BBC Buiness Articles text exploration and benchmark bulding for Automatic Topic Modeling evaluation

Tommaso Di Vincenzo, 801487* Riccardo Maganza, 808053†

22 aprile 2018

Abstract

Nel paper viene condotta un'analisi su un corpus di 510 articoli di business in lingua inglese. Il fine è quello di confrontare le performance di riconoscimento dei topics individuati dal modello *Latent Dirichlet Allocation* (LDA) con quelle derivanti da una suddivisione per argomenti creata manualmente a partire da un clustering ottenuto con l'algoritmo dei k-medoidi. Viene successivamente analizzata la coerenza semantica dei topics individuati attraverso la Normalized Google Distance (NGD). I risultati sono incoraggianti: le performance dell'algoritmo automatico sono molto vicine al benchmark ed in alcuni casi lo superano. Si discutono inoltre le applicazioni metodologiche di questi risultati. [1]

1 Introduzione

L'estrazione di parole chiave da un testo è un'abilità innata per gli umani, che sono in grado di distinguere con molta chiarezza ciò che è importante da ciò che può passare in secondo piano.
In tempi moderni la quantità di dati disponibili sotto forma di documenti non strutturati è aumentata esponenzialmente, e con essa la difficoltà nel conservare questi testi ed indicizzarli. Se precedentemente categorizzare ogni documento manualmente in fase di caricamento, in un qualche tipo di database, poteva essere una possibilità, oggigiorno ciò sembra alquanto proibitivo.
Il cosiddetto *topic modeling* è un concetto alquanto recente, essendosi sviluppato a partire dal 1999 all'interno della ricerca sull'*information retrieval*. [**1999**] Esso mira ad estrarre automaticamente delle *keywords* da un elenco di documenti e ad associare ognuna di queste *keywords* a uno o più macroargomenti.
A partire dal modello *Latent Semantinc Indexing*, ne sono state proposte alcune

*t.divincenzo@campus.unimib.it
†r.maganza@campus.unimib.it
[1]Tutto il codice è disponibile su https://gitlab.com/xelmagax/TextMiningBBCBusiness

estensioni, quali il *Probabilistic Latent Semantic Analysis* e il *Latent Dirichlet Allocation*, che risulta ad oggi il modello probabilmente più usato nell'ambito. Recentemente sono state proposte anche alcune generalizzazioni del modello LDA quali la *Pachinko Allocation*, che consente di modellizzare anche la correlazione fra i vari *topics*, ma ai fini di questo lavoro ciò è stato reputato non essenziale, ed è stato considerato il modello LDA. [**survey**]
L'obiettivo ultimo del lavoro è quello di comparare le performance di estrazione di *topics* del modello LDA con un benchmark creato ad hoc ed individuare una corretta metrica di confronto semantico per valutare la bontà del modeling del metodo automatizzato.
Generalmente la coerenza semantica fra due o più parole si misura osservandone la presenza e la vicinanza in un corpus di testi molto ampio e rappresentativo quale, per esempio, il corpus di articoli di Wikipedia, e calcolando alcune misure di associazione.[**ReadingTeaLeaves**] Non avendo a disposizione un corpus adeguato allo scopo, è stata utilizzata la *Normalized Google Distance* (NGD), una metrica basata sulla co-occorrenza di una coppia di parole in una ricerca su Google. [**NGD**] I risultati ottenuti con questa metrica sono stati sorprendenti: in alcuni casi il modello LDA è riuscito ad estrarre una collezione di keywords semanticamente più coerente rispetto ad una divisione in *topics* generata manualmente. Verrà inoltre menzionato un metodo appartenente allo stato dell'arte nell'ambito del *Natural Language Processing*: il framework *Natural Language Understanding* (NLU) del supercomputer Watson di IBM. [**watson**]

2 Materiali e metodi

2.1 Materiali

I 510 documenti oggetto di analisi provengono da una collezione di 2225 articoli della BBC degli anni 2004-2005.[**Greene2005**] Gli articoli erano provvisti di labels che identificavano 5 aree: *business*, *entertainment*, *politics*, *sports* e *tech* e l'obiettivo dell'analisi originale era di suddividere efficacemente ed in maniera automatica i documenti fra questi macroargomenti.[**GreeneOriginal**] I documenti considerati corrispondono a quelli che in questo dataset possedevano la label *business*.
È da evidenziare immediatamente come il clustering su questi testi abbia presentato non pochi problemi: infatti, essendo tutti appartenenti alla stessa area semantica, presentavano differenze fini l'uno con l'altro e molti criteri, che potevano essere discriminanti nel dataset completo originale, qui non sono stati applicabili.

