Data Mining Project Report

Matteo Baldanza 826018 – Federico Legnani 829961 – Davide Luperi 826249 – Guglielmo Muoio 826029

Abstract

Nel presente rapporto, abbiamo affrontato il problema di elaborare, indagare ed analizzare una collezione di 261 documenti (tratti da “U.S. National Library of Medicine National Institutes of Health - NCBI”) attraverso tecniche di text mining apprese a lezione: abbiamo prima cercato di raggruppare i documenti in base ai termini presenti per poi raggruppare i termini con tecniche di clustering gerarchico e di clustering partizionale, per cercare di estrarre informazioni utili ed interessanti. Nella fase successiva, abbiamo cercato di classificare i documenti attraverso classificatori classici. Infine, abbiamo provato a eseguire una Sentiment Analysis elementare, al fine di comprendere se ci fossero sentimenti che potessero trasparire dai documenti.

1 Introduzione

Negli ultimi anni, c’è stato un crescente interesse circa l’estrazione di rappresentazioni strutturate in grado di catturare collegamenti utili alle analisi di interesse. Quando si fa riferimento a testi non strutturati (articoli, rassegne stampa, etc.) si ricorre al “Text Mining” che consiste nell’applicazione di tecniche di Data Mining allo scopo di classificare i documenti in categorie, evidenziare collegamenti e associazioni fra gli stessi, individuare gruppi semantici etc. In particolare, in ambito biomedico, estrazioni di informazioni dai testi, mirano a migliorare l’accesso alla conoscenza degli argomenti trattati automatizzando gli aspetti dell’elaborazione della letteratura. Nel report corrente, si vuole quindi cercare, partendo da citazioni biomediche, di estrarre informazioni di fondo col fine di comprendere gli argomenti rilevanti trattati.

2 Materiali e metodi

2.1 Materiali

Il corpus di riferimento per la nostra analisi comprende 261 documenti tratti da “U.S. National Library of Medicine National Institutes of Health - NCBI”, che raggruppa oltre 30 milioni di citazioni per la letteratura biomedica da MEDLINE, riviste di scienze della vita e libri online. Le citazioni possono anche includere collegamenti a contenuti full-text da PubMed Central e siti Web di editori.

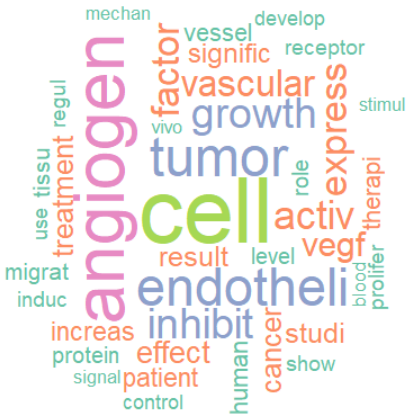
2.2 Metodi

Siamo partiti dalla fase di pre-processing del corpus, dunque di pulizia delle varie citazioni biomediche. Come da procedimento standard abbiamo eliminato i caratteri speciali, la punteggiatura, i numeri e le stopwords più comuni, anche quelle selezionate manualmente da noi ritenute poco significative per il nostro lavoro. Infine, abbiamo trasformato tutte le maiuscole in minuscole.

Successivamente abbiamo costruito la Document Term Matrix (serve per la clusterizzazione dei documenti), la più comune rappresentazione di un insieme di testi nel text mining, che ha sulle righe gli identificativi dei documenti (document è il termine generico che indica un insieme di documenti), sulle colonne tutte le singole parole presenti in tutti i documenti considerati (term indica la parola singola o brevi composizioni “data mining”), ed è composta da valori binari o conteggi, che possono rappresentare la frequenza di apparizione delle parole (nelle celle si ha quante volte è apparsa la singola parola nello specifico documento).

Abbiamo specificato quindi un ulteriore matrice che considerava solo parole composte da 2 a 15 caratteri presenti in almeno 26 testi (~10%). Visualizzando quest’ultima a schermo abbiamo così rimarcato quali fossero le parole più presenti nei documenti evidenziando termini fra loro simili (aggettivi, nomi e avverbi riferiti al medesimo significato). Questo procedimento ci è risultato utile per la fase di stemming dove abbiamo raggruppato parole frequenti sotto la stessa voce (ex. angiogen, angiogenesis, antiangiogen o tumor, tumour) e dove abbiamo sostituito alcuni termini con la loro forma base. Sempre grazie all’utilizzo della Term Document Matrix abbiamo rimosso ulteriori termini frequenti che abbiamo categorizzato come stop words, non avendo alcun significato per la nostra analisi (ex. “one”, “whereas”, “however”). Abbiamo concluso questa parte visualizzando un word cloud dei termini più frequenti che abbiamo considerato nelle fasi successive di analisi dei documenti che comprendono:

1. **Clustering dei documenti**
2. **Clustering dei termini**
3. **Classificazione dei documenti**
4. **Sentiment Analysis**



1. Clustering dei documenti

Senza conoscere la natura dei documenti (un corpus di 261 abstract di PubMed sull'angiogenesi, lo sviluppo di nuovi vasi sanguigni da quelli esistenti. Il dominio coinvolge un processo a livello di tessuto/organo che è strettamente associato al cancro e ad altre patologie a livello di organismo) abbiamo fatto riferimento ad un metodo non supervisionato (non potendo applicare supervisionati in questo contesto) caratterizzato dall’assenza della variabile risposta: in questo caso non è noto il numero di classi/gruppi in cui vogliamo dividere i documenti, che vengono raggruppati sulla base delle loro caratteristiche statistiche, come misure di distanza o misure di similarità, e non è necessaria la presenza del training set.

Abbiamo considerato il clustering; un metodo non supervisionato che raggruppa osservazioni che condividono caratteristiche simili sulla base di precisi criteri, quando non sono disponibili esempi di come i

dati dovrebbero essere raggruppati, e che individua patterns o strutture di interesse tra i dati. Esso si può definire come una divisione dei dati in gruppi, con una partizione tale che le osservazioni all'interno di ciascun gruppo sono simili una con l'altra, mentre i gruppi sono piuttosto differenti uno dall'altro. Siamo partiti utilizzando un clustering gerarchico, testando tre metriche diverse:

- Distanza di Euclidea $d_{eucl}(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$
- Distanza di Manhattan $d_{manh}(P,Q) = \sum_{i=1}^{n}|p_i - q_i|$
- Distanza di Minkowski (di ordine k) $d_{mink}(P,Q) = [\sum_{i=1}^{n}|p_i - q_i|^k]^{\frac{1}{k}}$

Abbiamo scelto queste distanze arbitrariamente di nostra spontanea volontà. Per confrontare tutte le singole distanze ci avremmo messo troppo, cosi abbiamo preso le più usate.

