Scuola pubblica... al *testoscopio*

di SILVIA BENVENUTI

26 Febbraio 2011, congresso dei Cristiano Riformisti, Roma: un Presidente del Consiglio in stato di grazia si produce in una girandola di esternazioni a ruota libera: no a «coppie gay e adozioni dei single», abbasso Fini e i «comunisti», viva «tutti i giovani presenti, [...] così simpatici e belli che vi invito tutti al bunga bunga!», abbasso l'opposizione che non collabora al bene comune, viva Tremonti, abbasso la patrimoniale e le intercettazioni e, ciliegina sulla torta, abbasso la scuola pubblica, i cui «insegnanti inculcano ai ragazzi valori diversi da quelli delle loro famiglie»

Affermazioni, applaudite dal pubblico presente, destinate ad avere una notevole eco nei quotidiani nazionali. Inevitabile inoltre la reazione di una delle categorie mazziate, gli insegnanti, che organizzano per il 12 marzo una manifestazione nazionale a sostegno della scuola pubblica, ad affiancare quella già prevista in difesa della Costituzione, con conseguente copertura mediatica.

Morale: se selezionate, negli archivi del quotidiano *la Repubblica* e, per *par condicio*, in quelli de *il Giornale*, tutti gli articoli usciti tra il 27 febbraio e il 15 marzo e contenenti la stringa "scuola pubblica", totalizzerete 527.492 caratteri, 97.877 parole, 7192 righe e 687 paragrafi, corrispondenti a ben 148 pagine. Leggendovele tutte, potete farvi un'idea del resoconto che dell'intera vicenda forniscono le due testate, inquadrando i fatti da due prospettive politiche notoriamente molto diverse.

Ma l'idea potete farvela anche non leggendole, le 148 pagine, a patto di disporre di un buon *software* per l'analisi testuale. Ed ecco che viene fuori la matematica, che di per sé con il capo del governo c'entra fortunatamente ben poco. Programmi di questo tipo, infatti, utilizzano strumenti matematici e statistici (oltre che, ovviamente, linguistici) per analizzare testi più o meno complessi e trasformarli in tabelle, elenchi, grafici e mappe, fornendo così una sorta di "rappresentazione geometrica" dei testi stessi, che consente in qualche senso di leggerli... senza leggerli!

L'ipotesi alla base di un qualunque *software* per l'analisi testuale è la presenza di un legame profondo tra la *struttura lessicale* di un testo (cioè i lemmi più ricorrenti, le associazioni tra lemmi, ecc.) e la sua *dimensione semantica* (cioè il significato che il testo vuole veicolare). In altri termini, effettuare l'analisi di un testo con un programma quale Alceste[1], DBT[2], T-Lab[3] ed altri ha senso se si crede che un'analisi mirata delle parole usate possa consentire di fare inferenze a proposito del messaggio che il testo stesso vuole trasmettere.

Illustrazione di Aria Querques

Ma come funziona, in pratica, un *software* per l'analisi testuale? Premettiamo che, per uno qualunque di questi programmi, il testo da analizzare non è altro che una lunga sequenza di parole, ciascuna delle quali non ha un senso in sé, ma acquista un significato in base alle sue relazioni con le altre parole, cioè in base alla distribuzione delle sue occorrenze all'interno del *corpus* in esame. Di conseguenza, quando diamo in pasto a uno di questi *software* un testo più o meno lungo, questo viene innanzi tutto trasformato in un *data base* che codifica in maniera opportuna tutte le parole, contando quante volte e ricordando in che punto ognuna di loro compare all'interno del testo stesso.

T-Lab, per esempio, costruisce in prima battuta due elenchi: il primo è costituito da tutte le parole che compaiono nel testo (escluse congiunzioni, articoli, ecc...) e il secondo da sottoinsiemi del testo, detti *contesti elementari*, che possono essere singole frasi, paragrafi o frammenti più o meno lunghi, determinati in base a criteri stabiliti dall'utilizzatore. Disponendo poi dei due elenchi, $(p_1, p_2, \dots p_n)$ e $(c_1, c_2, \dots c_m)$, e di una serie di tabelle allegate, che racchiudono tutte le informazioni rilevanti relative all'organizzazione del testo, T-Lab è in grado di produrre, su richiesta dell'utente, matrici, elenchi, istogrammi, diagrammi a torta, grafici e mappe di vario genere.