2.2 Metodi

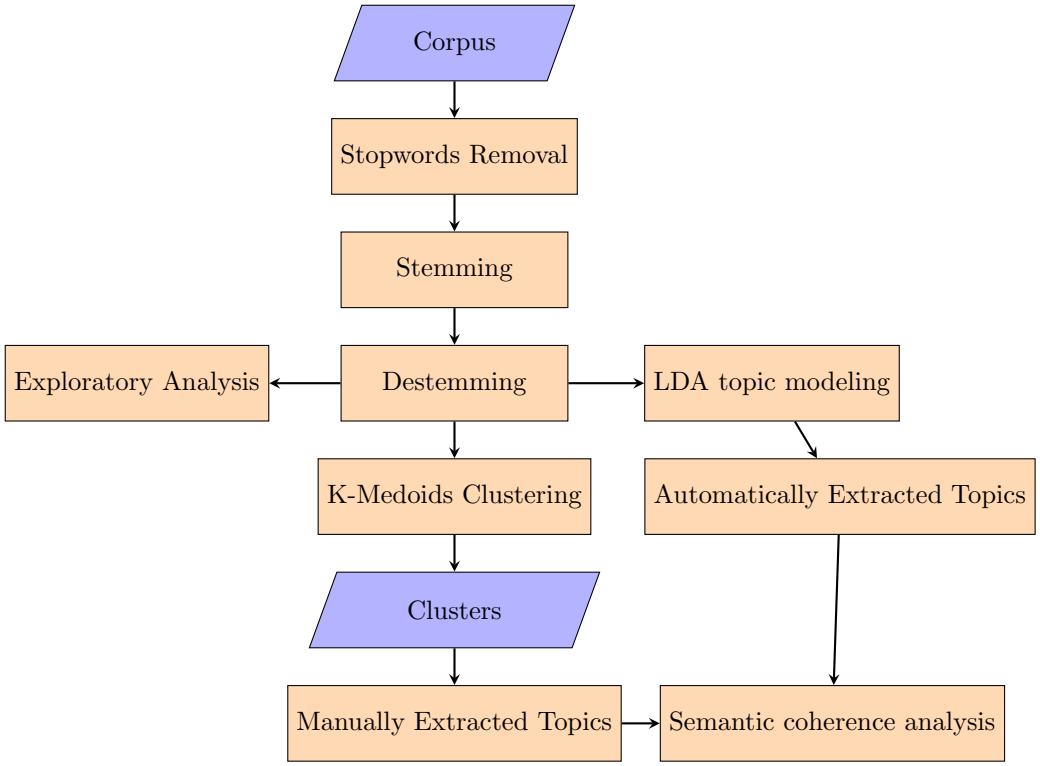

Figura 1: Diagramma di flusso del processo di analisi

2.2.1 Pre-Processing

Il corpus di articoli è stato preliminarmente trattato eliminando punteggiatura, simboli particolari e parole poco esplicative allo scopo dell'analisi.
Sono state rimosse le stopwords, utilizzando un dizionario sufficientemente ampio creato per il sistema *SMART*[2], ed è stato attuato lo stemming, rimuovendo le desinenze da tutte le parole e riconducendole a una radice comune. Questa procedura in particolare è stata affinata manualmente, poiché il sistema di riconoscimento utilizzato non è riuscito ad effettuare lo stemming di tutte le parole ed inoltre presentava, come logico, molti problemi nel riconoscere i nomi di aziende e compagnie, che nei testi in esame erano ovviamente molto presenti. Successivamente, dato che molte parole senza suffisso sono risultate di difficile comprensione, è stato applicato un destemming che ha riportato le parole ad una forma originale condivisa attraverso un criterio euristico che ha completato

