

# Information Retrieval

Giuseppe Magazzù

2021 - 2022

# Contents

|          |                                      |          |
|----------|--------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                  | <b>1</b> |
| 1.1      | Definizioni . . . . .                | 1        |
| 1.1.1    | Document . . . . .                   | 1        |
| 1.1.2    | Terms . . . . .                      | 1        |
| 1.1.3    | Stop Words . . . . .                 | 1        |
| <b>2</b> | <b>Text Processing</b>               | <b>2</b> |
| 2.1      | Tokenization . . . . .               | 2        |
| 2.2      | Normalization . . . . .              | 2        |
| 2.3      | Stemming and Lemmatization . . . . . | 2        |
| 2.4      | Stop Words Removal . . . . .         | 3        |
| <b>3</b> | <b>Text Representation</b>           | <b>4</b> |
| 3.1      | Bag Of Words . . . . .               | 4        |
| 3.2      | Zipf's Law . . . . .                 | 5        |
| 3.3      | Luhn's Analysis . . . . .            | 5        |

# 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

### 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 o simboli. Ogni token è un candidato a essere un termine significativo (index).

Issues:

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

La **Tokenization** si può effettuare tramite espressioni regolari o metodi statistici.

### 2.2 Normalization

Associare diversi possibili token alla stessa una parola. E' possibile ottenere le classi di equivalenza dei token rimuovendo punti, trattini, accenti.

- U.S.A. ⇔ USA
- anti-discriminatory ⇔ antidiscriminatory
- résumé ⇔ resume

Case folding: ridurre tutte le parole in lowercase a parte alcune eccezioni. Thesauri and soundex: ——— sinonimi e omonimi. in base a delle euristiche fonetiche è possibile ottenere delle classi di equivalenza per le misspelled

### 2.3 Stemming and Lemmatization

Lemmatization: Ridurre alla forma base ladies ⇒ lady, forgotten ⇒ forgot

Stemming: Ridurre una parola alla sua radice. automate, automation, automatic  
⇒ automat.

## 2.4 Stop Words Removal

Omettere parole molto frequenti inutili per la rappresentazione del testo. (Queste parole essendo presenti in più testi non portano informazione utili per distinguere i testi.) Esistono delle liste di Stop Words che possono essere usate per la rimozione. I Web Search Engine non effettuano la rimozione delle Stop Words perché sono necessarie per alcune ricerche.

N-Gram Le unità di testo che sono state estratte con la Tokenization possono essere unite in n-grammi.

N-grams: a contiguous sequence of N tokens from a given piece of text

## 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    | Rappresentazione di Doc1                            |
| Term2 | 0    | 1    | 1    | 1    | $R1 = \{\text{Term1}, \text{Term2}, \text{Term3}\}$ |
| Term3 | 0    | 0    | 1    | 0    | $R1 = \langle 1, 0, 0 \rangle$                      |

**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  | Rappresentazione di Doc1           |
| Term2 | 4    | 34   | 17   | 92   | $R1 = \langle 157, 4, 232 \rangle$ |
| Term3 | 232  | 2    | 10   | 293  |                                    |

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.

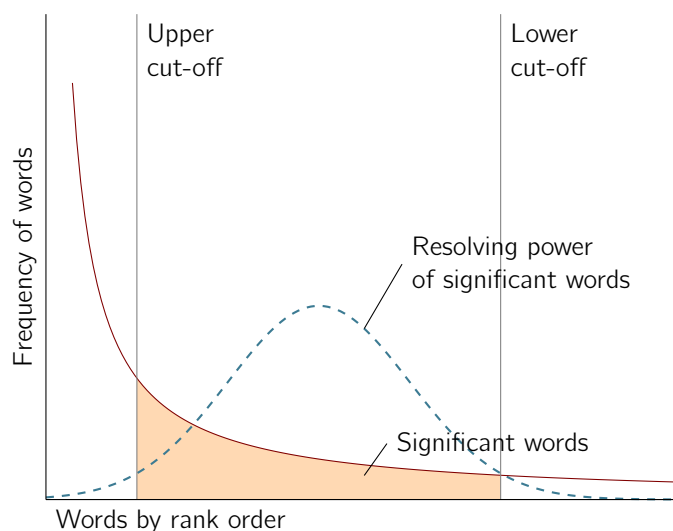
La frequenza di una parola  $w$ ,  $f(w)$  è proporzionale a  $1/r(w)$ .

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

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

## 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

$$w_{t,d} = tf_{t,d} / \max_{ti} tf_{ti,d} * \log(N/df_t)$$