Tra i linkage considerati:

- COMPLETE LINKAGE (maximal intercluster dissimilarity): la distanza tra clusters è determinata dalla più grande distanza di ogni elemento del gruppo 1 da ogni elemento del gruppo 2:

$$d(C_1, C_2) = \max\big(d\big(c_{1i}, c_{2j}\big) \in D\big)$$

- SINGLE LINKAGE (minimal intercluster dissimilarity): come distanza tra clusters si sceglie la minima distanza, cioè il valore più piccolo delle distanze di ogni elemento del gruppo 1 da ogni osservazione del gruppo 2:

$$d(C_1, C_2) = \min\big(d\big(c_{1i}, c_{2j}\big) \in D\big)$$

- GROUP AVERAGE LINKAGE (mean intercluster dissimilarity): la distanza tra clusters è determinata dalla distanza media tra tutte le coppie nei due clusters.

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|}\sum_{i=1}^{n}|C_1|\sum_{j=1}^{n}|C_2|d\big(c_{1i}, c_{2j}\big) \in D$$

- METODO DI WARD, che si basa sul concetto di devianza: ad ogni step si uniscono i due gruppi dalla cui fusione deriva il minimo incremento possibile della devianza "intra":

$$DEV_T = \sum_{s=1}^{p}\sum_{i=1}^{n}(x_{is} - \overline{x})^2$$

Per capire quale linkage sia il migliore, si utilizza il dendogramma (albero a testa in giù), dove le singole osservazioni sono le foglie, cioè i punti estremi, e i rami sono chiamati terminali, se non si dividono e portano alle osservazioni, o interni, se si dividono in altri due rami interni. In questo grafico la similarità tra i due oggetti è rappresentata dall'altezza del più basso nodo interno che dividono e in base all'altezza in cui si sceglie di porre il cut-off si determina il numero di gruppi che si avrà.

Abbiamo considerato come linkage definitivo Ward perchè risultava visibilmente migliore (i rami mostravano ampiezza maggiore) rispetto ai restanti che presentavano problematiche rimarcate, in particolare quello

Single per cui alla luce del grafico (Dendogramma) non aveva senso fare clustering. Per determinare il numero di clusters (cioè determinare il numero corretto di gruppi e capire quando un cluster è buono o no) abbiamo utilizzato la Silhouette, che mostra quali punti cadono nel cluster e quali invece occupano una posizione intermedia. Le osservazioni che hanno un'elevata silhouette, cioè con valore prossimo a 1, sono ben raggruppate e quindi il numero di clusters può essere determinato scegliendo il valore di k che porta al valore medio, della silouette per il clustering totale di ogni osservazione I ($sil_{av} = \sum_i \frac{sil_i}{N}$), più alto. Se il valore della silhouette è compreso tra 0.26 e 0.5 la suddivisione perde di significatività, mentre se è minore o uguale di 0.25 significa che non è stato individuato nessun pattern tra i gruppi. Nel nostro specifico caso abbiamo dovuto fare delle considerazioni molto più complesse in quanto il problema presentava diverse difficoltà.

Quindi ci siamo mossi utilizzando un approccio partizionale dove si fissa a priori il numero di clusters K (che si può interpretare come il punto di taglio nel dendogramma in un clustering gerarchico) e ogni osservazione è posizionata esattamente in uno dei K clusters sulla base di precise regole, migliorando ad ogni step la qualità dei gruppi e tale che l'intersezione tra due gruppi diversi fosse pari al vuoto. Anche in questo caso il numero dei K gruppi è scelto sulla base dei risultati della Silhouette. Come metodo partizionale abbiamo utilizzato K-Means, un approccio iterativo per partizionare i dati in K gruppi distinti e non sovrapposti che calcola per ogni iterazione i K centroidi. Il risultato che abbiamo ottenuto lo abbiamo rappresentato sul Cusplot visualizzando nel piano bidimensionale i K gruppi individuati ed eventuali outlier. Infine, per ciascun cluster individuato abbiamo visualizzato la relativa WordCloud.

2. Clustering dei termini

Seguendo le stesse linee guida precedentemente utilizzate per la clusterizzazione dei documenti, abbiamo proceduto alla clusterizzazione dei termini. In questa analisi abbiamo fatto riferimento alla Term-Document Matrix, che altro non è che la Document-Term Matrix trasposta. Siamo partiti utilizzando un metodo gerarchico per avvalerci del dendogramma per poter vedere come fossero divise le parole. Siamo passati quindi alla considerazione di un metodo partizionale applicando l'algoritmo K-Means individuando K gruppi per poi rappresentarli graficamente tramite Cusplot. In questa fase, data la presenza numerosa di parole, il dendogramma risultava impossibile da leggere. Abbiamo per cui deciso di considerare all'interno della Term Document Matrix solo le parole con una frequenza superiore a 60. È molto importante ricordarsi di questa considerazione quando si andranno a vedere i risultati

3. Classificazione dei documenti

Abbiamo assunto di sapere a priori che i documenti risultano divisi in due gruppi, così da poter classificare i documenti di analisi. Siccome questo chiaramente non era possibile abbiamo assunto per corretta la suddivisione tramite clustering. La classificazione prevede la conoscenza della variabile risposta che nel nostro caso fa riferimento ad una divisione fra articoli: quelli riferiti al fattore di crescita dell'endotelio vascolare ("Vegf") e quelli che trattano più generali attività a livello cellulare sempre in ambito di angiogenesi tumorale. Questa variabile target è stata aggiunta alla Term-Document -Matrix sottoforma di variabile dicotomica; 1 se il documento fa riferimento al "Vegf, 0 altrimenti. In questo modo è stato possibile effettuare una classificazione. Come metodi di classificazione abbiamo scelto il K-NN preferito a tutti i restanti poichè unico a non prevedere ipotesi sulla distribuzione dei dati di partenza. Viene infatti anche chiamato lazy learning algorithm proprio per il motivo sopra citato.

Partendo dal training data si deve definire una distanza tra osservazioni (nel nostro caso euclidea) per poi classificare le nuove osservazioni in base alla distanza minima tra la nuova osservazione e le K più vicine. Unico parametro che quindi può influenzare la complessità dell'algoritmo è il numero di osservazioni più vicine da considerare; più è alto più il classificatore risulterà avere varianza bassa ma bias alto, viceversa si rischia di incorrere nell'overfitting. Con questo algoritmo è stato inoltre possibile saltare la fase sul training+validation set, applicandolo direttamente sul test set. Si è reso necessario infine, costruire la curva di ROC per valutare se il classificatore fosse valido, calcolando anche sensitivity e specificity. Abbiamo poi provveduto alla stessa analisi su una classificazione in tre differenti cluster in quanto sebben la silhouette perdeva valore i raggruppamenti sembravano avere più senso.

