UNIVERSITY OF GENOA
DEPARTMENT OF INFORMATICS, BIOENGINEERING, ROBOTICS AND SYSTEMS ENGINEERING
MASTER'S PROGRAM IN BIOENGINEERING

Dibris

Thesis Project Form

Title (tentative): Progettazione, definizione e verifica di una pipeline di pre-elaborazione di note cliniche in italiano

Thesis advisor(s): Giacomini Mauro, Sara Mora, Daniele Roberto Giacobbe

E-mail: Mauro.Giacomini@dibris.unige.it

Address: Via Opera Pia 13

Phone: (+39) 010 33 56546

Description

Motivation and application domain

Questa tesi si concentra sulla gestione di dati prodotti in contesti medici. In questo ambito, spesso una buona parte di informazione Ã¨ espressa usando il linguaggio naturale in varie tipologie di documentazione, come in referti, note e report di laboratorio. Il corretto ed efficiente utilizzo di queste informazioni Ã¨ ritenuto fondamentale per moltissimi compiti in ambito di assistenza sanitaria. Ãˆ necessario sviluppare un algoritmo in grado di comprendere e gestire in maniera automatica le grandi moli di dati testuali prodotte, al fine di estrarre informazioni e manipolare correttamente il testo, senza la necessitÃ della supervisione da parte di personale. In particolare, la necessitÃ di concentrarsi sul linguaggio medico/tecnico italiano Ã¨ dovuta al fatto che la ricerca in questo ambito si Ã¨ concentrata principalmente sulla lingua inglese.

General objectives and main activities

Lo scopo della tesi Ã¨ quello di progettare e sviluppare presso la clinica di malattie infettive e tropicali dellâ€™IRCCS Policlinico San Martino un algoritmo di elaborazione di dati testuali raccolti nellâ€™ambito degli studi multicentrici del progetto MULTI-SITA. Il progetto si divide in:
- ricerca riguardante gli strumenti piÃ¹ recenti disponibili al fine di processare il linguaggio naturale in lingua italiana, con particolare focus posto sullâ€™implementazione, lâ€™utilizzo e il fine-tuning dei modelli Transformers di Hugging face (in particolare BERT e BART);
- sviluppo di una pipeline completa, ovvero un algoritmo in grado di gestire e comprendere correttamente dati testuali, quali report su pazienti presso il reparto di malattie infettive e tropicali, contenenti refusi e termini peculiari;
- verifica della congruitÃ e correzione delle criticitÃ dellâ€™algoritmo sviluppato, in modo di individuare criticitÃ in ognuno degli step selezionati per comporre la pipeline dello sviluppo e quindi proporre elementi di miglioramento del modello.

Training Objectives (technical/analytical tools, experimental methodologies)

Nel corso della tesi, lo studente utilizzerÃ e apprenderÃ le seguenti competenze:
- Progettazione e sviluppo in linguaggio Python (ambiente Anaconda Navigator)
- Strumenti di modifica e di interrogazione di database (Microsoft SQL Server)
- Utilizzo di algoritmi di machine learning quali gli ultimi modelli Transformers
- Uso di standard e vocabolari medici standardizzati

Place(s) where the thesis work will be carried out: DIBRIS, IRCCS Policlinico San Martino

Additional information

Maximum number of students: 1