

User Manual

Python code for hierarchical cluster analysis of detected R-strategies from rule-based NLP on 500 circular economy definitions

1. Introduction

This manual provides step-by-step instructions for using the rule-based NLP model to extract R-strategies and themes beyond, and performing Hierarchical Cluster Analysis (HCA) on the extracted strategies from 500 circular economy (CE) definitions. The approach is designed for computational and thematic analysis of CE definitions, enabling structured insights into how circular economy strategies are represented.

The rule-based NLP model utilizes 600 predefined keywords to classify CE definitions into 10R strategies (R0–R9) and broader sustainability themes, ensuring high interpretability and reproducibility. The HCA model applies Ward’s clustering method to explore relationships between strategies.

This Python code was optimized and debugged using ChatGPT-4o to ensure efficiency, accuracy, and clarity in implementation.

2. Software & Requirements

Before executing the code, ensure that the required software and libraries are installed.

2.1. Programming Environment

- **Google Colab (Recommended)** – No installation required; runs on cloud.
- **Jupyter Notebook** – For local execution using Anaconda or VS Code.
- **Python 3.7+** – Required for NLP and clustering libraries.

2.2. Required Python Libraries

Install the required Python libraries using the following command:

User Manual

```
python
CopyEdit
pip install spacy pandas matplotlib scipy ipywidgets
```

Libraries Used:

- **spaCy** → Natural Language Processing for keyword extraction.
- **pandas** → Data handling and structured processing.
- **matplotlib** → Visualizing clustering results.
- **scipy** → Hierarchical clustering using Ward's method.
- **ipywidgets** → Interactive UI elements for uploading and downloading files.

2.3. Downloading the NLP Model

Before running the **rule-based NLP model**, download the required **spaCy language model**:

```
python
CopyEdit
import spacy
spacy.cli.download("en_core_web_sm")
```

This ensures proper tokenization and linguistic processing.

3. Using the Rule-Based NLP Model

This section explains how to **extract R-strategies and broader themes from CE definitions**.

3.1. Input File Requirements

- The input file must be in **CSV format** with **two mandatory columns**:
 - **ID** → A unique identifier for each CE definition.
 - **Definition** → The circular economy definition text.
- Example **CSV format**:

ID Definition

- 1 "A circular economy aims to eliminate waste by designing products for longevity."
- 2 "Recycling and reuse are key strategies for reducing material consumption."

3.2. Running the NLP Model in Google Colab

1. Upload your **CSV file** using the interactive widget.

User Manual

2. The model detects and classifies **R0–R9 strategies** and **themes beyond**.
3. The processed results are displayed in a structured format.
4. You can download the **processed CSV file** containing extracted strategies.

3.3. Output File Format

The output file (`r0_r9_detected_definitions.csv`) contains:

- **Binary indicators (1/0) for each R-strategy** (e.g., R0–R9).
- **Extracted keywords per strategy** for interpretability.
- **Total keyword count per strategy** for quantitative analysis.

ID	R0	R1	R2	R3	R4	R5	R6	R7	R8	R9
1	1	0	1	0	0	0	0	0	0	0
2	0	0	1	1	0	0	0	0	1	0

4. Hierarchical Cluster Analysis of R-Strategies

This section explains how to perform **clustering on extracted strategies**.

4.1. Input File for Clustering

- Uses the **processed CSV file** (`r0_r9_detected_definitions.csv`) from the NLP model.
- This file contains **binary (1/0) indicators** of **R-strategies detected** in definitions.

4.2. Running the Hierarchical Clustering Model

1. Upload the **processed CSV file** from the NLP step.
2. The script converts binary values into a **distance matrix**.
3. **Ward's method** is applied to identify relationships among strategies.
4. A **horizontal dendrogram** is generated and displayed.
5. Download the clustering result as an **image file** (`horizontal_dendrogram.png`).

4.3. Output File: Dendrogram Image

- **Visualizes how R-strategies cluster together.**
 - **Key findings:** Unexpected placement of R0 clustering with R9 instead of R1.
-

5. Summary of the Workflow

User Manual

Step	Action	Output
1. Upload CE definitions CSV	Upload file to Google Colab	Input file ready
2. Extract R-strategies	Run rule-based NLP model	Processed CSV (r0_r9_detected_definitions.csv)
3. Upload processed CSV	Use file for clustering	Input for hierarchical analysis
4. Run hierarchical clustering	Generate dendrogram	Clustering results
5. Download results	Save processed CSV and dendrogram image	Ready for analysis

6. Key Features & Advantages

- ✓ **Expandable & Systematic** → The framework can be scaled to analyze larger datasets.
- ✓ **High Interpretability** → Rule-based NLP ensures transparent keyword classification.
- ✓ **Robust Methodology** → NLP + Hierarchical Clustering provides deep insights into CE definitions.
- ✓ **Optimized & Debugged** → ChatGPT-4o was used for **optimization and debugging**.
- ✓ **Comprehensive Analysis** → Covers **both R-strategies & broader sustainability themes**.

This **structured, reproducible, and scalable** framework lays the foundation for **future research**, enabling **longitudinal analyses of CE definitions and sustainability discourse**.