4. Sentiment Analysis

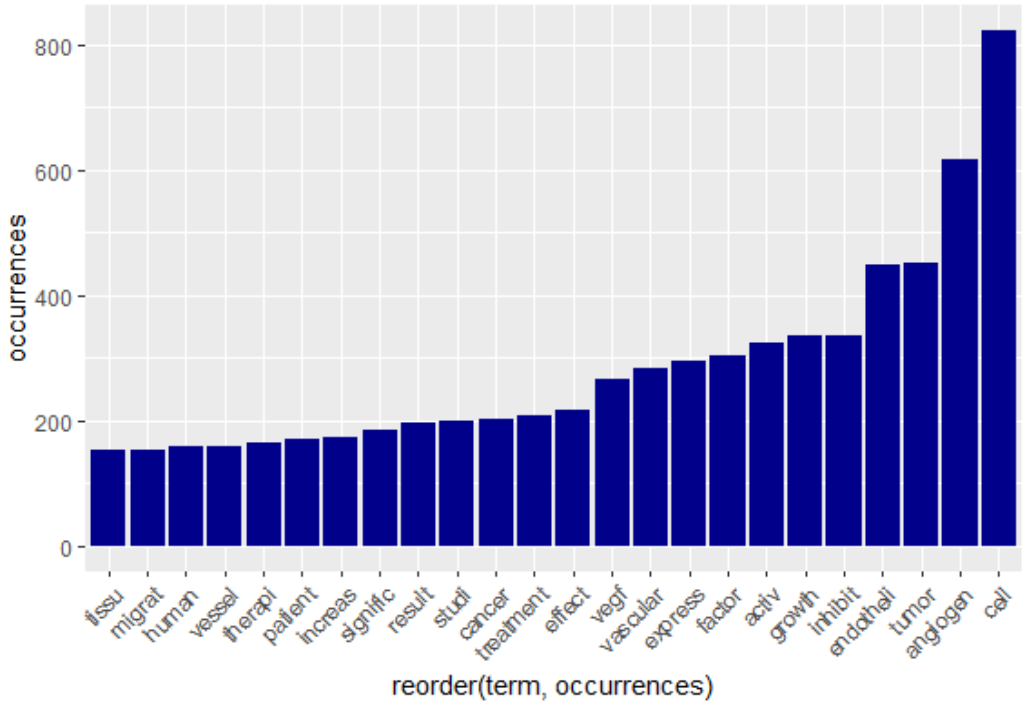
L'obiettivo principale della sentiment analysis è determinare la polarità generale di un documento (sia che si tratti di una recensione, che di un commento ad un post che di un insieme di parole), ossia classificare un documento o frase in positiva**,** negativa o neutrale. In r esistono diverse librerie e diversi dizionari. Abbiamo utilizzato nella nostra analisi il dizionario nrc che assegna 13901 termini a dieci categorie: positivo, negativo, rabbia, aspettativa, disgusto, paura, gioia, tristezza, sorpresa e fiducia. Non siamo andati ad analizzare tutte queste ma soltanto alcune. Non conoscendo la reale suddivisione degli articoli non ci siamo spinti in un'analisi sentimentale per ogni singolo raggruppamento. Il nostro elementare obbiettivo è stato solo quello di capire a carattere generale i sentimenti che traspaiono dal testo

3 Risultati

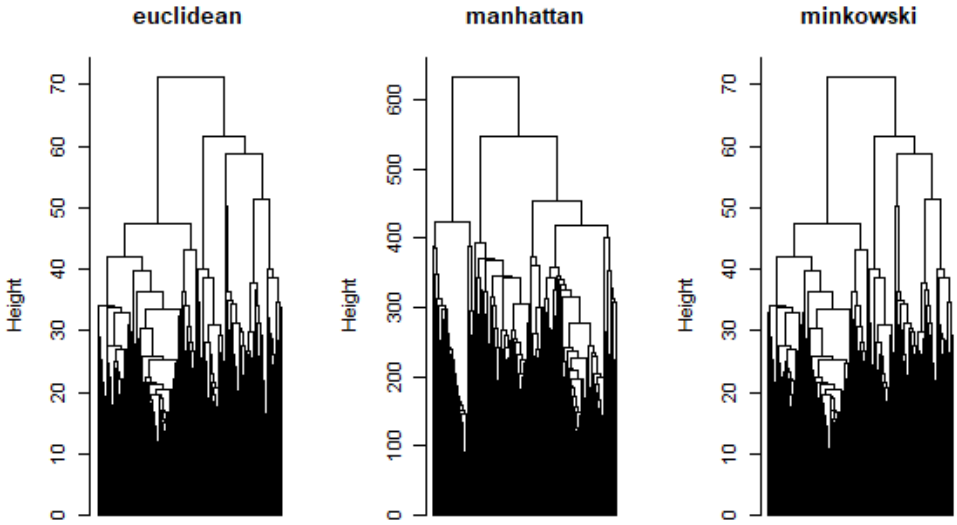
Per prima cosa abbiamo caricato i dati in R. Una volta importate tramite il comando "Corpus", che richiama il concetto di corpus del text mining, le 261 citazioni biomediche, oggetto del nostro lavoro, ci siamo mossi alla fase di pre-processing, ovvero di pulizia dei documenti. Per prima cosa abbiamo rimosso alcuni dei caratteri speciali presenti nei documenti. Tuttavia, dato che i nostri texts spesso trattano di nomi di proteine, enzimi etc... (ex. MMP-2 enzyme), abbiamo deciso di mantenere caratteri quali ("-", "/"), perciò non abbiamo usato "remove punteggiatura" come funzione preferendo procedere manualmente. Inoltre, parole rilevanti alla nostra analisi comprendono numeri (ex: Interleukin-1 receptor); anche in questo caso tralasciando la funzione "remove numbers" abbiamo rimosso solo numeri presenti singolarmente riferiti a date, percentuali e formule chimiche. Parole ininfluenti per il nostro lavoro come caratteri matematici ("+", "=") li abbiamo rimossi, così come le stopwords più comuni comprendenti anche quelle selezionate da noi.

