Analisi statistica

Il modello matematico

L'ipotesi di lavoro alla base del tipo di analisi statistica da noi applicata presuppone che il vocabolario totale sottostante un determinato corpus sia distribuito secondo una curva di distribuzione normale. Tale ipotesi è a sua volta giustificata dalle seguenti osservazioni:

- Un insieme di parole o occorrenze di un corpus raccolto secondo una determinata procedura costituisce soltanto un determinato "campione finito" dell'insieme di testi di un presumibile infinito insieme di testi o "corpus infinito" che potrebbe essere raccolto continuando indefinitivamente con la stessa metodologia.

- Nell'universo teorico delle occorrenze ogni lemma ha una probabilità vera. Matematicamente la probabilità vera è il limite tra il rapporto del numero di occorrenze di un dato lemma e il numero totale delle occorrenze quando quest'ultimo aumenta senza limiti; tale probabilità può assumere ogni valore fra 0 e 1. Chiaramente le parole più frequenti avranno una probabilità vera più elevata.

- Riferendoci al modello lognormal (Carroll et al., 1971), da noi utilizzato per l'analisi statistica riguardante la distribuzione delle frequenze, si suppone che vi sia un numero finito di lemmi rappresentati nell'universo infinito di occorrenze e che i logaritmi delle probabilità dei lemmi siano distribuiti secondo una determinata curva di distribuzione normale.

- Il modello lognormal era già stato precedentemente usato nell'implementazione degli algoritmi di analisi della distribuzione delle frequenze per la realizzazione del *"LESSICO ELEMENTARE (Dati statistici dell'Italiano scritto e letto dai bambini delle elementari)"* (Marconi L. et al., Zanichelli Ed., Bologna, 1994)

Analisi della distribuzione di frequenza

Determinati i sottosettori, maggiormente rappresentativi delle differenze interne del lessico del corpus di riferimento da noi considerato, 9 sottosettori per i ***quotidiani***, 12 sottosettori per i ***periodici*** e 13 sottosettori i ***libri,*** per ogni lemma e per ogni forma è stata calcolata la frequenza assoluta totale e parziale nei diversi sottosettori.

La **frequenza assoluta** di una parola in realtà è un dato assai ingannevole, poiché non tiene conto della diffusione della parola all'interno dei diversi sottosettori dell'intero corpus. Una determinata parola può avere una frequenza molto elevata in un solo settore e non essere affatto presente negli altri, tipico il caso dei linguaggi "tecnici", in cui alcuni termini compaiono con estrema frequenza ma non appartengono al linguaggio "comune" e la loro frequenza è determinata solo dal contesto in cui compaiono. Le tradizionali tecniche statistiche di linguistica computazionale adottate nei lessici di frequenza, prevedono la correzione della frequenza assoluta di una parola in base alla sua dispersione.

L'**indice di dispersione** è la misura della presenza del lemma nei diversi sottosettori dell'intero corpus ed è calcolato in modo tale che le parole con uguale frequenza assoluta totale che tendono a concentrarsi in pochi sottosettori risultino con frequenza relativa più bassa rispetto alle parole diffuse in più sottosettori. Tale indice varia da 0 a 1; tanto più un lemma è diffuso tanto più questo valore è elevato, cioè prossimo all'unità. Ne consegue che le parole più frequenti non sono quelle che hanno semplicemente la maggior frequenza assoluta ma quelle che hanno anche un alto indice di dispersione. Anche la dimensione dei sottosettori di un corpus può essere oggetto di "correzione", o normalizzazione poiché la diversa grandezza quantitativa può ovviamente influenzare i risultati.

La maggior parte dei lessici di frequenza prodotti nelle diverse lingue sono tratti da campioni costruiti con dei sottoinsiemi a dimensione uguale o sono sottoposti a un processo di

normalizzazione se di dimensioni diverse. La normalizzazione su sottoinsiemi di dimensioni diverse, almeno per quel che riguarda studi inerenti l'analisi di frequenza del lessico dell'italiano, generalmente è stata effettuata rispetto un unico valore costante. Nel nostro caso questo tipo di normalizzazione si dimostrava, a nostro avviso, inadeguato soprattutto per alcuni sottosettori in dove il numero delle occorrenze era molto basso; questa normalizzazione presentava pertanto il rischio di rappresentare inadeguatamente la probabilità di comparsa della parola soprattutto nei sottosettori più piccoli.

La metodologia adottata da Carroll nel Word Frequency Book per il calcolo della dispersione e dell'uso o frequenza corretta ci è parsa la più adeguata al nostro campione. Secondo tale metodologia, la normalizzazione necessaria per poter riunire le probabilità delle occorrenze di gruppi numerici diversi nei sottosettori considerati, effettuata sul calcolo della dispersione, tiene conto sia delle probabilità di una data parola considerata all'interno di ciascun sottosettore, sia della variabilità di tali probabilità all'interno delle classi o sottosettori. Tale procedura statistica viene sintetizzata per ogni parola oltre che negli indici di dispersione, di frequenza corretta o uso, di rango.