L'oggetto più immediato da descrivere è una grande matrice rettangolare, con n righe e m colonne, al cui posto ij T-Lab scrive il numero di volte che la parola p_i compare nel sottoinsieme c_j. In questo modo, ogni riga fornisce la distribuzione della parola corrispondente all'interno dei contesti in cui è frammentato il testo, mentre ogni colonna consente di stabilire quali parole compaiono, e quali no, nel corrispondente contesto elementare. Inoltre, dato l'elenco $(p_1, p_2, \dots p_n)$ delle parole, possiamo costruire un'altra matrice, detta matrice delle co-occorrenze, questa volta quadrata di ordine n, al cui posto ij troveremo il numero di contesti elementari in cui le parole p_i e p_j compaiono insieme, ovvero il valore di co-occorrenza delle parole p_i e p_j nel testo.

Supponiamo per esempio di avere un elenco di 6 parole, "BERLUSCONI", "GELMINI", "SCUOLA_PUBBLICA", "MANIFESTAZIONE", "INCULCARE" e "INSEGNANTI", considerate nell'ordine dato. Supponiamo poi di selezionare, tra tutti quelli scaricati, 10 articoli di uno dei due quotidiani, ciascuno dei quali costituisce un contesto elementare ed è quindi etichettato con un numero da 1 a 10. La prima matrice, quindi, sarà una matrice con 6 righe e 10 colonne, come la seguente:

1	2	0	0	0	2	1	0	0	0
0	3	0	0	0	0	0	2	1	0
1	1	2	0	0	0	0	3	0	1
0	1	1	3	0	1	0	6	0	0
2	1	0	2	0	8	0	0	0	1
1	3	2	0	1	0	0	2	8	1

La prima riga fornisce la distribuzione della prima parola, Berlusconi, all'interno degli articoli selezionati: precisamente, Berlusconi compare una volta nel primo e nel settimo articolo, due volte nel secondo e nel sesto e mai nei restanti. La prima colonna, invece, fornisce l'elenco delle parole che compaiono nel primo articolo: vi si menzionano, dunque, Berlusconi, la scuola pubblica, inculcare e gli insegnanti, mentre non vi compare nessuna delle altre parole in esame. Analogamente possiamo interpretare le altre righe e colonne. Con lo stesso elenco di parole, e con i dati riassunti nella prima matrice, potete divertirvi a costruire da soli la matrice quadrata delle co-occorrenze, che sarà 6x6 e diagonale (perché?):

4	1	2	2	3	2
1	3	2	2	1	3
2	2	5	3	3	5
2	2	3	5	3	3
3	1	3	3	5	3
2	3	5	3	3	7

I numeri che appaiono sulla diagonale corrispondono al numero di articoli in cui appare la parola p_i. Tenete presente che l'esempio appena fatto è puramente teorico: nel *corpus* che vogliamo analizzare, infatti, T-Lab individua più di 4000 parole e più di 600 contesti elementari: le matrici corrispondenti, quindi, sono un po' più grandine...

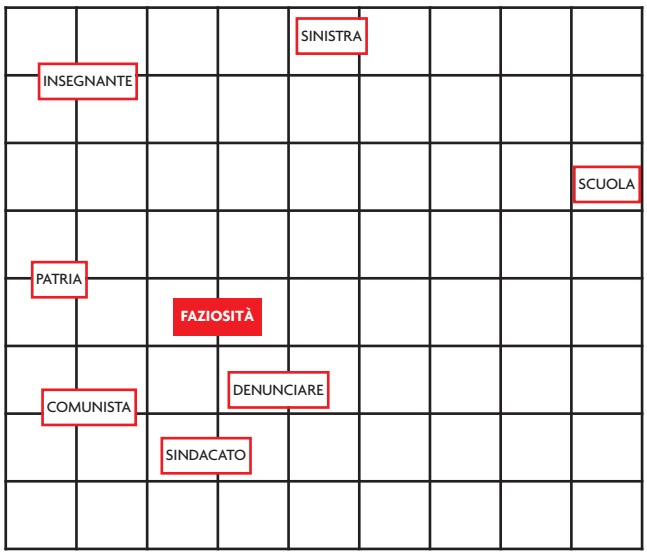
Ma veniamo al nostro problema specifico: abbiamo una raccolta di articoli provenienti da due quotidiani, selezionati per data e argomento, e vogliamo cercare di capire, utilizzando T-Lab, come i due giornali riportano lo stesso fatto di cronaca e come danno conto delle reazioni conseguenti. Ci interessa, cioè, effettuare una sorta di comparazione tra i due quotidiani, mirata a evidenziare la linea di pensiero di ciascuno.