[2] *System for the Mechanical Analysis and Retrieval of Text*

le parole troncate con i suffissi più frequenti nel corpus originale.
Si è proceduto quindi con la costruzione della Document-Term matrix imputando le frequenze attraverso il metodo *tf-idf:*

$$w_{ij} = \frac{f_{ij}}{max\{f_{1j,}, f_{2j}, ..., f_{|V|j}\}} \log \frac{N}{df_i}$$

dove V rappresenta il vocabolario del corpus, f_{ij} è la frequenza assoluta del termine i nel documento j, N è il numero totale di documenti e df_i il numero di documenti in cui appare il termine i.
È stata inoltre effettuata una analisi esplorativa del dataset, cercando di contestualizzare i risultati nel contesto socio-economico del periodo degli articoli in esame svolgendo, fra le altre cose, una *sentiment analysis* basata sull'*NRC Emotion Lexicon.*[**NRC**]

2.2.2 Clustering e Topic Extraction

È stata inizialmente calcolata la distanza fra le righe della Document-Term Matrix, utilizzando la metrica del coseno.

$$D(\mathbf{X}, \mathbf{Y}) = \frac{2}{\pi} \arccos(\frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|})$$

dove $\mathbf{X}$ e $\mathbf{Y}$ sono, in questo caso, due righe della DTM. [3] [**proxy**]

Successivamente è stato applicato l'algoritmo dei k-medoidi per $k \in \{2, \ldots, 15\}$ dando in input la matrice di distanze ottenuta al passo precedente. L'algoritmo sceglie inizialmente k data points casualmente come medoidi e associa ogni altro punto al medoide più vicino ad esso. Definendo una opportuna funzione di costo C si calcola per ogni oggetto non medoide O il costo $C(medoid, O)$ di scambio fra il medoide e l'oggetto. Se $C < 0$, si scambia il medoide con l'oggetto O, iterando fino a convergenza.
È stato scelto $k = 12$ poiché tale valore massimizzava la silhouette media del clustering.
Per la definizione del benchmark sono state selezionate, per ognuno dei 12 gruppi così ottenuti, le 5 parole che per la logica degli autori sono risultate essere maggiormente identificative.

Sono stati quindi estratti i topic utilizzando il modello *Latent Dirichlet Allocation* (LDA), che utilizza una logica di inferenza bayesiana [**survey**] e le cui specifiche vanno oltre lo scopo di questo articolo, richiedendo di ottenere 12 topic in output per poter avere risultati confrontabili. L'algoritmo genera in output una lista di topics con un indice di associazione di ogni parola con ciascun topic e uno di ogni topic con ciascun documento. Sono state considerate le 5 *keywords* più associate a ciascun topic come rappresentative dell'argomento.

I topics ottenuti dai due metodi sono riportati nella sezione successiva.

[3] Implementazione fornita dalla libreria *proxy* di R che permette all'indice di dissimilarità del coseno usuale di soddisfare la disuguaglianza triangolare e quindi di rispettare le proprietà che deve avere un indice di distanza.

E' stata poi utilizzata la Normalised Google Distance (NGD) [**NGD**]

$$NGD(x,y) = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - min\{\log f(x), \log f(y)\}}$$

dove $f(x)$ e $f(y)$ sono, rispettivamente, il numero di risultati ottenuti ricercando su Google le parole x e y singolarmente, $f(x,y)$ è il numero di risultati ottenuto da una query combinata delle due parole, e N è il numero totale di pagine indicizzate da Google nella lingua in esame [4]. Essa è stata utilizzata come indice di dissimilarità semantica, ed è stata calcolata per tutte le possibili combinazioni di coppie di parole per ogni gruppo. Ciascun gruppo, inerente sia i risultati manuali che quelli automatici, è stato sintetizzato attraverso la media campionaria.

In conclusione, è stata effettuata una analisi di un documento casuale nel corpus con il framework NLU di IBM Watson, per mostrarne le potenzialità.

