Compito 2: Esperimenti di Statistica Lessicale

4 novembre 2004

Introduzione

- Lo scopo di questo compito è duplice: da un lato permette di applicare le tecniche di manipolazione di corpora e liste di frequenza apprese fin qui, dall'altra fa toccare con mano alcune delle questioni trattate nel campo della *statistica lessicale*.
- Quasi tutto quello che facciamo in questo corso, e più in generale quasi tutta la linguistica computazionale, si basa in un modo o nell'altro sulla raccolta di frequenze da un corpus.
- Dunque, prima di gettarci a capofitto nello studio di tecniche che sfruttano statistiche basate sulla frequenza di parole e altre unità nei corpora, è bene farsi un'idea del tipo di fenomeni che abbiamo di fronte.
- La statistica lessicale studia le proprietà statistico-distribuzionali di morfemi, parole e altre unità.
- Vedi: Harald Baayen, *Word Frequency Distributions*, Kluwer, 2001.
- Data di consegna: venerdì 12 novembre, a lezione.
- Punti: 4+2, per un totale di 6/4.

1 Spettri di frequenze [2 punti]

- Lo spettro di frequenze è una lista in cui, per ciascuna frequenza attestata in un corpus, riportiamo il conto di quante parole hanno quella frequenza.
- Per esempio, se in un corpus ci sono cinque parole che capitano sette volte, lo spettro di frequenze riporterà che la frequenza 7 ha una frequenza di 5.
- Lo spettro di frequenze, dunque, altro non è che una lista di frequenze di frequenze.

- Per chiarezza, invece di dire cose come "la frequenza 7 ha una frequenza di 5", possiamo usare il termine *livello di frequenza*, definito come la frequenza la cui frequenza stiamo raccogliendo; dunque: "il livello di frequenza 7 ha una frequenza di 5".
- Per costruire uno spettro di frequenze, partiamo da una lista di frequenza e contiamo, per ciascuna frequenza, quante parole hanno tale frequenza.
- Considerate questo mini-corpus:

```
il cane abbaia e il gatto miagola mentre la famiglia discute
se abbandonare sia il gatto che il cane per la strada
```

- Lista di frequenza:

```
1 abbaia          1 abbandonare
2 cane            1 che
1 discute         1 e
1 famiglia        2 gatto
4 il              2 la
1 mentre          1 miagola
1 per             1 se
1 sia             1 strada
```

- Osserviamo che ci sono tre livelli di frequenza in questa lista: 1, 2 e 4.
- Contiamo quante parole hanno frequenza 1, quante parole hanno frequenza 2 e quante parole hanno frequenza 4.
- Otteniamo dunque una lista di frequenza di livelli di frequenze, cioè uno spettro di frequenze:

```
12 1
3  2
1  4
```

- Questa lista ci dice che ci sono 12 parole che capitano una volta, 3 parole che capitano 2 volte e 1 parola che capita 4 volte.
- Adesso, **costruite lo spettro di frequenze del Brown e delle novelle di Pirandello.**
- Per esempio, una spettro di frequenze potrebbe avere il seguente formato:

```
18209 1
 6675 2
 3627 3
 2366 4
```

```
1607 5
  ...
  ...
   1 17884
   1 18348
   1 20000
   1 21591
   1 26498
```

- La prima riga ci dice che ci sono 18209 parole con frequenza 1, la seconda riga che ci sono 6675 parole con frequenza 2, e l'ultima riga che c'è una sola parola con frequenza 26498 (notate che in questo caso conviene ordinare per la seconda colonna, cioè quella che contiene i livelli di frequenza, e in ordine crescente).
- Pensateci bene – partendo dalle liste di frequenza con un po' di `gawk`, `sort` e `uniq` non è difficile ricavare liste di questo genere.
- **Riportate i comandi usati per costruire gli spettri di frequenze.**
- La distribuzione di molti fenomeni naturali (e non) ha un andamento "a campana", in cui i valori intermedi sono molto frequenti, mentre più ci si allontana dal centro più si incontrano valori rari.
- Per esempio, se dividiamo le studentesse della SSLMIT in cinque categorie in base all'altezza: molto basse, basse, medie, alte, molto alte, troveremo quasi di sicuro che ci sono poche studentesse molto basse o molto alte, abbastanza studentesse basse o alte, e molte studentesse d'altezza media.
- Un andamento di questo tipo (frequenze che diminuiscono man mano che ci si allontana dai valori centrali) viene approssimato in statistica dalla cosiddetta *distribuzione normale* (la classica curva a campana).
- Osservando gli spettri di frequenza, **vi sembra che la distribuzione dei livelli di frequenza nei corpora segua un andamento normale** (valori che crescono man mano che ci si avvicina al centro dello spettro, e poi diminuiscono di nuovo)?
- **In caso contrario, che tipo di andamento si osserva invece negli spettri di frequenze?**

2 Frequenza degli n-grammi [2 punti]

- **Quante parole (unigrammi), bigrammi e trigrammi capitano una sola volta nel Brown corpus? E in proporzione** (rispetto al numero totale di parole, bigrammi e trigrammi *distinti*)?[1]

[1] NB: Il numero di n-grammi distinti si ottiene contando le righe di una lista di frequenza, in cui a ciascuna riga corrisponderà un n-gramma distinto.

- **Quante parole, bigrammi e trigrammi capitano 10 o più volte nel Brown corpus? E in proporzione?**

3 Frequenza media [2 punti]

- La frequenza media di un corpus si ottiene dividendo il numero totale di parole (tokens) nel corpus per il numero di parole distinte (types).
- Dato un corpus tokenizzato e una lista delle parole distinte che capitano nel corpus (per esempio, una lista di frequenza), il numero di tokens e il numero di types si ottengono facilmente con `wc`.[2]
- Per esempio, il Brown contiene 1,008,057 tokens e 49,613 types; dunque la frequenza media di una parola nel Brown è di 20.32 tokens.
- (Questo significa che nel Brown c'è un gran numero di parole che hanno una frequenza di (più o meno) 20 tokens?)
- Il file *brown.ran.tok* contiene il Brown tokenizzato e randomizzato (cioè con le parole disposte in ordine casuale).[3]
- A partire da questa versione del Brown, create 4 sottocorpora: uno che contenga le prime 200,000 parole, uno che contenga le prime 400,000 parole, uno che contenga le prime 600,000 parole e uno che contenga le prime 800,000 parole.
- **Calcolate la frequenza media nei 4 sottocorpora** (quella per il corpus intero ve l'ho data io qui sopra).
- **Cosa osservate?**
- Uno studioso deve decidere se due corpora di scritti anonimi sono stati prodotti dallo stesso scrittore. Il corpus A contiene 250,000 parole. Il corpus B contiene 1,000,000 di parole.
- Lo studioso verifica che la frequenza media nel corpus A è molto inferiore di quella nel corpus B, e dunque conclude che i due corpora devono essere stati prodotti da due autori diversi, con caratteristiche stilistiche distinte.
- **La conclusione dello studioso è legittima? Perché?**

[2] E se la lista di parole distinte non è già disponibile, la si può creare dal corpus tokenizzato con `sort` e `uniq`.

[3] Siccome vi chiedo di creare dei sottocorpora, ho randomizzato l'ordine delle parole in modo che in tutti i sottocorpora capitino un po' di parole da tutte le sezioni del Brown.