Illustrazione di Aria Querques

Volendo confrontare tra loro due (o più) testi, la prima domanda da porsi è se esistano, e in caso quali siano, i lemmi che li caratterizzano: nel software T-Lab risponde a tale quesito la funzione *specificità*, che individua le parole "tipiche" e quelle "esclusive" di ogni testo. Le prime sono quelle che compaiono in entrambi i testi ma con frequenza significativamente maggiore in uno dei due. dove l'avverbio "significativamente" ha un significato ben preciso in termini di *chi quadro*, un indice statistico che serve per testare se le frequenze osservate coincidono o meno con quelle "attese" (se ne è parlato nel n. 20 di *XlaTangente*). Le seconde, invece, sono le parole che compaiono esclusivamente in uno dei testi, con maggiore o minore frequenza: nel nostro caso, per esempio, "travisare" e "faziosità" sono specificità esclusive di uno dei due quotidiani analizzati (quale?), mentre l'altro ha come specificità esclusive "corteo" e "mobilitazione". "Strumentalizzare" e "indottrinamento" sono invece parole tipiche del primo, mentre "tagli", "precario" e anche "scuola pubblica" lo sono del secondo. Il che, in effetti, ci dice già qualcosa, sia sull'interpretazione dell'episodio di apertura data dai due quotidiani che sulla copertura mediatica e l'interpretazione politica della manifestazione conseguente.

Tramite le analisi descritte fino a questo momento, riusciamo a farci un'idea di quali siano le *parole chiave*, presenti in entrambi i documenti o tipiche/esclusive dell'uno o dell'altro. Lo strumento dell'*associazione di parole* consente adesso di capire meglio come si sviluppa il discorso nei due contesti diversi. Per ogni parola che reputiamo significativa, infatti, il suo grafico delle associazioni permette di visualizzare tutte le parole che le sono più frequentemente associate all'interno del testo selezionato.

Consideriamo ad esempio la parola "faziosità", che come si diceva è esclusiva degli articoli di uno dei due quotidiani. La funzione "associazioni di parole", applicata a questi articoli, produce un grafico (vedi figura seguente), in cui il lemma selezionato ("faziosità", appunto) è evidenziato in rosso e gli altri lemmi, in nero, sono distribuiti attorno, a distanza inversamente proporzionale al grado di associazione di ognuno con la parola "faziosità".



Facendo doppio click su ogni lemma nero che compare nel grafico, si apre una finestra che elenca tutti i paragrafi in cui quel lemma compare associato a quello rosso (le cosiddette co-occorrenze): in altri termini, con un semplice click sui box relativi possiamo verificare se l'interpretazione "denunciare la faziosità della sinistra comunista" sia quella corretta o sia solo una interpretazione maligna e, appunto, faziosa. La distanza tra due lemmi neri, invece, non ha alcun significato: le relazioni significative, cioè, sono del tipo uno-a-uno, tra il lemma rosso e ciascuno degli altri. In altri termini, il grafico non significa affatto che negli articoli del quotidiano in esame ci sia un accostamento insistente tra insegnante e sinistra, né tra patria e comunista.
Per poter costruire grafici di questo tipo, T-Lab deve stabilire, dato un qualunque lemma presente nel testo (chiamiamolo lemma B), a quale distanza collocarlo rispetto al lemma rosso (lemma A). Tale distanza si calcola attraverso il cosiddetto *coefficiente del coseno*, la cui formula è la seguente:

