Progetto di Basi di Dati II
La legge di Zipf

Luca Cinti Lorenzo Masetti Paolo Panconi

10 luglio 2002

Ho chiesto a Lotaria se ha già letto alcuni miei libri che le avevo prestato. M'ha detto di no, perchè qui non ha a disposizione un elaboratore elettronico.

M'ha spiegato che un elaboratore deliberatamente programmato può leggere un romanzo in pochi minuti e registrare tutti i vocaboli contenuti nel testo, in ordine di frequenza.

- Posso così disporre subito d'una lettura già portata a termine, - dice Lotaria, - con un'economia di tempo inestimabile. [...] La lettura elettronica mi fornisce una lista delle frequenze, che mi basta scorrere per farmi un'idea dei problemi che il libro propone al mio studio critico. Naturalmente alle frequenze più alte sono registrate delle sfilze di articoli, pronomi, particelle, ma non è là che soffermo la mia attenzone. Punto subito sulle parole più ricche di significato, che mi possono dare un'immagine del libro abbastanza precisa.

Italo Calvino, *Se una notte di inverno un viaggiatore.*

Indice

1 Introduzione

1.1 Zipf e la sua legge

George Kingsley Zipf (1902-1950), professore di linguistica presso l'Università di Harvard per circa venti anni, fu personaggio alquanto inusuale e controverso secondo quello che riportano molte delle fonti che trattano dei suoi lavori e dei suoi studi. Già filologo, ma anche statistico, matematico, economista, addirittura pioniere di quella scienza che oggi tutti riconoscono nella definizione di computer science, può semplicemente essere definito un *statistical human ecologist*, come lui stesso amava presentarsi. Alcuni documenti rintracciabili in rete affiancano il suo nome a quello del contemporaneo Turing, ricordando Zipf soprattutto per aver lasciato una traccia indelebile nella storia della scienza del ventesimo secolo, e per lo straordinario eclettismo culturale tipico del genio.

Del suo *Human Behavior and the Principle of Least Effort* (1949-1965), si legge "...è uno di quei libri in cui lampi di genio, che si proiettano in molte direzioni, sono continuamente intramezzati da una miriade di nozioni selvagge e da una stravaganza gratuita. Si possono trovare parti in cui, discorrendo delle forme degli organi sessuali, viene giustificata l'annessione dell'Austria alla Germania in quanto produrrebbe una migliore aderenza ad una formula matematica, altre in cui complesse e innovative teorie statistico-matematiche vengono brillantemente esposte e sostenute" (*The Fractal Geometry of Nature* Benoit Mandelbrot, 1983).

Ciò che ha reso il professor Zipf famoso è indubbiamente la formulazione della legge che regola la frequenza delle occorrenze di un qualche evento (P_i) in funzione del rango (i), dove quest'ultimo è determinato in base all'ordinamento decrescente sulla frequenza stessa. Tale legge, conosciuta appunto come la *legge di Zipf*, afferma che

$$f(P_i) = \frac{c}{i} \tag{1}$$

dove $f(P_i)$ è il numero delle occorrenze osservate dell'evento P_i, i è il posizionamento di P_i nell'ordinamento in base alla frequenza e c una costante reale positiva. Più in generale la precedente può essere riformulata come

$$f(P_i) = \frac{c}{i^{(1-\theta)}} \tag{2}$$

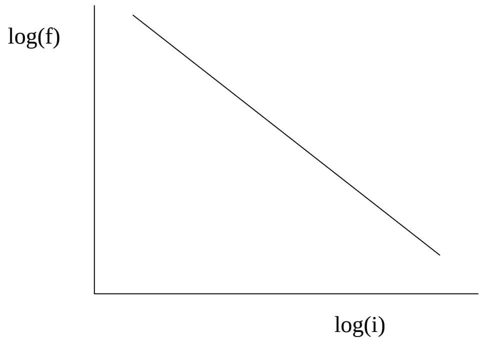

Figura 1: L'andamento lineare su doppia scala logaritmica della legge di Zipf.

