Gene-Disease association analyzing scientific literature

Salvatore Calderaro

Università degli Studi di Palermo

Sommario

Introduzione

Nell'ambito di questo progetto è stato realizzato un software che dato l'ID del gene in input - tramite tecniche di *text mining* e *sentiment analisys* - analizzando la letteratura scientifica restituisce in output le malattie associate al gene.
Il software è stato scritto in *Python* e le principali librerie utilizzate per lo sviluppo sono:

- pyspark:
- biopython;
- nltk;
- spacy;
- scispacy
- etc.

Caricamento delle informazioni inerenti il gene

Una volta inserito l'ID del gene, il sistema tramite *Entrez* verifica sul databaese associato *Gene* quale è il gene associato all'ID - se esistente - e ne ricava alcune informazioni e le memorizza in un dataframe.

```
Inserisci l'ID del gene--->3569
+-------------+----+--------------+-----------------+
|TaxonomyName|  ID|OfficialSymbol|OfficialFullName|
+-------------+----+--------------+-----------------+
|Homo sapiens|3569|           IL6|   interleukin 6|
+-------------+----+--------------+-----------------+
```

Estrazione della letteratura scientifica

Entrez effettua una query per verificare la presenza di *link* inerenti l'ID fornito in input. Ricavata tale lista di riferimenti si procede all'estrazione da *Pubmed* di titolo ed abstract (se disponibile) dei 200 articoli più rilevanti. Le informazioni cosi ottenute vengono memorizzate alll'interno di un dataframe.

Title	Abstract
Interleukin-6 in ...	The role of inter...
Role of Interleuk...	COVID-19 is viral...
EBV Rta-induced I...	Rta, a transactiv...
Elevated levels o...	Coronavirus disea...
Prognostic value ...	The inflammatory ...
IL-6 produced by ...	Trichomonas vagin...
Association betwe...	We aimed to compa...
Association of IL...	The -174G>C (rs18...
Interleukin-6 gen...	Several studies h...
IL-6 is present i...	IL-6 is a pro-inf...
IL-6 mediated JAK...	We investigated t...
Does serum interl...	The diagnosis of ...
Association of -1...	Interleukin-6 (IL...
Baseline Interleu...	The objective of ...
Interleukin-6 med...	The phosphoinosit...
Association betwe...	The association b...
Association of <i...	Autoimmune thyroi...
The Role of Inter...	Studies have show...
Association betwe...	
Association of Va...	Lung cancer is kn...

only showing top 20 rows

Pulitura dei dati

In questa fase - per ogni articolo presente all'interno del dataframe - si eseguono i seguenti step:

1. tokenizzaziome;
2. rimozione dei simboli di punteggiatura;
3. rimozione delle stopwords;
4. lemmatizzazione.

Tokenizzazione

In questo software per effettuare la tokenizzazione del testo si è utilizzato *RegexpTokenizer* di *nltk*. Quest'ultimo effettua la tokenizzazione del testo in base all'espressione regolare che viene fornita in input la quale fungerà da delimitatore tra una parola ed un'altra. In questa fase vengono inoltre eliminati tutti i simboli di punteggiatura presenti.

```
TESTO:
A unique subset of low-risk Wilms tumors is characterized by loss of function of TRIM28 (KAP1), a gene crit
ical in early renal development: A Children's Oncology Group study


TESTO DOPO LA RIMOZIONE DELLA PUNTEGGIATURA E TOKENIZZATO:
['A', 'unique', 'subset', 'of', 'low', 'risk', 'Wilms', 'tumors', 'is', 'characterized', 'by', 'loss', 'of'
, 'function', 'of', 'TRIM28', 'KAP1', 'a', 'gene', 'critical', 'in', 'early', 'renal', 'development', 'A',
'Children', 's', 'Oncology', 'Group', 'study']
```