$$\frac{\langle occ(A), occ(B)\rangle}{\|occ(A)\| \cdot \|occ(B)\|}$$ [4]

dove con *occ*(X) indichiamo il vettore delle occorrenze del lemma X nei contesti elementari del testo in esame e ‹ , › e || || rappresentano rispettivamente il prodotto scalare e la norma euclidea. Il coefficiente del coseno è quindi un numero compreso tra 0 e 1, che è nullo se e solo se i termini non compaiono *mai* insieme nel testo, ed è esattamente uno se e solo se le due parole compaiono *sempre* insieme. Fornisce, quindi, una misura della "vicinanza" tra il lemma B e il lemma A in tutto il testo in esame, e la distanza tra il lemma A e il lemma B nel grafico è sostanzialmente il suo reciproco: sono infinitamente lontani da A (e dunque non compaiono nel grafico) i termini B con coefficiente del coseno 0, cioè quelli che non compaiono mai con A, e sono progressivamente più vicini quelli il cui coefficiente si avvicina a 1, cioè quelli che, quando compaiono, compaiono spesso insieme ad A.
Se, ad esempio, presi in esame 10 contesti elementari, il lemma A compare nei primi cinque e nell'ultimo, mentre il lemma B compare nel primo, sesto e settimo, i due vettori delle occorrenze sono

$$occ(A) = [1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1]$$

e

$$occ(B) = [1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0]$$

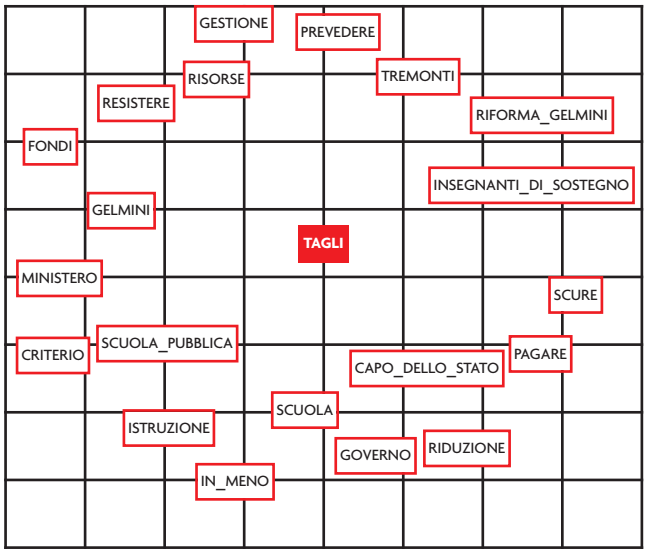
e il coefficiente del coseno è quindi

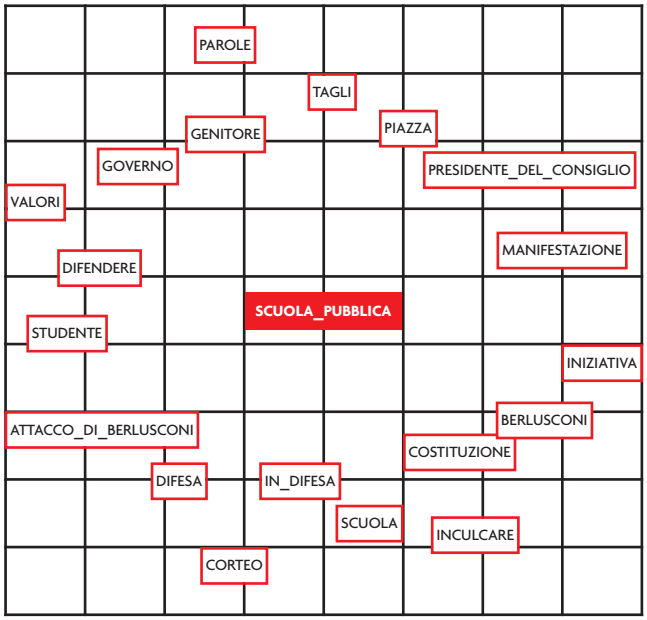
$$\frac{1\cdot1+1\cdot0+1\cdot0+1\cdot0+1\cdot0+0\cdot1+0\cdot1+0\cdot0+0\cdot0+1\cdot0}{\sqrt{1+1+1+1+1+1}\cdot\sqrt{1+1+1}} =$$

