Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2013/2014

Corso di Laurea Magistrale in Editoria e Giornalismo

Analisi di testi

Efficacia e... efficienza (1)

- Le misure per la valutazione dei SRI quantificano l'efficacia dei sistemi di IR
- Ma se prendiamo in considerazione grandi collezioni di documenti anche l'efficienza è importante
  - La capacità di elaborare le interrogazioni usando il minor quantitativo di risorse
    - Tempo
      - Chi aspetterebbe più di pochissimi secondi per avere la risposta ad una ricerca?

Efficacia e... efficienza (2)

- Le pagine in tutto il WEB non si contano
  - Una statistica dice che a settembre 2013 Google "conosceva" circa 40 miliardi di pagine web
- Quanto pensate ci metta Google ad analizzare tutte queste pagine per rispondere ad una interrogazione?
- Per un'interrogazione, elaborare un documento ogni
  - millisecondo significa attendere 463 giorni circa
  - milionesimo di secondo significa attendere 11 ore circa
  - … noi aspettiamo molto meno!!!
- I motori di ricerca attuali elaborano centinaia di milioni di documenti e migliaia di interrogazioni al secondo

Raggiungere l'efficienza

- Per aumentare l'efficienza dei SRI si utilizzano gli indici
  - Struttura dati costruita a partire dal testo dei doc
  - Capire come un indice è fatto e viene costruito ci dice diverse cose sul funzionamento dei SRI
- Un indice
  - viene prodotto tramite un processo detto “indicizzazione”
  - È costruito sulla base di un insieme di termini detto vocabolario o linguaggio di indicizzazione
    - I termini possono essere singole parole o frasi

Composizione del vocabolario

- Ma quali termini usare per costruire il vocabolario?
  - Le singole parole che compongono i documenti
  - Ma sorgono diversi problemi
    - Indipendenti dalla lingua
    - Dipendenti dalla lingua

Problemi nella scelta dei termini (1)

- Come considerare
  - Web e WEB??
  - UE e U.E.??
  - Italiano e Italiani??
  - Hewlett-Packard o San Francisco??
  - 01/01/2013 e 1 gennaio 2013??
  - 0458027103 e 045 8027103??
  - leggo e lessi??
  - autoveicolo e automezzo??
  - …

Problemi nella scelta dei termini (2)

- In cinese non vi sono gli spazi

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- 和尚
  - Può essere interpretato come
    - Singola parola con significato di “scimmia”
    - Due parole con significato di “e” e “ancora”

Problemi nella scelta dei termini (3)

- Parole composte
  - Tedesco:
    - Computerlinguistik → Computer + Linguistik
  - Molte altre lingue hanno parole composte
- In giapponese si usano 4 diversi alfabeti
  - Cinese, Hiragana, Katakana e latino
- In arabo si scrive da destra verso sinistra

Problemi nella scelta dei termini (4)

- Accenti e segni diacritici
  - balbùzie vs balbuzie
    - Omissione degli accenti
  - Universität vs Universitæt
    - Sostituzione con altre sequenze di caratteri

Soluzioni ai problemi (1)

- Criterio: come scrivono normalmente gli utenti?
  - Tendenzialmente senza accenti dove non necessari
    - balbùzie → balbuzie
    - Eliminiamo gli accenti normalmente non utilizzati
  - Tutto in minuscolo
    - WEB → web
    - Riduciamo tutte le parole in minuscolo
    - In fondo quante sono le parole/sigle uguali ma con significato diverso solo perché scritte in minuscolo piuttosto che in maiuscolo?
- Consideriamo “equivalenti” tutte le possibilità nelle forme ambigue
  - Sia 01/01/2013 sia 1 gennaio 2013

Soluzioni ai problemi (2)

- Stemming:
  - Riduzione di una forma flessa di una parola alla sua forma radice (tema)
  - Rimozione sintattica del suffisso delle parole in modo da ottenerne solo la radice
  - Esempio: “riproducibile”, “riprodotto”, “riproduzione” si riducono tutti a “riprod”
  - Dipende dalla lingua in uso
  - Esistono più regole per calcolare lo stemming
    - Ottengono radici diverse

