# Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2014/2015

Corso di Laurea Magistrale in
Editoria e Giornalismo

# Esempio completo di analisi di testi, indicizzazione, ricerca e ranking

# Cosa vedremo

- Un esempio completo di
  - Indicizzazione
    - Costruzione del vocabolario di indicizzazione
    - Indicizzazione della collezione di documenti
  - Ricerca
  - Ranking dei risultati
    - Con modello booleano esteso



# La collezione di documenti

- D1:
  Pirro vince! È vera vittoria? Non si sa, ma intanto ha vinto.
- D2:
  Pirro stupisce tutti e vince. Torna la pace?
- D3:
  I romani sconfitti da Pirro
- D4:
  Pirro nasce nel 318a.C. e muore ad Argo nel 272a.C.

# Costruzione del vocabolario
# 1. divisione in parole

- La lettura dei documenti porta ad ottenere le seguenti parole

| | | |
|---|---|---|
| 272 | la | sconfitti |
| 318 | ma | si |
| a.C. | muore | stupisce |
| ad | nasce | Torna |
| Argo | nel | tutti |
| da | Non | una |
| e | pace | vince |
| grande | Pirro | vinto |
| ha | romani | vittoria |
| l | sa | |



# Costruzione del vocabolario
# 2. riduzione al maiuscolo

| | | |
|---|---|---|
| 272 | la | sconfitti |
| 318 | ma | si |
| a.c. | muore | stupisce |
| ad | nasce | torna |
| argo | nel | tutti |
| da | non | una |
| e | pace | vince |
| grande | pirro | vinto |
| ha | romani | vittoria |
| i | sa | |

# Costruzione del vocabolario
# 3. applicazione della stop-list

| | | |
|---|---|---|
| 272 | ~~la~~ | sconfitti |
| 318 | ~~ma~~ | si |
| a.c. | muore | stupisce |
| ~~ad~~ | nasce | torna |
| argo | ~~nel~~ | tutti |
| ~~da~~ | ~~non~~ | ~~una~~ |
| ~~e~~ | pace | vince |
| grande | pirro | vinto |
| ha | romani | vittoria |
| ~~i~~ | sa | |



# Costruzione del vocabolario
# 4. gestione accenti

- Nulla da fare

# Costruzione del vocabolario
## 5. stemming

| | | | |
|---|---|---|---|
| 272 | pirro | vinto | stup |
| 318 | romani | vittoria | torn |
| a.c. | sa | grand | tutt |
| argo | sconfitti | muor | vinc |
| grande | si | nasc | vint |
| ha | stupisce | pac | vittor |
| muore | torna | pirr | |
| nasce | tutti | roman | |
| pace | vince | sconfitt | |

Snowball stemmer, by Dr Martin Porter. http://snowball.tartarus.org/

# Costruzione del vocabolario
## 6. lemmatizzazione

| | | |
|---|---|---|
| 272 | romani | avere |
| 318 | sa | morire |
| a.c. | sconfitti | nascere |
| argo | si | romano |
| grande | stupisce | sapere |
| ha | torna | sconfitto |
| muore | tutti | stupire |
| nasce | vince | tornare |
| pace | vinto | tutto |
| pirro | vittoria | vincere |

TreeTagger, by Helmut Schmid. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

# Vocabolario di indicizzazione

| | | | |
|---|---|---|---|
| 272 | sa | nasc | avere |
| 318 | sconfitti | pac | morire |
| a.c. | si | pirr | nascere |
| argo | stupisce | roman | romano |
| grande | torna | sconfitt | sapere |
| ha | tutti | stup | sconfitto |
| muore | vince | torn | stupire |
| nasce | vinto | tutt | tornare |
| pace | vittoria | vinc | tutto |
| pirro | grand | vint | vincere |
| romani | muor | vittor | |



# Indicizzazione

- Una volta costruito il vocabolario si passa alla vera e propria indicizzazione
  - Viene costruito un indice inverso
  - I termini vengono associati ai documenti in cui appaiono

# Indice inverso (I)

| | | |
|---|---|---|
| 272 | → | 4 |
| 318 | → | 4 |
| a.c. | → | 4 |
| argo | → | 4 |
| avere | → | 1 |
| grand | → | 1 |
| grande | → | 1 |
| ha | → | 1 |
| morire | → | 4 |
| muor | → | 4 |
| muore | → | 4 |

| | | |
|---|---|---|
| nasc | → | 4 |
| nasce | → | 4 |
| nascere | → | 4 |
| pac | → | 2 |
| pace | → | 2 |
| pirr | → | 1 → 2 → 3 → 4 |
| pirro | → | 1 → 2 → 3 → 4 |
| roman | → | 3 |
| romani | → | 3 |
| romano | → | 3 |
| sa | → | 1 |