Le formule utilizzate sono:

D misura della dispersione delle frequenze negli n sottosettori considerati

$$D = \frac{\log\left(\sum_{i=1}^{n} p_i\right) - \frac{\sum_{i=1}^{n} p_i \log(p_i)}{\sum_{i=1}^{n} p_i}}{\log(n)}$$

dove

n numero totale dei sottosettori, cioè 34 nel nostro caso

i numero dell'i-esimo sottosettore, con i = 1, 34

p_i probabilità di una occorrenza nell'iesimo sottosettore, calcolata come la frequenza del lemma nell' i -esimo sottosettore rapportata al numero di occorrenze in quel sottosettore;

ovvero

$$p_i = \frac{f_i}{s_i} \qquad P = \sum_{i=1}^{n} p_i$$

$$p_i \log(p_i) = 0 \quad \text{per} \quad p_i = 0$$

f_i frequenza nell' i -esimo sottosettore

s_i numero di occorrenze nell' i -esimo sottosettore

D è costituito essenzialmente dalla somma algebrica di due termini, il primo, che tiene conto della sommatoria delle probabilità calcolate all'interno di ciascun sottosettore diviso per il numero di sottosettori, rappresenta una media delle probabilità; il secondo rappresenta una misura della variabilità o entropia della probabilità all'interno dei sottosettori; anche il secondo termine non è altro che un valor medio di tale variabilità poiché diviso per la sommatoria delle probabilità e per i sottosettori considerati. Il valore di D, che riflette appunto la dispersione delle frequenze nei sottosettori considerati, ha valore pressoché nullo quando tutte le occorrenze di una parola si trovano in un singolo sottosettore indipendentemente dalla frequenza e ha valore uno se le frequenze sono distribuite negli otto sottosettori in maniera proporzionale al numero totale di occorrenze nei sottosettori.

U frequenza corretta secondo i valori di D e di f min

$$U = (fnorm / N)FD + (1 - D)f_{\min}$$

dove

F frequenza assoluta della parola nel corpus

D indice di dispersione

f min

$$f_{\min} = \sum_{i=1}^{34} f_i t_i = \sum_{i=1}^{34} f_{\min_i}$$

fnorm fattore di normalizzazione per sopperire all'esigenza di confrontabilità con corpora di diverse dimensioni[1]; la normalizzazione è effettuata quindi rispetto a un corpus di 1.000.000 di occorrenze (es. corpus del "*Lessico Elementare*"); si ha quindi *fnorm=1.000.000* per il calcolo della frequenza corretta nel corpus totale[2]

f_i frequenza nell' i-esimo sottosettore

t_i rapporto fra il numero di occorrenze nell' i-esimo sottosettore e il totale delle occorrenze

$$t_i = s_i / N$$

s_i numero di occorrenze nell' i esimo sottosettore

N numero totale delle occorrenze del corpus

La frequenza corretta o uso così calcolata considera i valori relativi alla frequenza assoluta, all'indice di dispersione e al valore di f_{min} . Il valore f_{min} in particolare rappresenta una misura non solo della diffusione qualitativa, o presenza del lemma nei diversi sottosettori, ma anche della sua dimensione quantitativa all'interno di ciascun sottosettore, in relazione al peso differenziale del sottosettore in rapporto all'intero corpus. Il valore f_{min} infatti è la sommatoria degli $f\ min_i$ relativi a ogni sottosettore ed è calcolato rispetto al numero di frequenze di un lemma in un determinato sottosettore per la percentuale dell'incidenza della dimensione di quel sottosettore all'interno dell'intero corpus. Il valore di U è uguale a F, se la dispersione D è uguale a uno; per valori di D minori di uno il valore di U è minore di F; quando D è prossimo al valore nullo U ha un valore minimo f_{min} che riflette la media pesata delle probabilità di trovare quella parola nei sottosettori, il valore di U calcolato con questa metodologia, dovrebbe rispecchiare meglio il valore della frequenza che si potrebbe determinare in un corpus di dimensioni estremamente grande.

R rango di una parola nelle liste di frequenza, per frequenza decrescente:

$$R = [(n-1) + (n+y)]/2$$

dove:

y numero delle parole con la stessa frequenza relativa

n numero progressivo della prima di queste parole.

---

[1] La formula utilizzata da Carroll riguardante l'uso o frequenza corretta prevede un fattore moltiplicativo di normalizzazione che era il risultato del rapporto di 1.000.000 di occorrenze rispetto alle 5.088.721 occorrenze del corpus AHI. Questo termine di normalizzazione era dovuto nel loro caso all'esigenza di confrontabilità con altri corpus analoghi (Lorge Magazine Count di 4.591.122 occorrenze e il Brown University Corpus di 1.014.232 occorrenze; la normalizzazione in CoLFIS è stata effettuata quindi rispetto a un corpus con dimensioni più piccole. Si suggerisce l'applicazione, qualora un lettore fosse interessato ad effettuare confronti con altri lessici esistenti dell'italiano diversi da 1.000.000 di occorrenze, di una normalizzazione rapportando i valori relativi all'uso U ricavati a una costante moltiplicativa calcolata in maniera da tener conto delle dimensioni (relativamente al numero di occorrenze) del corpus con cui si vuole fare il raffronto divisa per N1=1.000.000. Ad esempio se il corpus di raffronto ha dimensioni NR=500.000 il valore U_{norm} = (NR/1.000.000) * U.

[2] Il valore Fnorm è Fnorm = 500.000 per il calcolo della UQ uso o frequenza corretta dei quotidiani, UP uso o frequenza corretta dei periodici, UL uso o frequenza corretta dei libri; tale normalizzazione è stata effettuata per poter fare un raffronto interno ai tre sottosettori costituenti CoLFIS.