Soluzioni ai problemi (3)

- Lemmatizzazione:
  - Riduzione di una forma flessa di una parola alla sua forma/variante canonica (lemma)
  - Esempio: “leggo” e “lessi” si riducono a “leggere”
  - Esempio: “car”, “cars”, “car's” e “cars'” si riducono tutte a “car”
- Tema e lemma non è detto che coincidano
- Esempio concreto:
  - Morph-it
    - Sviluppato all'Università di Bologna
    - Usato anche in wordreference.com

Soluzioni ai problemi (4)

- Stop list:
  - Eliminazione delle parole estremamente comuni che sono di scarso aiuto nel discriminare i documenti rispetto ad una interrogazione di un utente
    - Stop word
  - Congiunzioni, articoli, preposizioni, …
  - Ma come trattare ad esempio “to be or not to be”?
    - Vedremo quando parleremo della ricerca basata su frasi

Soluzioni ai problemi (5)

- Thesaurus:
  - Riduzione di una parola ad un termine che la rappresenta in quanto semanticamente equivalente
  - Esempi:
    - automezzo vs autoveicolo
    - thesaurus vs tesauro
  - Discorso lungo, lo approfondiremo fra un po'

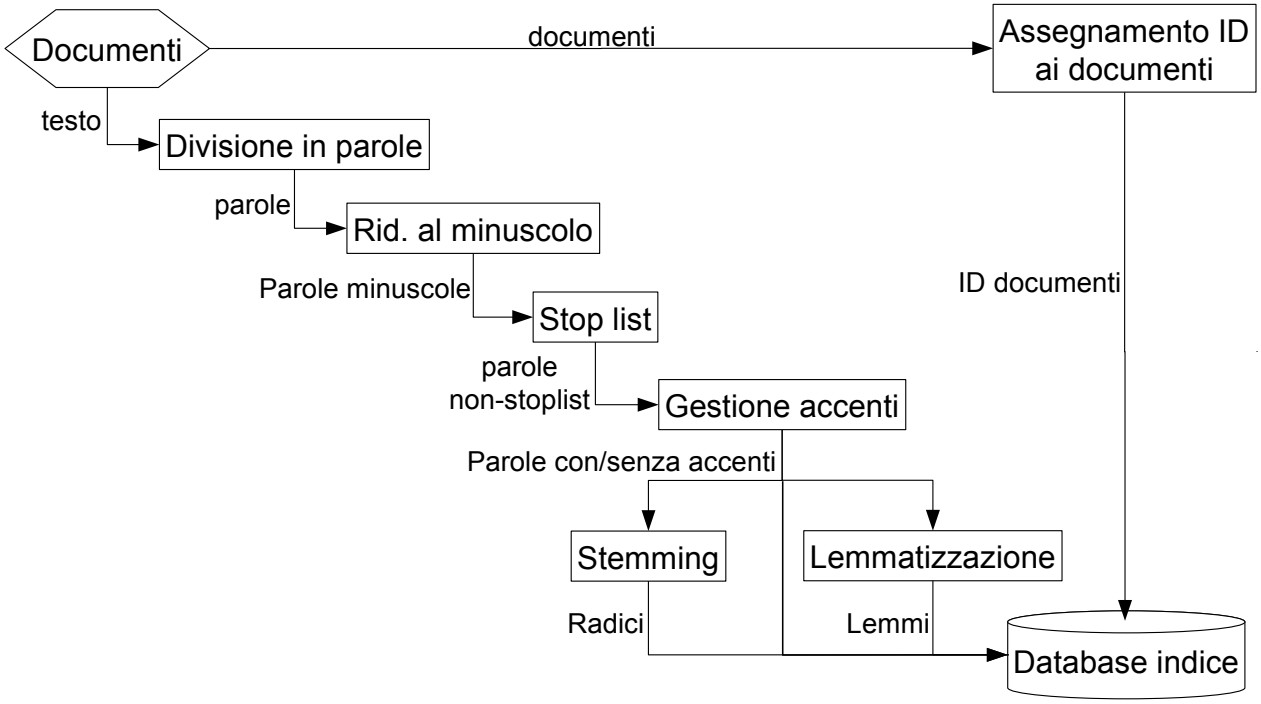
Nella realtà

- Come si comporta Google rispetto a tutti i problemi precedenti?
  - Stop word
  - Stemming
  - Lemmatizzazione
  - Gestione di maiuscole/minuscole
  - Caratteri non latini
  - Accenti e segni diacritici
  - Parole composte
  - Numeri e date