con θ costante positiva prossima a 0 (per $\theta = 1$ si rientrerebbe nella distribuzione casuale, cioè nell'equiprobabilità). Se tale legge viene riportata in grafico con doppia scala logaritmica, si ottiene una retta con pendenza vicina a -1. I campi di applicabilità della legge di Zipf sono svariati, e sono anche esistite tendenze di pensiero che l'hanno addirittura proposta come la controparte della distribuzione gaussiana nell'ambito della scienze sociali. In realtà come le distribuzioni nelle scienze naturali non seguono sempre un andamento gaussiano, tanto meno le distribuzioni delle scienze sociali o degli eventi correlati all'attività umana, possono essere fortemente inquadrate negli andamenti iperbolici della legge di Zipf. Esistono tuttavia casi in cui i risultati previsti dalla legge di Zipf sono fortemente verificati dai dati osservati. Fra i più famosi senz'altro vi sono quelli che considerano le frequenze delle parole negli scritti, o quelli che analizzano la distribuzione della popolazione nelle varie città di uno stato. Esempi questi ultimi approfonditamente analizzati e documentati dallo stesso Zipf nei suoi più famosi scritti. In campo economico la legge di Zipf si adatta bene a rappresentare la situazione di produttività delle varie compagnie che operano in un medesimo settore. Se si rappresenta la produttività di una azienda in funzione del sua posizione di classifica si ottiene l'andamento iperbolico della legge di Zipf. Questo fatto era stato, del resto, già osservato alla fine del'800 dall'economista Pareto.

1.2 Finalità del progetto

In questo elaborato ci siamo limitati a verificare la legge per quello che riguarda la frequenza delle parole in alcuni testi letterari. In particolare vengono raccolte tutte la parole che compogono un testo, e ordinate decrescentemente in base alla loro frequenza di comparizione. Per gestire le parole che presentano un ugual numero di occorrenze si è proceduto con due strategie differenti. Secondo la prima, a cui ci si riferisce con il termine *rango non bloccato*, il numero d'ordine viene aumentato anche per i termini con uguale frequenza, secondo l'ordinamento lessicale. In base alla seconda, denominata *rango bloccato*, si mantiene il medesimo rango per tutte le parole con uguale frequenza.

La legge di Zipf, che può anche essere espressa nella forma

$$f_i \propto i^{-\alpha} \tag{3}$$

risulta essere verificata da quasi tutta la letteratura di diverse lingue. Inoltre risulta particolarmente accurata quando ci si limita a considerare le parole che hanno un rango non troppo alto nè troppo basso, cioè se ci concentra sulle frequenze intermedie tralasciando i valori estremi. Si può osservare inoltre, che per testi particolarmente ampi, ci può essere un forte discostamento dalla prima potenza, cioè dal valore di $\alpha = 1$, e che quando si vanno a considerare i valori di rango alto c'è una tendenza verso la potenza seconda, cioè meglio si adatterebbe un andamento con $\alpha \sim 2$.

L'evoluzione dei calcolatori ha senz'altro agevolato lo studio statistico dei testi, rilevando la stretta aderenza alla distribuzione di Zipf; tuttavia tale caratteristica rimane ancora inspiegata all'interno di un modello del comportamento umano, anche se risulta evidente che la legge di Zipf vale per le forme di comunicazione umana ed è quindi legata al fatto che il testo considerato abbia un significato (la legge di Zipf ovviamente non vale per un testo casuale), ed è tuttora impossibile generalizzare tale legge al fine di caratterizzare e differenziare le scritture umane da altre forme di testo.

2 Realizzazione dell'analizzatore di testi

Il primo problema da risolvere per verificare la legge di Zipf è stato quello di implementare un programma che contasse le frequenze delle parole e ordinasse questi valori in base alla frequenza per poi determinare il rango.

Abbiamo realizzato un analizzatore che scandisce un file di testo isolando le singole parole. L'analizzatore individua le parole leggendo un carattere alla volta e fermandosi quando incontra un carattere che appartiene a un insieme specificato di *stop characters* (chiamati così per analogia con le *stop words*). L'insieme dei caratteri da ignorare comprende lo spazio e tutti i segni di punteggiatura.

Ogni parola così trovata viene inserita (senza distinguere tra lettere minuscole e maiuscole) in una tabella hash che associa a ogni parola la sua frequenza. Ogni volta che si scandisce una parola si effettua una ricerca nella tabella hash e, se la parola non era presente viene inserita con frequenza 1, altrimenti la sua frequenza viene incrementata.