Rimozione delle stopwords

Dopo aver effettuato la tokenizzazione del testo si procede con la rimozione delle *stopwords* ovvero quelle parole comunemente usate che non portano nessuna informazione utile al testo come avverbi, preposizioni, pronomi, etc.

```
TESTO DOPO LA RIMOZIONE DELLE STOPWORDS:
['unique', 'subset', 'low', 'risk', 'Wilms', 'tumors', 'characterized', 'loss', 'function', '
TRIM28', 'KAP1', 'gene', 'critical', 'early', 'renal', 'development', 'Children', 'Oncology',
 'Group', 'study']
```

Lemmatizzazione

Questo processo permette di ridurre le parole dalla loro forma flessa alla loro forma canonica, che viene detta giustappunto *lemma*. Per effettuare la lemmatizzazione si è utilizzato la classe *WordNetLemmatizer* di *nltk*.

```
TOKEN DOPO LA LEMMATIZZAZIONE:
['unique', 'subset', 'low', 'risk', 'wilms', 'tumor', 'characterized', 'loss', '
function', 'trim28', 'kap1', 'gene', 'critical', 'early', 'renal', 'development'
, 'child', 'oncology', 'group', 'study']
```

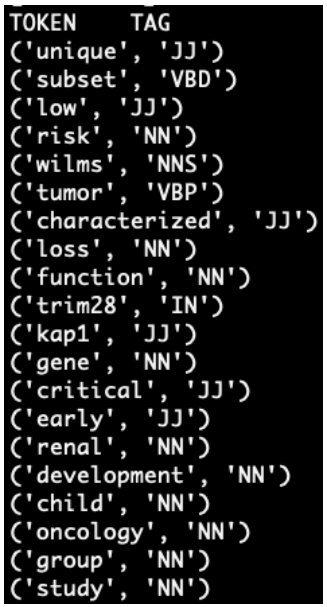
Salvataggio dei dati

Gli articoli cosi processati vengono memorizzati all'interno di un nuovo dataframe il quale conterrà sia per il titolo che per l'abstract una lista di token.

```
+--------------------+--------------------+
|               Title|            Abstract|
+--------------------+--------------------+
|[interleukin, 6, ...|[role, interleuki...|
|[role, interleuki...|[covid, 19, viral...|
|[ebv, rta, induce...|[rta, transactiva...|
|[elevated, level,...|[coronavirus, dis...|
|[prognostic, valu...|[inflammatory, re...|
|[il, 6, produced,...|[trichomonas, vag...|
|[association, ser...|[aimed, compare, ...|
|[association, il,...|[174g, c, rs18007...|
|[interleukin, 6, ...|[several, study, ...|
|[il, 6, present, ...|[il, 6, pro, infl...|
|[il, 6, mediated,...|[investigated, ch...|
|[serum, interleuk...|[diagnosis, persi...|
|[association, 174...|[interleukin, 6, ...|
|[baseline, interl...|[objective, study...|
|[interleukin, 6, ...|[phosphoinositol,...|
|[association, int...|[association, pla...|
|[association, il6...|[autoimmune, thyr...|
|[role, interleuki...|[study, shown, si...|
|[association, il,...|                  []|
|[association, var...|[lung, cancer, kn...|
+--------------------+--------------------+
only showing top 20 rows
```

Part of speech tagging

Il *part of speech tagging* è una tecnica che permette di identificare la parte del discorso (*part of speech* (POS)) di una determinata parola all'interno di un testo come nomi, aggettivi, avverbi etc. In questo software è stato utilizzato *Averaged Perceptron Tagger* unitamente alla funzione *pos_tag* di *nltk*. Per la lingua inglese il tagger che viene utilizzato sfrutta il tagset *Penn Treebank*.



