Dipartimento di Scienze
Politiche e della Comunicazione

UNIVERSITÀ DEGLI STUDI
DI SALERNO

Classificazione Automatica di Testi Medici Basata sulla Terminologia di Dominio

Alessandro Maisto

Dipartimento di Scienze Politiche e della Comunicazione - Università degli Studi di Salerno

Il problema del Linguaggio Medico

- La classificazione automatica dei testi rappresenta un task estremamente comune all'interno della Linguistica Computazionale
- I modelli di classificazione automatica di tipo supervisionato o semi-supervisionato sono in grado di raggiungere alte prestazioni in questo tipo di operazione

- Il dominio specialistico della **Medicina**, tuttavia, rappresenta un'importante eccezione sia per la classificazione automatica come per altri task tipici dell'NLP

La classificazione Automatica

La classificazione automatica è caratterizzata da due filoni principali:

- Metodi supervisionati o semi-supervisionati, dove l'algoritmo 'impara' come classificare un testo da un set di documenti manualmente classificati (training set)[1]

- Metodi non-supervisionati, che cercano di raggruppare i testi sulla base della loro somiglianza, vettorializzando i documenti sulla base di dati statistici e/o derivati da pattern testuali specifici[2]

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) 34(1), 1–47 (2002)
2. Grimmer, J., Stewart, B.: Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 21(3), 267–297 (2013)

La classificazione Automatica

- Le principali differenze tra i metodi non-supervisionati riguardano il modo in cui il testo viene convertito in vettore numerico:
  - La metodologia più utilizzata prevede la costruzione delle cosiddette bag-of-words, gruppi di Keywords che rappresentano il testo[3]
  - L'estrazione di queste parole può essere affidata a metodi statistici (tf-idf, frequenza…) o a metodologie rule-based (dizionari terminologici, estrazione basata su pattern)
  - La presenza di Parole Composte Terminologiche (Multi-Word Expressions, MWE) può essere fondamentale per la classificazione di testi specialistici. Oltre a non essere ambigue, esse rappresentano circa il 90% delle parole che caratterizzano un linguaggio specialistico[4]

3. Miller, T., Dligach, D., Savova, G.: Unsupervised document classification with informed topic models. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp. 83–91 (2016)
4. Elia, A., Cardona, G.R.: Discorso scientifico e linguaggio settoriale. un esempio di analisi lessico-grammaticale di un testo neuro-biologico. Quaderni del Dipartimento di Scienze della Comunicazione–Università di Salerno, Cicalese A., Landi A., Simboli, linguaggi e contesti (2) (2002)

Il dominio medico

Cosa caratterizza il Dominio Medico?

- Una terminologia specialistica più vasta rispetto ad altri domini
- Una frequenza di occorrenza di ciascun termine molto ridotta proprio a causa della presenza di numerosi sotto-domini
- Difficile e costoso reperimento di risorse terminologiche del dominio

Il dominio medico

Un grande numero di termini specialistici del dominio medico possono essere definiti:

Rare Events[5]

5. Möbius, B.: Rare events and closed domains: two delicate concepts in speech synthesis. Int. J. of Speech Technol. 6(1), 57–71 (2003)

