News API Services Explanation

Overview

The News API (https://newsapi.org/) is a simple HTTP REST API for searching and retrieving live articles from all over the web. It provides access to millions of articles from over 150,000 news sources and blogs.

Available Endpoints

1. Everything Endpoint (`/v2/everything`)

Purpose: Search through millions of articles from over 150,000 large and small news sources and blogs. **Use Case**: Article discovery and analysis.

Key Parameters:

- `apiKey` (required): Your API key
- `q`: Keywords or phrases to search for
- `searchIn`: Fields to search (title, description, content)
- `sources`: Specific news sources (max 20)
- `domains`: Specific domains to search
- `excludeDomains`: Domains to exclude
- `from`/`to`: Date range (ISO 8601 format)
- `language`: 2-letter language code
- `sortBy`: Sort order (relevancy, popularity, publishedAt)
- `pageSize`: Results per page (max 100)
- `page`: Page number for pagination

Response Fields:

- `status`: Request success status
- `totalResults`: Total available results
- `articles`: Array of article objects containing:
  - `source`: Source identifier and name
  - `author`: Article author
  - `title`: Article headline
  - `description`: Article snippet
  - `url`: Direct article URL
  - `urlToImage`: Article image URL
  - `publishedAt`: Publication date/time (UTC)
  - `content`: Article content (truncated to 200 chars)
  - **Note**: The News API only provides truncated content (200 characters max). For full text articles, you need to scrape the article URLs directly.

2. Top Headlines Endpoint (`/v2/top-headlines`)

Purpose: Provides live top and breaking headlines for countries, categories, or specific sources. **Use Case**: News tickers and headline displays.

Key Parameters:

- `apiKey` (required): Your API key
- `country`: 2-letter country code (e.g., 'us')

- `category`: News category (business, entertainment, general, health, science, sports, technology)
- `sources`: Specific news sources
- `q`: Keywords to search for
- `pageSize`: Results per page (max 100)
- `page`: Page number

Response Fields: Same as Everything endpoint

3. Sources Endpoint (`/v2/top-headlines/sources`)

Purpose: Returns available news publishers for top headlines. **Use Case**: Tracking available publishers and source discovery.

Key Parameters:

- `apiKey` (required): Your API key
- `category`: Filter by news category
- `language`: Filter by language
- `country`: Filter by country

Response Fields:

- `status`: Request success status
- `sources`: Array of source objects containing:
  - `id`: Source identifier
  - `name`: Source name
  - `description`: Source description
  - `url`: Homepage URL
  - `category`: News category
  - `language`: Source language
  - `country`: Source country

Search Capabilities

Advanced Search Features

- **Exact phrases**: Use quotes ("exact phrase")
- **Required terms**: Use + symbol (+bitcoin)
- **Excluded terms**: Use - symbol (-bitcoin)
- **Boolean operators**: AND, OR, NOT with parentheses
- **URL encoding**: Required for special characters

Supported Languages

Arabic (ar), German (de), English (en), Spanish (es), French (fr), Hebrew (he), Italian (it), Dutch (nl), Norwegian (no), Portuguese (pt), Russian (ru), Swedish (sv), Chinese (ud), Chinese (zh)

Supported Countries

Over 50 countries including US, UK, Canada, Australia, Germany, France, etc.

Rate Limits and Usage

- Free tier available for development
- API key required for all requests
- Rate limits apply based on subscription plan
- Results are paginated for large datasets

Full Text Content Capabilities

API Limitations

- **News API Content**: Limited to 200 characters maximum
- **No Full Text**: The API does not provide complete article content
- **URL Access**: Full articles are available via the provided URLs

Getting Full Text Content

To access full text articles, you need to:

1. **Use the article URLs** provided by the News API
2. **Web scraping techniques** to extract content from the original sources
3. **Respect robots.txt** and website terms of service
4. **Handle different website structures** (each news site has different HTML layouts)

Implementation Example

```python
import requests
from bs4 import BeautifulSoup

def fetch_full_text(url):
    headers = {'User-Agent': 'Mozilla/5.0...'}
    response = requests.get(url, headers=headers)
    soup = BeautifulSoup(response.content, 'html.parser')

    # Remove scripts and styles
    for script in soup(["script", "style"]):
        script.decompose()

    # Find article content
    content = soup.find('article') or soup.find('.article-content')
    return content.get_text() if content else None
```

Considerations

- **Legal compliance**: Check website terms of service
- **Rate limiting**: Don't overwhelm servers with requests
- **Content quality**: Scraped content may include ads and navigation
- **Reliability**: Websites may block automated access

Use Cases

1. **Content Discovery**: Find articles on specific topics
2. **Trend Analysis**: Track mentions over time
3. **News Aggregation**: Build news tickers and feeds
4. **Brand Monitoring**: Track company/product mentions
5. **Research**: Analyze news patterns and sources

6. **Full Text Analysis**: Extract complete articles for detailed analysis