NL Taxonomy Mapper V3

Executive Summary

Document Date: January 08, 2026

Overview

NL Taxonomy Mapper V3 is an automated content classification system that maps URLs from semantic carriers to a hierarchical taxonomy structure. The application uses advanced fuzzy string matching algorithms to intelligently categorize web content based on extracted keywords, with specialized support for Dutch language variations.

Business Problem

Organizations often maintain large collections of web content that need to be systematically categorized into predefined taxonomy structures. Manual classification is time-consuming, inconsistent, and difficult to scale. The NL Taxonomy Mapper addresses this challenge by automating the classification process while maintaining high accuracy rates.

Solution

The application processes URLs with their associated keywords and matches them against a comprehensive taxonomy structure using intelligent fuzzy matching algorithms. The system includes:

• Configurable similarity thresholds (50-100%) for precision control
• Built-in Dutch language synonym expansion for improved recall
• Automatic deduplication to prevent redundant classifications
• Hierarchical segment auto-addition to preserve taxonomy structure
• User-friendly GUI with real-time progress tracking

Key Capabilities

Capability	Description
Processing Volume	Handles 200+ URLs with 10 keywords each
Taxonomy Coverage	50+ hierarchical taxonomy topics
Match Accuracy	~92% successful classification rate
Processing Speed	30-60 seconds for full dataset
Language Support	Dutch with synonym expansion

Technical Architecture

The system is built on Python with a clean separation between core matching logic and user interface. It uses the Levenshtein distance algorithm for fuzzy string matching, pandas for data processing, and Tkinter for the GUI. The architecture supports both command-line and graphical interfaces.

Input/Output Specifications

Inputs:
• Semantic Carriers List: Excel file containing URLs and extracted keywords (Keyword 1-10)
• Taxonomy Structure: Excel file with hierarchical taxonomy (Product → Domain → Segment → Topics)

Output:
• Classification Results: Excel file mapping each URL to matched taxonomy topics
• Unmapped URLs: Flagged for manual review with Domain='UNMAPPED'
• Multiple matches per URL supported for comprehensive categorization

Business Benefits

Efficiency: Reduces manual classification time from hours to minutes
Consistency: Applies uniform classification rules across all content
Scalability: Easily handles growing content volumes
Accuracy: Achieves 90%+ match rates with configurable precision
Transparency: Identifies unmapped content for continuous improvement
Flexibility: Adjustable thresholds and extensible synonym dictionary

Typical Use Cases

• Content management system taxonomy alignment
• Knowledge base article categorization
• Help documentation organization
• SEO content classification
• Information architecture validation

Getting Started

The application includes pre-built template files for both input formats. Users simply populate the templates with their URLs and taxonomy structure, launch the GUI application, select their input files, adjust the similarity threshold, and click 'Start Matching'. Results are automatically exported to Excel for review and further processing.

For technical documentation, see CLAUDE.md in the project repository.
For questions or support, contact the development team.