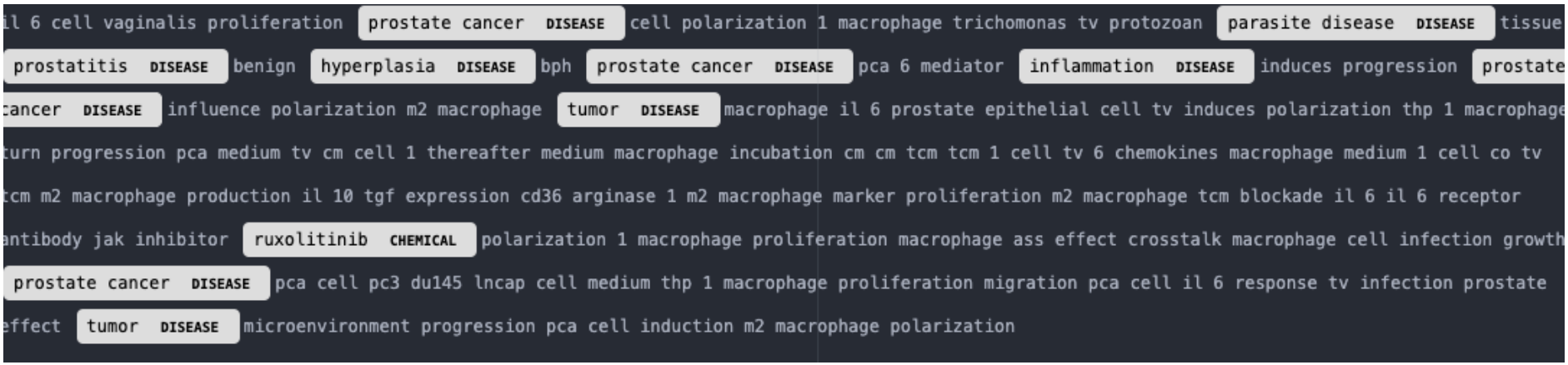
Rimozione delle POS non essenziali

In questa fase vengono conservate solamente quelle POS che risultano utili per gli scopi finali del programma come: sostantivi singolari e plurali, simboli, parole straniere, numeri e cardinalità. Fatto ciò, i dati cosi manipolati vengono salvati all'interno di un dataframe.

```
+--------------------+--------------------+
|               Title|            Abstract|
+--------------------+--------------------+
|[6, rheumatoid, a...|[role, interleuki...|
|[role, interleuki...|[covid, 19, respi...|
|[ebv, rta, 6, pro...|[rta, transactiva...|
|[level, il, 6, cr...|[coronavirus, dis...|
|[value, 6, c, pro...|[inflammatory, re...|
|[il, 6, cell, vag...|[trichomonas, tv,...|
|[association, ser...|[concentration, s...|
|[association, 6, ...|[174g, rs1800795,...|
|[6, gene, polymor...|[study, associati...|
|[il, 6, beta, alp...|[il, 6, cytokine,...|
|[il, 6, jak, stat...|[change, il, 6, e...|
|[association, per...|[b, aim, b, impai...|
|[serum, 6, guide,...|[diagnosis, infec...|
|[association, 174...|[6, 6, protein, c...|
|[baseline, 6, sed...|[study, baseline,...|
|[6, mediates, res...|[phosphoinositol,...|
|[association, 6, ...|[association, pla...|
|[evolution, cytok...|[cytokine, chemok...|
|[association, il6...|[autoimmune, dise...|
|[role, 6, differe...|[study, concentra...|
+--------------------+--------------------+
only showing top 20 rows
```

Applicazione dells Named Entity Recognition

Per ricavare le malattie associate al gene viene utilizzata la *Named Entity Recognition* (NER) un processo utilizzato per identificare la classe di appartenenza di una parola all'interno di un certo documento. Per far ciò è stata utilizzata la libreria *scispacy* dalla quale si è scelto il modello *en_ner_bc5cdr_md* addestrato sul corpus di testo *BC5CDR*. Tale modello riesce ad identificare malattie e composti chimici.