Una volta analizzato l'intero file e costruita la relativa tabella, si scandisce la tabella hash inserendo la parola e la sua frequenza in un albero ordinato in base alla frequenza. Visitando l'albero in ordine è così possibile ottenere la lista delle parole con relativa frequenza ordinate in base alla frequenza e quindi associare a ogni parola un rango.

Questa parte del progetto è scritta in Java, la scelta di questo linguaggio è stata dettata dalla semplicità di trattare in questo linguaggio queste strutture dati complesse.

Questo è il codice java che realizza l'aggiornamento della tabella hash ogni volta che si inserisce una parola.

---

addword - *Per aggiornare la frequenza di una parola una parola nella tabella hash. La tabella hash associa a una stringa un oggetto di classe* Counter *che contiene un intero (inizializzato a 1 al momento della creazione dell'oggetto) che viene incrementato ogni volta che la parola viene incontrata*

```java
public void addWord(String word) {
    if ( _hash.containsKey(word) )
        ( (Counter) _hash.get(word) ).i++;
    else
        _hash.put(word,new Counter());
}
```

Questo metodo riempie l'albero con gli elementi della tabella hash.

---

sortByFreq - *Scandisce la tabella hash e inserisce gli elementi nell'albero ordinati per frequenza. L'albero contiene oggetti di classe* CountedWord *che sono formati da una stringa e da un intero che memorizza la frequenza.*

```java
public void sortByFreq() {
    _tree.clear();
    Iterator i = _hash.entrySet().iterator();
    while (i.hasNext()) {
        Map.Entry e = (Map.Entry) i.next();
        _tree.add(new CountedWord((String) e.getKey
            (),((Counter) e.getValue()).i));
    }
```

}

Ad esempio nella tabella 2 sono riportati i risultati relativi alle 30 parole più frequenti ne "La coscienza di Zeno". La legge di Zipf nella sua formulazione originale prevede che il prodotto $rango \cdot frequenza$ (riportato nell'ultima colonna) sia costante. Da notare che, dato che questa analisi non richiede l'uso di *stop words*, le 30 parole più frequenti sono articoli, preposizioni, pronomi e non parole con un significato rilevante.

3 Analisi

3.1 Testi analizzati

Abbiamo utilizzato il programma java presentato precedentemente per analizzare tre testi letterari, uno in italiano antico (la *Divina Commedia*), uno in italiano moderno (*La Coscienza di Zeno*) ed uno in in inglese (*Ivanhoe*).

I tre testi hanno le caratteristiche descritte nella seguente tabella

Testo	Numero Totale di Parole	Parole Distinte
Divina Commedia	105688	13182
La Coscienza di Zeno	145444	13799
Ivanhoe	195142	13236

3.2 Metodologia di analisi

Per valutare l'aderenza dei dati sperimentali alle due leggi abbiamo utilizzato il metodo dei minimi quadrati.

In particolare per la legge di Zipf classica, (1), si può facilmente ricavare qual è il valore della costante che minimizza i quadrati delle distanze. Infatti, detti x_i i valori del rango e y_i i valori delle frequenze, la quantità da minimizzare è

$$\sum_{i=1}^{n}(y_i - \frac{c}{x_i})^2$$

derivando rispetto a c si ottiene

$$\sum_{i=1}^{n} -\frac{2}{x_i}(y_i - \frac{c}{x_i})$$

Rango	Parola	Frequenza	Prodotto
1	di	4992	4992
2	che	4232	8464
3	e	3412	10236
4	non	3019	12076
5	la	2904	14520
6	il	2319	13914
7	a	2150	15050
8	per	2029	16232
9	un	1918	17262
10	mi	1735	17350
11	io	1614	17754
12	in	1437	17244
13	era	1376	17888
14	una	1253	17542
15	si	1172	17580
16	ma	1162	18592
17	con	1155	19635
18	l	1108	19944
19	le	1105	20995
20	da	1092	21840
21	mia	875	18375
22	piú	842	18524
23	mio	827	19021
24	me	719	17256
25	del	709	17725
26	anche	696	18096
27	ch	693	18711
28	perché	656	18368
29	della	617	17893
30	essa	614	18420

