ANALISI LESSICALE DEI TESTI

Carlo Poli e Giorgio Carboni, Aprile 1998

I N D I C E

PRESENTAZIONE

Questa volta ci occupiamo di un argomento molto diverso da quelli consueti per Fun Science Gallery: l'analisi dei testi. Devo confessare che la linguistica non è un campo nel quale mi muovo agevolmente. Allora vi chiederete perché mi ci sono avventurato. Infatti c'è un motivo che spiega questo imprudente sconfinamento. A parte il naturale interesse per il linguaggio che abbiamo più o meno tutti, almeno a livello di curiosità, quello che mi ha spinto in questa direzione è stato un motivo professionale. Infatti, per lavoro ho dovuto compiere numerose ricerche bibliografiche.

Il problema principale nell'interrogare le banche dati sta nel mettere a punto una domanda tale da ottenere i documenti che si cercano e solo quelli. Quello che invece avviene normalmente è di rimanere sepolti sotto una valanga di articoli che hanno poca o nessuna attinenza con quanto cercato. Se non abbiamo sbagliato proprio tutto, in mezzo a questi articoli, possono esserci anche quelli che cerchiamo, ma individuarli manualmente è praticamente impossibile. Trovare le parole chiave adatte a ottenere soltanto i documenti che ci interessano fra i milioni che sono conservati dalle banche dati è dunque un'impresa tutt'altro che facile! Per avere successo in queste ricerche, occorre avere preparato una domanda efficace. Anche le ricerche su Internet non sono facili da compiere e, con la rapida crescita del cyberspazio, questo problema probabilmente si aggraverà.

E' stato dunque nel corso di queste ricerche bibliografiche che ho sentito la necessità di programmi che mi aiutassero ad individuare le parole chiave più adatte. Normalmente, quando ci si appresta a compiere una ricerca bibliografica, non si parte proprio dal nulla, ma si possiedono alcuni articoli, frutto di una ricerca compiuta con metodi manuali presso la propria biblioteca, oppure ottenuti da colleghi. Questi articoli possono essere il punto di partenza per individuare parole chiave. Per potere fare questo con l'aiuto di programmi, è necessario acquisire il loro testo in formato elettronico, per esempio per mezzo di uno scanner. Come fare per individuare le parole chiave per mezzo di programmi informatici? Generalmente, una **parola chiave** è un termine che ha una frequenza elevata nel testo in esame e bassa nella lingua comune. Disponendo della frequenza dei termini della lingua comune, per un computer è facile estrarre i termini specialistici di un testo. Un altro sistema per isolare parole chiave è quello di individuare le **locuzioni ripetute**. Infatti, molti termini specialistici sono formati da più parole consecutive, o locuzioni. Isolare le locuzioni e mettere in evidenza quelle maggiormente ripetute ci può fornire importanti parole chiave per le nostre ricerche bibliografiche. Con questi metodi automatici vengono a galla termini ai quali non avevamo pensato e che possono contribuire parecchio al successo delle nostre interrogazioni.

E' stato in base a queste considerazioni che ho parlato con un mio collega, esperto di informatica, proponendogli di realizzare i programmi necessari. Abbiamo lavorato insieme per oltre due mesi, ma alla fine abbiamo ottenuto programmi semplici e nello stesso tempo potenti. Infatti, abbiamo realizzato questi programmi in ambiente ACCESS, sfruttandone la potenza delle routines di gestione dei database (archivi elettronici) e di gestione grafica delle finestre. Access è il "modulo" di gestione dei database presente nel pacchetto Office della Microsoft. Probabilmente è possibile realizzare programmi di questo tipo anche con altri applicativi.

Ho parlato di queste attività con un'insegnante di un Liceo di Bologna. Questa insegnante ha trovato interessante il lavoro e si è convenuto di utilizzarlo per compiere un'analisi di testi di letteratura, da presentare durante la Settimana della Cultura Scientifica e tecnologica (23-29 marzo 1998) promossa dal Ministero dell'Università e della Ricerca Scientifica e Tecnologica (MURST). Questo impegno ci ha spronato a migliorare i programmi e soprattutto ad adattarli per un uso un po' più generale in campo linguistico e non soltanto per preparare ricerche bibliografiche. L'interazione con gli allievi e i loro insegnanti è stata utile per collaudare i programmi e per definirli meglio. Il risultato è dunque un insieme di programmi che possono essere utilizzati per numerose operazioni di analisi e confronti di testi e per ricerche linguistiche oltre che per la messa a punto di ricerche bibliografiche.

ATTENZIONE: potete scaricare il programma necessario per queste analisi lessicali cliccando qua: **Analisi.zip**.

Ma di che programmi si tratta e che cosa fanno esattamente questi programmi? Ecco una loro brevissima descrizione:

Normalizza	inserisce un file di testo in una tabella
Frequenze	calcola la ricorrenza e la frequenza dei termini di un testo
A - B	ricava i termini del testo A non presenti nel testo B
A x B	ricava i termini comuni ai testi A e B
A + B	somma i termini dei testi A e B
A => Thesaurus	aggiunge i termini del testo A al Thesaurus indicato
A <= Ripristina	toglie i termini del testo A al Thesaurus indicato
Testo - Gramma	toglie a un testo i termini grammaticali
Locuzioni	associa le parole di un testo in sequenze
Leggibilità	determina la leggibilità di un testo
Periodi	produce una tabella di periodi
Paragrafi	produce una tabella di paragrafi
Punteggiatura	produce una tabella dei segni di punteggiatura

E' probabile che prima di passare alla lettura del manuale vogliate avere un'idea di quello che si può fare con questi programmi. Come abbiamo visto con il liceo di cui vi abbiamo parlato, questi programmi possono essere utilizzati in maniere diverse, a cui non avevamo pensato neppure noi che li abbiamo realizzati. In ogni caso, ecco alcuni esempi di possibili utilizzazioni:

- determinare la frequenza dei termini di testi di letteratura o di un proprio componimento
- confrontare la frequenza dei termini di testi di autori diversi allo scopo di analizzarne lo stile
- confrontare la ricchezza lessicale, intesa come la quantità di termini impiegati in testi della stessa lunghezza
- confrontando due testi, si possono isolare i termini usati solo nel primo (o solo nel secondo):
    - se si tratta di due testi di letteratura di cui uno antico e l'altro moderno, si possono isolare i termini antichi, oppure quelli moderni, infine quelli comuni che si sono conservati nel tempo
    - se si tratta di un testo scientifico e di uno letterario, si possono isolare i termini scientifici (che possono poi essere usati per ricerche bibliografiche)
- confrontando più documenti, si possono comporre dizionari dei termini:
    - grammaticali
    - più comuni
    - colti
    - specialistici (per i vari argomenti)
- sommando più testi, si possono ottenere dizionari:
    - della frequenza dei termini della letteratura dell'800, etc
    - della frequenza dei termini della lingua italiana
- si possono isolare le locuzioni ripetute in un testo:
    - esaminando un testo letterario si possono mettere in evidenza le locuzioni grammaticali, i modi di dire, le forme idiomatiche, infine i luoghi comuni
    - esaminando un testo specialistico si possono ricavare i termini specialistici composti, molto utili per le ricerche bibliografiche
- un successivo gruppo di programmi permette di determinare la leggibilità dei testi, di isolare i periodi più lunghi, di isolare i paragrafi più lunghi. Ciò facilita la loro verifica e la loro eventuale sistemazione. Infine un ultimo programma permette di ottenere dati statistici sulla punteggiatura.

Mi auguro che questo nostro lavoro possa rivelarsi utile a qualcuno, per esempio a scuole e ad amanti di studi linguistici.
(g.c.)

REQUISITI E LIMITAZIONI

REQUISITI
I programmi che vi abbiamo fornito, richiedono l'applicativo di gestione dei database ACCESS, nella versione che fa parte di Microsoft Office 97 o successive.

LIMITAZIONI DI LINGUA
Con questi programmi, potete trattare testi scritti in caratteri latini, più precisamente, con i primi 256 caratteri ASCII. Quindi, potete esaminare testi in lingua italiana, inglese, tedesca, spagnola, francese, etc. Tuttavia, le nostre prove si sono limitate a testi in **lingua italiana** e in **lingua inglese**. Se volete utilizzare questi programmi per analizzare testi in lingua inglese, servitevi del database che trovate nella versione inglese di questo articolo. Se vorrete utilizzare questi programmi per altre lingue, per ciascuna di esse dovrete realizzare un nuovo database. A tale scopo, vi conviene fare una copia del nostro database ed adattarlo alla nuova lingua. Il codice dei programmi è accessibile e potete modificarlo.

