

# Google News RSS Ingestion Pipeline - File Structure

```
/rss_collector/
├── config/
│   ├── feeds.json          # Keywords configuration
│   └── settings.json       # System settings (proxy, schedule, etc.)
├── logs/                   # Log directory (auto-populated)
│   └── .gitkeep
├── feeds/                  # Output directory (auto-populated)
│   └── .gitkeep
├── src/                   # Source code
│   ├── __init__.py         # Package initialization
│   ├── main.py             # Entry point script
│   ├── config_manager.py   # Configuration loading & validation
│   ├── rss_fetcher.py      # HTTP client for RSS feeds
│   ├── rss_parser.py       # RSS parsing & normalization
│   ├── storage_manager.py  # File I/O and deduplication
│   ├── scheduler.py        # Job scheduling
│   └── utils/              # Utility modules
│       ├── __init__.py
│       ├── logging_utils.py # Logging setup
│       ├── proxy_utils.py   # Proxy configuration
│       └── helpers.py      # Misc helper functions
├── tests/                 # Unit and integration tests
│   ├── __init__.py
│   ├── test_config_manager.py
│   ├── test_rss_fetcher.py
│   ├── test_rss_parser.py
│   ├── test_storage_manager.py
│   ├── test_scheduler.py
│   └── fixtures/          # Test fixtures
│       ├── sample_rss.xml
│       └── sample_feeds.json
├── run.py                 # Main execution script
├── setup.py               # Package installation script
├── requirements.txt       # Dependencies
└── README.md              # Project documentation
```

Note: Each Python module is designed to be focused on a specific responsibility and kept under 400 lines of code. The modular structure allows for easy testing and maintenance.