# Indice inverso (II)

| | | |
|---|---|---|
| sapere | → | 1 |
| sconfitt | → | 3 |
| sconfitti | → | 3 |
| sconfitto | → | 3 |
| si | → | 1 |
| stup | → | 2 |
| stupire | → | 2 |
| stupisce | → | 2 |
| torn | → | 2 |
| torna | → | 2 |
| tornare | → | 2 |

| | | |
|---|---|---|
| tutt | → | 2 |
| tutti | → | 2 |
| tutto | → | 2 |
| vinc | → | 1 → 2 |
| vince | → | 1 → 2 |
| vincere | → | 1 → 2 |
| vint | → | 1 |
| vinto | → | 1 |
| vittor | → | 1 |
| vittoria | → | 1 |

# Vettori TF-IDF (I)

| | 272 | 318 | a.c. | argo | avere | and | andare | ha | morire | muore | nascere | nascere | pacce | pire | piro | roman | romanio | romano |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D1** | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,25 | 0,25 | 0 | 0 | 0 |
| **D2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0,25 | 0,25 | 0 | 0 | 0 |
| **D3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,25 | 0,25 | 1 | 1 | 1 |
| **D4** | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0,25 | 0,25 | 0 | 0 | 0 |



# Vettori TF-IDF (II)

| | sa | sapere | sconfitti | sconfitto | si | stup | stupire | stupisce | torna | tornaa | tornare | tutti | tutto | vinc | vince | vincere | vint | vinto | vittoor | vittoria |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D1** | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,5 | 0,5 | 1 | 1 | 1 | 1 | 1 |
| **D2** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | 0,5 | 0,5 | 0 | 0 | 0 | 0 |
| **D3** | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **D4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# Possibile thesaurus (I)

- Come detto, non esiste un unico thesaurus e non esiste una regola precisa per costruirne uno
- Nel nostro esempio consideriamo di avere il seguente piccolo thesaurus



# Possibile thesaurus (II)

- pace
  UF guerra
- vittoria
  UF sconfitta
- morire
  UF nascere
- sconfitto
  UF vinto
- roma
  RT romano
- vittoria
  RT vincere
- vittoria
  RT sconfitto
- città
  NT1 roma
  NT1 argo

# Esempi di ricerche
## con modello booleano esteso

- L'effettivo algoritmo (processo) per rispondere ad una ricerca dipende dal software/applicazione/motore di ricerca
  - Qui ne vedremo solo uno plausibile
- I termini della ricerca subiscono lo stesso “trattamento” delle parole nei documenti
  - Stop-list, stemming, lemmatizzazione
  - Così ANCHE i termini “correlati” a quelli inseriti vengono usati per cercare i documenti nell'indice inverso
- Useremo un modello booleano esteso con sottinteso l'operatore AND
  - Come fa Google



# vittoria di Pirro (I)

- Riduzione al minuscolo e stop-list: vittoria pirro
- vittoria
  - È contenuto solo in D1 (con peso 1)
  - La sua radice e lemma sono ancora contenuti solo in D1, quindi sono “inutili”
  - Nel thesaurus è correlato a
    - vincere (RT): è contenuto anche in D2 (con peso 0,5)
    - sconfitta (UF): non compare in nessun documento
    - sconfitto (RT): compare in D3 (con peso 1)

# vittoria di Pirro (II)

- pirro
  - È contenuto in D1, D2, D3, D4 (sempre con peso 0,25)
  - La sua radice e lemma non portano ad altri documenti
  - Nel thesaurus non è correlato a nessun altro termine



# Ranking

- Ricapitolando
  - D1 ha peso totale
    1 + 0,25 = 1,25
  - D2 ha peso totale
    0,75 × 0,5 + 0,25 = 0,625
  - D3 ha peso totale
    0,75 × 1 + 0,25 = 1
  - D4 ha peso totale
    0,25
- Si ottiene il ranking finale:
  D1, D3, D2, D4

0,75 è un fattore che serve a tenere conto che usando il thesaurus in qualche modo ci "allontaniamo" dai termini della query e quindi quello che troviamo "vale" meno

# Risultato della ricerca: vittoria di Pirro

(D1) Pirro vince! È vera vittoria? Non si sa, ma intanto ha vinto.

(D3) I romani sconfitti da Pirro

(D2) Pirro stupisce tutti e vince. Torna la pace?

(D4) Pirro nasce nel 318a.C. e muore ad Argo nel 272a.C.