Il dominio medico

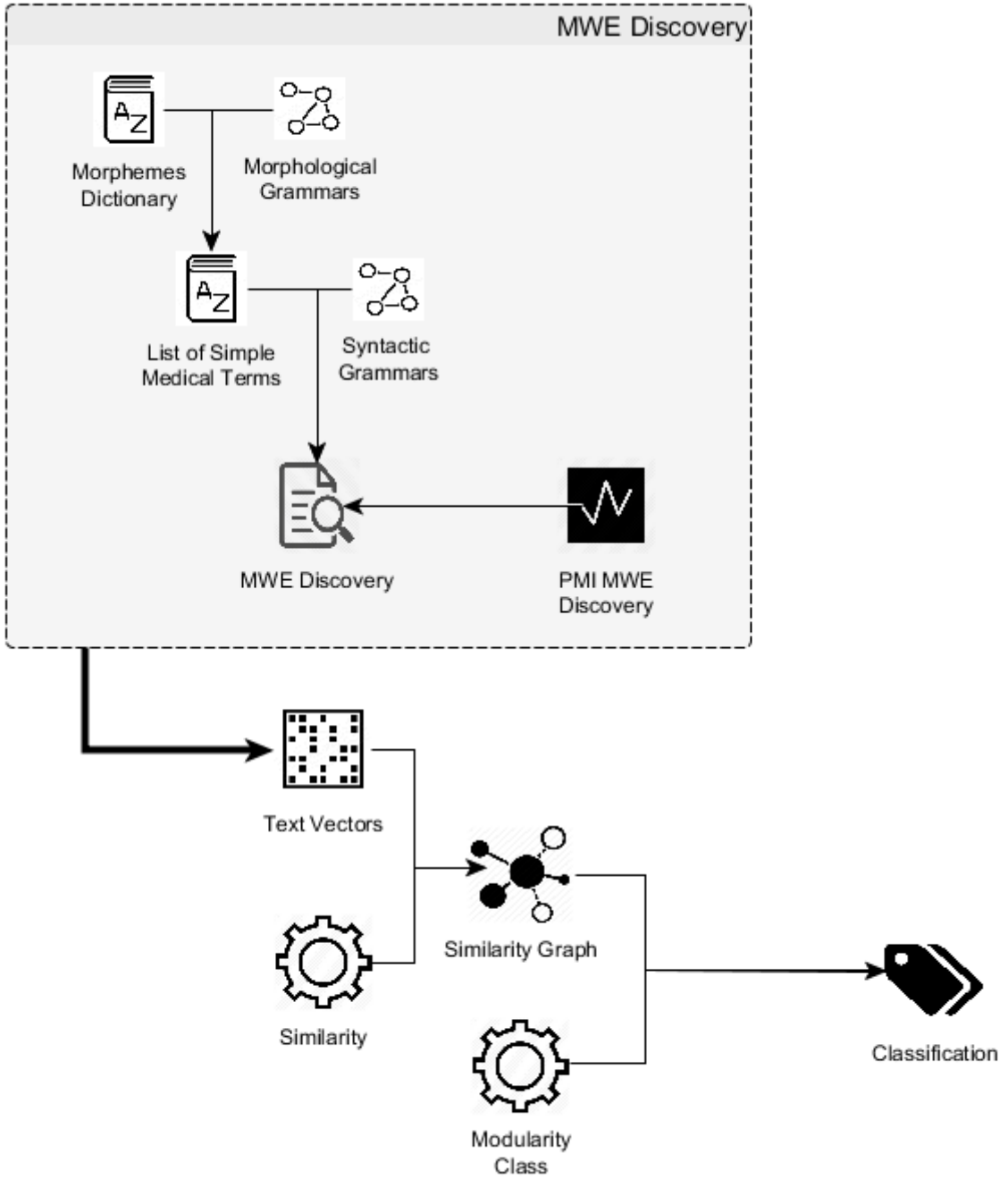
- La presenza di numerosi sotto-domini ed i contorni sfumati che i documenti medici spesso presentano rendono complesse le operazioni di costruzione di **training set** adeguati

- La terminologia specialistica presente nei documenti medici spesso differisce enormemente anche tra testi appartenenti allo stesso sotto-dominio, rendendo poco precisi anche metodologie non supervisionate basata esclusivamente sulle cosiddette **bags-of-words**

Metodologia proposta

- La metodologia proposta prevede un approccio non-supervisionato basato sulla MWE:
  - I testi vengono vettorializzati sulla base delle MWE e delle parole semplici di dominio che contengono
  - L'estrazione delle MWE viene affrontata in maniera ibrida
  - Ad un approccio statistico (PMI) viene affiancata una ricerca basata su regole
  - Il problema della terminologia medica viene affrontato tramite un approccio Morfo-semantico

Metodologia proposta



Approccio Morfo-semantico

- L'idea di base è quello di utilizzare un dizionario ristretto di morfemi di origine Greca o Latina con altra frequenza nella terminologia medica[6-7]
- Garantisce discreti risultati a partire da un set ristretto di elementi
- Permette una descrizione analitica del significato delle parole appartenenti ad una stessa 'famiglia morfologica'