- Provare per capire/credere

Il vocabolario

- Tutte le tecniche viste finora permettono, a partire dalle parole contenute nei documenti nella collezione, di ottenere i termini che compongono il linguaggio di indicizzazione
- I documenti nella collezione vengono letti parola per parola
  - Si ottengono tutti i termini nel vocabolario
- Il vocabolario viene utilizzato per la costruzione di un indice che permetta una rapida ricerca dei termini nei documenti

Processo di indicizzazione



Indice

- Obiettivo degli indici è permettere di sapere in quali documenti si trovano i termini del vocabolario
- Una prima soluzione è usare una matrice termine-documento

Matrice termine-documento (1)

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	1	1	0	0	0	1	
BRUTUS	1	1	0	1	0	0	
CAESAR	1	1	0	1	1	1	
CALPURNIA	0	1	0	0	0	0	
CLEOPATRA	1	0	0	0	0	0	
MERCY	1	0	1	1	1	1	
WORSER	1	0	1	1	1	0	
...							

- 1: il termine occorre nel documento
  - Brutus occorre in Hamlet
- 0: il termine non occorre nel documento
  - Anthony non occorre in Hamlet

Matrice termine-documento (2)

- Come rispondere alla query "Brutus AND NOT Anthony"?
  - Prendiamo la riga (vettore) di Brutus
  - Prendiamo la riga di Anthony
  - Cerchiamo le posizioni nelle righe dove c'è un 1 per Brutus e uno 0 per Anthony
    - Nei computer si può fare efficientemente
  - Risultato: Hamlet

