

AI Impact on IT Employment - Project README

PROJECT OVERVIEW

This project analyzes discussions from Reddit and V2EX technical communities regarding the impact of Large Language Models on IT industry employment.

RESEARCH QUESTIONS

1. How do IT practitioners view AI's impact on their careers?
2. What are the differences between Chinese and English tech communities?
3. How has practitioners' attitude evolved over time?
4. Which topics receive the most attention?

DATA SCALE

- Total Posts: 24
- Total Comments: 885
- Time Range: December 2022 - November 2025
- Data Sources: Reddit (English), V2EX (Chinese)

PROJECT STRUCTURE

```
ai_impact_data_collection/
|-- data/
|   |-- raw/           # Raw collected data
|   |-- processed/     # Cleaned data
|   |-- analysis/      # Analysis results
|   |-- visualizations/ # Visualization charts
|-- scripts/          # Analysis scripts
|-- reports/          # Analysis reports
|-- src/               # Source code modules
|-- requirements.txt    # Dependencies
```

ANALYSIS METHODS

1. Data Preprocessing
 - Data cleaning: Remove special characters, unify encoding
 - Date standardization: Unified to YYYY-MM-DD format
 - Language identification: Distinguish Chinese/English content
2. Keyword Analysis
 - Word segmentation (separate for Chinese/English)
 - Stop word filtering
 - Word frequency statistics

3. Sentiment Analysis

- Dictionary-based sentiment analysis
- Calculate positive/negative/neutral ratios
- Compute average sentiment scores

4. Topic Analysis

- Predefined topic categories
- Keyword matching classification
- Topic distribution statistics

MAIN FINDINGS

1. Sentiment Trend: From strongly negative (-1.0) in 2022 to -0.4 in 2025
2. Regional Difference: Chinese community anxiety higher than English
3. Skill Consensus: Continuous learning is key to addressing AI disruption
4. Layered Impact: Junior positions more impacted, senior relatively safe

HOW TO RUN

```
# Install dependencies
```

```
pip install -r requirements.txt
```

```
# Data merge and clean
```

```
python scripts/data_merge_and_clean_v3.py
```

```
# Text analysis
```

```
python scripts/text_analysis.py
```

```
# Generate visualizations
```

```
python scripts/generate_visualizations.py
```

VISUALIZATION CHARTS

- platform_distribution.png - Platform distribution pie chart
- year_distribution.png - Year distribution bar chart
- sentiment_analysis.png - Sentiment distribution comparison
- topic_distribution.png - Topic distribution horizontal bar
- keyword_frequency.png - Top 20 keyword frequency
- time_trend.png - Sentiment trend timeline
- overview_dashboard.png - Comprehensive dashboard
- keyword_cloud.png - Keyword bubble cloud

REPORT FILES

- reports/analysis_report.pdf - Complete analysis report

LIMITATIONS

1. Data only from Reddit and V2EX, potential sample bias
2. Sentiment analysis based on dictionary method, limited accuracy
3. Cannot distinguish professional practitioners from non-professionals
4. Some comments may contain sarcasm, affecting sentiment judgment

Document Update: November 28, 2025