Dopodiché abbiamo eliminato gli spazi bianchi presenti nei documenti. Sempre come parte del pre-processing, abbiamo affrontato la fase di stemming. Prima di quest'ultima è risultato necessario visualizzare tramite un barplot quali fossero i termini con frequenza maggiore presenti in tutti i nostri dati così da capire quali potessero essere le parole da raggruppare sotto un'unica voce in fase di stemming. Sempre in fase di stemming si è mostrato efficacie trasformare parte dei termini riportandoli alla loro forma base. Dopo la fase di text cleaning, abbiamo costruito la Document-Term Matrix per trasformare il corpus in una tabella di conteggio, dove ciascuna cella indica la frequenza assoluta del termine di riferimento. Considerando la matrice è stato necessario focalizzarsi solo su quei termini che avessero una frequenza di apparizione relativamente "alta" tralasciandone altri da considerarsi come "rari". Così facendo abbiamo analizzato solo quelle parole che comparivano almeno in 26 articoli (~10%) composte da un minimo di tre caratteri. Abbiamo quindi valutato le parole con frequenza minima pari a 50 considerando anche le correlazioni presenti tra loro.

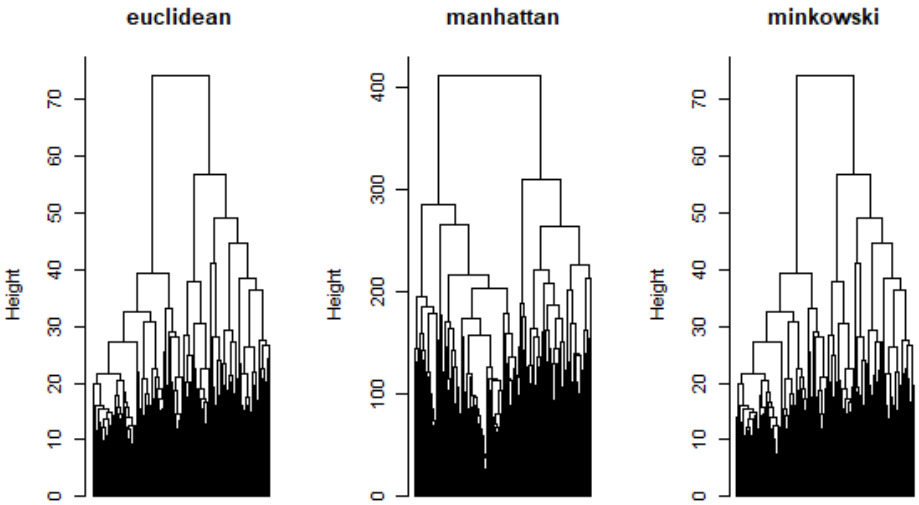
I risultati ottenuti da una breve analisi sull'associazione di parole erano quelli che ci aspettavamo di trovare, considerando ad esempio la parola piu frequente "angiogen" questa aveva correlazione intorno allo 0.3 con parole quali "inhibit" o "growth". Abbiamo quindi deciso di rappresentare tramite ggplot e wordcloud i termini più frequenti per averne un corrispettivo grafico (le prime tre sono "Cell", "Angiogen" e "Tumor"). Da una breve analisi visiva si può notare che i primi termini altro non sono che i temi rilevanti dei nostri documenti.



Siamo quindi passati ad analizzare tramite clustering i vari texts, procedendo in ordine con clustering gerarchico su documenti e quindi sulle parole. Per quanto riguarda quello gerarchico per evidenziare pattern tra i documenti, ove presenti, abbiamo considerato tre distanze: Minkowski, Euclidea e Manhattan. Aiutandoci con una rappresentazione grafica, il "Dendogramma", abbiamo ritenuto utile considerarle utilizzando il metodo di Ward come linkage dato che è risultato visibilmente migliore rispetto agli altri ("complete", "single, "average") su ciascuna delle tre metriche. Considerando quindi i "Dendogrammi" con metodo Ward considerati su ciascuna metrica abbiamo optato per quella di Manhattan in quanto, il grafico appare migliore su una eventuale scelta di 2/3 cluster. Migliore in quanto le distanze in quello di Manhattan ci permettono graficamente una suddivisione

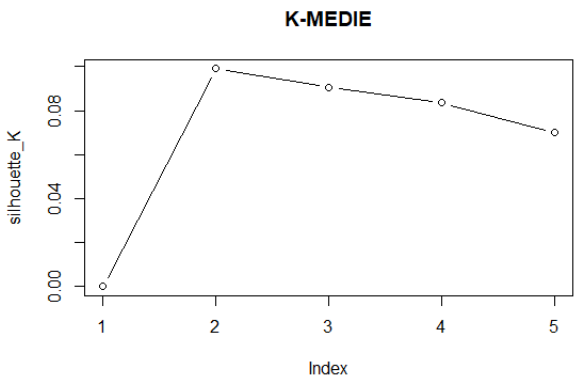


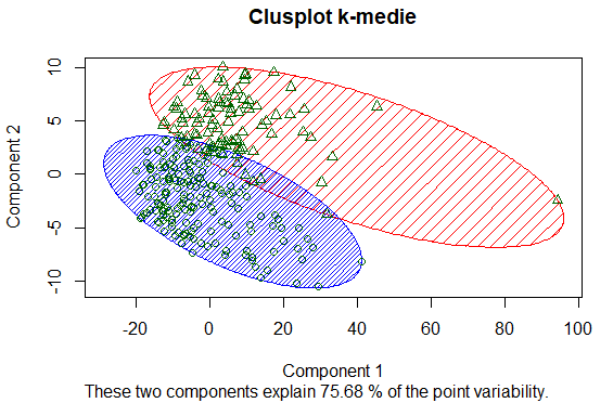
Per avere una conferma empirica abbiamo analizzato anche i risultati delle Silhouette che, seppure poco incoraggianti, ci consigliavano di analizzare il clustering tra documenti considerando linkage=" Ward", metrica=" Manhattan". Non soddisfatti dei risultati, la silhouette riportava valori con due gruppi pari a 0.01 e 0.05, abbiamo deciso di prendere come oggetto di analisi la Matrice Sparsity, che considera solo i termini con sparsity >95 per le matrici delle distanze, per cercare di migliorare il nostro risultato. Anche in questo caso abbiamo considerato "Ward"-"Manhattan" che risultava la scelta migliore (si vede che un taglio con due gruppi è migliore rispetto agli altri metodi), tuttavia i valori della silhouette hanno evidenziato ancora una volta come la divisione in cluster fosse poco rappresentativa. Ma è davvero così?



Abbiamo intuito come il risultato poco soddisfacente della silhouette non fosse dovuto a qualche tipo di problema nelle metriche da noi usate ma bensì a qualcosa di più profondo. Ricordando che la tecnica di clustering è un raggruppamento che forza tutte le osservazioni in gruppi anche se magari queste sono outlier oppure diverse dal cluster fatto, viene facile credere che all'interno di qualche ramo siano presenti articoli non rilevanti per la nostra analisi e/o molti outlier. Considerando inoltre che molti articoli usano le medesime parole, in quanto parlano sempre delle stesse cose, risulta facile pensare che l'analisi abbia problemi.