3 Risultati

La Document-Term matrix dei testi ripuliti contiene 2831 termini unici e un indice di sparsità del 96%. Applicare metodi di clustering più complessi avrebbe potuto causare problemi dovuti a una sparsità così elevata, come suggerito in [**GreeneOriginal**] e si raccomanda di tenere ciò a mente in ottica di sviluppi futuri. Il corpus completo viene ben sintetizzato dalla seguente Wordcloud[5]:

[4] Poiché questo numero non è né noto né conoscibile, esso è stato stimato con il numero di risultati ottenuti ricercando la parola "the". Questo porterà ad avere dei risultati in cui la NGD è decisamente sovrastimata, ma ciò non sarà un problema mantenendo N costante per tutti i confronti.

[5] La wordcloud è stata generata in Python con l'ausilio della libreria Wordcloud.

Figura 2: Wordcloud del corpus completo

Le parole più evidenti ben identificano il periodo storico degli articoli. Il periodo pre-crisi era caratterizzato da un clima di fiducia (profit, growth..) e dall'ascesa dei BRICS[6] (China, India..). Le possibilità per analisi esplorative nel dettaglio sono infinite, ma vengono rimandate a successivi sviluppi per non deviare dal focus dell'articolo.
Si è scelto di concentrarsi su un tema molto presente negli articoli: lo scandalo Yukos[7] [**Yukos**]: la parola "Yukos" è associata al 64% con la parola *evasion*, e al 63% con la parola *court*, per esempio.

Verrà inoltre eseguita una sentiment analysis fra i documenti del cluster individuato corrispondenti allo scandalo Yukos, confrontandola con quella di un altro cluster.

I 12 cluster individuati, sono per l'appunto ben definiti: l'algoritmo dei k-medoidi unito alla dissimilarità del coseno ha prodotto ottimi risultati. I topics di ciascun cluster, individuati dagli autori per il confronto con il metodo

[6] Brasile, Russia, India, Cina, Sudafrica

[7] Nel 2003 Mikhail Khodorkovsky, CEO della Yukos, una delle più importanti aziende petrolifere russe di epoca post-sovietica, viene arrestato con le accuse di frode fiscale e bancarotta fraudolenta. Da quel momento al 2007, l'azienda viene liquidata.

automatizzato, sono riportati nella seguente tabella, unitamente alla loro NGD media:

Cluster	Keywords	Mean NGD
Cluster 1	{profit, sale, game, share, earn}	0.6419
Cluster 2	{dollar, crude, deficit, bush, barrell}	0.4823
Cluster 3	{yukos, russia, gazprom, court, action}	0.8301
Cluster 4	{fiat, italy, saab, opel, motor}	0.7555
Cluster 5	{economy, growth, house, unemployed, inflation}	0.6056
Cluster 6	{china, yuan, japan, israel, islam}	0.7376
Cluster 7	{lanka, disaster, people, indonesia, tsunami}	0.8734
Cluster 8	{airline, india, quantas, airbus, lufthansa}	0.6679
Cluster 9	{börse, deutsche, euronext, takeover, shareholder}	0.7223
Cluster 10	{retail, sale, store, christmas, lvmh}	0.8551
Cluster 11	{ebbers, fraud, verizon, qwest, lawyer}	0.7972
Cluster 12	{insurance, marsh, investigation, pension, plead}	0.4112

Tabella 1: Topic estratti manualmente

È stata effettuata una sentiment analysis relativa ai documenti del primo cluster, ben rappresentativo dell'intero corpus, e una relativa ai documenti del terzo cluster, già citato in precedenza. Le differenze (e le similarità) sono notevoli:

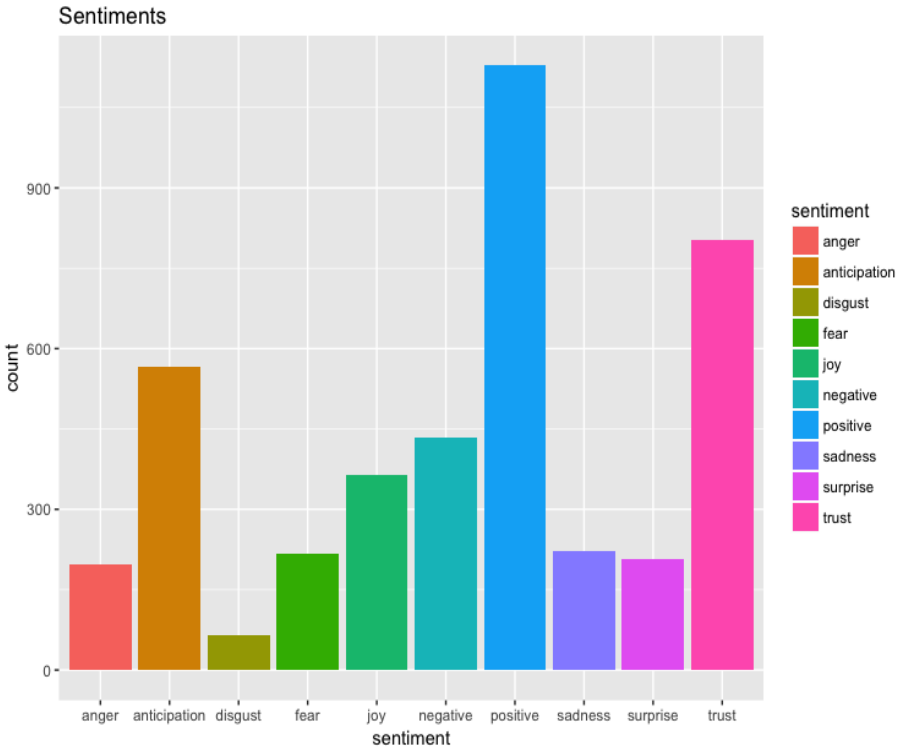

Figura 3: Sentiment Analysis Cluster 1

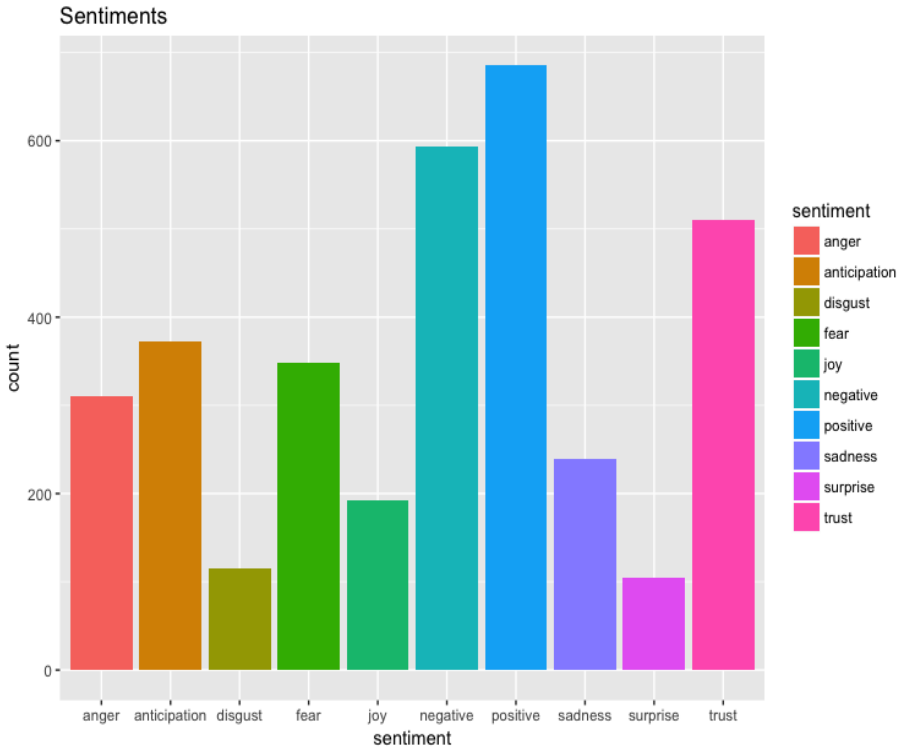

Figura 4: Sentiment Analysis Cluster 3

In entrambi i cluster, il sentimento prevalente è positivo, il che riflette il periodo generale pre-crisi, caratterizzato da una incrollabile fiducia nel sistema economico.
È pero evidente come nel cluster 3 ci sia molta più negatività, trattandosi comunque di documenti che trattavano temi quali evasione fiscale, e bancarotta.