- Il dizionario è composto da 537 morfemi estratti dal *Dizionario della Lingua Italiana De Mauro*, suddivisi in:
  - 493 **Confissi** (CFX-CFXS) (acusia-, mico-, gastro-, -cefalia, ecc…)
  - 15 **Prefissi** (PFX) (ipo-, iper, ecc…)
  - 29 **Suffissi** (SFX) di cui 10 di dominio (-oma, -ite, ecc…), e 19 indicanti la classe sintattica (-abile, -ale, -oso, ecc…)

6. Pacak, M.G., Norton, L.M., Dunham, G.S.: Morphosemantic analysis of-ITIS forms in medical language. Methods Inf. Med. 19(2), 99–105 (1980)
7. Wolff, S.: The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. Methods Inf. Med. 23(4), 195–203 (1984)

Approccio Morfo semantico

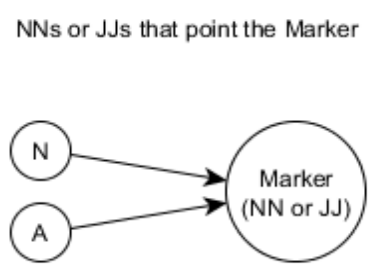
- Il dizionario dei morfemi indica, per ogni entrata, un **tag semantico** riferito alla sotto-classe medica di riferimento.
- Le sotto-classi prese in considerazione sono 27 e corrispondono a discipline mediche classiche quali *neurologia*, *pneumologia*, *cardiologia*, *traumatologia*, *urologia*, ecc…
- I morfemi che non afferiscono a nessuna sotto-classe in particolare sono stati etichettati come *generici*.

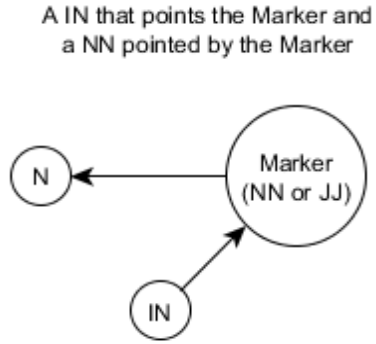
Approccio Morfo semantico

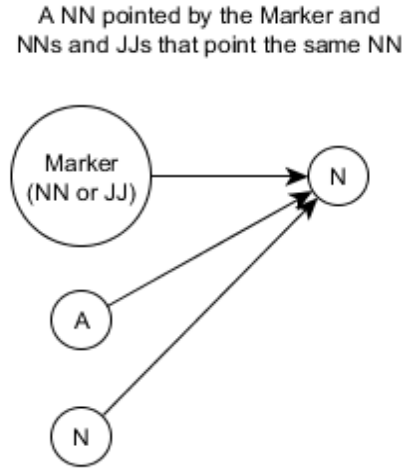
- Il dizionario dei morfemi è associato ad una serie di pattern applicati ai singoli lemmi presenti nel testo:
  - CFX-CFXS-SFX (i.e. Athero-scler-osis)
  - CFX-CFX (i.e. Cephalo-pathy)
  - CFX-CFX-CFXS-SFX (i.e. Bronco-pneumo-path-ic)
  - PFX-CFX (i.e. Hypo-acusia)
  - PFX-CFX-CFX (i.e. Peri-cardio-pathy)
  - PFX-CFX-CFXS-SFX (i.e. Hemi-megal-enceph-aly)
  - CFXS-SFX (i.e. encephal-itis)

Approccio Morfo semantico

- Successivamente i termini afferenti al dominio medico sono inclusi in una ricerca per pattern sintattici in grado di estratte Parole Composte Terminologiche (rule-based MWE discovery)
- I pattern per l'inglese sono i seguenti:







1. Interventional cardiology
2. Treatment of cardiopathy
3. Coronary artery bypass

Approccio Morfo semantico

- L'approccio morfo-semantico ottiene una precisione stimata inclusa tra l' 85 ed il 95%
- La recall, tuttavia, presenta valori estremamente variabili e spesso bassi