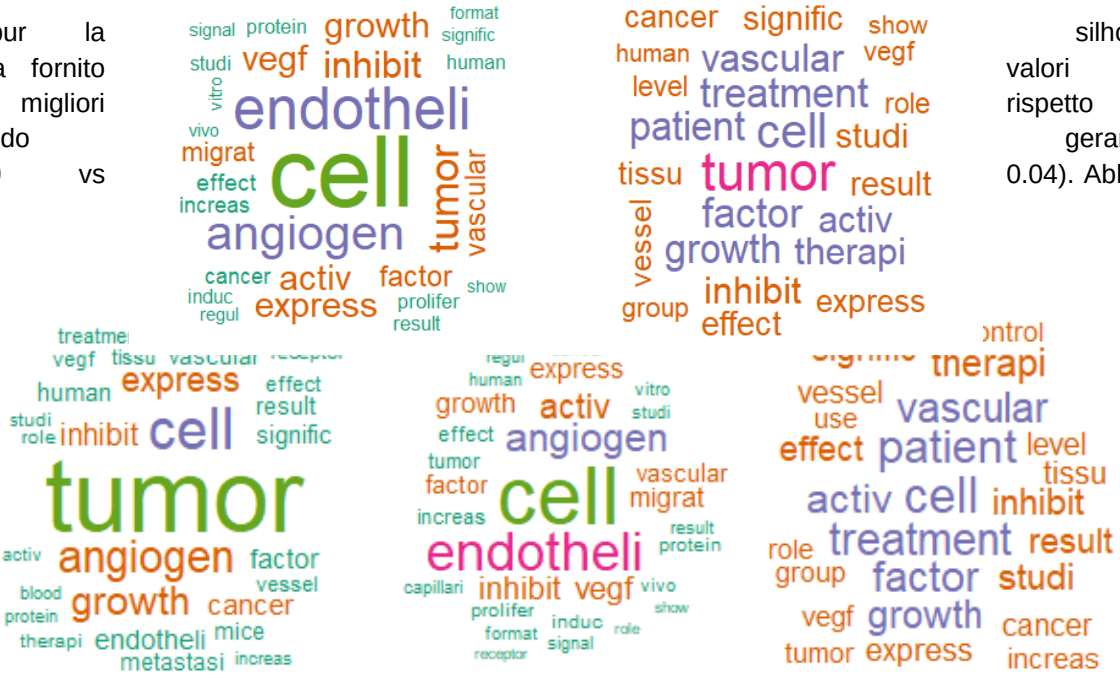
Non ottenendo quindi risultati soddisfacenti siamo andati avanti considerando un clustering partizionale, nella speranza che questa suddivisione potesse essere migliore (facile pensare che ci saranno gli stessi problemi). Abbiamo provato a vedere se col K-Means si riuscisse a migliorare il nostro lavoro di analisi; abbiamo utilizzato la matrice di distanze euclidea in quanto migliore rispetto agli altri metodi. La scelta del K, cruciale per la nostra analisi, è stata fatta prendendo a riferimento il valore di K che massimizzasse l'average silhouette width, ottenendo K=2. Considerando sempre come punto di riferimento empirico i valori della silhouette la situazione non è migliorata. Tuttavia, procedendo con il cusplot (grafico rappresentativo del k-means, che considera le prime due componenti principali) per vedere come venissero raggruppati i nostri testi, siamo arrivati a delle conclusioni interessanti: i testi raggruppati nella zona centrale del grafuco sono anche quei testi difficilmente classificabili in un cluster piuttosto che in un altro perchè riguardanti i "macroargomenti" comuni a tutti.





Il risultato tramite wordcloud sulle k-means tramite una divisione in due gruppi. Risulta il seguente:

Seppur la silhouette abbia fornito valori bassi, sono migliori rispetto al metodo gerarchico (0.10 vs 0.04). Abbiamo



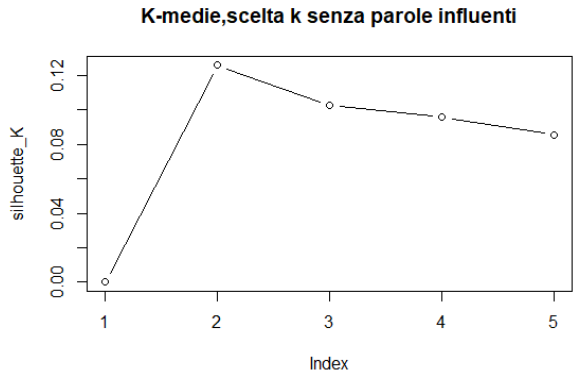
deciso, dato il valore medio della silhouette simile di tenere nelle nostre considerazioni anche una suddivisione in 3 gruppi. Ricordiamo che in questo caso il suo valore non è così importante (spiegato prima) e quindi bisogna capire se conviene tenere più gruppi con un valore più basso o meno gruppi con valore leggermente più alto (differenza di valori, 0.10 vs 0.0909…). Ora con k=3 abbiamo questo risultato:

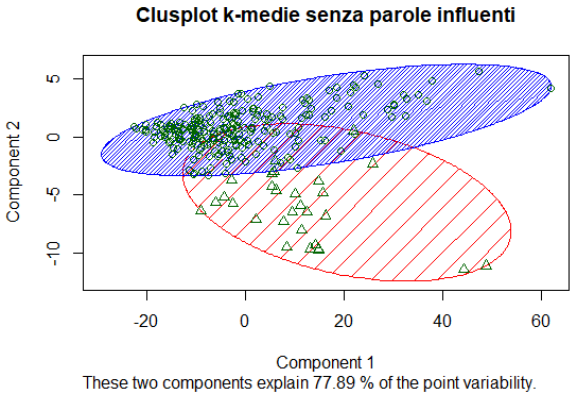
Una possibile suddivisione logica (seppur quasi inesistente) appare sia nel taglio con 2 cluster sia nel taglio con 3. Ad esempio, possiamo intuire che nella suddivisione con k=2 una parte di articoli parlerà nello specifico dell'Angiogenesi mentre l'altra parte più dei trattamenti di tale malattia. Nella suddivisione con k=3 invece possiamo raggruppare gli articoli suddividendo rispetto al k=2 quelli che parlano del tumore e quelli relativi al trattamento. Molte parole però sono in comune sia nel taglio dei 2 sia nel taglio dei 3 gruppi, e

questo rende difficile trarre qualche significato dai wordcloud.

