Information Retrieval

Giuseppe Magazzù

2021 - 2022

Contents

Chapter 1

Introduction

1.1 Definizioni

1.1.1 Document

Metadata

1.1.2 Terms

I termini sono dei descrittori che vengono associati al testo

1.1.3 Stop Words

I termini che non sono significativi per la rappresentazione del testo (articoli, particelle, ...)

Chapter 2

Text Processing

Il text processing è una fase necessaria per pulire e preparare il testo.

2.1 Tokenization

La tokenization consiste nell'identificare e separare all'interno di un testo delle unità chiamate token. I token possono essere parole, frasi, simboli o n-grammi. Ogni token è un candidato a essere un termine significativo (index).

> e.g. "Text mining is to identify useful information"
>
> **Tokens**: "Text", "mining", "is", "to", "identify", "useful", "information"

Problemi:

- parole composte ("Hewlett-Packard" $\rightarrow$ "Hewlett", "Packard")
- numeri, date ("Mar. 12, 1991", "12/3/1991", "(800) 234-2333")
- problemi linguistici (parole composte, assenza di spazi, ...)

I token possono essere raggruppati in sequenze contigue di N elementi chiamate N-grammi.

> e.g. "Corpus is the collection of text documents."
>
> **Bigrammi**: "Corpus is", "is the", "the collection", "collection of", "of text", "text documents", "documents ."

La tokenization si può effettuare tramite espressioni regolari o metodi statistici.

2.2 Normalization

Ad un parola possono essere associati diversi token. La normalizzazione consiste nell'ottenere le classi di equivalenza dei token rimuovendo punti, trattini, accenti.

U.S.A. ⇔ USA
anti-aliasing ⇔ antialiasing
résumé ⇔ resume
15/10/2021 ⇔ 15 Ott 2021

Lemmatization

Le parole vengono ridotte alla loro forma base (lemma) tenendo in considerazione l'intero vocabolario della lingua e analizzando la parte del discorso.

e.g. "ladies" ⇒ "lady", "forgotten" ⇒ "forgot"

Stemming

Le parole vengono ridotte a una radice (stem) rimuovendo le flessioni tramite l'eliminazione dei caratteri non necessari.

e.g. "automate(s)", "automation", "automatic" ⇒ "automat"

Case folding

Tutte le parole vengono convertite in lowercase a parte alcune eccezioni.

Thesaurus and Soundex

Un thesaurus (tesauro) è una risorsa linguistica generata manualmente da essere umani in cui è possibile esprimere relazioni tra parole (e.g. gerarchie, sinonimi, ...).

Soundex è un algoritmo fonetico che permette di rappresentare correttamente diverse parole omofone nonostante differenze di ortografia usando delle euristiche fonetiche.

2.3 Stop Words Removal

Le **stop words** sono le parole più frequenti all'interno di un testo che possono essere rimosse senza perdere il significato. Queste parole essendo presenti in più documenti non portano informazioni utili per distinguerli.

Esistono delle liste di **stop words** definite in base alla lingua che possono essere usate per la rimozione.

I web search engine non effettuano la rimozione delle **stop words** perché sono necessarie per alcune ricerche.

Chapter 3

Text Representation

In un sistema di information retrieval i documenti devono essere rappresentati in un formato interno e ordinati per essere indicizzati.

3.1 Bag Of Words

Un modo semplice per rappresentare un testo è una matrice in cui sulle righe ci sono termini estratti dal corpus (vocabolario) e sulle colonne i documenti.

La **Bag Of Words (BOW)** è una rappresentazione del testo che descrive le occorrenze di parole in un documento.

Incidence Matrix: specifica la presenza di un termine in un ogni documento.

Ogni documento può essere rappresentato da un insieme di termini o da un vettore binario.

	Doc1	Doc2	Doc3	Doc4
Term1	1	1	1	0
Term2	0	1	1	1
Term3	0	0	1	0

Rappresentazione di Doc1

$R1 = \{\text{Term1}, \text{Term2}, \text{Term3}\}$
$R1 = \langle 1, 0, 0 \rangle$

Count Matrix: specifica la numero di occorrenze di un termine in ogni documento.

Un documento viene rappresentato da un vettore di occorrenze.

	Doc1	Doc2	Doc3	Doc4
Term1	57	57	71	133
Term2	4	34	17	92
Term3	232	2	10	293

Rappresentazione di Doc1

$R1 = \langle 157, 4, 232 \rangle$

Le rappresentazioni vettoriali non considerano l'ordine delle parole nel testo.

Bag Of Words con N-grammi
Pro: cattura le dipendenze locali e l'ordine
Contro: incrementa la frequenza delle parole

3.2 Zipf's Law

Descrive la frequenza di un evento (parola) in un insieme in base al suo rank.

rank: posizione di un termine nell'ordine decrescente di frequenza dei termini in tutta la collezione.

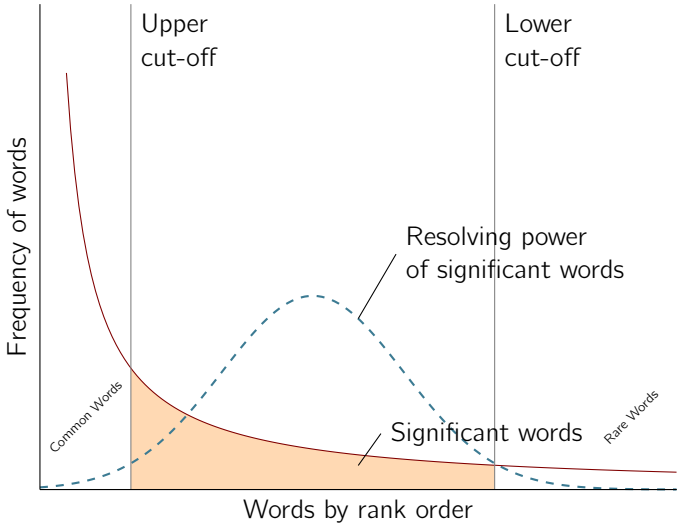
La frequenza di una parola w, f(w) è proporzionale a 1/r(w).

$f \propto \frac{1}{r} \quad \Rightarrow \quad f \cdot r = k$ (costante)

$P_r = \frac{f}{N} = \frac{A}{r}$ probabilità del termine di rank r, $A = \frac{k}{N} \approx 0.1$

3.3 Luhn's Analysis

Generalmente termini con frequenza molto alta e molto bassa sono inutili per discriminare i documenti.



L'abilità delle parole di discriminare il contenuto di un documento è massimo nella posizione tra i due livelli di cut-off.

Vogliamo assegnare dei pesi ai termini.

- corpus-wide: alcuni termini portano più informazione riguardo al documento
- document-wide: non tutti i termini sono ugualmente importanti

TF (Term Frequency): within document
IDF (Inverse Document Frequency): whole collection

Il peso di un termine deve essere proporzionale a TF e inversamente proporzionale a IDF

$tf_{t,d}$: numero di occorrenze del termine t nel documento d

$w_{t,d} = tf_{t,d}/max_{ti}tf_{ti,d}$

df_t document frequency: numero di documenti che contiene t

$df_t \leq N$

definiamo la inverse document frequency (idf)

$idf_t = log(N/df_t)$

tf-idf weight
$w_{t,d} = tf_{t,d}/max_{ti}tf_{ti,d} * log(N/df_t)$