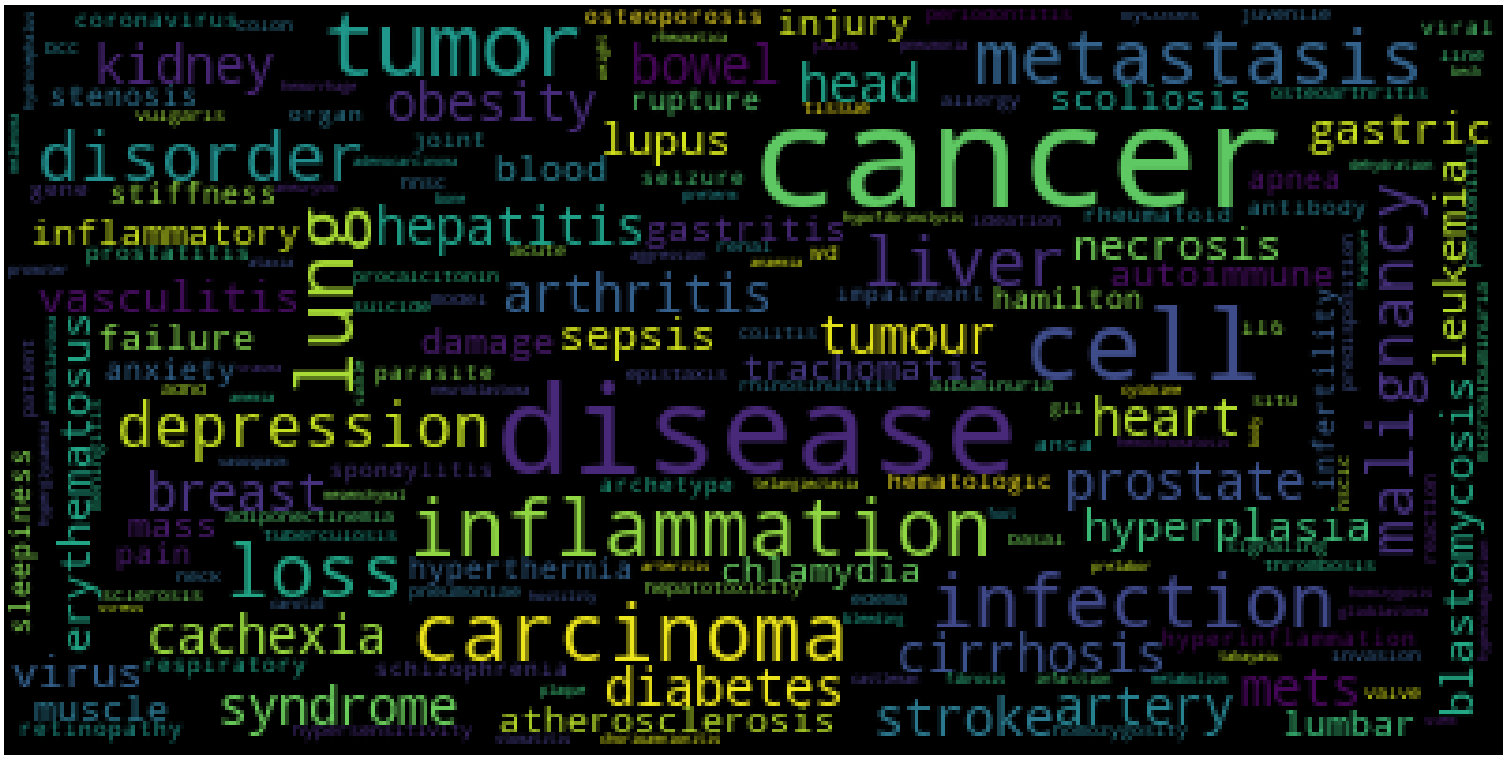
Creazione lista delle malattie

Dopo aver applicato la NER si inseriscono all'interno di una lista tutte le entità etichettate come malattie. La lista cosi ottenuta viene manipolata in modo da rimuovere duplicati, termini non inerenti a malattie, stringhe che contengono parole duplicate.

```
MALATTIE TROVATE ANALIZZANDO LA LETTERATURA SCIENTIFICA:
rheumatoid arthritis    respiratory infection   tumor   carcinoma       coronavirus disease     viral hyperinflammation prostate cancep
arasite disease prostatitis     hyperplasia     mets    cancer  cancer  liver disease   diabetes        cachexia        gastric lung br
east cancer     schizophrenia   impairment inflammation joint infection obesity spondylitis     tumor necrosis  hematologic malignancys
eizure  archetype organ disease sepsis  il6 gene lung cancer    lung cancer disease     cancer inflammation     lung cancer     tumourt
umour wd syndrome mets  cell leukemia   retinopathy     vasculitis      antibody anca vasculitis        hepatitis virus carcinoma     o
steoporosis     hypersensitivity        allergy acute stroke    stroke  blastomycosis   mycoses blastomycosis   lupus erythematosus   t
umor malignancy colon cancer    colitis cancer  head neck cancer        osteoarthritis  liver cancer    head cell carcinoma hnsc      a
lbuminuria      microalbuminuria        diabetes disease obesity        adiponectinemia mets diabetes   kidney inflammation     renal c
ell tumor       arthritis       depression      inflammatory disease    signaling disorder      chlamydia trachomatis   heart disease a
therosclerosis disease  chlamydia pneumoniae infection  atherosclerosis heart valve disease     trachomatis infection   cancer cell rea
ction   necrosis        vulgaris disorder       leukemia        liver cirrhosis blood patient liver cirrhosis   cirrhosis       hepatit
is autoimmune   hepatitis       liver damage    tissue damage   gastric cancer  gastritis       gastritis cancer        cancer cachexia
 loss   muscle mass cancer      6 loss cancer   cancer cachexia disease adhd disorder   prostate hyperplasia    lumbar disease  heart f
ailure  periodontitis   disorder suicide ideation       syndrome injury artery disease  carcinoma situ  breast cancer   breast carcinom
