Master Universitario in: "ANALISI DATI PER LA BUSINESS INTELLIGENCE"

A.A. 2015-2016

Titolo della tesi: Estrazione dei temi e classificazione di testi di Wikipedia con l'apprendimento automatico

Autore: Berta Mirko

Abstract

Il presente lavoro ha come obiettivo lo sviluppo di un modello per l'estrazione di informazioni contenute in documenti testuali scritti in lingua italiana tramite l'utilizzo di strumenti di apprendimento automatico. Successivamente all'estrazione dei contenuti il modello mette a confronto diverse tecniche algoritmiche di classificazione dei testi utilizzando algoritmi di machine learning. Si utilizzano due approcci differenti per quanto concerne la classificazione cercando così di scoprire e valutare i benefici di un approccio non tradizionale al problema della classificazione documentale. Come strumento concettuale si utilizza il calcolo e la rappresentazione dei termini e dei documenti in uno spazio astratto rappresentato in forma matriciale. Il progetto comprende la creazione del dataset contenete i testi di Wikipedia, l'estrazione automatica delle parole che descrivono meglio una determinata categoria di testi e la successiva classificazione dei testi nella rispettiva categoria. Si è utilizzato la decomposizione della matrice in numeri positivi per cercare di scoprire quali fossero i termini che descrivevano meglio una determinata categoria. I risultati sperimentali dimostrano che il problema risiede molto dalla qualità dei testi, dalla quantità di parole all'interno dei testi e dalla numerosità del campione di ogni categoria.