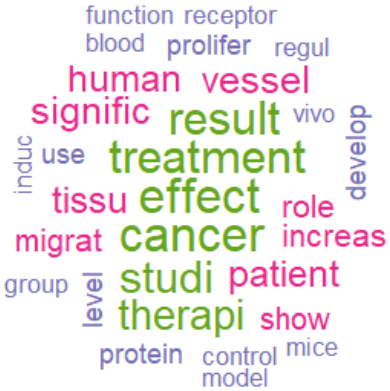
Dall'ultimo risultato ottenuto siamo quindi passati ad una analisi togliendo alcune parole influenti in comune tra i due/tre cluster nonché anche quelle più frequenti nei diversi testi. Questo per vedere se ci fosse qualche tipo di collegamento tra gli articoli rimasto nascosto a causa della presenza di queste parole. Abbiamo quindi "eliminato" alcuni termini: "cell", "angiogen", "tumor", "growth", "factor", "vascular", "endotheli", "inhibit", "express", così da riproporre una analisi con k-means, il metodo più performante visto in precedenza, considerando sempre la Matrice Sparsity. Abbiamo quindi scelto nuovamente il K che massimizasse la distanza euclidea con lo stesso procedimento precedente, trovando sempre come valore migliore quello pari a K=2(come spiegato già in precedenza porteremo avanti anche K=3, differenze silhouette 0.12-0.10).

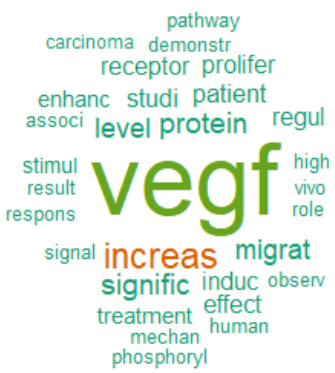
Notiamo un miglioramento della silhouette seppur molto debole dato che non supera quei valori intorno 0.25, per cui la struttura sostanziale possa considerarsi non totalmente assente (problema già spiegato in precedenza). Anche qui rappresentando con cusplot quanto ottenuto possiamo dedurre che non vi è una netta distinzione tra cluster. Ma appare buona la varianza spiegata, circa il 77%.





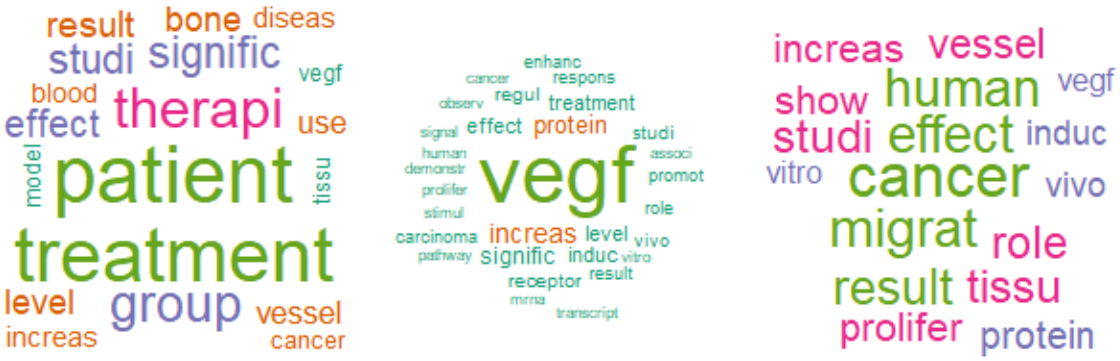
Abbiamo quindi optato per una rappresentazione tramite wordcloud dei termini più rilevanti divisi per cluster riuscendo molto soggettivamente ad individuare due (forse tre) temi distinti. Uno ruota intorno al termine "Vegf" che abbiamo scoperto essere un noto fattore di crescita





dell'endotelio vascolare, l'altro riguardante termini collegabili a risulati, trattamenti e terapie, cancro e comprendente anche termini riferiti ad una più generale attività cellulare.

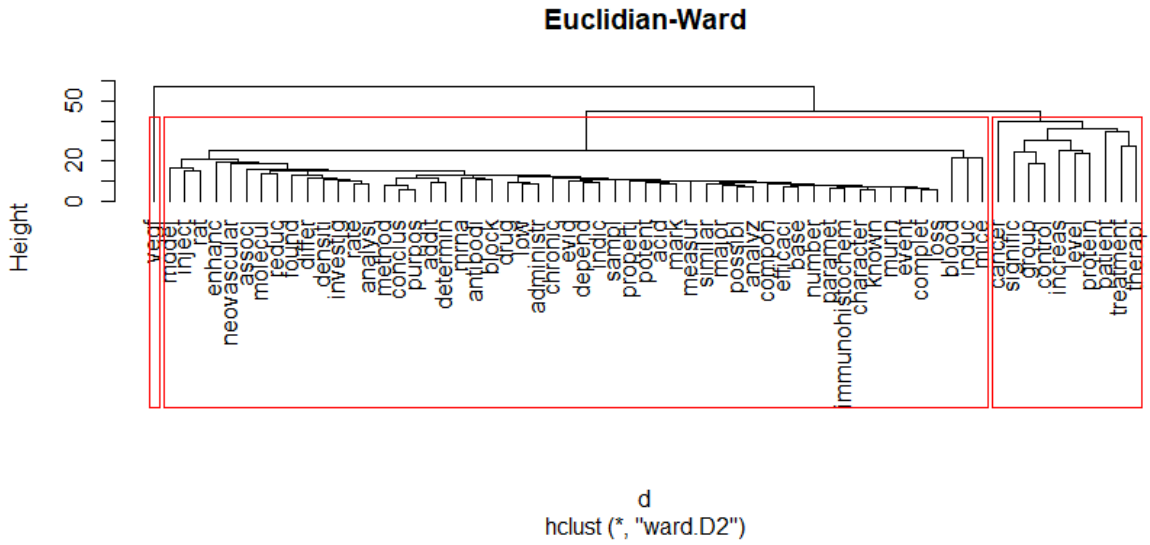
Abbiamo quindi poi considerato anche wordcloud di termini divisi in tre cluster. Da qui si è evidenziato come



il secondo cluster precedente fosse composto da due cluster più significativi dal punto di vista interpretazionale; si hanno ora da un lato termini collegati a terapie risultati e analisi sul cancro dall'altro attività cellulari più generali, individuando sempre il ruolo del "Vegf" a sè stante in un unico cluster. Un rilevante problema è dato dallo sbilanciamento fra le classi in quanto il cluster 1 che ruota intorno al termine "Vegf" faccia riferimento a solo 20 di 261 dei documenti (~13%), per cui sarebbe una suddivisione davvero debole quella con 2 cluster seppur con maggior silhouette.