NOZIONI PRELIMINARI
Per l'uso più efficace di questi programmi, **è necessario avere una conoscenza minima di Access**:

gestione finestre: - apertura, chiusura, spostamento, ridimensionamento delle finestre (come in Windows)
gestione tabelle: - copia, spostamento, cancellazione di tabelle (simile alla gestione dei files); - creazione di una tabella; - ordinamento dei dati; - immissione di dati; - cancellazione di record e di colonne; - visualizzazione della struttura delle tabelle ed

esecuzione di modifiche
altro: - esecuzione di una query e trasformazione in tabella; - ricavo di diagrammi.

PREPARAZIONE DEL COMPUTER

Installate il programma:	Access (Office 97 o versione successiva)	
Scaricate il file:	**Analisi.zip**	
Create le directory:	C:\Lessico\	
	C:\Lessico\Manuale	(per tenere questa pagina)
	C:\Lessico\Testi	(per i testi da analizzare)
	C:\Lessico\Varie	(per le vostre relazioni)
Decomprimete il file:	Analisi.zip	
Inserite Analisi.mdb in:	C:\Lessico\	
Seguendo le nostre **istruzioni**, scaricate questa pagina in: (I.Explorer creerà una sottocartella per le immagini.)	C:\Lessico\Manuale\	(per tenere una copia del nostro manuale)
Inserite i testi originali (.zip .doc .html .txt) in:	C:\Lessico\Testi\	

NOMI DEI FILES

Nome.zip	testo compresso
Nome.doc	testo in formato Word
Nome.html	testo in formato ipertestuale (HTML)
Nome.txt	testo in formato "testo", completo di punteggiatura
Analisi.mdb	database. Contiene tutte le tabelle ed i programmi

NOMI DELLE TABELLE

Nome	tabella di tutte le parole
Nome_freq	tabella delle frequenze
NomeA-NomeB_freq	tabella dei termini presenti in A, ma non in B (frequenze)
NomeAxNomeB_rapp	tabella dei termini comuni fra A e B (rapporti di frequenza)
NomeA+NomeB_freq	tabella somma dei termini di A e di B (frequenze)
Gramma	tabella di termini grammaticali, verbi ausiliari, etc
Nome-G	tabella sottratta dei termini presenti in Gramma
Nome_n	tabella con le parole prese in sequenze di n parole
Th_Nome-freq	thesaurus di termini e loro frequenze
Th_Nome_lista	elenco dei testi inseriti nel thesaurus (titoli)
Statistica	tabella prodotta da Leggibilità. Contiene il nome dei documenti esaminati e i dati ricavati
ZZ_...	tabelle di sistema (modelli di tabelle, elenchi, etc)

TIPI DI TABELLE

Possiamo distinguere le tabelle in:

tabelle semplici
tabelle di frequenze
altre tabelle

TABELLE SEMPLICI

Nome
Nome-G
Nome_n

Nome_peri
Nome_para
Nome_punti
Gramma

Per tabelle semplici intendiamo quelle che hanno i campi: ID, Parole, Campo1. Sono tabelle semplici, quelle che derivano dai programmi Normalizza, Testo-Gramma, Locuzioni, Periodi, Paragrafi, Punteggiatura. Queste tabelle sono adatte a essere trattate con il programma Frequenze. Le tabelle prodotte da Normalizza sono adatte ad essere trattate anche con Testo-Gramma.

TABELLE DI FREQUENZE

Nome_freq
Nome_rapp
Th_Nome_freq

Le tabelle di frequenze possono avere più campi (o colonne), ma fra di essi devono esserci i seguenti: Termini, Ricorrenze, Frequenze. Sono tabelle di frequenze quelle che derivano dai programmi: Frequenze, A-B, AxB, A+B, A=>Thesaurus. Sono adatte ad essere trattate dai programmi: A-B, AxB, A+B, A=>Thesaurus.

ALTRE TABELLE
Le tabelle Statistica e Th_Nome_Lista contengono dati riassuntivi. Le tabelle del tipo "ZZ_" sono di sistema. Queste tabelle non sono adatte ad essere trattate dai programmi come lo sono le tabelle semplici o quelle di frequenze.

AVVERTENZE
Non cancellate mai le tabelle Gramma, Statistica e tutte quelle del tipo ZZ_...
Potete modificare le dimensioni delle colonne o delle righe a tutte le tabelle.
Potete cancellare o aggiungere record a tutte le tabelle ad esclusione di quelle tipo ZZ_...

NOMI DEI PROGRAMMI

Normalizza	normalizza un file di testo e lo inserisce in una tabella (le parole mantengono l'ordine originale)
Frequenze	calcola la ricorrenza e la frequenza dei termini di una tabella
A - B	ricava i termini della tabella A non presenti nella tabella B
A x B	ricava i termini comuni delle tabelle A e B e ne calcola il rapporto delle frequenze
A + B	somma i termini delle tabelle A e B e ne ricalcola le frequenze
A => Thesaurus	aggiunge i termini della tabella A al Thesaurus indicato, ricalcola le frequenze e aggiunge il nome della tabella alla lista
A <= Ripristina	toglie i termini della tabella A dal Thesaurus indicato e corregge la lista (da usare in caso di errore)
Testo - Gramma	toglie i termini presenti in Gramma da una tabella di testo normalizzato
Locuzioni	associa le parole di un testo in sequenze
Leggibilità	determina la leggibilità di un testo
Periodi	produce una tabella di periodi
Paragrafi	produce una tabella di paragrafi
Punteggiatura	produce una tabella dei segni di punteggiatura

TERMINOLOGIA

Parole	parole del testo, anche ripetute
Termini	parole prese una volta sola
Tp	totale delle parole
Tt	totale dei termini
Ric	ricorrenza: numero di volte che un termine è presente nel testo
Freq	frequenza: rapporto fra la ricorrenza e il totale delle parole. Freq = Ric/Tp
Rapp	Rapporto di frequenza che lo stesso termine ha in due testi diversi. Rapp = Freq(A)/Freq(B)
Thesaurus	elenco di termini. Nel nostro caso, i thesauri contengono anche le frequenze dei termini
Maschera	finestra di dialogo di un programma
Testi/Tabelle	poichè abbiamo inserito i testi in tabelle, fare operazioni sulle tabelle significa farle sui testi

RICERCA DEI TESTI

Procuratevi alcuni testi in formato elettronico. Potete ottenerli nei seguenti modi:
- cercandoli in rete
- acquisendoli con lo scanner + OCR
- inserendoli nel computer tramite tastiera

SITI DAI QUALI SCARICARE TESTI O ARTICOLI

ATHENA	Testi in formato elettronico da tutto il mondo anche in lingua italiana (cliccare su "Books") (http://un2sg4.unige.ch/athena/html/athome.html) (Univ. di Ginevra)
Biblioteche elettroniche	Guida a cura dell'Univ. di Bologna (http://www.economia.unibo.it/dipartim/stoant/rassegna1/bibliot.html)
Project Gutenberg	Progetto Gutenberg (cliccare su "Etext Listings") (http://www.promo.net/pg/)
Progetto Manuzio	Testo di libri italiani e latini (http://www.liberliber.it/)
Antologia della letteratura italiana	Testo di libri italiani (http://www.crs4.it/HTML/Literature.html)
Letteratura italiana	Letteratura italiana (http://wwwmedea.clio.it/scuola/letter2.htm)
Libri e giornali	Libri e giornali inglesi (via Gopher) (gopher://gopher.tc.umn.edu:70/11/Libraries/)
OUSIA	Di tutto per il Bibliofilo (http://www.freshnet.de/user/will/ousia/Bibliophile.html)
NlightN service	Ricerche su oltre 300 DB di libri, riviste, articoli (anche www) (http://www.nln.com/)
News and Media	Giornali e riviste da tutto il mondo (http://www.yahoo.com/News/)
News and Media, Indices	Indici di link a giornali e riviste da tutto il mondo (http://www.yahoo.com/News_and_Media/Indices/)
Yahoo Electronic Literature, Indices	Indici Yahoo di libri in formato elettronico (http://www.yahoo.com/Arts/Humanities/Literature/Electronic_Literature/Indices/)

Allo scopo di imparare ad usare questi programmi, procuratevi testi come i seguenti:

- due testi di letteratura contemporanea
- un testo di letteratura antica
- un testo scientifico o tecnico
- un vostro componimento (almeno 2000 parole)

Nelle nostre prove, abbiamo impiegato questi testi:

I. Calvino	*Il cavaliere inesistente*	1959
F. Tozzi	*Bestie*	1913
I. Svevo	*Senilità*	1898
A. Manzoni	*I promessi sposi*	1827
D. Compagni	*Cronica delle cose occorrenti ne' tempi suoi*	1312
FSG	*Microscopio a sfera di vetro*	1996

Nel passaggio da un formato all'altro, le **lettere accentate** di questi testi possono risultare alterate. Verificate che appaiano correttamente ed eventualmente correggetele. Per esempio: "&egrave" sostituitelo con "è".