Tabella 1: Le 30 parole più frequenti de ”La Coscienza di Zeno”

ed eguagliando tale derivata a 0 risulta

$$c = \frac{\sum \frac{y_i}{x_i}}{\sum \frac{1}{x_i^2}} \tag{4}$$

Per stimare i parametri della legge di Zipf generalizzata (2) abbiamo utilizzato un metodo iterativo fornito dal Matlab (la function *fminsearch*) per trovare il minimo di una funzione in più variabili. In particolare il metodo dei minimi quadrati prevede, per la legge di Zipf generalizzata, di trovare il minimo della funzione

$$f(c, \theta) = \sum_{i=1}^{n} (y_i - \frac{c}{x_i^{1-\theta}})^2 \tag{5}$$

Il metodo iterativo ha bisogno di una stima iniziale dei parametri: abbiamo quindi posto inizialmente c uguale al valore di c nella legge di Zipf classica stimato con la (4) e $\theta = 0$ in modo che il metodo iterativo faccia convergere la legge di Zipf classica a quella generalizzata.

Per stimare l'incertezza sui dati o, in altre parole, quanto i dati si discostano dalle curve delle due leggi abbiamo utilizzato le stime delle deviazioni standard σ_y che sono, rispettivamente per la legge di Zipf classica e per quella generalizzata

$$\sigma_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \frac{c}{x_i})^2 \tag{6}$$

$$\sigma_y^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \frac{c}{x_i^{1-\theta}})^2 \tag{7}$$

I fattori $(n-1)$ e $(n-2)$ nelle definizioni (6) e (7) sono giustificati dal fatto che inserendo al posto dei parametri c e θ le loro stime si riduce il numero di gradi di libertà rispettivamente a $n-1$ e a $n-2$. Comunque per valori di n alti (molte decine di migliaia) questa differenza è in pratica trascurabile.

function fitminimiquadr - *Calcola i coefficienti della legge di Zipf e della legge di Zipf generalizzata e restituisce le deviazioni standard*

```
% function [dev_zipf, dev_zipf_gen,czipf,c,theta]=
    fitminimquadr(xy,high,titolo)
%
% Prende i dati sperimentali in xy, l'ascissa massima
    del
% grafico e il titolo del grafico
```

```
%
% Calcola le approssimazioni secondo la legge di zipf e
% la legge di zipf generalizzata
% Restituisce le deviazioni standard
%dei dati rispetto alle due
% diverse leggi
function [dev_zipf, dev_zipf_gen,czipf,c,theta]=
   fitminimquadr(xy,high,titolo)
% calcola c secondo la formula dei minimi quadrati
quoz = xy(:,2) ./ xy(:,1);
unosuxiquad = 1./(xy(:,1).*xy(:,1));
czipf = sum(quoz) / sum(unosuxiquad);

% Approssima la legge di Zipf generalizzata
% partendo da c uguale alla legge di Zipf
% e da theta = 0
[c,theta] = approssima(xy,czipf,0);
% la function disegna restituisce i punti delle
% curve di zipf (z) e zipf generalizzata (zg)
% valutati nelle ascisse
[z,zg]=disegna(xy,c,theta,czipf,high,titolo);
% Calcolo delle deviazioni standard
N = length(xy(:,1));

errz = ((xy(:,2)-z).^2)/(N-1);
errzg = ((xy(:,2)-zg).^2)/(N-2);
dev_zipf = sqrt(sum(errz));
dev_zipf_gen= sqrt(sum(errzg));
```

3.3 Risultati dell'analisi dei dati

Riportiamo i risultati dell'analisi svolta sui tre testi citati precedentemente. Per tutti e tre la legge di Zipf è discretamente verificata. I valori di θ (riportati nelle legende dei grafici) nella legge di Zipf generalizzata (2) non sono lontani da 0: sono compresi tra $0,13$ e $0,26$.