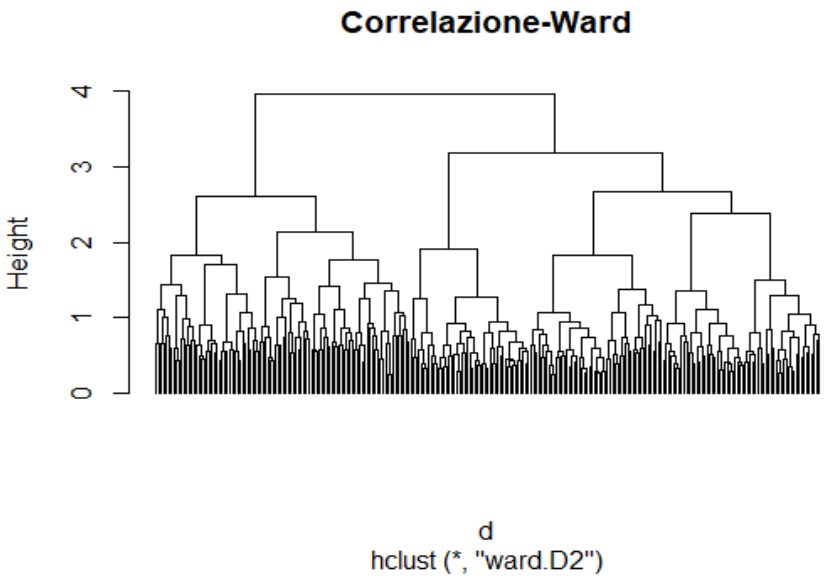
Quindi abbiamo considerato il clustering per le parole totale indipendentemente dai testi. Le aspettative sono quelle di notare due cluster sbilanciati fra di loro per il risultato visto in precedenza. Siccome presenti troppe parole decidiamo di rimuovere quelle meno influenti per evitare problemi di lettura del dendogramma. Anche in questo caso il metodo Ward è risultato il migliore per la nostra analisi, a differenza di prima, tuttavia, abbiamo preso a rifermento la distanza euclidea. Analizzando i dendogramma,

considerando parallelamente un cut in corrispondenza di K=2 e K=3, abbiamo ritrovato i termini corrispondenti ai gruppi individuati in preferenza clusterizzando sui documenti. Abbiamo quindi fatto sempre riferimento alla silhouette per trovare un corrispettivo empirico, ma i valori dei gruppi minori, nonche composti da un solo termine non sono risultati rappresentativi, così come il cusplot. Si potrebbero considerare questi gruppi composti da un solo termine (cluster 1: "Vegf" come outlier, tuttavia si commetterebbe un errore in quanto tali parole risultano il tema principale dei clustering fatti nella prima analisi)

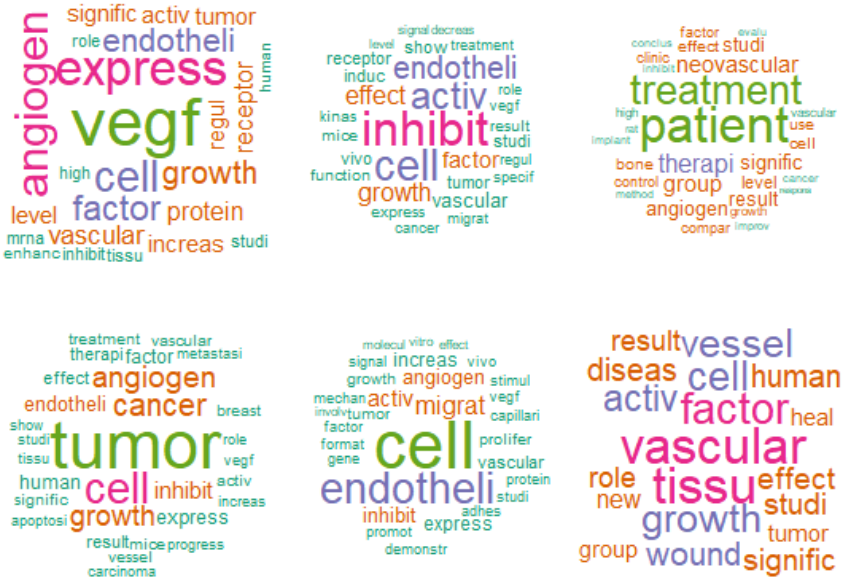


.

Prima di procedere verso una possibile classificazione, dati i risultati deboli delle analisi precedenti, abbiamo deciso di provare un metodo di dissimilarità ovvero quello della correlazione per vedere se migliorasse qualcosa. L'abbiamo testata con la matrice iniziale senza aver tolto nessuna parola. Come prima analisi il dendogramma appare molto bello:



Abbiamo optato sia per valori di silhouette, sia per maniera visiva un taglio a 6 che comunque massimizza la silhouette media. Il risultato tramite wordcloud è stato il seguente:



Con questo metodo di distanza si riescono a scovare 6 gruppi con differenti parole centrali, cosa che prima non era accaduta. Seppur buona questa suddivisione, sia in termini empirici sia in termini grafici procediamo con la soluzione trovata prima.

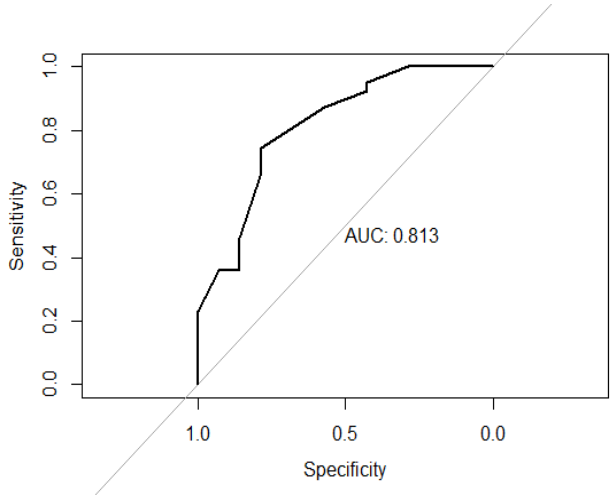
I nostri testi e citazioni biomediche non sono risultate suddivisibili in gruppi evidenti. Questo può essere dovuto al fatto che le tematiche che trattano sono molto simili tra loro. Inoltre, solo dai wordcloud si può evincere una suddivisione che trova valore solo se considerata in maniera soggettiva, infatti, i risultati

riportati da metodi oggettivi quali ad esempio la Silhouette hanno fatto propendere per una assenza di gruppi significativi. Si potrebbe analizzare più nel dettaglio tale problema andando a capire quali siano i testi outlier facendo una procedura top-down (il contrario di quello che abbiamo fatto)

Un altro passo della nostra analisi è stato quello di optare per una classificazione K-NN. Abbiamo assunto di sapere a priori che i documenti risultassero divisi in due gruppi; uno riguardante il fattore di crescita dell'endotelio vascolare, l'altro comprendente argomenti di vario genere a livello cellulare, nonchè i due gruppi individuati dal cluster sui documenti. Abbiamo quindi cercato di vedere se K-NN fosse efficace per la nostra classificazione delle due tipologie. Una volta aggiunta una colonna alla matrice Document-Term, con variabile dicotomica (0,1) dove 1 sta ad indicare gli articoli legati a "Vegf", abbiamo diviso in training e test la matrice stessa. Nella nostra analisi è importante ricordare che non va considerato il validation test.