Salvate i documenti in formato testo: Nome.txt
I documenti che troverete in rete saranno probabilmente in formato HTML. Al proprio interno hanno numerosi Tag (comandi come i seguenti:). Per eliminarli è sufficiente aprire il documento con un browser e salvarlo in formato testo.

AVVIO DEL DATABASE

Un database (DB) è un insieme di tabelle, query, programmi, etc. Quello che abbiamo preparato contiene già i programmi necessari. Per avviare il DB fate come segue:

Da Risorse del computer	
portatevi in:	C:\Lessico\
Cliccate su:	Analisi.mdb

NORMALIZZAZIONE

Lo scopo della normalizzazione è quello di inserire le parole di un testo in una tabella, in modo che esso possa venire esaminato dai programmi che abbiamo preparato. Per fare questo, ci serviamo del programma **Normalizza**. Il documento da sottoporre a questo programma deve essere in formato testo. Normalizza elimina tutti i caratteri diversi da quelli alfabetici (punteggiatura, simboli matematici, etc), trasforma le maiuscole in minuscole. Per evitare che una stessa parola venga distinta in base al tipo di accento, se non altrimenti indicato, Normalizza converte le lettere con accento acuto in gravi e trasforma la î in i normale. Come già detto, alla fine il programma crea una tabella di tutte le parole del testo trattato. In questa tabella, le parole mantengono l'ordine di lettura (figura 2).

Verificate che le accentate appaiano in modo corretto nel testo da trattare. Se necessario correggete e salvate il file.

Con Access, aprite **Analisi.mdb**
andate alla cartella **Maschere**
avviate il programma **Normalizza**
scegliete il **file di testo** da normalizzare
cliccate sul pulsante **Avvia**
(il programma creerà una tabella contenente tutte le parole del testo, mantenendone l'ordine originario. Il nome di questa tabella non avrà suffissi)

promessi : Tabella

	ID	Parole	Campo1
	1	quel	
	2	ramo	
	3	del	
	4	lago	
	5	di	
	6	como	
	7	che	
	8	volge	
	9	a	
	10	mezzogiorno	

Fig. 2 - Prime parole della tabella normalizzata ricavata da "I promessi sposi". Si noti come le parole mantengano l'ordine di lettura.

Chiudete il programma Normalizza
Aprite la tabella per verificarla
ordinatela in senso alfabetico
eliminate gli eventuali record vuoti
riordinate la tabella secondo l'ordine di lettura (col. ID)
salvate e chiudete la tabella

CALCOLO DELLE RICORRENZE E DELLE FREQUENZE

Applicato a una tabella di testo normalizzato, il programma Frequenze determina quante volte ogni termine ricorre nel testo e ne calcola la frequenza (figure 3 e 4).

Lanciate il programma **Frequenze**
Scegliete la tabella da trattare
Cliccate su Avvia
(Il programma genererà una tabella con suffisso _freq)

senilita_freq : Tabella

	Termini	Ricorrenze	Frequenze
	a	892	1.34E-02
	abbacinare	1	1.50E-05
	abbacinato	1	1.50E-05
	abbadò	1	1.50E-05
	abbandonare	7	1.05E-04
	abbandonarglisi	1	1.50E-05
	abbandonarla	1	1.50E-05
	abbandonarlo	1	1.50E-05
	abbandonarono	1	1.50E-05
	abbandonarsi	2	3.00E-05

Fig. 3 - Tabella delle frequenze dei termini di "Senilità" (ordine alfabetico).

senilita_freq : Tabella

	Termini	Ricorrenze	Frequenze
	di	2088	3.13E-02
	e	1507	2.26E-02
	che	1502	2.25E-02
	la	1408	2.11E-02
	non	1391	2.08E-02
	il	1257	1.88E-02
	a	892	1.34E-02
	era	878	1.32E-02
	un	868	1.30E-02
	per	817	1.22E-02

Fig. 4 - Tabella delle frequenze dei termini di "Senilità" (ordine di frequenza).

DETERMINAZIONI: Determinate la ricorrenza e la frequenza dei termini di alcuni fra i seguenti documenti:

- un vostro componimento
- un testo di letteratura
- un testo scientifico o tecnico
- Promessi sposi

ESAME DEI DATI OTTENUTI: Ordinando la tabella in senso alfabetico, potete osservare quali termini sono stati utilizzati e le relative forme flesse. Potete verificare se sono stati utilizzati sinonimi.

Le tabelle di frequenze hanno la colonna dei "Termini", "Ricorrenze" (numero di volte che il termine figura nel testo), "Frequenze" (la frequenza di un termine è calcolata come segue: Freq = Ric / Tp. Il valore ottenuto viene scritto in forma esponenziale in base 10. Per esempio, il termine "a" della Figura 3, ha una frequenza di 1,34 x 10^-2. Questo valore può essere scritto anche 0,0134).

Ordinando la tabella in base alla frequenza, potete vedere quali siano i termini più usati e quelli invece usati una volta sola. Prendete nota dei termini più frequenti e ricavatene delle vostre osservazioni. Fate altrettanto con i termini meno usati. Nell'esaminare le frequenze dei termini, è utile distinguere fra quelli grammaticali (articoli, preposizioni, pronomi, etc.) e quelli di contenuto (nomi, aggettivi, verbi).

Se vi interessa, potete determinare la ricchezza lessicale (RL), intesa come rapporto fra il numero dei termini (tt) ed il numero delle parole (tp):

$$RL = tt / tp$$

Il valore di questo rapporto va da zero a 1. Tenete presente che RL tende a diminuire con l'aumentare della dimensione del testo, per cui vanno confrontati testi di dimensione analoga, per esempio 2000 parole. Se un testo è più lungo, ricavatene una parte di 2000 parole per effettuare questa determinazione.

OPERAZIONI LOGICHE

Per mezzo di operazioni logiche compiute fra due testi (figura 5), possiamo ottenere:

- i termini presenti solo nel testo A **A - B**
- i termini presenti solo nel testo B **B - A**
- i termini comuni ad A e B **A x B**
- somma dei termini presenti in A e in B **A + B**

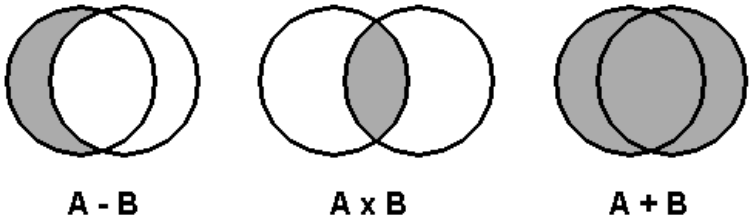

Fig. 5 - Le tre operazioni logiche, o booleane.

TERMINI PRESENTI SOLO IN UN DOCUMENTO FRA DUE (A - B)

L'esame dei termini presenti solo in un documento e non in un altro permette di ottenere interessanti indicazioni.
Se si tratta di due documenti di cui uno recente e l'altro antico, si possono evidenziare i termini moderni oppure quelli antichi.
Se si tratta di due documenti di cui uno letterario e l'altro scientifico o tecnico, si possono evidenziare i termini specialistici. Comunque, anche un testo letterario può possedere termini propri e caratteristici (figura 6).