$$= \frac{1}{\sqrt{18}} \approx 0{,}238.$$

Otteniamo un valore basso che corrisponde al fatto che i due termini compaiono relativamente poco assieme (solo in 1 contesto su 10), mentre la prima parola appare spesso ma da sola (6 contesti su 10). Nel grafico quindi le due parole sono distanti.

Vediamo adesso i grafici relativi a due delle parole tipiche dell'altro quotidiano, "tagli" e "scuola_pubblica":





Certo, di fronte a grafici come questi la tentazione di fare inferenza è molto forte: guardandoli, in effetti, l'argomentare del quotidiano sembra delinearsi in modo inequivocabile. Tuttavia un buon giro di click sulle varie caselle è assolutamente indispensabile al fine di verificare l'ipotesi interpretativa formulata. Guardare solo il grafico, infatti, può portare a conclusioni paradossali: nel nostro caso, per esempio, in modo del tutto inaspettato, i grafici centrati sulla stringa "scuola pubblica" sono molto simili, nel primo e nel secondo quotidiano. Possiamo concluderne che *il Giornale* e *la Repubblica* ce la raccontano, sulla scuola pubblica, sostanzialmente allo stesso modo? Possibile?? Su un tema tanto caldo??? Beh, sarebbe un discreto *scoop*, per un bimestrale di matematica come *XLaTangente*... ma non è così, e il *software* ce lo dimostra, se lo utilizziamo nel modo corretto. A ben vedere, cari lettori, vi abbiamo già fornito un buon indizio per spiegare l'arcano, qualche riga più sopra: avete già indovinato quale? E voi, Ezio Mauro e Alessandro Sallusti, voi che siete in fondo i più coinvolti in questa faccenda, ve la sentite di azzardare una spiegazione? Come in ogni mistero che si rispetti, la risposta è... nel prossimo numero!

Illustrazione di Aria Querques

Note

1. **Alceste è stato messo a punto dal CNRS, il CNR francese.**
2. **DBT, acronimo di Data Base Testuale, è un sistema di trattamento informatico di dati testuali sviluppato da Eugenio Picchi presso l'Istituto di Linguistica Computazionale del CNR di Pisa.**
3. **T-Lab è stato sviluppato da Franco Lancia, http://www.tlab.it/it/presentation.php**
4. **Il prodotto scalare è una maniera di associare a due vettori v e w un numero $\langle v,w\rangle$ con la caratteristica che $\|v\| = \sqrt{\langle v,v\rangle}$ rappresenta la lunghezza del vettore v e $\frac{\langle v,w\rangle}{\|v\| \cdot \|w\|}$ rappresenta il coseno dell'angolo tra v e w.**
   In coordinate, se $v = (v_1, \dots, v_n)$ e $w = (w_1, \dots, w_n)$ sono vettori di $\mathbb{R}^n$, allora $\langle v,w\rangle = v_1w_1 + \dots + v_nw_n$ e $\|v\| = \sqrt{v_1^2 + \dots + v_n^2}$.

Silvia Benvenuti
Dopo la laurea e il dottorato di ricerca in Matematica conseguiti all'Università di Pisa, ha frequentato il Master in Comunicazione della scienza della SISSA di Trieste. Attualmente è ricercatrice in geometria presso l'Università di Camerino. Il suo campo di ricerca è la topologia in dimensione bassa: teoria dei nodi, delle superfici e delle 3-varietà.
Ha un'esperienza didattica pluriennale a livello universitario e collabora con diverse case editrici alla stesura di testi per le scuole superiori e l'università. È autrice di un libro sulle geometrie non euclidee edito da Alphatest nella collana Gli Spilli e del libro *Insalate di matematica 3. Sette variazioni su arte, design e architettura*, edito da Sironi.