a       prostate tumor  metastasis muscle       tumor   edema   homozygosity    rhinosinusitis  tuberculosis    hepatotoxicity  pcos  m
alignancy       sepsis  basal cell carcinoma bcc        gli     predisposition sclerosis        metastasis      metastasis malignancy p
rocalcitonin    meningitis      scoliosis       hyperthermia    hyperthermia cancer     epistaxis       thrombosis      juvenile arthri
tis     lung cancer tumour      lung cancer nsclc       peritonitis     cancer cell invasion    cancer 6 blood  cancer cell line modelm
etastasis cancer        cancer metastasis       adenocarcinoma  neuroblastoma   foot infection  hemochromatosis homozygosis     stress
inflammation    melanoma        chorioamnionitis        rupture preterm prelabor rupture        beck depression trauma  hamilton depres
sion    hamilton anxiety        autoimmune disease tumor        mesenchymal cell tumor  glioblastoma    ataxia telangiectasia   bone me
tabolism bowel disease  inflammatory bowel disease      plaque  lupus erythematosus disease     dehydration     body mass loss  hyperco
agulation       hypofibrinolysis        apnea obesity   virus infection hostility       aggression      hemorrhage      aneurysm      v
asospasm        hydrocephalus   infarction      bleeding        pneumonia       castleman disease       carotid artery stenosis promote
r stroke        artery stenosis inflammation depression anemia  cytokine loss stiffness stiffness loss  cancer cell lung        pain  b
owel syndrome disorder pain     kidney disease  hyperglycaemia  fibrosis        rheumatica      takayasu arteritis      kidney injury t
umor metastasis ameloblastoma   cancer stem     fracture        head cell carcinoma     weight loss     stomatitis      sickle cell ana
emia    infertility     infertility     craniopharyngioma       calcification   failure bladder cancer  anxiety prognosis disease     a
denoma  lumbar scoliosis        sleepiness inflammation apnea   daytime sleepiness      hypertension    inflammation dysfunction
```

Rappresentazione grafica della lista di malattie

La lista contenente le malattie oltre ad essere salvata su file CSV, è stata utilizzata per generare una *wordcloud*.