Punto fondamentale è stato quello di decidere il valore di K che abbiamo testato in un intervallo da 1 a 20 in quanto essendo le classi molto sbilanciate la scelta di un K alto altererebbe i risultati. La decisione migliore è stata quella di un K pari a 16 in corrispondenza del quale l'accuracy è massima. Abbiamo infine rappresentato la curva di ROC, significativa a livello grafico, prendendo a riferimento la Confusion Matrix che ha portato a risultati soddisfacenti. La accuracy si avvicina al 85%, con una sensitivity e una specificity corrispondentemente del 100% e 83% mentre l'AUC anch'esso elevato è pari a 0.813.

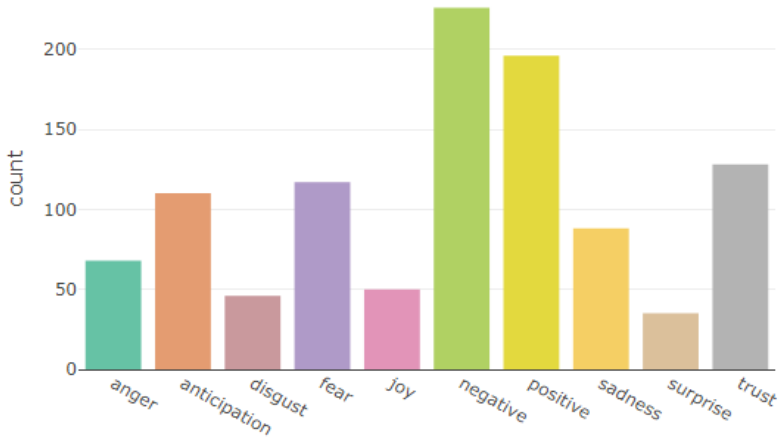
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  6  8
         1  0 39

               Accuracy : 0.8491
                 95% CI : (0.7241, 0.93
    No Information Rate : 0.8868
    P-Value [Acc > NIR] : 0.85996

                  Kappa : 0.5247

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 1.0000
            Specificity : 0.8298
```



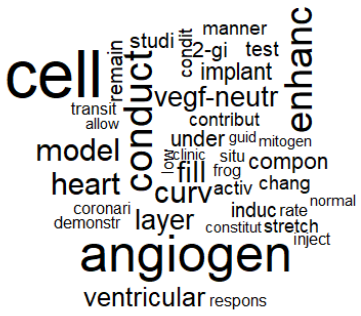
Abbiamo portato in parallelo lo stesso procedimento considerando anche la divisione degli articoli nei 3 gruppi differenti precedentemente trattati. La colonna aggiuntiva alla matrice Document-Term risulta in questo caso tricotomica (0,1,2) dove si ha 1 in corrispondenza degli articoli riguardanti il "Vegf", e 2 gli articoli legati ai trattamenti e alle analisi. Svolgendo gli stessi procedimenti precedentemente descritti siamo arrivati ad ottenere in questo caso, con un K ottimo testato pari 11, una accuracy pari al 60% (molto bassa) circa ma ottimi valori di sensitivity e specificity.

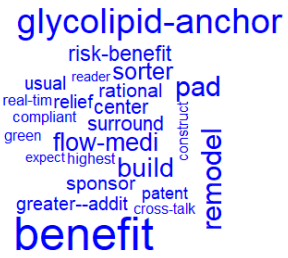
Ultimo punto considerato, più per diletto che per fini rappresentativi di rilevanza, è stato quelli di procedere con una Sentimental Analysis sui nosti termini. Come prima cosa abbiamo cercato di capire se nei nostri testi fossero presenti più parole positive, negative o rappresentative di altri tipi di emozioni, aspettandoci chiaramente una prevalenza di termini classificati come negativi. Abbiamo valutato dunque una distribuzione dei sentimenti dove si è notato come i termini negativi fossero in maggioranza ma che ha

evidenziato comunque la presenza di diversi termini positivi, superiori alle aspettative.



Non soddisfatti abbiamo indagato il fatto accorgendoci come molte parole classificate come “positive” nei nostri documenti avessero valenza piuttosto negativa. Rappresentando i termini più frequenti tramite wordcloud differenziati per classi dati i diversi sentimenti di riferimento, abbiamo notato appunto come certi termini tra i quali spiccava “Intervention” fossero caratterizzati come positivi, distorcendo i risultati. Abbiamo quindi dedotto che il dizionario spesso associa in malo modo molti termini che nel nostro text mining hanno valenze diverse. Si sarebbe potuta fare un’analisi riguardante nello specifico le parole più influenti oppure guardando solo gli articoli relativi a diversi gruppi ma non essendo stato un nostro obiettivo non ce ne siamo interessati.





4 Discussioni

Abbiamo affrontato un problema di text mining relativamente a una collezione di articoli medici, cercando di trarne fuori qualcosa di significativo. Abbiamo visto che metodi oggettivi e analitici, quali clustering gerarchico e k-medie, non risultano validi quando la quantità di dati da trattare risulta ampia e piena di insidie (vedasi parole comuni in quasi tutti i testi). Ci siamo quindi spinti ad un approccio più soggettivo cercando qualche escamotage per tirare fuori dai dati qualche risultato utile. Siamo riusciti quindi ad extrapolare dei wordcloud previa eliminazione di tutte quelle parole appartenenti alla maggior parte dei file arrivando ad una suddivisione significativa degli articoli (in 2/3 cluster). Abbiamo analizzato il metodo della correlazione grazie al quale abbiamo scoperto una suddivisione più logica tra articoli seppur con un maggior numero di clustering. Il nostro problema rilevante è stato la non conoscenza della classificazione dei vari articoli che siamo riusciti a superare tramite quanto detto in precedenza. Ci siamo inoltre a classificare gli articoli KNN supponendo che la suddivisione tramite i nostri cluster fosse “reale”. Abbiamo concluso l’analisi scoprendo che come da aspettativa dai testi traspaiono sentimenti negativi.