Python Environment Setup for CIRF Research

Option 1: Using Conda (Recommended)

1. Install Anaconda/Miniconda

Download from: https://docs.conda.io/en/latest/miniconda.html

2. Create CIRF Research Environment

bash

```
# Create new environment with Python 3.11
conda create -n cirf-research python=3.11 -y

# Activate environment
conda activate cirf-research

# Install core data science packages
conda install pandas numpy scipy scikit-learn jupyter matplotlib seaborn -y

# Install database packages
conda install sqlite postgresql psycopg2 sqlalchemy -y

# Install NLP packages
conda install spacy nltk -y

# Install web scraping packages
conda install beautifulsoup4 requests selenium -y

# Install additional useful packages
conda install tqdm python-dotenv openpyxl xlsxwriter -y
```

3. Install Additional Packages via pip

bash

```
# Still in cirf-research environment
pip install newspaper3k
pip install scholarly
pip install crossref-commons
pip install python-semantic-scholar
pip install textblob
pip install wordcloud
pip install plotly
pip install dash

# Download spaCy language models
python -m spacy download en_core_web_sm
python -m spacy download en_core_web_lg
```

Option 2: Using pip and venv

1. Create Virtual Environment

```bash
# Create project directory
mkdir cirf-research
cd cirf-research

# Create virtual environment
python -m venv cirf-env

# Activate environment (Windows)
cirf-env\Scripts\activate

# Activate environment (Mac/Linux)
source cirf-env/bin/activate
```

2. Install Required Packages

```bash
# Upgrade pip
pip install --upgrade pip

# Install all required packages
pip install -r requirements.txt
```

Verify Installation

Test Script to Verify Setup

python

```
# test_environment.py
import sys
print(f"Python version: {sys.version}")

try:
    import pandas as pd
    print(f"✓ Pandas {pd.__version__}")
except ImportError:
    print("✗ Pandas not installed")

try:
    import numpy as np
    print(f"✓ NumPy {np.__version__}")
except ImportError:
    print("✗ NumPy not installed")

try:
    import sklearn
    print(f"✓ Scikit-learn {sklearn.__version__}")
except ImportError:
    print("✗ Scikit-learn not installed")

try:
    import spacy
    print(f"✓ spaCy {spacy.__version__}")
    # Test language model
    nlp = spacy.load("en_core_web_sm")
    print("✓ English language model loaded")
except ImportError:
    print("✗ spaCy not installed")
except OSError:
    print("✗ English language model not downloaded")

try:
    from bs4 import BeautifulSoup
    print("✓ BeautifulSoup installed")
except ImportError:
    print("✗ BeautifulSoup not installed")

try:
    import requests
    print(f"✓ Requests installed")
except ImportError:
    print("✗ Requests not installed")

try:
    import sqlite3
    print("✓ SQLite available")
except ImportError:
    print("✗ SQLite not available")

try:
    import sqlalchemy
    print(f"✓ SQLAlchemy {sqlalchemy.__version__}")
except ImportError:
    print("✗ SQLAlchemy not installed")

print("\nEnvironment setup verification complete!")
```

IDE Setup Recommendations

VS Code Extensions

- Python
- Jupyter
- SQLite Viewer
- Python Docstring Generator
- GitLens

PyCharm Setup

- Enable Scientific Mode
- Install Database Tools plugin
- Configure Python interpreter to use your virtual environment

Environment Variables Setup

Create a `.env` file in your project root:

```
bash
```

```
# Database Configuration
DATABASE_URL=sqlite:///cirf_research.db
POSTGRES_URL=postgresql://username:password@localhost:5432/cirf_research

# API Keys (add as needed)
CROSSREF_EMAIL=your.email@domain.com
SEMANTIC_SCHOLAR_API_KEY=your_api_key_here

# Project Configuration
PROJECT_ROOT=/path/to/your/project
DATA_DIR=/path/to/your/data
LOGS_DIR=/path/to/your/logs

# Web Scraping Configuration
USER_AGENT=CIRF-Research-Bot/1.0 (your.email@domain.com)
REQUEST_DELAY=1
MAX_RETRIES=3

# NLP Configuration
SPACY_MODEL=en_core_web_lg
BATCH_SIZE=1000
```

Troubleshooting Common Issues

spaCy Model Download Issues

bash

```bash
# If language model download fails
python -m spacy download en_core_web_sm --user
python -m spacy download en_core_web_lg --user
```

SSL Certificate Issues

bash

```bash
# For corporate networks with SSL issues
pip install --trusted-host pypi.org --trusted-host pypi.python.org --trusted-host files.pythonhosted.org <package_name>
```

Memory Issues with Large Models

- Start with smaller spaCy model (en_core_web_sm)
- Increase system memory allocation
- Use batch processing for large datasets

Next Steps After Setup

1. Run the test script to verify installation
2. Create project directory structure
3. Initialize git repository
4. Set up database schema
5. Begin building data collection infrastructure