	Zipf classica	Zipf generalizzata
Rango non bloccato	31,54	18,76
Rango Bloccato	249,64	135,29

Tabella 2: Deviazioni Standard per i dati sperimentali relativi alla ”Divina Commedia”

Divina Commedia

Nelle figure 2 e 3 abbiamo riportato l’andamento dei dati sperimentali relativi alle frequenza delle parole nella *Divina Commedia* e le due curve che approssimano i dati. In figura 2 il rango non è bloccato per le parole con la stessa frequenza, in figura 3 il rango è bloccato.

Il grafico in figura 1 riporta solo la prima parte della curva (le 250 parole più frequenti) ma il rango arriva fino a 13182 (il numero di parole distinte nel testo). Considerando il rango bloccato invece il rango massimo è 204 e il grafico riporta tutta la curva.

La deviazione standard risulta ovviamente minore per la legge di Zipf generalizzata che ottimizza due parametri.

Da notare che la definizione di rango ”non bloccato” risulta migliore, questo perché le parole con frequenze basse e rango molto alto sono approssimate molto bene dalla curva approssinante.

La Coscienza di Zeno

I dati sono rappresentati in figura 4. Si considera solo il rango non bloccato. La parte di grafico disegnata riguarda valori del rango da 1-200.

Per questo testo abbiamo inoltre verificato l’affermazione che la legge di Zipf è meglio verificata per le parole che non hanno frequenze troppo alte o troppo basse. Nel romanzo di Italo Svevo la parola con frequenza più alta è “di” con 4992 occorrenze, la seconda (“che”) è presente 4232 volte; si vede quindi che per questi primi dati la legge di Zipf non è ben verificata (la frequenza della seconda parola avrebbe dovuto essere la metà della prima). Se invece consideriamo solo le parole con frequenze che variano da 10 a 25 (il cui rango è compreso tra 600 e 1493) si ottiene un’approssimazione molto migliore e valori delle deviazioni standard molto bassi. La situazione è illustrata in figura 5. Le linee orizzontali evidenziano l’andamento dei dati quando si usa il rango non bloccato: parole con la stessa frequenza hanno valori del rango appartenenti a un intervallo di valori successivi.

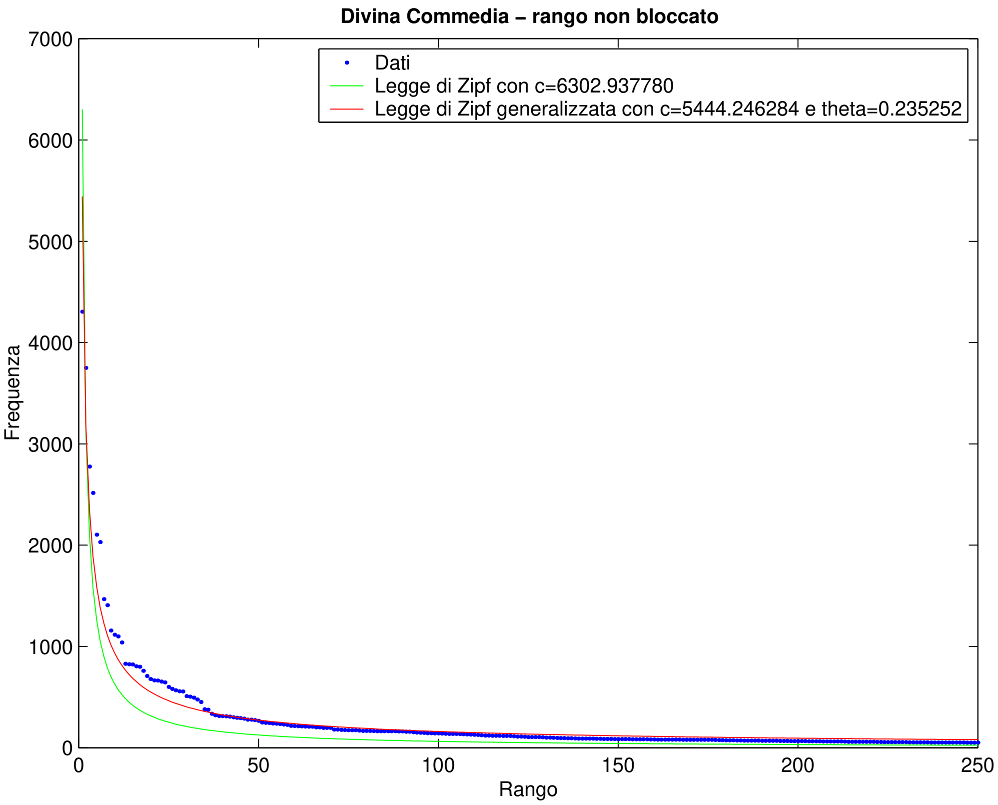

Figura 2: Grafico rango-frequenza delle parole della Divina Commedia. Rango non bloccato. Parte iniziale della curva (rango nell'intervallo 1-250)