Cavaliere-promessi_freq : Tabella

	Termini	Ricorrenze	Frequenze
	agilulfo	170	5.09E-03
	gurdulù	75	2.24E-03
	torrismondo	54	1. 2E-03
	carlomagno	38	1.14E-03
	bradamante	36	1.08E-03
	sofronia	33	9.87E-04
	priscilla	32	9.57E-04
	paladini	31	9.27E-04
	gral	28	8.38E-04
	elmo	25	7.48E-04
	scudiero	24	7.18E-04

Fig. 6 - Termini del "Cavaliere inesistente" non comuni con "I promessi sposi".

Avviate il programma **A-B**
Scegliete la tabella A (deve essere una tabella di frequenze)
Scegliete la tabella B (deve essere una tabella di frequenze)
Premete il tasto Avvia
(il programma realizzerà una nuova tabella di nome: NomeA-NomeB_freq)

La tabella realizzata dal programma A-B contiene i termini presenti soltanto in A. I valori delle ricorrenze e delle frequenze, sono gli stessi della tabella A.
Questo programma può essere utilizzato anche per sottrarre i termini di un thesaurus da una tabella di frequenze, come quando si vogliono isolare i termini antichi o quelli specialistici di un documento (figure 7 e 8).

DETERMINAZIONI:
Confrontate due documenti letterari di cui uno antico e l'altro moderno (figura 7).
Confrontate due documenti di cui uno scientifico o tecnico e l'altro di letteratura (figura 8).

cronica-Th_Letteratura_freq : Tabella

	Termini	Ricorrenze	Frequenze
▶	fiorentini	74	1.98E-03
	ghibellini	70	1.87E-03
	aveano	66	1.76E-03
	feciono	66	1.76E-03
	imperadore	64	1.71E-03
	donati	61	1.63E-03
	guelfi	56	1.50E-03
	nimici	53	1.42E-03
	fusse	51	1.36E-03
	pistoia	47	1.25E-03
	sanza	41	1.09E-03
	priori	40	1.07E-03
	giano	39	1.04E-03

Fig. 7 - Termini di italiano antico ricavati sottraendo a "Cronica delle cose occorrenti ne' tempi suoi" (D. Compagni 1312) i termini letterari più recenti.

SFERA-Th_Letteratura_freq : Tabella

	Termini	Ricorrenze	Frequenze
▶	microscopio	32	7.93E-03
	obiettivo	30	7.44E-03
	cellule	13	3.22E-03
	leeuwenhoek	11	2.73E-03
	mm	11	2.73E-03
	vetrini	10	2.48E-03
	fig	10	2.48E-03
	diametro	8	1.98E-03
	vetrino	8	1.98E-03
	coprioggetti	7	1.74E-03
	realizzazione	6	1.49E-03
	tessuti	6	1.49E-03
	dimensioni	6	1.49E-03
	microcosmo	6	1.49E-03

Fig. 8 - Termini tecnici ricavati sottraendo i termini letterari all'articolo del microscopio a sfera di vetro della nostra galleria.

ESAME DEI DATI:
Osservate quali termini sono presenti solo nel documento antico, o solo in quello moderno.
Osservate quali termini sono presenti solo nel documento scientifico, o solo in quello letterario.
Distinguete fra due tipi di termini:
- grammaticali (danno indicazioni sullo stile)
- di contenuto (danno indicazioni sul contenuto)

Più avanti, quando parleremo dei thesauri, descriveremo un metodo più efficace per estrarre i termini specifici di un documento.

TERMINI COMUNI A DUE DOCUMENTI (A x B)

I termini comuni fra due documenti danno indicazioni su caratteristiche comuni di autori diversi. Se si tratta di scrittori appartenenti ad epoche diverse, verranno messi in evidenza i termini antichi che si sono conservati. In ciascun documento, questi termini possiedono una propria frequenza. Il rapporto delle frequenze di ciascun termine, fornisce indicazioni sulla tendenza ad usare maggiormente un termine in tempi passati o al contrario recenti, quindi sulla tendenza ad abbandonarlo o ad acquisirlo.

Questo esame, può affiancare quello compiuto sui termini presenti solo in uno dei due documenti, secondo lo schema:

A AxB B

Intendendo l'esame dei termini presenti solo in A, poi quelli comuni, poi quelli presenti solo in B.

Per determinare i termini comuni a due documenti, utilizziamo il programma **AxB**, il quale calcolerà anche il rapporto delle frequenze che lo stesso termine ha nei due documenti.

Lanciate il programma **AxB**
Scegliete la tabella A (deve essere una tabella di frequenze)
Scegliete la tabella B (deve essere una tabella di frequenze)
Premete il tasto Avvia
(il programma realizzerà una nuova tabella di nome: NomeAxNomeB_rapp)
(anche questa è una tabella di frequenze)

La tabella realizzata dal programma AxB contiene i termini presenti sia in A che in B, ma non quelli presenti solo in A o solo in B. Essa è il risultato dell'operazione booleana di intersezione, indicata anche con l'operatore logico AND.

Questo programma determina anche il rapporto della frequenza con cui i termini comuni sono stati usati nei due documenti in esame:

Rapp (A/B) = freq (A) / freq (B)

Il rapporto delle frequenze è molto utile per evidenziare i termini comuni utilizzati più in A che non in B, o viceversa. E' utile anche per potere osservare i termini comuni utilizzati con uguale frequenza (rapporto = 1). Per facilitare queste operazioni, ordinate la colonna dei rapporti di frequenze, poi osservate i termini con rapporto maggiore, quelli con rapporto minore e quelli con rapporto vicino a uno.

cronicaxpromessi_rapp : Tabella

	Termini	Ricorrenze	Frequenze	Ric A	Ric B	freq A	freq B	Rapporto
	incontro	38	1.97E-04	6	32	1.60E-04	1.44E-04	1.11E+00
	sportello	19	9.84E-05	3	16	8.01E-05	7.19E-05	1.11E+00
	costoro	38	1.97E-04	6	32	1.60E-04	1.44E-04	1.11E+00
	male	146	7.56E-04	23	123	6.14E-04	5.53E-04	1.11E+00
	anni	83	4.30E-04	13	70	3.47E-04	3.15E-04	1.10E+00
	conte	103	5.33E-04	16	87	4.27E-04	3.91E-04	1.09E+00
	sotto	155	8.02E-04	24	131	6.41E-04	5.89E-04	1.09E+00
	amico	84	4.35E-04	13	71	3.47E-04	3.19E-04	1.09E+00
	caccia	13	6.73E-05	2	11	5.34E-05	4.94E-05	1.08E+00
	chiamata	13	6.73E-05	2	11	5.34E-05	4.94E-05	1.08E+00
	aprire	13	6.73E-05	2	11	5.34E-05	4.94E-05	1.08E+00
	cadere	13	6.73E-05	2	11	5.34E-05	4.94E-05	1.08E+00
	anno	52	2.69E-04	8	44	2.14E-04	1.98E-04	1.08E+00

Fig. 9 - Alcuni dei termini presenti con la stessa frequenza nella "Cronica delle cose occorrenti ne' tempi suoi" (1312) e ne "I promessi sposi" (1827).

Esaminando due testi, uno scientifico o tecnico (A) e uno letterario (B), fra i termini più frequenti di A avremo principalmente termini specialistici, mentre fra quelli più frequenti di B ne avremo molti letterari.

Esaminando due testi, uno moderno (A) e uno antico (B), fra i termini più frequenti di A avremo i termini usati più recentemente, mentre fra i termini più frequenti di B avremo i termini usati più anticamente. In posizione intermedia, avremo i termini la cui frequenza si equivale e il cui uso si è mantenuto costante nel tempo (figura 9). Questi termini daranno inoltre indicazioni sulle analogie nello stile e nei contenuti.

DETERMINAZIONI:
Confrontate due documenti letterari di cui uno moderno l'altro antico.
Confrontate due documenti di cui uno di letteratura e l'altro scientifico o tecnico.

ESAME DEI DATI:
Osservate quali sono i termini comuni più frequenti, o meno frequenti, in entrambi i documenti. Le differenze di frequenza danno indicazioni sull'uso dei diversi termini.

Esaminate:
- i termini più frequenti in A (più propri ad A)
- i termini più frequenti in B (più propri a B)
- i termini con uguale frequenza (rapp = 1 circa)

Distinguete fra due tipi di termini:
- grammaticali (danno indicazioni sullo stile)
- di contenuto (danno indicazioni sul contenuto)

NB: l'esame delle ricorrenze di ciascun termine può portare a conclusioni errate se la lunghezza dei due documenti è differente. A tale scopo, è necessario effettuare il confronto delle frequenze, espresso dal loro rapporto.

