Ricolingua
Progetto di Laboratorio di Programmazione
Aprile 2008

1 Analisi testuale mediante n-grammi

In molte applicazioni di manipolazione testuale è importante riuscire a determinare in modo automatico e con buona affidabilità in che lingua è redatto un testo. Fra le tecniche esistenti, una delle più popolari è quella basata sugli n-grammi, e verrà brevemente descritta in questa sezione.

Nel seguito, per *carattere* intendiamo un carattere alfabetico minuscolo o uno spazio; i caratteri alfabetici maiuscoli sono preventivamente convertiti in minuscoli, qualunque carattere non alfabetico viene interpretato come uno spazio, e la presenza di due o più spazi consecutivi viene considerata equivalente a un singolo spazio. Ad esempio, il testo

```
La mamma di Carlo, lo sai, ha comprato 5 mele.
```

viene considerato come se fosse:

```
la mamma di carlo lo sai ha comprato mele
```

Un *n-gramma* è una sequenza di n caratteri consecutivi che compaiono nel testo. Ad esempio, i 2-grammi che compaiono in `La mamma di Carlo, lo sai, ha comprato 5 mele.` sono indicati in Figura 1.

Indichiamo con G_n l'insieme dei possibili n-grammi: siccome un n-gramma è una sequenza di n caratteri, e i caratteri possibili sono 27 (le 26 lettere dell'alfabeto più 1 spazio), risulta $|G_n| = 27^n$.

Dato un testo T, e fissato un intero positivo n, considerate la funzione $f_T : G_n \to [0, 1]$ che associa a ogni n-gramma $g \in G_n$ la sua frequenza relativa nel testo T (cioè, il numero di volte che g compare in T, divisa per il numero complessivo di n-grammi in T). Naturalmente, per definizione,

$$\sum_{g \in G_n} f_T(g) = 1.$$

La funzione f_T viene detta *distribuzione degli n-grammi nel testo T*.

2-gramma	numero di occorrenze
la	1
a-	3
-m	2
ma	2
am	1
mm	1
-d	1
di	1
i-	2
-c	2
ca	1
ar	1
rl	1
lo	1
o-	2
-s	1
sa	1
ai	1
-h	1
ha	1
co	1
om	1
mp	1
pr	1
ra	1
at	1
to	1
me	1
el	1
le	1

Figura 1: Frequenza assoluta dei 2-grammi nel testo "La mamma di Carlo, lo sai, ha comprato 5 mele". Per leggibilità indichiamo lo spazio con un trattino. Tutti gli altri possibili 2-grammi che non compaiono nella tabella hanno frequenza 0.

Se T è abbastanza lungo e n è abbastanza grande, f_T risulterà di fatto indipendente dal testo, e dipenderà essenzialmente solo dalla lingua in cui il testo è scritto.

La cosa è già evidente nel semplice caso $n = 1$: in tal caso f_T misura la frequenza delle singole lettere che compaiono nel testo. Ora, è abbastanza evidente che, per esempio, in un testo italiano $f_T(y)$ sarà molto piccolo (poiché la lettera "y" non compare quasi mai nella lingua italiana), mentre in un testo inglese $f_T(y)$ sarà significativa.

Supponete di avere k testi $T_1, T_2, \ldots, T_k$ di ciascuno dei quali conoscete la lingua (p.es.: sapete che T_1 è un testo inglese, T_2 un testo italiano, T_3 un testo danese ecc.), e supponete di avere un testo T scritto in una lingua ignota (una delle k precedenti).

Calcolate i valori $d_1, d_2, \ldots, d_k$ dove

$$d_i = \sum_{g \in G_n} \left(f_T(g) - f_{T_i}(g)\right)^2 .$$

In pratica, d_i è la somma dei quadrati delle differenze fra il valore della distribuzione degli n-grammi in T_i e in T, sommata su tutti gli n-grammi possibili. Ci aspettiamo che d_i sia tanto più piccolo quanto più sono simili le lingue in cui T e T_i sono scritti.

Quindi, concluderemo che T è scritto nella lingua i per cui d_i è minimo.

2 Il programma

Si chiede di scrivere un programma di riconoscimento linguistico che utilizzi precisamente la tecnica descritta. Il programma deve avere l'esatto comportamento seguente:

- supponete di avere un certo numero di file di testo, ciascuno dei quali è fatto come segue:
  - sulla prima riga, contiene una stringa (di al massimo 100 caratteri) che rappresenta il nome della lingua in cui il file è scritto;
  - le restanti righe contengono del testo arbitrario scritto in quella lingua;
- il programma viene lanciato passandogli sulla riga di comando, in quest'ordine: il numero n (cioè, la dimensione degli n-grammi da considerare), una sequenza di nomi di file di esempio, in cui l'ultimo file, pur avendo la stessa struttura dei precedenti, è scritto in una lingua sconosciuta (e porta, sulla prima riga, la stringa "Sconosciuta");
- ad esempio, se il programma si chiama `progr`, un possibile esempio di esecuzione sarà:

```
./progr 3 italiano.txt inglese.txt danese.txt sconosciuto.txt
```

dove il file `italiano.txt` potrebbe essere fatto così:

```
Italiano
Nel mezzo del cammin di nostra vita
mi ritrovai per una selva oscura,
ché la diritta via era smarrita.
```

e gli altri file, mutatis mutandis, hanno una struttura del tutto simile;

- il programma deve stampare unicamente il nome della lingua più simile calcolata come discusso sopra; a parità di similarità, dovrà stampare la prima lingua più simile (nell'ordine in cui i file sono specificati sulla riga di comando).

Scegliete come strutturare il programma, tenendo conto dei seguenti vincoli:

- potete assumere che $n < 5$;
- non potete fare alcuna assunzione sul numero dei file di esempio (ma potete supporre che siano almeno due);
- non potete fare alcuna assunzione sulla lunghezza dei file, né sulla lunghezza delle righe (eccettuata la prima riga, che è al massimo lunga 100 caratteri);
- potete assumere che le lingue in cui sono scritti i file di esempio siano tutte distinte (cioè, ad esempio, non ci possono essere due file che portano sulla prima riga la stringa "Italiano");
- i file possono contenere caratteri alfabetici maiuscoli e/o caratteri non alfabetici, che dovranno essere trattati come specificato nel testo; potete comunque assumere che ogni file contenga almeno n caratteri alfabetici.

Sul sito sono forniti alcuni file di esempio, scritti in varie lingue, con cui potete provare i vostri programmi.

3 Modalità di consegna

Il programma:

- dovrà essere scritto in ANSI C (cioè, dovrà compilare senza errori con l'opzione -ansi);
- dovrà essere costituito da un singolo file sorgente;

- dovrà essere opportunamente commentato;
- dovrà essere accompagnato da un documento (PDF) di descrizione del programma.

La consegna del programma dovrà avvenire come segue:

- dovrete inviare un'e-mail a `boldi@dsi.unimi.it` avente come Subject (titolo): "Consegna LabProgr NOMEGRUPPO" (dove, al posto di NOMEGRUPPO, dovrete scrivere il nome con cui il vostro gruppo si è preiscritto);
- l'e-mail dovrà avere *come allegati* il file sorgente (non compresso) e il file PDF con la descrizione del programma.

Consegne che non rispettino queste modalità non verranno prese in considerazione. Per qualunque domanda, scrivete un'e-mail a `boldi@dsi.unimi.it` avente come Subject (titolo): "Chiarimento per LabProgr NOMEGRUPPO".