Midterm Research Plan: Employment & Unemployment (Baidu Index & Baidu Encyclopedia)

Chenxi Zhang, Haowen Shi, Haotian Zhou

Course: Data Science and AI

Nov 27, 2025

Executive Summary

Aim: Use Baidu Index (search interest) and Baidu Encyclopedia to identify employment market trends, influencing factors, and potential patterns.
Scope: China, past five years (2019–2025).
Deliverables: Clean datasets, descriptive analytics, and clear visuals.

Pipeline

- Data Collection (Baidu Index & Baidu Encyclopedia)
- Data Cleaning (completeness, missingness, duplicates, outliers)
- Descriptive Analysis (distribution, correlation, trends)
- Visualization (time trends, regional comparisons, correlations)
- Conclusions & Outlook

0. Motivation and Significance

Why search data?

- Search interest provides a high-frequency proxy for public concern and job-market sentiment.
- Compared to official statistics (monthly/quarterly), it reacts faster to shocks and policy events.

Practical value

- Early-warning signals of labor-market stress (graduation season, layoffs, policy shifts).
- Support regional employment monitoring and policy evaluation.

1. Research Background and Objectives

Background: Employment and unemployment are core issues affecting social stability and economic development.
Objective: Use Baidu Index and Baidu Encyclopedia to identify employment-market trends, influencing factors, and potential patterns.

Key questions

- What do search-interest dynamics reveal about temporal and regional variations?
- How do related concepts (e.g., unemployment rate, job hunting) co-move?
- Which policy topics or definitions appear most frequently in encyclopedia entries?

1.1 Related Work and Our Contribution

Prior evidence

- Search data has been used to nowcast macro variables (consumption, finance, epidemics).
- Labor-market studies show online attention correlates with job-search intensity.

Our contribution

- Combine *Index (behavior)* + *Baike (definition/policy text)*.
- Provide interpretable descriptive results by region/keyword/event.
- Build a reproducible dataset pipeline for future forecasting extensions.

2. Data Collection — 2.1 Baidu Index

Keywords: Select terms related to “employment” and “unemployment” (e.g., “job hunting”, “unemployment rate”).
Method: Use Baidu Index self-service tools to set time window and regional scope, then export data.

Expected fields

- `date`, `region`, `keyword`
- `search_index_total`, `pc_index`, `mobile_index`
- frequency (daily/weekly), notes

Coverage: National and major provinces.

2.1.1 Keyword Selection Strategy

Principles

- **Coverage:** include both outcome terms (“unemployment”) and behavior terms (“job hunting”).
- **Specificity:** avoid overly broad words; prefer policy/market-relevant phrases.
- **Robustness:** add synonyms and colloquial variants to reduce wording bias.

Candidate buckets

- Employment: “employment”, “recruitment”, “campus hiring”, “graduate jobs”
- Unemployment: “unemployment”, “unemployment rate”, “layoff”, “jobless”
- Job-search behavior: “job hunting”, “resume”, “interview”, “offer”

2. Data Collection — 2.2 Baidu Encyclopedia

Targets: Entries like “employment policies” and “unemployment types.”
Extraction: Policy details, industry employment data, unemployment causes.

Typical fields

- entry title, abstract, section text
- key dates (publication/revision), policy highlights
- target groups, administrative level, references, URL

Method: Web scraping with standard HTML parsing; store as structured JSON.

2.2.1 Baike Text Processing Plan

Goals

- Extract policy/definition themes and their evolution over time.

Planned steps

- Cleaning: remove boilerplate, unify punctuation, Chinese tokenization.
- Keyword/phrase mining: TF-IDF or simple frequency counts.
- Topic sketching (descriptive): cluster sections by similarity to find theme groups.

Output

- Policy-topic frequency table by year.
- Co-occurrence network of policy terms (descriptive).

3. Data Cleaning (Part 1)

Initial Review: Check data completeness and accuracy.
Missing Values: Fill with mean/linear interpolation, or delete invalid entries; flag imputed points.
Duplicates: Remove repeated records and keep an audit trail.