	Zipf classica	Zipf generalizzata
Rango non bloccato	45,88	26,65
Rango non bloccato, parole con frequenze intermedie	0,471	0,365

Tabella 3: Deviazioni Standard per i dati sperimentali relativi a "La Coscienza di Zeno"

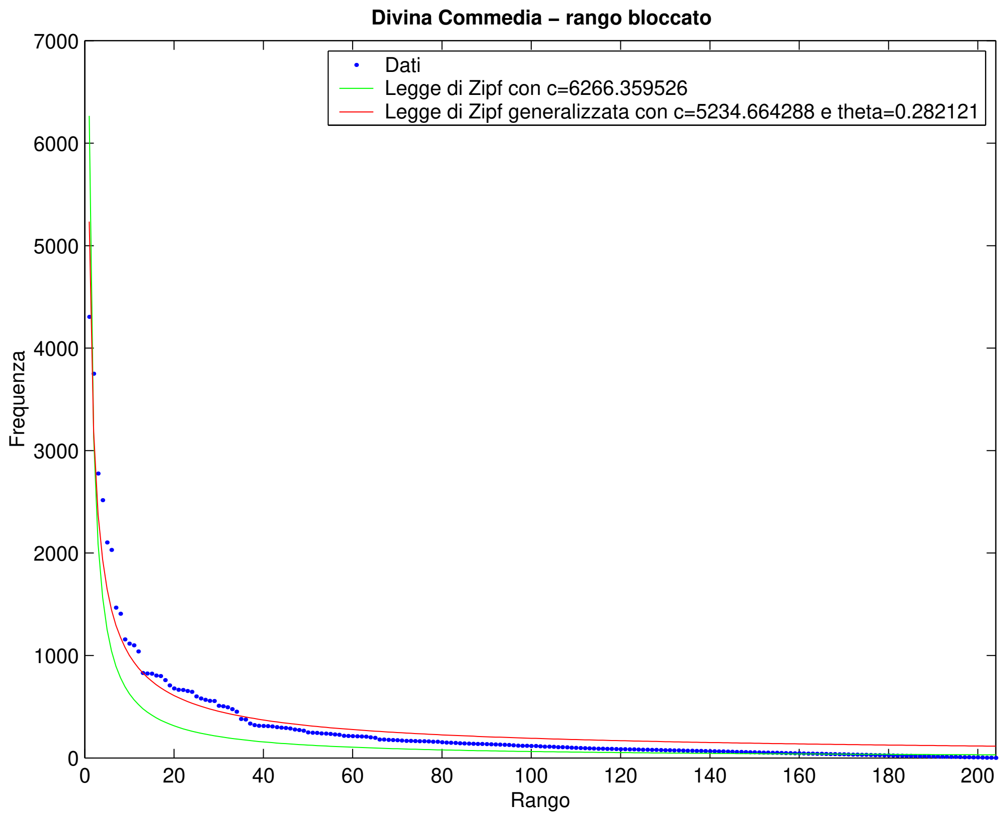

Figura 3: Grafico rango-frequenza delle parole della Divina Commedia. Rango bloccato. Curva completa (rango nell'intervallo 1-204)