Oltre che in base al rapporto di frequenza, i termini comuni fra **due documenti** possono essere ordinati anche in base alla **frequenza**. Normalmente, quelli più frequenti sono di tipo grammaticale. Quelli via via meno frequenti sono termini di uso comune. Mano a mano che la frequenza diminuisce, i termini comuni tendono ad acquistare una funzione di contenuto.

SOMMA DI DUE TABELLE (A + B)

Date due tabelle A e B di termini, ricorrenze e frequenze, è possibile sommarle in una tabella C. Per sommare due tabelle, utilizziamo il programma "**A+B**", il quale ricalcolerà i valori delle ricorrenze e delle frequenze, come se i termini provenissero da un unico documento.

Lanciate il programma **A+B**
Scegliete la tabella A (deve essere una tabella di frequenze)
Scegliete la tabella B (deve essere una tabella di frequenze)
Premete il tasto Avvia
(il programma realizzerà una nuova tabella di nome: NomeA+NomeB_freq)

Sommando successivamente più tabelle, il nome della tabella risultante si allunga sempre più. Potete rinominare la tabella. Questo programma potrebbe essere usato per creare thesauri, ma per fare questo abbiamo preparato un programma più adatto.

CREAZIONE DI THESAURI

I thesauri sono semplici elenchi di termini, senza definizione. Nel nostro caso, i thesauri sono corredati dalla frequenza dei termini e sono tabelle di frequenze.

A => Thesaurus

Per la realizzazione di thesauri, utilizzate il programma **A=>Thesaurus** che è più adatto di A+B. Questo programma aggiunge i termini della tabella A al thesaurus indicato (es: Th_Letteratura_freq), mantiene aggiornata una tabella (es: Th_Letteratura_lista) contenente

l'elenco dei testi aggiunti al thesaurus (figura 10). Le ricorrenze della tabella di partenza vengono sommate a quelle del thesaurus e alla fine vengono ricalcolate le frequenze di tutti i termini.

Th_Letteratura_lista : Tabella

Titolo	Tot Parole	Tot Termini	Aggiunto
promessi_freq	222459	19430	27/03/98 11.14.42
senilita_freq	67573	8479	27/03/98 11.16.44
bestie_freq	15624	3698	27/03/98 11.19.59
Th_Letteratura_freq	305656	24366	27/03/98 11.19.59

Fig. 10 - Esempio di lista di testi inseriti in un thesaurus (Th_Letteratura_freq).

Ecco alcuni esempi di thesauri che possono essere creati:

- Thesaurus dei termini grammaticali
- Thesaurus dei termini comuni (termini di uso comune, non grammaticali)
- Thesaurus dei termini colti (termini non comuni, ma di uso generale: es: termini astratti e filosofici)
- Thesauri dei termini specialistici (es: thesaurus di biologia)
- Thesauri gergali (es: burocratichese, politichese)
- Thesauri dei termini letterari (eventualmente limitati all'800, al 900, contemporanei)
- Thesaurus generale della lingua italiana (ottenuto sommando parecchi testi di ogni tipo)
- Thesauri in altre lingue

A <= Ripristina

Nel caso in cui per errore si sia aggiunto un testo ad un thesaurus, questo programma consente di toglierlo, ripristinando la situazione precedente.

THESAURI DELLA LINGUA ITALIANA

Per l'individuazione di termini specialistici e di parole chiave, è necessario disporre di un thesaurus di riferimento come ad esempio un **thesaurus dei termini letterari** (figure 11 e 12). Per realizzare questo thesaurus, si dovranno raccogliere solo testi di letteratura. In questo modo, si avranno a disposizione i termini di uso comune e tradizionali. Il confronto, effettuato con il programma **A-B**, di un testo tecnico con un thesaurus così compilato mette in evidenza i termini non letterari che quel documento possiede: in generale termini specialistici (figure 7 e 8). Operando in questo modo, si ha l'inconveniente che i termini specialistici, che per qualche ragione sono presenti anche nel thesaurus, non figureranno fra quelli raccolti. Vedremo fra poco come sia possibile evitare questo problema.

Th_Letteratura_freq : Tabella

Termini	Ricorrenze	Frequenze
a	5862	1.92E-02
abate	1	3.27E-06
abati	1	3.27E-06
abbacinare	1	3.27E-06
abbacinati	1	3.27E-06
abbacinato	1	3.27E-06
abbado	1	3.27E-06
abbadò	1	3.27E-06
abbagliaron	1	3.27E-06
abbagliato	1	3.27E-06
abbagliava	1	3.27E-06
abbaia	1	3.27E-06

Fig. 11 - Primi termini del thesaurus Th_Letteratura_freq (ordine alfabetico).

Th_Letteratura_freq : Tabella

Termini	Ricorrenze	Frequenze
e	10177	3.33E-02
di	8835	2.89E-02
che	8576	2.81E-02
a	5862	1.92E-02
il	5470	1.79E-02
la	5425	1.77E-02
non	5021	1.64E-02
un	4990	1.63E-02
in	4209	1.38E-02
per	3793	1.24E-02
si	2889	9.45E-03
con	2789	9.12E-03

Fig. 12 - Alcuni termini del thesaurus Th_Letteratura_freq (ordine di frequenza)

Un altro modo per individuare i termini specialistici e per compiere altre determinazioni come l'indice di leggibilità di un testo, richiede il confronto con un thesaurus che non comprenda solo i temini letterari, ma anche tutti gli altri. Potremmo definire questo, il thesaurs delle frequenze dei termini della lingua italiana scritta.

Tale thesaurus sarebbe quella tabella che si otterrebbe sommando tutti i documenti scritti fino ad ora nella nostra lingua e calcolando le frequenze di tutti i termini. Per evidenti ragioni pratiche, non possiamo ottenere questa tabella. Tuttavia, con un ragionevole impegno, sommando un certo numero di testi, è ugualmente possibile ottenere una tabella delle frequenze di una grande quantità di termini. Le frequenze dei termini di questa tabella non si discosteranno in modo significativo da quelle ideali, quindi il thesaurus così ottenuto risulterà ancora utile a numerosi scopi pratici e potremo utilizzarlo con successo per individuare i termini specialistici di un testo. Definiamo questo thesaurus reale: **thesaurus generale** per indicare che può comprendere tutti i termini senza distinzioni.

Nel realizzare questo thesaurus, è necessario evitare di inserire un numero eccessivo di testi che parlano dello stesso argomento. Se per esempio si sommassero prevalentemente testi di architettura, i termini di questa disciplina figurerebbero con una frequenza più alta che nell'uso generale. Se invece si aggiungono testi eterogenei, i termini specialistici di ognuno assumeranno una frequenza bassa, più vicina al valore "vero", mentre i termini di uso comune manterranno una frequenza alta.

INDIVIDUAZIONE DI TERMINI SPECIALISTICI PER MEZZO DI THESAURI

METODO 1 Sottraete il testo in esame al thesaurus dei termini letterari (progr. A-B) Ordinate la colonna delle frequenze della tabella risultante
METODO 2 (consigliato) Sommate il testo al Thesaurus Generale (progr. A=>Thesaurus) Calcolate il rapporto delle frequenze fra i termini del testo e quelli del Thesaurus Generale (progr. AxB) Ordinate la colonna dei rapporti della tabella risultante

Anche con un thesaurus generale possiamo isolare i termini specialistici di un documento. In questo caso, è necessario **sommare il documento al thesaurus generale prima di cominciare le nostre analisi** (programma A => Thesaurus). Poi, per mettere in evidenza i termini specialistici, useremo il programma **AxB** che cerca i termini comuni. Ma se tutti i termini del nostro documento sono comuni con quelli del thesaurus (li abbiamo aggiunti al thesaurus prima di cominciare) come faremo per mettere in evidenza quelli specialistici? Per fare questo ci basiamo sul fatto che i termini specialistici hanno una **frequenza relativamente elevata nel documento, ma bassa nel thesaurus**. Il programma AxB calcola il **rapporto della frequenza che ogni termine ha nei due documenti**. Ora, ordinando in senso decrescente la colonna del rapporto delle frequenze, i termini specialistici di quel documento verranno a galla.

