USPTO Patent Data Collector

Overview

Downloads and extracts patent data files from the USPTO (United States Patent and Trademark Office) database. Supports both grant and application patents with automated downloading and extraction of XML files.

Installation

Prerequisites

- Python 3.x
- Required packages:

```
pip install requests argparse tqdm
```

Setup

1. Clone the repository:

```
git clone https://github.com/repository_name/prophetic_patents.git
cd prophetic_patents
```

2. Install dependencies:

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

3. Ensure write permissions in target directory

Usage

Basic Example

```
python scripts/collector.py
```

Advanced Examples

```
# Download grant patents for 2015
python scripts/collector.py --kind grant --year 2015
```

```
# Use custom download path
python scripts/collector.py --kind grant --year 2015 --path "D:\patents"
```

Command Line Arguments

Argument	Description	Default	Required
`--kind`	Patent type (`grant` or `application`)	`grant`	No
`--year`	Year to download (1976-2025)	2015	No
`--path`	Custom download directory	`data`	No

Directory Structure

```
data/
├── temp/
│   └── zipped_{kind}_files_{year}/    # Temporary zip files
└── patent_{kind}_{year}/              # Extracted XML files
```

Features

Input Validation

- ✓ Year validation (1976-2025)
- ✓ Patent type validation (grant/application)
- ✓ Path validation and creation

Download Management

- ✓ Automatic USPTO file downloads
- ✓ Progress tracking
- ✓ Error handling
- ✓ Temporary file management

File Extraction

- ✓ Automatic unzipping
- ✓ Organised output structure
- ✓ Error handling

API Reference

Functions

`validate_path(path: str) -> str`

Validates and creates directories if they don't exist.

Parameters:

- `path`: Directory path to validate

Returns:

- Validated path or raises `ArgumentTypeError`

`main()`

Main function orchestrating download and extraction process.

Error Handling

The script handles:

- ⚠ Invalid year ranges
- 🌐 Network connection issues
- 📁 File system permissions
- ❌ Invalid patent types
- 📥 Download failures
- 📦 Extraction failures

Example Output

```
Downloading XML files from USPTO for 2015
[Download progress...]
Download complete. Unzipping files...
[Extraction progress...]
Unzip complete
```

Dependencies

Module	Description
`utilities.utils_clean.download_patents_pto`	USPTO downloads
`utilities.utils_clean.unzip_files`	File extraction
`utilities.utils_clean.validate_year`	Year validation
`utilities.utils_clean.validate_kind`	Patent type validation

Best Practices

System Requirements

- Disk Space: 500MB per year minimum

- Memory: 4GB RAM recommended
- Internet: Stable connection required

Performance Tips

1. Download one year at a time
2. Use SSD for faster extraction
3. Close other applications during processing

Troubleshooting

Common Issues

1. **Download Failures**
    - Check internet connection
    - Verify USPTO server status
2. **Extraction Errors**
    - Ensure sufficient disk space
    - Check file permissions
3. **Path Errors**
    - Use absolute paths
    - Avoid special characters