Robust DCInside Stock Ticker Analysis Prompt

Task Overview

Scrape and analyze trending U.S. stock tickers from the DCInside "미국주식 갤러리 (stockus)" board over the last 7 days using a **data-driven discovery approach** to ensure no popular tickers are missed.

Critical Requirements

1. TWO-PASS TICKER DISCOVERY (MANDATORY)

Pass 1: Data-Driven Discovery

- Extract ALL potential ticker candidates from post titles using pattern matching
- Capture:
  - Capitalized 2-5 letter sequences (e.g., TSLA, NVDA, BYND, RGTI)
  - Korean stock-related terms that appear 3+ times (e.g., 비욘드, 테슬라, 엔비디아)
  - Mixed patterns (e.g., "비욘드미트", "테슬람")
- Build a frequency-ranked candidate list from the actual scraped data
- **Do NOT use a pre-defined ticker list in this phase**

Pass 2: Validation & Mapping

- For each candidate, determine if it's a valid U.S. stock ticker:
  - Check against known ticker databases
  - Validate via financial API (Yahoo Finance, etc.)
  - Map Korean terms to their ticker equivalents
- Create final validated ticker list with Korean nickname mappings
- Document any ambiguous or rejected candidates

2. Comprehensive Data Collection

Navigate to: https://gall.dcinside.com/mgallery/board/lists?id=stockus

For each post from the last 7 days, capture:

- Post ID
- Timestamp in KST (Korea Standard Time)

- Timestamp converted to EST (Eastern Time)
- Author/username
- Post title (full text)
- Number of comments
- Number of up-votes/recommendations
- Number of views
- First 200 characters of post body (if accessible without additional requests)

Scraping Strategy:

- Iterate through pages until reaching posts older than 7 days
- Handle pagination properly (50 posts per page typically)
- Implement rate limiting (0.3-0.5 second delays between requests)
- Log any blocked/inaccessible pages

3. Ticker Extraction from Posts

For each post, extract ticker mentions from:

- Post title (primary source)
- Post body snippet (if available)

Extraction Rules:

- Count multiple mentions of the same ticker in one post
- Match both English ticker symbols AND Korean nicknames
- Case-insensitive matching for ticker symbols
- Record the context of each mention (which post, when, engagement metrics)

4. Aggregation & Analysis

For each ticker, calculate:

- **Total mentions** across all posts
- **Unique post count** (how many different posts mentioned it)
- **Total comments** on posts mentioning this ticker
- **Total recommends/up-votes** on posts mentioning this ticker
- **Total views** on posts mentioning this ticker
- **Engagement score** = comments + recommends
- **Hourly time series** of mentions (6-hour buckets acceptable if hourly is too granular)

5. Small-Move Potential Flagging

For each ticker, attempt to fetch:

- Market capitalization (USD)

- Shares outstanding
- Float (publicly traded shares)
- Short interest % (if available)
- Borrow fee rate (if available)

Flagging Criteria:

- Market cap < $2 billion USD: **Yes**
- Float < 100 million shares: **Yes**
- Short interest > 10% OR borrow fee > 5%: **Bonus flag**

Flag Values:

- "Yes" = Meets market cap AND float criteria
- "Yes*" = Meets all criteria including short interest/borrow fee
- "No" = Does not meet criteria
- "N/A" = ETF or leveraged product
- "Unknown" = Unable to fetch market data

Handling Missing Data:

- If API data unavailable, use publicly known estimates for well-known tickers
- Document which tickers have estimated vs. actual data
- Do NOT skip tickers just because market data is unavailable

6. Excel Output Requirements

Create file: DCInside_StockUS_7day_trending.xlsx

Sheet 1: Summary

- Columns: Ticker | Total Mentions | Unique Posts | Comments | Recommends | Views | Engagement Score | Market Cap | Float | Small-Move Potential | Notes
- Sort by Total Mentions (descending)
- Highlight "Yes" flags in green
- Format numbers with thousand separators

Sheet 2: Report

- Executive summary with key statistics
- Top 10 tickers table
- Small-move potential tickers detail
- Key insights and trends