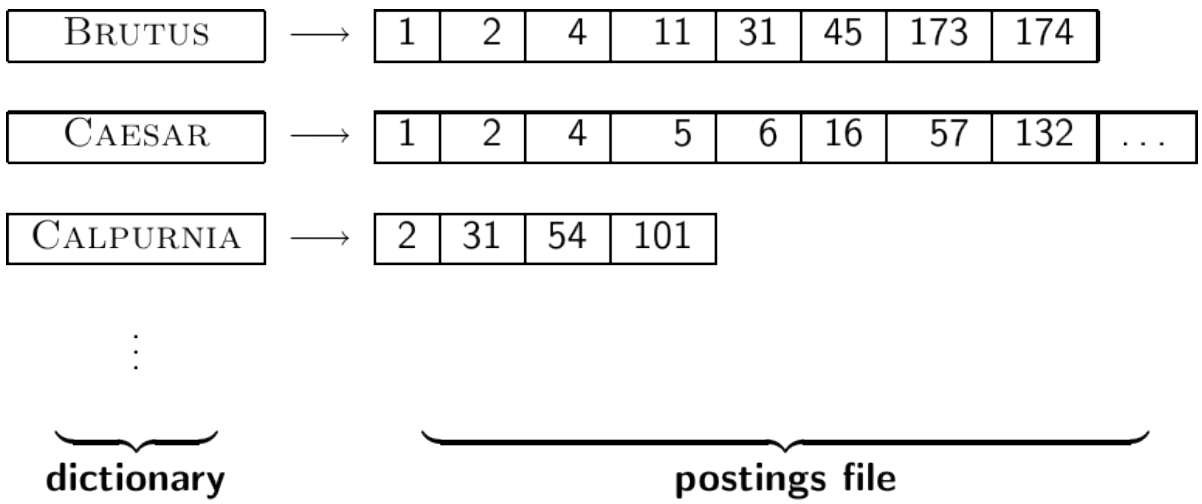
Problema delle matrici

- Supponiamo di avere 1 milione di documenti
- Supponiamo che questi documenti portino ad ottenere 200.000 termini
- Otteniamo una matrice con 200 miliardi di celle con 0 e 1
  - Anche rappresentando 0 e 1 con un solo bit la matrice occuperebbe 23,3 GB
    - Tanto (troppo) per così poca informazione
  - Probabilmente con non più di un miliardo di 1
- Le matrici sono troppo poco efficienti
  - La loro dimensione cresce velocemente

Indice inverso (1)

- Quale potrebbe essere una migliore rappresentazione?
  - Memorizziamo solo gli 1!!!
- Indice inverso
  - Per ogni termine memorizziamo una lista di tutti i documenti in cui occorre

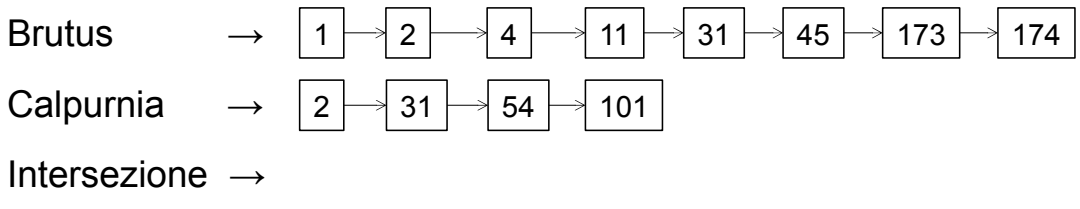
Indice inverso (2)



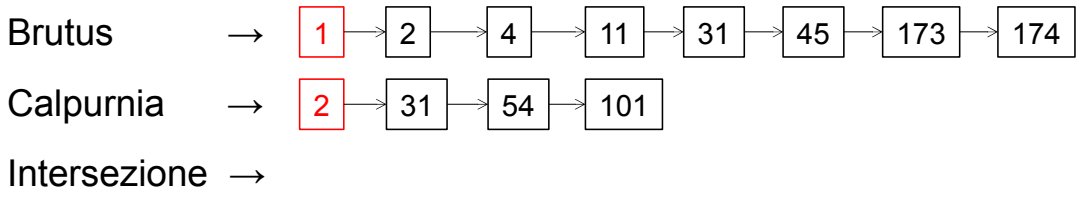
Esempio: indice inverso in una query booleana (1)

- Consideriamo la query:
  BRUTUS AND CALPURNIA
- Usiamo l'indice inverso per trovare i documenti che rispondono alla query:
  - Recuperiamo la postings list relativa a BRUTUS
  - Recuperiamo la postings list relativa a CALPURNIA
  - Intersechiamo le due postings list
  - Il risultato dell'intersezione viene restituito all'utente

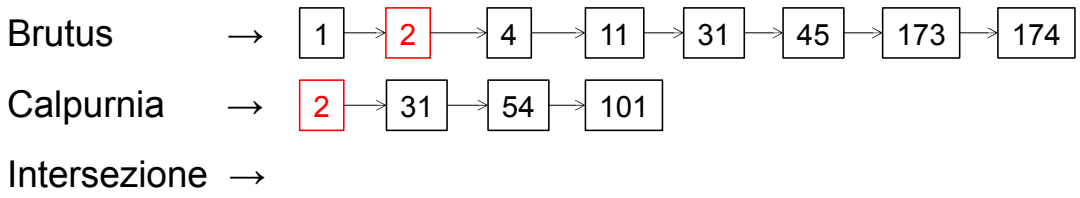
Esempio: indice inverso in una query booleana (2)



