MASTER di I LIVELLO in ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE

ANNO ACCADEMICO 2024/2025

Loccioni Gabriele

Abstract

Il presente lavoro di tesi descrive la progettazione e l'implementazione di un sistema avanzato per la normalizzazione degli indirizzi stradali sviluppato in un progetto di collaborazione tra Metis Ricerche e CSI Piemonte. L'obiettivo principale è la sostituzione dei sistemi legacy deterministici (rule-based) con una soluzione open source flessibile, capace di riconciliare dati eterogenei e degradati verso un golden record autoritativo. L'architettura proposta si basa su un approccio ibrido che vede l'integrazione di due strumenti tecnologici complementari. L'efficienza statistica del sistema è garantita da libpostal, una libreria scritta nativamente in linguaggio C, specializzata nella scomposizione strutturale degli indirizzi su scala globale attraverso modelli probabilistici. Tale componente viene potenziata dalla precisione semantica di spaCy, una libreria Python per il Natural Language Processing, all'interno della quale è stato sviluppato un modello di Named Entity Recognition sottoposto a fine-tuning specifico sui dataset della toponoma□stica regionale piemontese. Grazie a questa sinergia, il sistema introduce un'innovativa strategia di matching gerarchico "a imbuto" che genera embedding numerici basati sulla confidenza del fuzzy matching ai diversi livelli territoriali di Provincia, Comune e Via. I risultati sperimentali confermano l'efficacia della metodologia: l'introduzione della logica probabilistica ha permesso un incremento del 9,4% nella capacità di riconciliazione dei comuni rispetto ai metodi tradizionali. Inoltre, l'analisi del modello decisionale ha mostrato una contrazione dell'albero di decisione sul solo parametro del numero civico, dimostrando come la complessità del problema sia stata risolta a monte dalla struttura gerarchica dei dati. Il sistema ha raggiunto un F1-Score di 1.0 sui dataset di test, garantendo la totale assenza di falsi positivi e assicurando la massima integrità del patrimonio informativo dell'Ente.