UNIVERSITÀ
DEGLI STUDI
DI PALERMO

Gene-Disease association analyzing the scientific literature

Salvatore Calderaro

Indice

1 Introduzione

In questo progetto è stato implementato un sistema che dato in input un gene (il suo ID), controlla se l'ID inserito appartiene a un gene realmente esistente, se il controllo va a buon fine vengono memorizzate all'interno di un dataframe una serie di informazioni inerenti il gene, in particolare: la tassonomia, l'ID, il simbolo e il nome ufficiale completo. Fatto ciò si procede con l'estrazione dalla piattaforma *PubMed* - mediante *web scraping* - dei duecento articoli più rilevanti (se disponibili) in cui il gene è stato studiato. Di quest'ultimi vengono estratti e caricati all'interno di un dataframe titolo ed abstract. Per ogni articolo presente nel dataframe viene effettuato un pre-processing: eliminazione delle stop words, della punteggiatura e altre tecniche di natural language processing che verranno descritte più approfonditamente nella prossima sezione. Fatto ciò si procede con la *Named Entity Recognition* (NER), per stabilire all'interno di un testo quali parole o insiemi parole possono essere etichettate come malattie. La lista di malattie cosi ottenuta viene filtrata in modo tale da eliminare eventuali duplicati o parole che si hanno a che fare con l'ambito biomedico, ma che in realtà non sono nomi riconducibili a malattie. Infine per valutare la bontà dei risultati ottenuti, la lista di malattie viene confrontata, mediante *fuzzy string matching* con la lista di malattie che sono associate al gene che si sta studiando estratta dal database *DisGenNet*. In output verrà restituita: la percentuale di malattie esatte trovate, la lista della malattie e una wordcloud per rappresentare graficamente i risultati. Il codice sorgente del software è reperibile alla seguente repository GitHub: https://github.com/Calder10/Gene-Desease-Association.

2 Descrizione del software

Il linguaggio di programmazione utilizzato per l'implementazione del software è *Python*. Come IDE per effettuare lo sviluppo è stato scelto *Atom*. Le librerie utilizzate per l'implementazione del software sono le seguenti:

- *pyspark*:
- *nltk*;
- *biopython*;
- *entrez*;
- *scispacy*;
- *spacy*;
- *textblob*;
- *fuzzywuzzy*;
- *wordcloud*;
- *matplotlib*.

Di seguito verranno descritte in dettaglio le varie fasi della progettazione del software.

2.1 Estrazione delle informazioni inerenti il gene

Il sistema una volta che l'utente inserisce in inputl l'ID del gene,