Appunti Di Information Retrieval

Massimo Meneghello

March 2018

1 Introduzione

Per poter reperire dei documenti è necessario prima di tutto poter rappresentare il loro contenuto informativo in modo conciso. Questo passo è reso possibile dall'**analisi del testo** che deve poter essere effettuata in modo **automatico**, **rapido**, **affidabile** e **consistente**.
Esistono due principali approcci all'analisi del testo:

- l'approccio statistico;
- l'approccio linguistico.

Il primo di questi approcci è stato quello più utilizzato e studiato fin dagli albori della disciplina. Le prime osservazioni riguardanti un'analisi statistica dei testi sono state proposte da *Hans Peter Luhn*, prima, e in seguito formalizzate da *George Kingsley Zipf*.
Luhn si era reso conto infatti che:

- la distribuzione delle parole non è uniforme in un testo;
- poche parole compaiono molto di frequente;
- molte parole compaiono raramente.

La distribuzione è quindi **asimmetrica** - *skew distribution*. A Zipf si deve la formulazione di varie leggi empiriche che mettono in relazione la **frequenza** di una parola con la sua **forma** e il suo **significato**. La più nota è la legge che porta il suo nome, la **Legge di Zipf** appunto.

$$r \times f = costante \tag{1}$$

dove f è il valore della frequenza di una parola in un testo (o in un campione di testi) mentre r è il rango di quella parola dopo che tutte le parole sono state ordinate per frequenze decrescenti. Si ottiene quindi un andamento iperbolico della frequenza sul rango.
Si può anche pensare di riscrivere la legge di Zipf in termini probabilistici (in questo caso le frequenze assolute con cui una parola compare diventano le probabilità che una parola ha di comparire all'interno di un testo).

La scelta dei descrittori che permettono di **discriminare meglio il contenuto informativo di un testo** - *resolving power* - si basa sulle osservazioni di Luhn e sulla legge di Zipf.
Una volta fissate una soglia inferiore di cut-off - *lower cut-off* - e una soglia superiore di cut-off - *upper cut-off* - si considerano solamente i descrittori che hanno rango compreso tra queste due soglie. Operando in questo modo si escludono:

- i descrittori che hanno frequenze troppo elevate, chiamati anche *stop words*, che portano poca informazioni sul contenuto dei singoli documenti;
- i descrittori che hanno frequenze troppo basse, che invece costituiscono rumore.

2 Indicizzazione

3 Valutazione Dei Sistemi Di Reperimento

Per poter valutare un IRS è necessario disporre dei **giudizi di rilevanza**, ovvero dei valori associati a ciascuna coppia documento-topic. Esistono più metodi per la creazione di questi giudizi:

- **giudizi completi**, per ogni documenti si giudica la sua rilevanza relativamente ad ogni topic (lavoro troppo oneroso);
- **campionamento casuale**, i giudizi vengono assegnati soltanto ad un campione casuale di documenti, per ogni topic;
- **campionamento basato sugli esperimenti dei partecipanti**, è il metodo utilizzato da TREC ed ormai divenuto stardard per tutte le campagne di valutazione, noto anche come **pooling**.

Per poter dare una definizione formale del pooling è necessario introdurre alcuni concetti. Si definiscono:

$$D = \{d_1, ..., d_n\} \text{un insieme di documenti} \tag{2}$$
$$T = \{t_1, ..., t_m\} \text{un insieme di topic} \tag{3}$$

Dato un numero naturale $N \in \mathbb{N}^+$ detto *lunghezza della run*, una **run** è definita come

$$R : T \to D^N \tag{4}$$
$$t \mapsto \mathbf{r_t} = (d_1, ..., d_N) \tag{5}$$