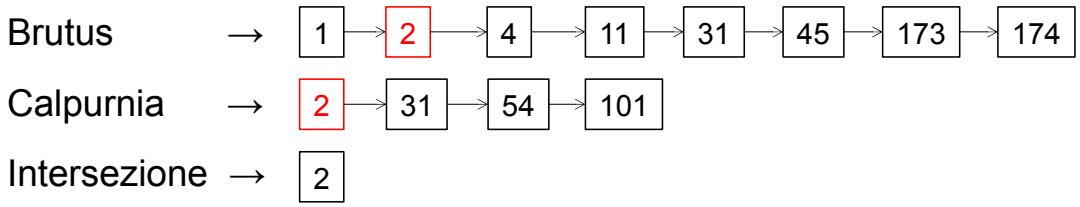
Esempio: indice inverso in una query booleana (2)



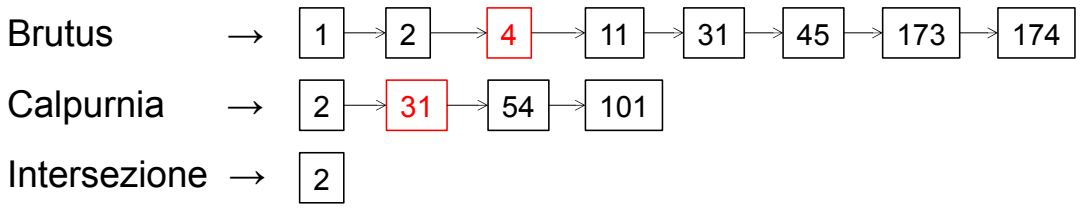
Esempio: indice inverso in una query booleana (2)



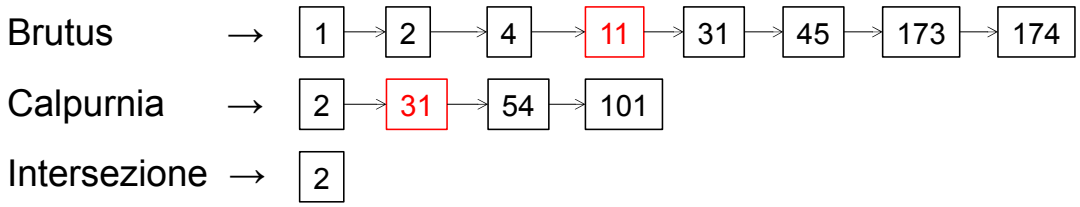
Esempio: indice inverso in una query booleana (2)



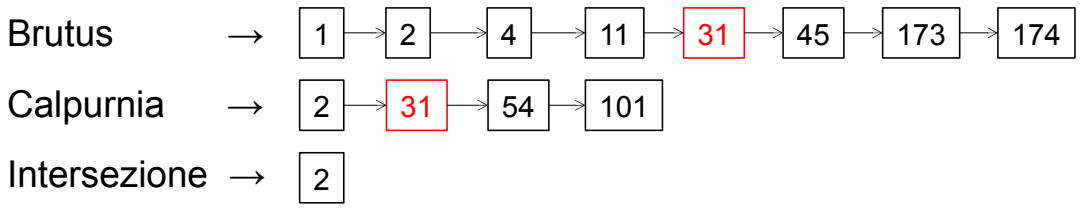
Esempio: indice inverso in una query booleana (2)



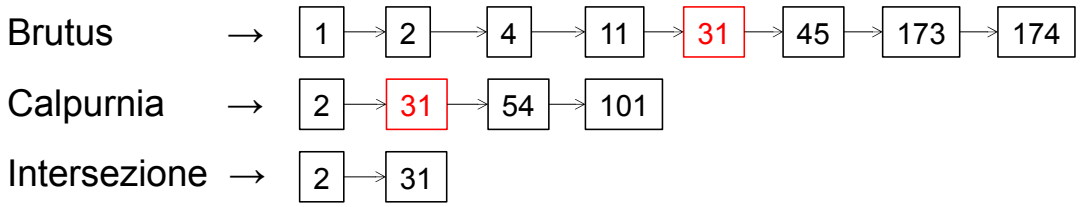
Esempio: indice inverso in una query booleana (2)



