Patent Test Dataset Creator Documentation

Overview

This script creates a test dataset for a specified year using Freilich dataset and USPTO XML patent files. It supports downloading patent data directly from USPTO if needed.

Prerequisites

- Python 3.x
- Required packages:

```
pip install argparse requests xlwings BeautifulSoup lxml
```

- Access to Freilich dataset (`.xlsb` format)

Usage

Basic Usage

```
python test_dataset_creator.py --year 2015 --freilich-path
Freilich.Data.Compressed.xlsb --xml-path patent_grants_2015
```

Download and Process

```
python test_dataset_creator.py --year 2015 --freilich-path
Freilich.Data.Compressed.xlsb --download
```

Arguments

Argument	Description	Default
`--year`	Year to analyse (1976-2025)	2015
`--freilich-path`	Path to Freilich dataset	Freilich.Data.Compressed.xlsb
`--xml-path`	Directory containing XML files	patent_grants_2015
`--download`	Download XML files from USPTO	False

Features

Year Validation

- Accepts years between 1976 and 2025
- Automatically validates input year format
- Raises clear error messages for invalid years

File Path Validation

- Verifies existence of input files
- Validates both Freilich dataset and XML directory paths
- Provides clear error messages for missing files

USPTO Download Support

1. Downloads patent grant files for specified year
2. Automatically unzips downloaded files
3. Shows progress indicators during download and unzip
4. Creates organised directory structure:
    - Downloaded files: `zipped_files_[YEAR]`
    - Extracted files: `patent_grants_[YEAR]`

Output

- Displays processing status and progress
- Shows number of patents extracted
- Lists sample document numbers
- Provides error messages for any failures

Error Handling

- Handles download failures gracefully
- Manages unzip operation errors
- Validates input files and paths
- Catches and reports dataset creation errors

Directory Structure

```
project_root/
│
├── test_dataset_creator.py
├── Freilich.Data.Compressed.xlsb
├── zipped_files_[YEAR]/         # Created when using --download
└── patent_grants_[YEAR]/        # Contains extracted XML files
```

Example Output

```
Processing year: 2015
Using Freilich data from: Freilich.Data.Compressed.xlsb
Using XML files from: patent_grants_2015
```

```
Number of patents extracted: 1234
Sample document numbers: ['US123456', 'US234567', ...]
```

Dependencies

- test_dataset_utils.py: Contains `create_test_dataset_from_freilich()`
- utils_clean.py: Contains `download_patents_pto()` and `unzip_files()`