decision for engineering students. Traditional guidance methods rely on anecdotal advice or static job reports, which often lag behind real-world changes. Online search behavior reflects learning intent, curiosity, and early adoption trends, making it a valuable signal for anticipating future skill demand.

This project investigates whether Google search trends can be systematically analyzed and validated to support student decision-making. By integrating data analytics, NLP, and RAG, the project moves beyond visualization and builds an interactive decision-support system.

---

2. Problem Statement

How can online search behavior be analyzed to anticipate future skill demand and help students make informed decisions about which technologies to learn?

---

3. Objectives

- Collect and analyze Google search trend data for technical skills
- Identify long-term growth, volatility, and saturation patterns
- Validate trends using job demand indicators
- Analyze regional adoption of skills
- Group skills using NLP-based semantic clustering
- Build a RAG-based question answering system for decision support

---

4. Data Sources

4.1 Google Trends Data

Google Trends provides normalized search interest values (0–100) over time. This data was used to analyze public interest in selected technical skills over a five-year period.

4.2 Job Demand Proxy

A relative job demand score was manually curated based on aggregated industry reports and job portal trends. This dataset serves as an external validation signal.

---

5. Methodology

The project was implemented in multiple structured phases.

Phase 1: Data Acquisition and Setup

5.1 Library Imports

Python libraries such as `pandas`, `numpy`, `matplotlib`, `seaborn`, and `pytrends` were imported to support data handling, visualization, and API access.

5.2 Google Trends Connection

A `TrendReq` object was initialized to establish communication with Google Trends.

5.3 Skill Selection

Skills were grouped into categories:

- Programming Languages
- Data & AI
- Cloud & DevOps

This controlled scope ensured meaningful comparison.

5.4 Time Window Definition

A fixed analysis period (2019–2024) was defined to maintain consistency across skills.

Phase 2: Time-Series Trend Analysis

6.1 Data Collection

Search interest over time was fetched for each skill using Google Trends and combined into a single time-indexed DataFrame.

6.2 Smoothing

A rolling mean was applied to reduce short-term noise and highlight long-term patterns.

6.3 Growth Rate Calculation

Growth rate was computed using the first and last valid smoothed values to avoid rolling-window NaN bias.

6.4 Volatility Measurement

Standard deviation of smoothed trends was used to quantify stability.

6.5 Growth vs Stability Visualization

A scatter plot of growth rate versus volatility was created to classify skills as:

- Emerging
- Stable
- Saturated
- Hype-driven

Phase 3: Job Market Validation

7.1 Job Demand Dataset

Each skill was assigned a relative job demand score.

7.2 Correlation Analysis

Search growth was compared with job demand to evaluate alignment between public interest and market needs.

7.3 Demand vs Growth Visualization

A scatter plot was used to identify:

- High demand & high growth skills
- High growth but low demand (hype)
- High demand but low growth (mature skills)

Phase 4: Region-Wise Analysis

8.1 Regional Interest Extraction

Google Trends region-wise interest was fetched for selected high-impact skills.

8.2 Normalization

Regional values were normalized to allow fair cross-skill comparison.

8.3 Heatmap Visualization

A heatmap revealed geographic concentration and early adoption patterns, particularly in technology-driven regions.

Phase 5: NLP-Based Skill Clustering

9.1 Skill Corpus Creation

Each skill was represented using a descriptive sentence to provide semantic context.

9.2 Embedding Generation

Sentence embeddings were generated using the `all-MiniLM-L6-v2` transformer model.

9.3 Similarity Analysis

Cosine similarity was computed to analyze conceptual closeness between skills.

9.4 Clustering

Agglomerative clustering grouped skills into:

- Core Programming Languages
- DevOps & Containers
- Data & AI

- Cloud Platforms

This clustering was fully data-driven, not manually assigned.

Phase 6: Cluster-Level Analysis

10.1 Metric Aggregation

Growth rate, volatility, and job demand were averaged at the cluster level.

10.2 Interpretation

- Core languages showed saturation
- Data & AI exhibited strong growth
- Cloud and DevOps clusters showed stable expansion

Phase 7: RAG-Based Question Answering System

11.1 Knowledge Base Construction

Analytical results were converted into short textual documents representing skills and clusters.

11.2 Vector Store

Document embeddings were stored using FAISS for efficient similarity search.

11.3 Retrieval

User queries were embedded and matched against stored analytical insights.

11.4 Generation

A lightweight transformer model (`t5-small`) summarized retrieved evidence into coherent answers, ensuring no hallucination.

Phase 8: Gradio Interface

12.1 Interface Design

A Gradio interface was created with:

- Query input box
- Insight output box

12.2 User Interaction

Students can ask natural language questions such as:

- “Which skills should I focus on for the future?”
- “Which skills are hype-driven?”

12.3 Feedback Logging

Gradio’s flagging mechanism enables optional collection of user queries for future refinement.

---

13. Results and Insights

- Data & AI and Cloud skills show sustained growth and strong demand
- Core programming languages are saturated but essential
- Some skills exhibit hype-driven volatility
- Regional analysis highlights early adoption patterns
- NLP clustering reveals meaningful skill groupings

---

14. Limitations

- Google Trends data is relative, not absolute
- Search intent may not always represent learning intent
- Job demand scores are proxies, not scraped data

- LLM summarization depends on retrieved context quality

---

15. Future Work

- Use real job postings data
- Add time-series forecasting
- Expand skill coverage
- Deploy system on Hugging Face Spaces
- Improve answer personalization

---

16. Conclusion

This project demonstrates that online search behavior, when analyzed rigorously and validated with external indicators, can serve as a meaningful signal for anticipating future skill demand. By integrating analytics, NLP, and RAG into a single interactive system, the project moves beyond visualization and offers actionable insights for students. The methodology and system design are extensible, explainable, and suitable for real-world decision support.

---

17. Technologies Used

- Python
- Google Trends (PyTrends)
- Pandas, NumPy
- Matplotlib, Seaborn
- Sentence Transformers
- FAISS
- Hugging Face Transformers
- Gradio