Valutazione dei risultati ottenuti

Per valutare l'attendibilità dei risultati ottenuti si è utilizzato il database *DisGenNet* il quale contiene 1.134942 associazioni fra geni e malattie (in particolare contiene associazioni tra 21.671 geni e 30.170 malattie) unitamente al *fuzzy string matching* per paragonare la lista delle malattie ottenuta con la lista delle malattie corrette ottenute da questo database.

Caricamento del database DisGenNet

Dopo aver caricato il database DisGenNet viene ricavato un dataframe contenente le associazioni fra gene e malattia ed una lista contenente le malattie che servirà successivamente per il confronto.

geneId	diseaseName
3569	Abdominal Pain
3569	Spontaneous abortion
3569	Abortion, Tubal
3569	Abscess
3569	Acanthosis Nigricans
3569	Acidosis, Lactic
3569	Acne Vulgaris
3569	Acquired Immunodeficiency Syndrome
3569	Acute alcoholic liver disease
3569	Acute pancreatitis
3569	Acute periodontitis
3569	Acute-Phase Reaction
3569	Acute vascular insufficiency of intestine (disorder)
3569	Addison Disease
3569	Adenocarcinoma
3569	Adenoma
3569	Agammaglobulinemia
3569	Osteoporosis, Age-Related
3569	Primary Myelofibrosis
3569	AIDS Dementia Complex

only showing top 20 rows

Confronto delle liste tramite fuzzy string matching

Per calcolare la percentuale di malattie corrette della nostra lista, le stringhe delle due liste vengono confrontate a coppie mediante *fuzzy string matching*. Per fare ciò si è utilizzato la funzione *token_set_ratio* di *fuzzywuzzy*. La funzione restituisce un valore indicativo sul grado di similarità delle due stringhe prese in input.

```
('rheumatoid arthritis', 'Arthritis', 100)
('respiratory infection', 'Respiratory Tract Infections', 86)
('tumor', 'Malignant tumor of colon', 100)
('carcinoma', 'Carcinoma', 100)
('prostate cancer', 'Prostate cancer recurrent', 100)
('parasite disease', 'Parasitic Diseases', 88)
('prostatitis', 'prostatitis', 100)
('hyperplasia', 'Angiolymphoid hyperplasia', 100)
('cancer', 'Hypopharyngeal Cancer', 100)
('liver disease', 'Acute alcoholic liver disease', 100)
('diabetes', 'Alloxan Diabetes', 100)
('cachexia', 'Cachexia', 100)
('gastric lung breast cancer', 'Early gastric cancer', 82)
('schizophrenia', 'Schizophrenia', 100)
('impairment inflammation', 'Inflammation', 100)
('joint infection', 'Prosthetic joint infection', 100)
('obesity', 'Obesity', 100)
('spondylitis', 'Spondylitis', 100)
('tumor necrosis', 'Necrosis', 100)
('seizure', 'Jacksonian Seizure', 100)
('sepsis', 'Sepsis', 100)
('lung cancer disease', 'Chronic lung disease', 82)
('cancer inflammation', 'Inflammation', 100)
('lung cancer', 'Progression of non-small cell lung cancer', 100)
('tumour', 'Tumour inflammation', 100)
('tumour wd syndrome mets', 'Job Syndrome', 80)
('cell leukemia', 'leukemia', 100)
('retinopathy', 'Nonproliferative diabetic retinopathy', 100)
('vasculitis', 'Vasculitis', 100)
('antibody anca vasculitis', 'Vasculitis', 100)
('hepatitis virus carcinoma', 'Carcinoma', 100)
('osteoporosis', 'Osteoporosis, Age-Related', 100)
```