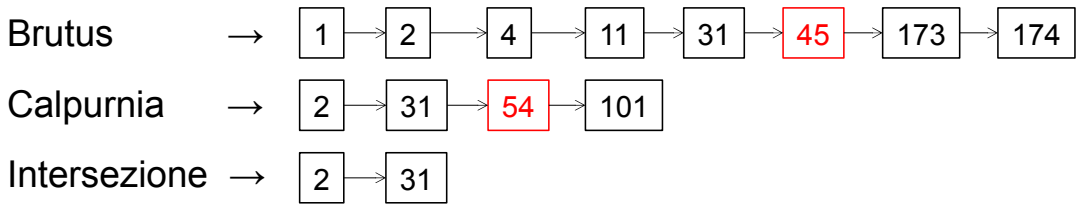
Esempio: indice inverso in una query booleana (2)



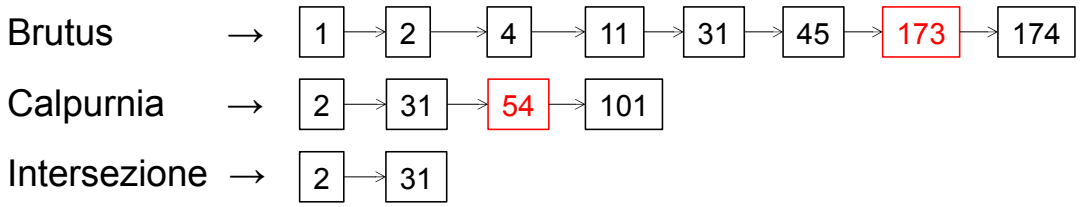
Esempio: indice inverso in una query booleana (2)



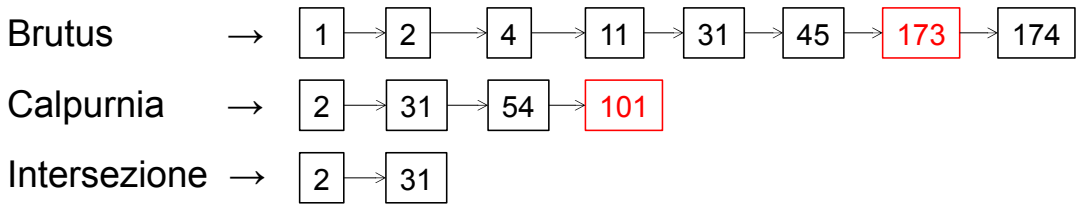
Esempio: indice inverso in una query booleana (2)



Esempio: indice inverso in una query booleana (2)



Esempio: indice inverso in una query booleana (2)



Query basate su frasi

- E se volessimo rispondere alla query “sistemi distribuiti” ?
  - Un documento contenente “sistemi informativi distribuiti” non dovrebbe essere recuperato
- Circa il 10% delle interrogazioni web fanno uso di frasi
- Un indice inverso come l'abbiamo visto finora non è sufficiente
  - Non è sufficiente memorizzare solo gli id dei documenti in cui i termini appaiono
    - Non sappiamo dove i termini appaiono
    - Non sappiamo vicino a quali altri termini un dato termine appare
- Idee per risolvere questo problema??

Indice posizionale

- In un indice non-posizionale le posting list memorizzano solo l'id dei documenti
- In un indice posizionale le posting list memorizzano l'id dei documenti e una lista delle posizioni in cui il termine appare
- be:

  ⟨ 1: ⟨ 17, 25 ⟩;
  4: ⟨ 17, 191, 291, 430, 434 ⟩;
  5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Esempio di indice posizionale

Query: “to_1 be_2 or_3 not_4 to_5 be_6”

to:

⟨ 1: ⟨ 7, 18, 33, 72, 86, 231⟩;
2: ⟨ 1, 17, 74, 222, 255 ⟩;
4: ⟨ 8, 16, 190, 429, 433 ⟩;
5: ⟨ 363, 367 ⟩;
7: ⟨ 13, 23, 191 ⟩; ... ⟩

be:

⟨ 1: ⟨ 17, 25 ⟩;
4: ⟨ 17, 191, 291, 430, 434 ⟩;
5: ⟨ 14, 19, 101 ⟩; ... ⟩

Il documento 4
è un candidato!!