PMI Multiword Expression Discovery

- Per aumentare il numero di risultati e migliorare la descrizione vettoriale del testo è stata prevista, in parallelo, un'estrazione basata sulla frequenza degli n-grams
- Sono stati estratti Bi-grams e Tri-grams ed è stato calcolato il rapporto di associazione[8] per ognuno di essi

$$\mathrm{I}(\mathrm{x},\mathrm{y}) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad (1) \qquad \mathrm{I}(\mathrm{x},\mathrm{y},\mathrm{z}) = \log_2 \frac{P(x,y,z)}{P(x)P(y)P(z)} \quad (2)$$

8. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. 16(1), 22–29 (1990)

PMI Multiword Expression Discovery

- La precisione della MWE discovery tramite PMI si attesta tra il 46 ed il 60%
- In totale, la precisione della metodologia utilizzata per l'estrazione delle MWE è di circa il 60%

Classificazione dei Documenti

- Il risultato della precedente fase di analisi è una lista di parole (MWE)
- Le dimensioni del vettore saranno dunque uguali al numero di MWE estratte più il numero di tag semantici relativi ai sotto-domini medici.
- Ogni dimensione del vettore corrisponderà al numero di occorrenze di ciascuna MWE o tag nel testo considerato
- La similitudine tra i testi è stata calcolata comparando i vettori tramite Similarità di Coseno

Classificazione dei Documenti

- Il risultato è una tavola dove a ciascuna coppia di documenti corrisponde un valore di similitudine.
- Questa tavola può essere considerato un grafo pesato in cui i documenti corrispondono a nodi ed i valori di similitudine corrispondono al peso della relazione tra i nodi.
- L'individuazione delle categorie è stata affidata ad una funzione di *Modularity Class*[9], Il cui valore quantifica la qualità della divisione della rete in moduli o comunità

9. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 2008(10), P10008 (2008)