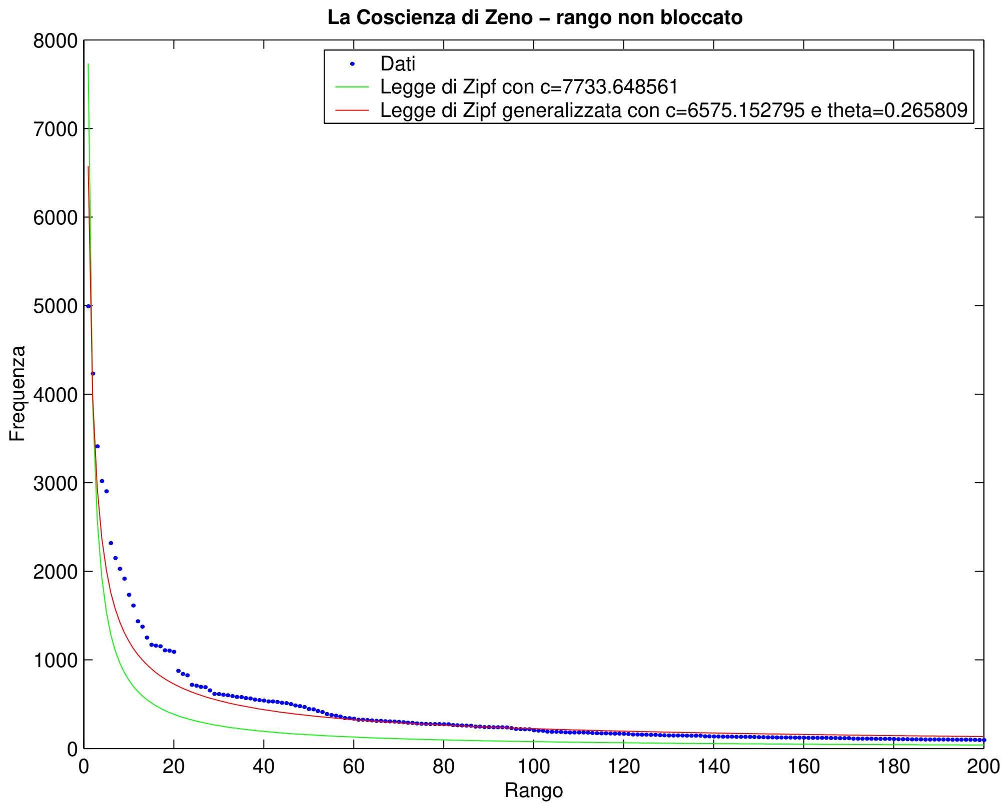

Figura 4: Grafico rango-frequenza delle parole de "La Coscienza di Zeno". Rango bloccato. Parte iniziale della curva (rango nell'intervallo 1-200)