Outputs

- Cleaned index table(s) aligned by date, region, and keyword
- Structured encyclopedia table(s) for downstream analysis

3. Data Cleaning (Part 2)

Outliers: Identify via 3σ, IQR fences, or boxplot indicators.
Format Conversion: Standardize date/numeric formats; ISO-8601 dates, normalized region codes.

Quality flags

- `impute_flag`, `outlier_flag`, `source_note`
- reproducible scripts/notebooks with deterministic results

4. Descriptive Analysis — Distribution

Goal: Analyze distributions across regions and time to find central tendencies and dispersion.

Examples

- Regional boxplots/violin plots for key keywords
- Temporal distribution summaries (by month/quarter/year)
- Heatmaps of average index levels or coefficients of variation

4. Descriptive Analysis — Correlation

Goal: Explore links between employment/unemployment search data and economic factors, and among related keywords.

Examples

- Correlation matrices (Pearson/Spearman)
- Scatter plots with trend lines (e.g., “layoff” vs “job hunting”)
- Optional alignment with macro indicators (GDP, CPI, youth unemployment)

4. Descriptive Analysis — Trends

Goal: Use time-series views to observe changes and discuss short-term signals.

Examples

- Line charts of keyword indices with moving averages
- Seasonal/holiday/graduation-season annotations
- Descriptive STL decomposition or change-point checks

4.1 Event Annotation and External Context

Why annotate events?

- Peaks are often driven by real-world shocks or policies.

Planned annotations

- Graduation seasons (Jun–Jul annually)
- Major policy releases (employment support, labor market reforms)
- Macro shocks (COVID waves, large-scale layoffs, tech/real estate cycles)

Use

- Compare pre/post-event search levels; describe regional heterogeneity.

5. Data Visualization

Tools: Python (Matplotlib, Seaborn) or BI tools (e.g., FineBI).
Charts: Line charts, bar charts, scatter plots, heatmaps, co-occurrence networks.

Design guidelines

- Consistent labels; figures include source & study window
- Clear legends/annotations; readable fonts for classroom screens
- Reproducible and parameterized code

5.1 Robustness and Limitations

Robustness checks (descriptive)

- Compare multiple synonyms to confirm consistent patterns.
- PC vs Mobile split to see user-group differences.
- Smoothing sensitivity (MA window length).

Limitations

- Search data is attention, not necessarily real employment outcomes.
- Media/viral events may cause temporary spikes.
- Baike entries may have editorial lag or incomplete regional detail.

6. Project Timeline (Next 4–5 Weeks)

- **Week 1:** finalize keyword list; collect Index & Baike data; build raw database.
- **Week 2:** cleaning pipeline; missing/outlier handling; unify region codes.
- **Week 3:** descriptive analysis (distribution/correlation/trends); event annotations.
- **Week 4:** visualization polishing; interpret results; draft report/slides.
- **Week 5 (if time):** optional extension: simple forecasting / policy comparison.

6.1 Team Roles and Work Allocation

- **Chenxi Zhang:** Baidu Index keyword design; data export; regional comparison analysis.
- **Haowen Shi:** scraping/structuring Baike text; text mining + topic sketches.
- **Haotian Zhou:** cleaning scripts; time-series visualization; robustness checks.

Collaboration

- shared Git repo; weekly merge + clear data/version logs.

6.2 Risks and Mitigation

Potential risks

- Data access limits / missing regions.
- Changes in Baidu Index export rules.
- Baike pages with inconsistent templates.

Mitigation

- Backup keyword sets; cache raw HTML locally.
- Use modular scrapers; log failures and retry.
- Fall back to partial-region analysis with clear caveats.

7. Conclusions and Outlook

Expected conclusions

- Summarize distribution/correlation/trend signals by keyword and region.
- Highlight data-driven value of search-interest proxies for labor-market research.

Outlook

- Expand sources (social media, job-posting platforms).
- Add short-term forecasting or deeper policy-event causal comparisons.