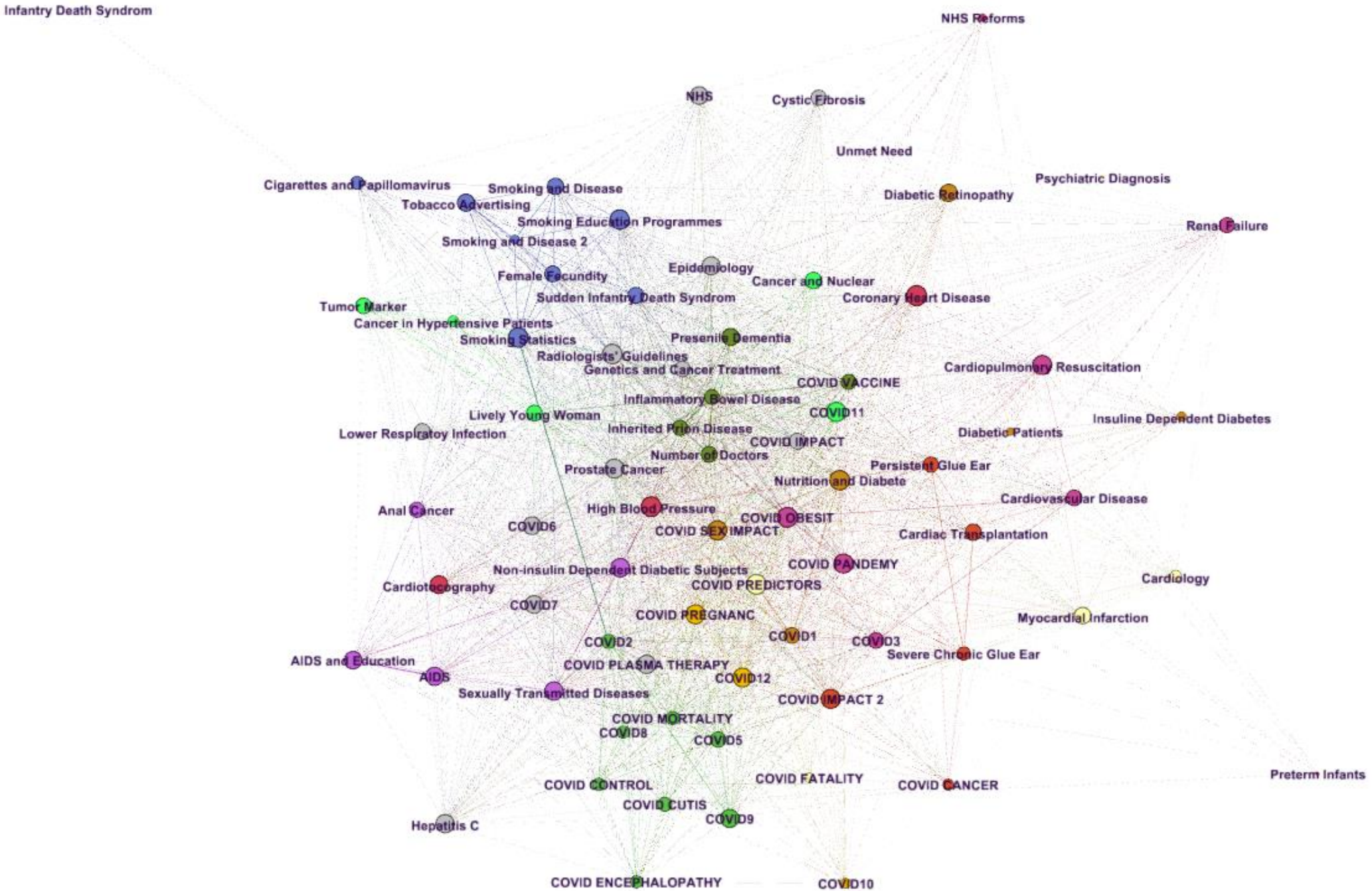
Esperimento

- L'esperimento è stato effettuato su un corpus di 74 articoli medici in Inglese
- La fonte principale da cui sono stati estratti gli articoli è il British National Corpus (BNC). Dal BNC sono stati estratti:
  - Articoli divulgativi (AIDS, effetti del fumo, salute in generale)
  - Articoli scientifici provenienti dal British Medical Journal
- La fonte secondaria è Google Scholar, da cui sono stati estratti articoli scientifici di libera consultazione a tema COVID-19

9. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 2008(10), P10008 (2008)

Esperimento

- La Modularity Class (Resolution 0.4) ha evidenziato la presenza di 15 classi
- Nel grafo, il titolo dell'articolo è stato rimpiazzato con una descrizione sintetica del suo contenuto
- I colori dei nodi corrispondono alle classi calcolate automaticamente
- La posizione dei nodi è calcolata automaticamente da un algoritmo di Layout definito ForceAtlas2 che tiene conto della attrazione/repulsione dei nodi in base alle misure di peso


Infantry Death Syndrom
NHS Reforms
NHS
Cystic Fibrosis
Unmet Need
Psychiatric Diagnosis
Cigarettes and Papillomavirus
Smoking and Disease
Diabetic Retinopathy
Tobacco Advertising
Smoking Education Programmes
Renal Failure
Smoking and Disease 2
Epidemiology
Female Fecundity
Cancer and Nuclear
Sudden Infantry Death Syndrom
Coronary Heart Disease
Tumor Marker
Cancer in Hypertensive Patients
Smoking Statistics
Presenile Dementia
Radiologists' Guidelines
Genetics and Cancer Treatment
Cardiopulmonary Resuscitation
COVID VACCINE
Inflammatory Bowel Disease
Lively Young Woman
COVID11
Insuline Dependent Diabetes
Lower Respiratoy Infection
Inherited Prion Disease
Diabetic Patients
COVID IMPACT
Number of Doctors
Persistent Glue Ear
Prostate Cancer
Nutrition and Diabete
Cardiovascular Disease
Anal Cancer
High Blood Pressure
COVID OBESIT
COVID6
COVID SEX IMPACT
Cardiac Transplantation
COVID PANDEMY
Non-insulin Dependent Diabetic Subjects
Cardiology
Cardiotocography
COVID PREDICTORS
COVID7
COVID PREGNANC
Myocardial Infarction
COVID2
COVID1
COVID3
AIDS and Education
Severe Chronic Glue Ear
COVID PLASMA THERAPY
AIDS
COVID12
Sexually Transmitted Diseases
COVID IMPACT 2
COVID MORTALITY
COVID8
COVID5
Preterm Infants
COVID FATALITY
COVID CONTROL
COVID CANCER
COVID CUTIS
COVID9
Hepatitis C
COVID ENCEPHALOPATHY
COVID10