Sheet 3: TimeSeries

- Rows: Tickers (top 50 by mentions)
- Columns: Time buckets (hourly or 6-hour intervals in EST)

- Values: Mention counts per time bucket
- Useful for identifying mention velocity and trending patterns

Sheet 4: All Posts

- Complete list of scraped posts
- Columns: Post ID | DateTime (KST) | DateTime (EST) | Title | Author | Comments | Recommends | Views
- Preserve original data for verification

Sheet 5: Ticker Mentions Detail

- Each row = one ticker mention in one post
- Columns: Ticker | DateTime (KST) | DateTime (EST) | Post ID | Post Title | Mention Count | Comments | Recommends | Views
- Allows drill-down analysis

Sheet 6: Discovery Log

- Document the discovery process
- Show all candidates found in Pass 1
- Show validation results
- Show Korean nickname mappings
- Helps verify completeness

7. Output Summary Requirements

Provide in your response:

- **Number of posts scraped** (exact count)
- **Number of unique tickers found** (after validation)
- **Top 3 tickers by mention count** (with counts and post counts)
- **Number of tickers flagged for small-move potential** (with names)
- **List of all small-move flagged tickers** with brief details
- **Any tickers that were discovered but couldn't be validated** (for manual review)

8. Quality Assurance Checks

Before finalizing, verify:

- ☐Discovery phase found at least 50+ ticker candidates
- ☐Final validated list includes at least 20+ unique tickers
- ☐Top 10 tickers account for at least 50% of total mentions
- ☐Korean nicknames are properly mapped (check 테슬라, 엔비디아, 비욘드, etc.)
- ☐Time series data covers the full 7-day period
- ☐All Excel sheets are properly formatted and readable

- ☐Small-move flags are applied consistently
- ☐No obvious popular tickers are missing (manual spot check against post titles)

Additional Instructions

Handling Errors

- If CAPTCHA or login wall encountered, report which pages were blocked
- Continue with accessible data and document limitations
- Do not fail silently - report all issues

Korean Language Handling

- Recognize common Korean stock terms: 주식, 매수, 매도, 상승, 하락, etc.
- Map Korean company names to tickers: 테슬라→TSLA, 애플→AAPL, etc.
- Handle spacing variations: "비욘드미트" vs "비욘드 미트"

Performance Optimization

- Scrape efficiently (parallel requests if possible)
- Cache API responses to avoid redundant calls
- Provide progress updates every 10-20 pages

Validation & Transparency

- Show sample of discovered candidates before validation
- Explain any ticker rejections
- Provide confidence levels for market data estimates
- Include data sources in notes

Example Discovery Output Format

```
PASS 1: DISCOVERY
```

```
================
Found 156 ticker candidates from 7,491 posts:
```

```
Top 20 by frequency:
1. TSLA (261 mentions) ✓ Valid ticker
2. 비욘드 (72 mentions) → Maps to BYND ✓
3. RGTI (215 mentions) ✓ Valid ticker
4. INTC (126 mentions) ✓ Valid ticker
5. GOOGL (110 mentions) ✓ Valid ticker
...

Rejected candidates:
- "USA" (common word, not a ticker)
- "NEW" (common word, not a ticker)
- "AI" (too ambiguous, could be C3.AI but not clear)

PASS 2: VALIDATION
==================
Validated 32 unique U.S. stock tickers
Mapped 15 Korean nicknames to tickers
Rejected 109 false positives

Korean Nickname Mappings:
- 테슬라, 테슬 → TSLA
- 비욘드, 비욘드미트 → BYND
- 엔비디아, 엔비 → NVDA
```

```
...
```

Success Criteria

The analysis is complete and successful when:

1. ✓ All posts from last 7 days are scraped
2. ✓ Ticker discovery is data-driven, not pre-defined
3. ✓ No popular tickers are missing (validated by spot-checking post titles)
4. ✓ Korean nicknames are properly captured and mapped
5. ✓ Excel file contains all 6 required sheets
6. ✓ Small-move potential flags are accurate and documented
7. ✓ Time series data enables trend analysis
8. ✓ All data is traceable back to source posts