Indici biword

- La ricerca di frasi può essere ottimizzata utilizzando insieme indici posizionali e indici biword
- Molte coppie di parole sono estremamente frequenti: “Michael Jackson”, “New York”, ecc...
- Includere le coppie di parole molto frequenti nel vocabolario e quindi negli indici

Complessità della ricerca di frasi

- Per i motori di ricerca la ricerca di frasi (usando indici posizionali) è molto più complessa delle semplici query booleane estese
- Esempio in Google:
  [government information and services available]
  → 1.740.000.000 risultati in 0,33s → $1{,}89 \times 10^{-10}$ s/ris
  [“government information and services available”]
  → 392.000 risultati in 0,29s → $7{,}39 \times 10^{-7}$ s/ris

  la ricerca della frase è circa 3900 volte più lenta dell'altra
- Più la frase è lunga più è complessa la ricerca
  - Provare per credere

Concludendo con gli indici

- Gli attuali motori di ricerca utilizzano tutte le tecniche viste finora, e altre...
- Nella realtà quindi si usano indici inversi posizionali biword

Ricerca per prossimità

- Gli indici posizionali sono usati per la ricerca di frasi
- Ma possono essere usati anche per la ricerca per prossimità
  - Esempio: luogo /4 lavoro
    trovare tutti i documenti in cui luogo e lavoro si trovano a non più di 4 parole di distanza
- Inefficienti se non utilizzati insieme ad altre tecniche e strutture dati

Torniamo un po' indietro...

- Nell'indicizzazione e nelle query come trattare le parole semanticamente simili?
  - automezzo vs autoveicolo
  - thesaurus vs tesauro
- Thesaurus:
  - Riduzione di una parola ad un termine che la rappresenta in quanto semanticamente equivalente

Tesauri

(Si veda anche l'apposita dispensa)

Definizione

- Secondo l'ISO (International Organization for Standardization)
  - il thesaurus è il vocabolario di un “linguaggio di indicizzazione” controllato, organizzato in maniera formale, in maniera cioè da rendere esplicite le relazioni “a priori” fra i concetti

Caratteristiche: controllato (I)

- Controllato:
  - relazione biunivoca fra termine e concetto, fra significante e significato
  - un termine esprime sempre uno ed un solo concetto, e che un concetto è sempre espresso da uno ed un solo termine
  - condizione tutt'altro che vera nel linguaggio naturale

Caratteristiche: controllato (II)

- Controllato, si ottiene tramite 2 accorgimenti

1) la raccolta e la collazione
   - dei sinonimi
   - dei quasi sinonimi: termini non sinonimi in senso proprio, ma considerabili tali ai fini della rappresentazione dei concetti dell'ambito coperto dal thesaurus
   - degli antonimi: cioè degli opposti, o meglio dei termini collocati in diversi punti dello stesso continuum semantico: non solo calore e freddo, ma anche tepore, fresco e così via

   atti a descrivere il medesimo concetto e la scelta di uno solo di questi termini:
   - il termine prescelto diventa nel vocabolario termine preferito, TP (o PT, Preferred Term)
     - Termine abilitato a descrivere quel determinato concetto;
   - tutti gli altri termini (termini non preferiti, NPT)
     - Termini che non possono essere assegnati ai documenti per esprimerne il contenuto concettuale, possono però vantaggiosamente entrare nel vocabolario controllato come punti di accesso che rinviano al termine preferito

Caratteristiche: controllato (III)