Dopo avere ordinato in questo modo la tabella, i termini si raggrupperanno in due strati:
1) i termini che prima di aggiungere il testo A al thesaurus erano **esclusivi di A**. Essi avranno tutti lo **stesso valore di frequenza** e si raccoglieranno tutti nella parte superiore della tabella
2) i termini che prima della somma di A al thesaurus erano **comuni ad entrambi**. Questi termini sono ordinati in tre parti:
2-1) i termini comuni **più frequenti in A** (rapporto > 1)
2-2) i termini comuni che hanno **una frequenza simile in entrambi i documenti** (rapporto = 1 circa)
2-3) i termini comuni **più frequenti nel thesaurus** (rapporto < 1).

Realizzazione dei due thesauri
Vi consigliamo di realizzare entrambi i thesauri descritti. Per fare questo, cominciate a comporre il thesaurus dei termini letterari inserendovi testi che troverete in Rete o che acquisirete con lo scanner. Quando questo thesaurus sarà divenuto abbastanza consistente (circa 300.000 parole, per almeno 20.000 termini), fatene una copia. Ora disponete di due thesauri identici. Riservate il primo ai soli testi letterari (**thesaurus dei termini letterari**), aggiungete invece al secondo anche testi scientifici, tecnici, giornalistici e avrete in questo modo un **thesaurus generale**. Nel database che vi abbiamo fornito, c'è già un thesaurus di termini letterari chiamato: Th_Letteratura_freq.

TABELLA DEI TERMINI GRAMMATICALI

Gramma : Tabella

	ID	Parole	Campo1
	748	trovo	
	749	tu	
	750	tua	
	751	tue	
	752	tuo	
	753	tuoi	
	754	tutt	
	755	tutta	
	756	tuttavia	
	757	tutte	
	758	tutti	
	759	tutto	
	760	tuttora	

Fig. 13 - Alcuni termini della tabella Gramma (ordine alfabetico).

Raccogliendo i termini comuni fra più documenti (programma AxB), si può comporre una tabella dei termini più comuni. In questa tabella, figureranno principalmente articoli, preposizioni, pronomi, aggettivi possessivi, avverbi, etc. quindi termini che hanno una funzione più grammaticale che di contenuto (figura 13). Questa tabella risulterà molto utile in operazioni che vedremo fra breve. Nel realizzarla, occorre impiegare testi non molto grandi perché altrimenti comparirebbero come termini comuni anche termini di contenuto. Non si devono impiegare neppure testi molto piccoli, altrimenti resterebbero fuori molti termini grammaticali. Comunque, dopo aver realizzato una simile tabella, occorre esaminarla per eliminare manualmente i termini di contenuto eventualmente presenti e per inserire quelli grammaticali non compresi.

Abbiamo preparato una tabella di termini grammaticali, comprendendo anche certi termini di uso molto comune come alcune voci dei verbi ausiliari. Questa tabella ha nome **Gramma** ed è un semplice elenco di termini. A questa tabella potete togliere o aggiungere parole, in modo da adattarla alle vostre esigenze.

Il programma **Testo - Gramma** sottrae la tabella Gramma a una tabella di parole (ottenuta con il programma Normalizza). Questa operazione serve a preparare la ricerca delle locuzioni ripetute, che verrà compiuta più avanti.

La tabella Gramma è una tabella semplice e come tale non può essere utilizzata con i programmi A-B, AxB, A+B. Se avrete bisogno di fare questo, ricavatene una tabella di frequenze per mezzo del programma Frequenze.

ESPRESSIONI RIPETUTE, O LOCUZIONI

Molto spesso, i termini specialistici sono formati da locuzioni (termini composti da più parole). Queste locuzioni sono molto utili per descrivere un documento e per effettuare ricerche bibliografiche perchè normalmente sono molto specifiche. Prese da sole, le singole parole di cui sono composte queste locuzioni non sono altrettanto specifiche. Per esempio la locuzione "parola chiave" è propria dell'ambito delle ricerche bibliografiche, mentre i termini parola e chiave presi da soli non sono affatto specifici di questo argomento. Lo scopo di questa procedura è quello di individuare le locuzioni ripetute in un documento e di mettere in evidenza quelle usate più frequentemente.

Utilizzato in modo opportuno, il programma di cui parleremo fra breve è capace di individuare anche le espressioni ripetute in campo letterario. Anche questa applicazione può risultare utile per analisi di tipo linguistico.

Per effettuare questa analisi, utilizziamo il programma **Locuzioni**. Questo programma associa le parole in sequenze. Queste sequenze possono essere lunghe da 2 a 5 parole, a seconda di quello che avrete indicato. Come vedete dalla figura 14, le parole di una sequenza sono congiunte dal segno "_" a formare un'unica stringa. A questo punto, il programma Frequenze, vede queste stringhe come parole e ne ricava la ricorrenza e la frequenza. Ordinando in senso decrescente la colonna delle ricorrenze o delle frequenze, le stringhe più ripetute si raccoglieranno in alto.

Come si è detto, nell'usare il programma bisogna indicare la lunghezza delle sequenze da individuare. Questo valore può andare da due a cinque parole. I testi da esaminare con questo programma possono essere completi o sottratti dei termini grammaticali.

Se si vogliono individuare le espressioni ripetute di un testo letterario, occorre esaminare un **testo completo**. In questo caso, le sequenze di due parole hanno uno scarso significato dal momento che verrebbero abbinati sostantivi con i loro articoli e la locuzione non sarebbe riconoscibile. In questo caso è meglio cercare sequenze composte da almeno tre parole (figura 14).

Al contrario, nella ricerca di locuzioni specialistiche la presenza dei termini grammaticali risulta inutile o dannosa. Il programma Testo-Gramma serve proprio a togliere i termini grammaticali da una tabella prima di sottoporla al programma Locuzioni. Nel caso in cui si esamini un **testo sottratto dei termini grammaticali**, è conveniente che la lunghezza delle sequenze da cercare sia di due sole parole (figura 15).

Come si è detto, per individuare le locuzioni ripetute, applicate alla tabella delle locuzioni il programma frequenze e ordinate la tabella risultante in ordine decrescente di ricorrenze: le locuzioni ripetute si raccoglieranno in alto. Potete eliminare le sequenze non ripetute (ricorrenza = 1).

bestie_4_freq : Tabella

	Termini	Ricorrenze	Frequenze
	di_quando_in_quando	3	1.92E-04
	la_mia_anima_è	3	1.92E-04
	il_chiaro_di_luna	3	1.92E-04
	non_c_era_più	3	1.92E-04
	con_la_mia_anima	3	1.92E-04
	io_m_ero_messo	3	1.92E-04
	tutte_le_volte_che	3	1.92E-04
	dall_altra_parte_della	2	1.28E-04
	in_punta_di_piedi	2	1.28E-04

Fig. 14 - Espressioni ripetute di "Bestie" (testo intero).

Caso1: ricerca delle locuzioni letterarie (testo completo)
Preparate una tabella normalizzata del testo da trattare
Aprite il programma **Locuzioni**
Scegliete la tabella da trattare (le parole devono essere nell'ordine di lettura)
Indicate il numero di parole da congiungere (3-5)
Avviate il programma
(otterrete la tabella: Nome_n, dove n è il n° di parole in ogni locuzione)

Sulla tabella ottenuta, applicate il programma **Frequenze**
(otterrete la tabella Nome_n_freq)
Ordinate la tabella secondo le ricorrenze
Cancellate la tabella: Nome_n

sfera-G_2_freq : Tabella

	Termini	Ricorrenze	Frequenze
▶	messa_fuoco	8	4.27E-03
	sistema_illuminazione	6	3.21E-03
	becco_bunsen	5	2.67E-03
	acqua_stagno	4	2.14E-03
	lama_metallica	4	2.14E-03
	mm_diametro	4	2.14E-03
	microscopio_leeuwenhoek	4	2.14E-03
	viaggio_microcosmo	3	1.60E-03
	sfera_vetro	3	1.60E-03
	midollo_sambuco	3	1.60E-03
	piano_vetrini	3	1.60E-03
	sede_conica	3	1.60E-03
	striscio_sangue	3	1.60E-03

Fig. 15 - Espressioni ripetute dell'articolo sul "Microscopio a sfera di vetro" della nostra galleria (testo sottratto dei termini grammaticali).