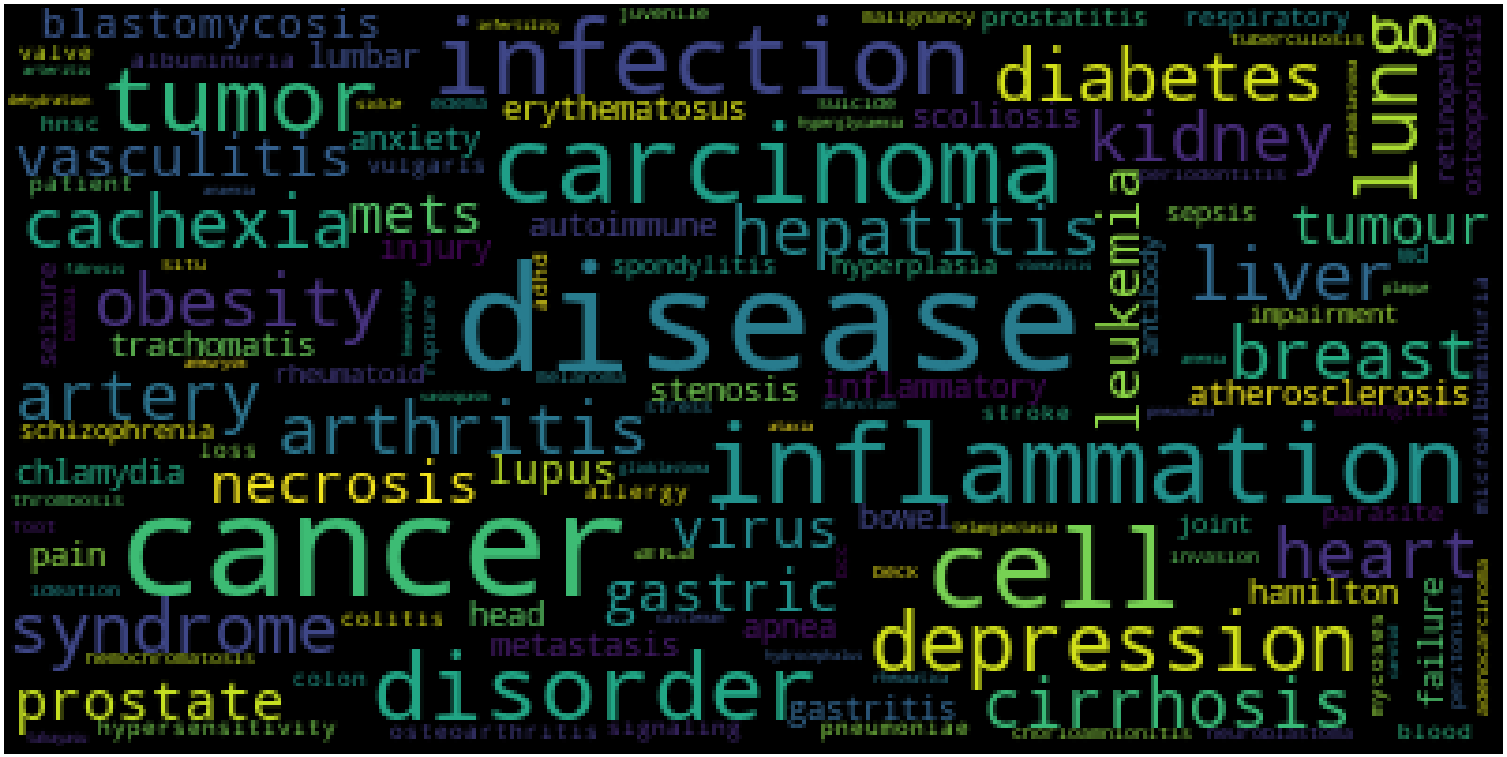
Creazione della lista finale delle malattie