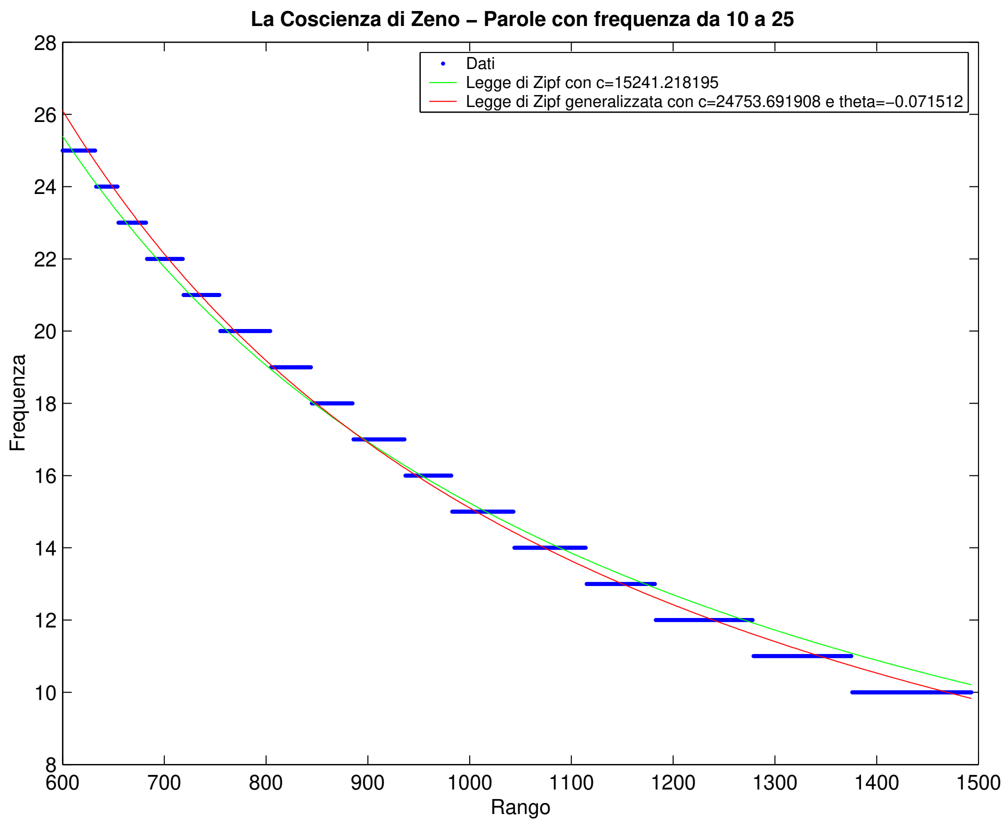

Figura 5: Grafico rango-frequenza dell'interpolazione delle parole con frequenze comprese tra 10 e 25 ne "La Coscienza di Zeno". Il rango varia da 600 a 1493.

	Zipf classica	Zipf generalizzata
Rango non bloccato	32,32	17,73

Tabella 4: Deviazioni Standard per i dati sperimentali relativi a ”Ivanhoe”

Ivanhoe

I dati sono rappresentati in figura 6, anche in questo caso abbiamo considerato solo il rango non bloccato. Il grafico è disegnato per le prime 100 parole più frequenti.

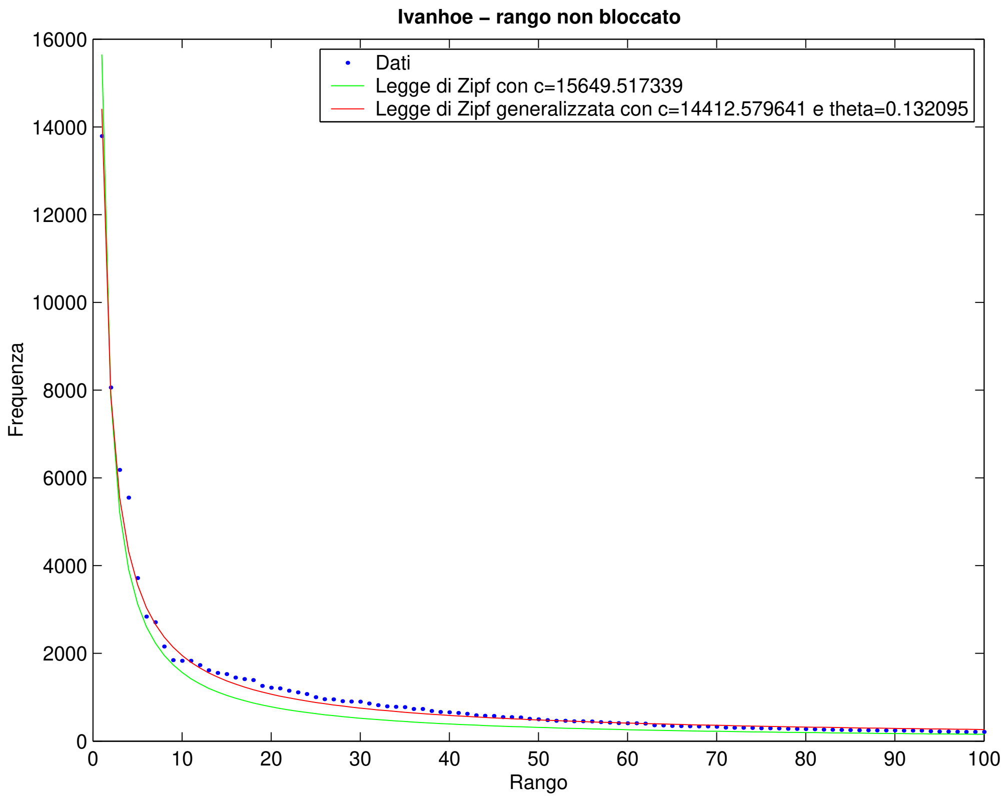

Figura 6: Grafico rango-frequenza delle parole di "Ivanhoe". Rango bloccato. Parte iniziale della curva (rango nell'intervallo 1-100)