Caso 2: ricerca di locuzioni specialistiche (testo sottratto dei termini grammaticali)
Preparate una tabella normalizzata del testo da esaminare
Con il programma **Testo-Gramma**, sottraetele i termini grammaticali
Aprite il programma **Locuzioni**
Scegliete la tabella da trattare (le parole devono essere nell'ordine di lettura)
Indicate il numero di parole da congiungere (2)
Avviate il programma
(otterrete la tabella: Nome-G_2)

Sulla tabella ottenuta, applicate il programma **Frequenze**
(otterrete la tabella Nome-G_2_freq)
Ordinatela secondo le ricorrenze
Cancellate le tabelle: Nome-G e Nome-G_2

ESAME DEI DATI:
Esaminate le coppie di termini di un testo specialistico o scientifico e valutate l'utilità di quelle più ripetute.
Cercate le espressioni ripetute di un articolo di giornale (distinguete le locuzioni "letterarie" da quelle relative alla notizia).

INDICE DI LEGGIBILITA'

Numerosi autori hanno cercato metodi per determinare il grado di difficoltà della lettura di un testo. A tale scopo, sono stati presi in considerazione parametri quali la lunghezza delle parole e la lunghezza dei periodi.

Il nostro programma **Leggibilità** tiene conto dei seguenti parametri:

- lunghezza periodi (da punto a punto)
- lunghezza paragrafi (a capo)
- familiarità dei termini impiegati (frequenza dei termini)

Per stabilire la lunghezza dei periodi e dei paragrafi di un documento, il programma esamina un **file di testo** (non una tabella) compreso della punteggiatura, conta il numero di segni di fine periodo (. ! ? a capo) non ripetuti, conta il numero di segni di fine paragrafo (a capo) non ripetuti, conta il numero di caratteri alfabetici del documento. Calcola la lunghezza in caratteri del periodo medio e del paragrafo medio. Per calcolare la familiarità, si serve della **tabella delle frequenze** ricavata da quel testo (quali termini sono presenti nel documento e quante volte sono usati) e di un **thesaurus delle frequenze** (valore della frequenza dei termini nella lingua italiana).

Il programma Leggibilità determina l'indice di leggibilità secondo una formula che tiene conto della lunghezza media dei periodi, della lunghezza media dei paragrafi e della familiarità dei termini impiegati.

Al termine, il programma mostra le proprie determinazioni nella **finestra di stato della maschera** e aggiunge una riga alla tabella **Statistica** con il nome del documento trattato, il thesaurus utilizzato e i dati ottenuti (figura 16).

Per compiere questa determinazione, potete utilizzare il **thesaurus dei termini letterari**. In questo caso, quando il programma trova un termine che non possiede, gli assegna una frequenza di 0,00001.

Un altro modo per compiere la stessa determinazione, è quello di **sommare il testo da esaminare al thesaurus prima di avviare il programma Leggibilità**. In questo caso utilizzeremo il **thesaurus generale**, quello che comprende anche i termini non letterari. Con questi due metodi si ottengono risultati leggermente diversi, con un indice di leggibilità un po' inferiore usando il thesaurus di termini letterari.

Prima di iniziare questa analisi, verificate che il testo vada **a capo** come definito dall'autore. Infatti, durante il passaggio del testo dal formato di partenza (.html, .doc) a quello "testo" (.txt), spesso la formattazione originale viene alterata. Se necessario, sistemate a mano una parte del testo: almeno 8000 caratteri (circa 3 pagine) e copiate questa parte in un nuovo documento. La determinazione dell'indice di leggibilità su questa parte del testo darà un risultato molto vicino a quello che si sarebbe ottenuto con il documento intero. Se il testo è un libro o un rapporto, è conveniente togliere l'indice e tutti i titoli.

Verificate gli a capo del file di testo
Togliete l'eventuale indice e salvate il file come testo
Il programma Leggibilità richiede la presenza di:
- il file di testo (Nome.txt)
- la relativa tabella di frequenze (Nome_freq)
- un thesaurus

Eventualmente aggiungete il testo al thesaurus
Aprite il programma **Leggibilità**
Indicate il nome del documento da esaminare
Indicate il thesaurus di riferimento per il calcolo della familiarità
Lanciate il programma

Esaminate i dati ottenuti
Gli stessi dati vengono inseriti nella tabella **Statistica**

Statistica : Tabella

	Titolo	IL	FFa	FPe	FPa	Thesaurus	Caratt
	bestie	93.9	58.4	26.0	9.6	Th_Letteratura_freq	69
	cavaliere	87.1	48.2	29.3	9.6	Th_Letteratura_freq	155
	sfera	78.1	42.4	28.0	7.8	Th_Letteratura_freq	21
	sfera	80.4	44.6	28.0	7.8	Th_Generale_freq	21
	dolomiti	42.4	34.9	0.8	6.8	Th_Generale_freq	5
		0.0	0.0	0.0	0.0		

Fig. 16 - La tabella "Statistica" raccoglie dati sulla leggibilità dei testi esaminati

DETERMINAZIONI: Determinate la leggibilità di alcuni testi, fra i quali anche uno vostro.

ESAME DEI DATI OTTENUTI: Confrontate e discutete i dati ottenuti con i diversi documenti. In particolare, confrontate la lunghezza dei periodi e dei paragrafi. Confrontate la familiarità dei termini impiegati. Discutete la formula.

FORMULA PER LA DETERMINAZIONE DELL'INDICE DI LEGGIBILITA'

IndiceL = FFa + FPe + FPa

Dove:

FFa = Cfa * (MFa - Fmin) funzione di familiarità
FPe = Cpe * MPe * Exp(-(MPe ^ 2) / (2 * Pei ^ 2)) funzione dei periodi
FPa = Cpa * MPa * Exp(-(MPa ^ 2) / (2 * Pai ^ 2)) funzione dei paragrafi

MFa = valore medio della frequenza dei termini
MPe = valore medio della lunghezza del periodo
MPa = valore medio della lunghezza del paragrafo

Fmin = 0.0042 familiarità minima
Pei = 60 lunghezza periodo ideale
Pai = 160 lunghezza paragrafo ideale
CFa = 32000 coefficiente di familiarità
CPe = 0,82 coefficiente del periodo (tale che CPe max = 30)
CPa = 0,103 coefficiente del paragrafo (tale che CPa max = 10)

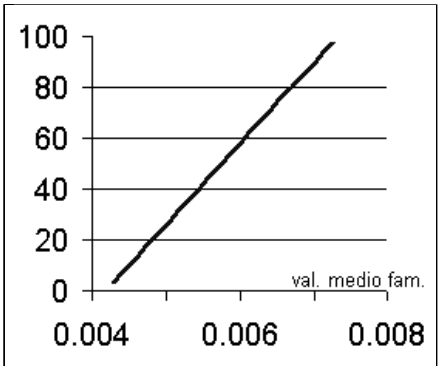

Fig. 17 - Andamento della funzione di familiarità per la determinazione dell'indice di leggibilità.

In questa formula, appaiono tre parametri.
Il primo è proporzionale al valor medio della familiarità dei termini impiegati (figura 17).
Il secondo e il terzo sono il risultato di una funzione che è massima in corrispondenza di un dato valore della variabile (lunghezza ideale del periodo o del paragrafo), e tende a zero quando la variabile si allontana molto da questo valore (figura 18). Questa funzione è un po' complicata, ma ha il vantaggio di evitare l'inconveniente della funzione lineare (retta) di penalizzare molto l'indice di leggibilità quando la lunghezza del periodo o del paragrafo è alta.