Esperimento

Classes	Common thread	No. Documents
0	Sexually transmitted disease	5
1	Infants + Glue ear	6
2	Covid-19	8
3	Cardiac and Pulmonary disease	4
4	Covid-19 Diffusion	3
5	Virus disease therapies	3
6	Family Inheritance	3
7	Cardiac disease + Health structures	5
8	Inheritance	5
9	Cancer	5
10	Diagnosis + mortality rate	4
11	Diabetes	6
12	Respiratory disease	3
13	Health structure	3
14	Smoking	9

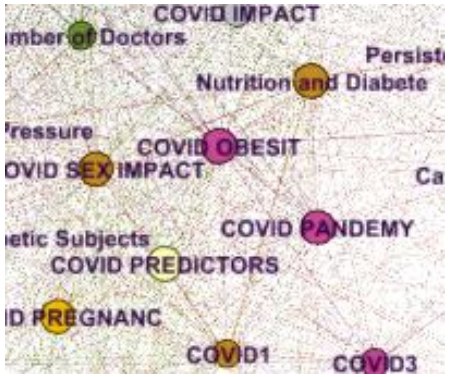
Esperimento

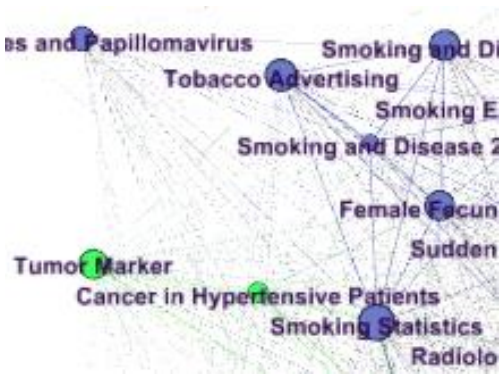
- La precisione dell'algoritmo (89%) è stata calcolata sulla base della presenza, all'interno delle classi, di documenti coerenti tra loro
- Le classi 0, 2, 5, 6, 7, 9, 10, 11, 13 e 14 risultano composte da documenti coerenti tra loro
- Nelle restanti classi è stata riscontrata la presenza di alcuni documenti non coerenti:
  - **Classe 1**: 3 documenti sono inerenti a diagnosi di Otite Siero Mucosa (Glue Ear), un articolo è incentrato sul trapianto cardiaco ed un altro sul rapporto tra Codiv e cancro
  - **Classe 3**: 5 articoli sono incentrati su problemi cardio-respiratori, mentre un articolo parla di insufficienza renale.
  - **Classe 4**: contiene 3 articoli a tema Covid-19 ma non coerenti nella sotto-classe trattata
  - **Classe 8**: 4 articoli riguardano malattie ereditarie, mentre un articolo riguarda il Vaccino per il Covid-19
  - **Classe 12**: 2 articoli riguardano malattie respiratorie ed un articolo riguarda il cancro della prostata

Conclusioni

- Dati i problemi nel trattamento dei testi di dominio medico, il nostro approccio fornisce due vantaggi principali:

    1. Permette di aumentare la precisione dei sistemi statistici di MWE discovery partendo da un dizionario dei morfemi di dimensioni estremamente ridotte
    2. La conversione del punteggio di similitudine in formato grafico permette di tenere traccia di articoli dove ad un argomento principale si affianca un argomento secondario





Conclusioni

- I prossimi passi riguarderanno:
  - la scelta di una metrica più efficace per la MWE Discovery non basata su regole
  - Migliorare il dizionario dei morfemi per incrementare la recall della MWE Discovery basata su regole
  - Sviluppo di un sistema di Labeling automatico per le classi
  - Adattamento del dizionario e dei pattern sintattici alla lingua italiana

Grazie Per l'attenzione

amaisto@unisa.it