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

Chenxi Zhang, Haowen Shi, Haotian Zhou

Course: Data Science and AI

Oct 30, 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, formatting)
- Descriptive Analysis (distribution, correlation, trends)
- Visualization (time trends, regional comparisons, correlations)
- Conclusions & Outlook

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?

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 collection tools to set the time window (e.g., past 5 years) and regional scope, then export search index data.

Expected fields

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

Coverage: National and major provinces.

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.

3. Data Cleaning (Part 1)

Initial Review: Check data completeness and accuracy.
Missing Values: Fill with mean/linear interpolation, or delete invalid entries; clearly 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 and handle via 3σ rule, IQR fences, or boxplot indicators.
Format Conversion: Standardize date and numeric formats; ensure ISO-8601 dates, normalized region names/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 (e.g., GDP), and among related keywords.

Examples

- Correlation matrices (Pearson/Spearman)
- Scatter plots with trend lines for pairs (e.g., unemployment-related vs. job-hunting terms)
- (If available) simple alignment with macro indicators for context

4. Descriptive Analysis — Trends

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

Examples

- Line charts of keyword indices with moving averages
- Seasonal/holiday/graduation-season annotations
- Optional: STL decomposition or simple change-point diagnostics (descriptive)

5. Data Visualization

Tools: Python (Matplotlib, Seaborn) or BI tools (e.g., FineBI).
Charts: Line charts for time trends, bar charts for regional comparisons, scatter plots for correlations.

Design guidelines

- Consistent color/labeling; all figures include source & study window
- Clear legends and annotations; readable fonts for classroom screens
- Keep code cells reproducible and parameterized

6. Conclusions and Outlook

Conclusions

- Summarize key findings from distribution, correlation, and trend analyses.
- Highlight data-driven value for employment research using search-interest proxies.

Outlook

- Expand data sources (e.g., social media posts, job postings).
- Consider short-term forecasting or deeper policy-event comparisons as next steps.