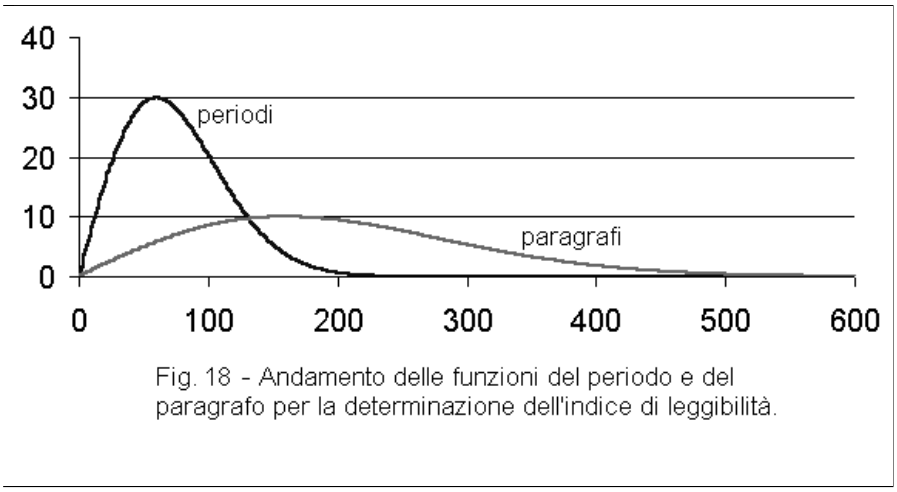

Fig. 18 - Andamento delle funzioni del periodo e del paragrafo per la determinazione dell'indice di leggibilità.

L'indice di leggibilità che determiniamo in questo modo tiene conto solo di pochi parametri, anche se sono in numero maggiore di quelli presi in considerazione da altri indici di cui siamo venuti a conoscenza. Tuttavia, la leggibilità "vera" dipende anche da altre componenti che riguardano funzioni intellettuali superiori fra le quali l'intelligenza e la sensibilità estetica. Per questo motivo, attualmente nessun programma può definire in modo preciso la leggibilità di un testo. In ogni modo, anche se limitato a pochi parametri, il nostro indice di leggibilità non è del tutto privo di senso, tanto è vero che in generale esso è in accordo con il nostro giudizio. Non solo, ma questa determinazione della leggibilità dei testi tiene conto anche della familiarità dei termini, mentre molti altri programmi di questo tipo non ne tengono conto.

VALUTAZIONE DI TESTI (Periodi, Paragrafi, Punteggiatura)

Abbiamo preparato alcuni programmi per aiutare l'esame dei testi. Essi possono essere usati anche per migliorare la leggibilità di un testo in corso di redazione. Questi programmi si chiamano:

Periodi
Paragrafi
Punteggiatura

Il programma **Periodi** esamina il testo indicato, ne estrae i periodi (delimitazione di periodo= .!? a capo) e produce una tabella (**Nome_peri**). In questa tabella ci sono i seguenti campi:

ID	indice di posizione della prima parola del periodo
Parole	lunghezza del periodo in parole
ASCII	contenuto del periodo (comprese punteggiatura e spazi)
Len	lunghezza del periodo in caratteri (comprese punteggiatura e spazi)
Campo1	necessario al sistema

Il programma **Paragrafi** esamina il testo indicato, ne estrae i paragrafi (delimitatore di paragrafo = a capo) e produce una tabella **Nome_para**.
In questa tabella ci sono gli stessi campi della tabella precedente, con i paragrafi al posto dei periodi.

Il programma **Punteggiatura** esamina il testo indicato e produce una tabella **Nome_punti**. In questa tabella ci sono i seguenti campi:

ID	indice di posizione della parola che precede il carattere o sequenza
Parole	carattere o sequenza, compreso fra i segni < >
ASCII	valore ASCII del carattere o sequenza
Len	lunghezza della sequenza, in caratteri
Campo1	necessario al sistema

La tabella dei periodi può essere ordinata in senso decrescente di lunghezza del periodo. In questo modo, verranno messi in evidenza i periodi più lunghi, i quali possono poi essere esaminati separatamente (figura 19). Altrettanto può essere fatto con la tabella dei paragrafi. Come si è detto, questo può essere utile per esaminare un documento in corso di preparazione, eventualmente allo scopo di migliorarne la leggibilità, per esempio riducendo la lunghezza dei periodi e dei paragrafi più lunghi.

Parole	ASCII	Len
110	Quello a cui con una frusta di salcio avevano fatto un nodo scorsoio e l'avevano lasciato lì ciondoloni; quello infilato, dal ventre, a una canna aguzzata: la canna riesciva dalla bocca, e il sangue colava più grosso e scuro; quello a cui avevano schiacciato con i sassi tutte e quattro le zampe; quello accecato con i tizzi della brace; quello sbudellato con un colpo di falcino; quello schiacciato dalle ruote del carro, a posta; quello lanciato in aria dando un colpo sopra una tavoletta messa in bilico; quello pestato dai due fidanzati; questi sono i rospi che ho visto morire, silenziosi, con quei loro occhi che di notte luccicano.	648
83	E vedevo i suoi orti squadrati entrare, con un angolo più alto degli altri, tra le case più rade; oppure, l'uno appresso all'altro, farsi largo e posto, ma fermati da una fila di cipressi la cui ombra oscurava il verde dell'erba; e qualche pesco fiorire e maturare accanto alle campane d'una chiesola, e qualche olivo chiamarsi dietro tutta la campagna soave, che impallidiva lontano, rasente i monti chiarissimi, talvolta più luminosa del sole; con una tenerezza che mi commoveva.	484

Fig. 19 - I due periodi più lunghi di "Bestie".

A tale proposito, occorre rilevare che non è sempre vero che un periodo lungo sia anche scarsamente leggibile. Infatti, se tale periodo descrive un elenco di cose, può risultare lungo, ma non per questo è di lettura difficoltosa. Non è neppure detto che un periodo corto sia sempre preferibile ad uno lungo. Infatti, esaminando i periodi più lunghi, vi capiterà di vederne alcuni logicamente ben strutturati, composti da frasi ben coordinate. E' vero che questi periodi richiedono al lettore un'attenzione superiore, ma possono risultare più belli da leggere. A questo punto, soprattutto se si tratta di un testo letterario, si può decidere di sacrificare la leggibilità alla bellezza. Se si tratta invece di un testo scientifico, si dovrà dare la precedenza alla leggibilità. Esistono infine periodi lunghi mal costruiti, nei quali più soggetti o complementi si contendono lo stesso verbo, nei quali la struttura logica non è facilmente rintracciabile. Questi periodi fanno venire in mente scene di lotta e ci costringono a riletture e a penose opere di ricostruzione e di interpretazione. Questi sono i periodi da sistemare, per migliorarne sia l'estetica che la leggibilità. Come conclusione, leggibilità e bellezza letteraria non sono la stessa cosa e non vanno sempre d'accordo. Si può decidere di sacrificare l'una a favore dell'altra, ma non si può rinunciare ad entrambe. Otterremo chiaramente i migliori risultati, quelle volte che riusciremo a combinare la bellezza dello scritto con la facilità di lettura.

Le tre tabelle _peri, _para, _punti possono essere esaminate così come sono prodotte dai rispettivi programmi, ma possono essere anche trattate con il programma Frequenze, in modo da ottenere tabelle con:

- il numero di periodi di lunghezza uguale
- il numero di paragrafi di lunghezza uguale
- il numero di segni di punteggiatura dello stesso tipo

CONCLUSIONE

Al termine di questi esami, vi consigliamo di cancellare le tabelle non più utili per evitarne la proliferazione e la conseguente confusione. Subito dopo, comprimete il database per ridurne le dimensioni (menù STRUMENTI/utilità database/compatta database).

La presente edizione di questo lavoro può essere considerata di prova, una specie di beta version. Siete invitati a inviarci le vostre osservazioni e suggerimenti. Se ci avete seguito fino a qui, soprattutto effettuando le prove che vi abbiamo indicato, avete imparato ad usare i programmi che vi abbiamo fornito. Ora potete utilizzarli per le vostre esigenze e secondo i vostri interessi.

Questi programmi possono essere utili alle scuole, specialmente a quelle che impartiscono un insegnamento umanistico, per effettuare ricerche di carattere linguistico; ai cybernauti che si trovano ad effettuare ricerche bibliografiche in Internet; ai redattori di documenti ipertestuali per indicarne i termini specifici ai motori di ricerca. Come avevamo detto all'inizio, questi programmi sono nati per facilitare la messa a punto delle interrogazioni di banche dati nel corso di ricerche bibliografiche, quindi possono essere di aiuto anche a ricercatori i quali potranno anche usarli per estrarre descrittori da associare a propri articoli o per comporre l'indice analitico di un libro. Possono infine servire a bibliotecari, nel loro lavoro di indicizzazione e di classificazione dei documenti. In ogni caso, sia che vengano usati per scopo amatoriale o professionale, ci auguriamo che questi programmi siano di vostro gradimento.

Invia i tuoi commenti sull'articolo