Verranno aggiunte alla lista finale delle malattie solo quelle stringhe che confrontate con le stringhe della lista che contiene le malattie corrette superano una certa soglia, in questo caso 80. Inoltre verrà riportata anche la percentuale di malattie corrette identificate nella nostra lista.

```
 DELLE MALATTIE IDENTIFICATE SOLO IL 71.63 % SONO RISULTATE CORRETTE:
rheumatoid arthritis    respiratory infection   tumor   carcinoma       prostate cancer parasite disease        prostatitis     hyperpl
asia    cancer  liver disease   diabetes        cachexia        gastric lung breast cancer      schizophrenia   impairment inflammatioj
oint infection  obesity spondylitis     tumor necrosis  seizure sepsis  lung cancer disease     cancer inflammation     lung cancer   t
umour   tumour wd syndrome mets cell leukemia   retinopathy     vasculitis      antibody anca vasculitis        hepatitis virus carcino
ma      osteoporosis    hypersensitivity        allergy stroke  blastomycosis   mycoses blastomycosis   lupus erythematosus     colon c
ancer   colitis cancer  osteoarthritis  head cell carcinoma hnsc        albuminuria     microalbuminuria        diabetes disease obesit
y       mets diabetes   kidney inflammation     arthritis       depression      inflammatory disease    signaling disorder      chlamyd
ia trachomatis  heart disease   atherosclerosis disease chlamydia pneumoniae infection  atherosclerosis heart valve disease     trachom
atis infection  necrosis        vulgaris disorder       leukemia        liver cirrhosis blood patient liver cirrhosis   cirrhosis     h
epatitis autoimmune     hepatitis       gastric cancer  gastritis       gastritis cancer        cancer cachexia loss    cancer cachexia
 disease        adhd disorder   lumbar disease  heart failure   periodontitis   disorder suicide ideation       syndrome injury artery
disease carcinoma situ  breast cancer   breast carcinoma        prostate tumor  edema   tuberculosis    malignancy      basal cell carc
inoma bcc       metastasis      meningitis      scoliosis       thrombosis      juvenile arthritis      peritonitis     cancer cell inv
asion   adenocarcinoma  neuroblastoma   foot infection  hemochromatosis stress inflammation     melanoma        chorioamnionitis      r
upture  beck depression hamilton depression     hamilton anxiety        autoimmune disease tumor        glioblastoma    ataxia telangie
ctasia  inflammatory bowel disease      plaque  lupus erythematosus disease     dehydration     apnea obesity   virus infection hemorrh
age     aneurysm        vasospasm       hydrocephalus   infarction      pneumonia       castleman disease       carotid artery stenosia
rtery stenosis  inflammation depression anemia  cancer cell lung        pain    bowel syndrome disorder pain    kidney disease  hypergl
ycaemia fibrosis        rheumatica      takayasu arteritis      kidney injury   tumor metastasis        ameloblastoma   head cell carci
noma    stomatitis      sickle cell anaemia     infertility     craniopharyngioma       calcification   failure anxiety adenoma lumbar
scoliosis       sleepiness inflammation apnea   hypertension    inflammation dysfunction
```

Rappresentazione grafica della lista di malattie

La lista contenente le malattie - dopo aver effettuato questo filtraggio - oltre ad essere salvata su file CSV, è stata utilizzata per generare una *wordcloud*.



Conclusioni

Questo software tramite tecniche di *Natural Language Processing* e *Text Mining* restituisce in output un insieme di malattie associate al gene. Nella maggior parte dei casi più del 50% delle malattie ottenute analizzando la letteratura scientifica corrisponde con quelle ottenute dal database DisGenNet.

Possibili miglioramenti

Nella lista finale delle malattie potrebbe capitare di riscontrare nomi - che pur essendo riconducibili a malattie o patologie - sono stati erroneamente inseriti nella lista (ad esempio nomi di due malattie concatenate, oppure due malattie che non hanno nulla in comune identificate come un'unica malattia).

Il software dunque, potrebbe essere migliorato utilizzando modelli che per applicare le tecniche utilizzate usino modelli addestrati prettamente su corpus di testo biomedici.

Grazie per l'attenzione !