Per sviluppi futuri, sarebbe interessante selezionare altre compagnie/aziende ed effettuare una sentiment analysis relativa ai testi che le nominano per comprendere meglio il sentimento del mercato verso le stesse.

Passando al topic modeling effettuato con il modello LDA, la lista delle keywords indivuate automaticamente con le NGD medie è riportata nella tabella successiva:

Topics	Keywords	Mean NGD
Topic 1	{dollar, deficit, euro, budget, trade}	0.5121
Topic 2	{bank, companies, firm, deal, financial}	0.2864
Topic 3	{companies, firm, worldcom, ebbers, telecom}	0.773
Topic 4	{yukos, russia, companies, court, firm}	0.6365
Topic 5	{airline, cost, report, fuel, india}	0.6619
Topic 6	{offer, deutsche, börse, share, london}	0.704
Topic 7	{price, house, market, china, mortgage}	0.9843
Topic 8	{economic, growth, rate, rise, figure}	0.3944
Topic 9	{club, unit, glazer, invest, argentinah}	0.8455
Topic 10	{economic, countries, govern, world, people}	0.6586
Topic 11	{profit, sale, share, market, companies}	0.6995
Topic 12	{companies, drug, call, firm, customer}	0.5654

Tabella 2: Topics estratti automaticamente

I topics estratti dal modello LDA in alcuni casi risultano sovrapponibili a quelli estratti manualmente o comunque chiari, mentre in altri sono un po' più fumosi. Per quanto riguarda la coerenza semantica, i 12 topics estratti manualmente presentano una NGD media di 0.7 mentre quelli automatici di 0.64.
Il modello automatico fornisce una coerenza semantica addirittura maggiore del modello manuale, mostrandosi però meno capace ad individuare topics sempre ben definiti. Ciò porta a chiedersi se la NGD catturi tutta l'informazione possibile circa la coerenza semantica di due parole. Si rimandano le discussioni sulla metrica utilizzata all'ultimo paragrafo.
Per ultimo, è stato utilizzato IBM Watson sul primo testo del corpus:

> Ad sales boost Time Warner profit
>
> Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.
>
> The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.
>
> Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.
>
> Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner

> is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.
>
> TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

Watson ha identificato come *keywords*: *Time Warner*, *Time Warner profit*, *higher advert sales*, *AOL Eruope*, fra le altre, riuscendo inoltre a riconoscere le compagnie in esame e le persone citate accompagnando il riconoscimento con un link alle informazioni di ciascuna.

4 Discussioni

Il riconoscimento automatico dei topics si è mostrato valido: per quanto ancora non riesca a sintetizzare in maniera "umana", la metrica utilizzata lo pone a un livello superiore alla sintesi manuale. Sarebbe interessante, da un lato, sviluppare il lavoro utilizzando metodi di topic extraction automatica più sofisticati, o con un approccio di tipo numerico come la *Non-negative Matrix Factorization* (NMF); dall'altro, per la parte di riconoscimento manuale, sarebbero interessanti metodi di clustering più adeguati .
Sarebbe inoltre necessario sviluppare una metrica di coerenza semantica migliore, in quanto l'utilizzo della Normalized Google Distance desta non pochi dubbi. Per farlo, servirebbe però un corpus di riferimento strutturato e molto ampio, ma non è stato possibile reperirlo.

Ci si chiede, inoltre, quale possa essere l'obiettivo ultimo della sommarizzazione di un documento: se essa debba essere *umanamente* comprensibile o se, come probabile, tale categorizzazione debba essere riconosciuta solamente da *macchine*, per esempio per la ricerca di documenti in banche dati testuali. A quel punto probabilmente non sarà nemmeno più necessario che esse mantengano una certa coerenza semantica, rendendo necessario lo sviluppo di un nuovo metodo di valutazione.
IBM Watson d'altro canto, fornisce uno strumento accessibile a tutti e decisamente sofisticato e all'avanguardia per l'estrazione di informazioni da documenti. Nonostante si sia quasi raggiunto l'apice nell'ambito dell'*information retrieval* con Watson, la sua natura proprietaria e privata rende comunque necessario uno sviluppo della ricerca nell'ambito dell'*automatic topic modeling*.