- Controllato, si ottiene tramite 2 accorgimenti
  2) la riduzione del contenuto semantico del termine preferito ad un solo significato
     - di solito il più tipico nell'ambito disciplinare del thesaurus
     - per cui ad esempio in un thesaurus di ornitologia il termine gru esprimerà il concetto corrispondente all'"uccello dei Ralliformi" e non alla "macchina dal braccio girevole per sollevare pesi"

Caratteristiche: formale

- Le relazioni rappresentate tra i termini sono formalizzate
  - Definite da una relazione esplicita e formalizzata
  - Usando un'apposita sintassi
  - Così ogni termine è inserito in una rete relazionale che
    - Ne chiarisce ulteriormente il contenuto semantico
    - Ne mostra la distanza semantica dagli altri termini

Caratteristiche: a priori

- Le relazioni trattate sono *a priori*
  - Appartengono all'ambito semantico, del significato, dei termini
  - Sono sempre valide in qualsiasi contesto

Concetti rappresentabili

- I concetti rappresentati dai termini di un tesauro possono appartenere a diverse categorie:
  - Entità concrete
    - oggetti e loro parti fisiche
    - Materiali
  - Entità astratte
    - azioni e avvenimenti
    - entità astratte e proprietà degli oggetti, dei materiali o delle azioni
    - discipline o scienze
    - unità di misura
  - Entità individuali o “classi di uno” analoghe a nomi propri

Relazioni

- Un tesauro può rappresentare tre grandi categorie di relazioni
  - Preferenziale
  - Gerarchica
  - Associativa
- Tutte le relazioni sono esplicitate e rappresentate da appositi simboli e sigle

Relazione preferenziale (I)

- Detta anche di equivalenza
- Deputata a rapportare uno o più NPT ad un PT che esprime lo stesso concetto o un concetto molto simile, che sarà rappresentato sempre univocamente dal TP
- Il gruppo di termini che si assume rappresentino lo stesso concetto, che si considerano, ai fini dell'indicizzazione, equivalenti, e fra i quali viene scelto il termine preferito, si definisce *gruppo di equivalenza*
- È l'unica a mettere in relazione PT con NPT
  - Tutte le altre relazioni sono solo fra PT

Relazione preferenziale (II)

- il rinvio dal TNP al TP viene indicato dal simbolo USE
  - `tesauro`
    `USE thesauro`
- la segnalazione dei TNP nel corredo semantico del TP è indicato dal simbolo UF
  - `thesauro`
    `UF tesauro`

Rel. preferenziale, classificazione

- Univoca: ad un NPT corrisponde un solo PT
- Biunivoca: ad un NPT rappresentante un concetto complesso corrispondono due distinti PT rappresentanti suoi concetti costitutivi più semplici, che devono essere usati obbligatoriamente insieme
  - Esempio: `sideremia`
    `USE Ferro AND Sangue`
  - Reciproco: `ferro`
    `UF+ sideremia`
    `sangue`
    `UF+ sideremia`

Rel. preferenziale univoca (I)

- Distinguiamo due sottocasi
  1. Sinonimia assoluta o accentuata
     - tra TP e TNP esiste sempre un rapporto sinonimico indipendentemente dall'area semantica, dal grado di analiticità del tesauro e da quale dei due termini viene definito come preferito e quale come non preferito
     - Tipologie
       - sinonimia vera (regola e norma)
       - variante ortografica (psicopedagogia e psico-pedagogia)
       - sigle e acronimi (CNR e Centro Nazionale Ricerche)
       - preferenza linguistica

Rel. preferenziale univoca (II)

- Distinguiamo due sottocasi
  2. Sinonimia relativa o convenzionale
     - la relazione tra due termini di significato vicino, appartenenti alla stessa area semantica, non è sinonimica in senso stretto, non verrebbe considerata tale nel linguaggio naturale, non è sempre considerata tale in tutti i tesauri
     - Tipologie:
       - quasi-sinonimia (punizione e sanzione)
       - upward posting o rinvio al superiore gerarchico (microformati e microfilm)
       - Antinomia (pace e guerra, malattia e salute, secchezza e umidità)

Relazione gerarchica (I)

- Esprime il concetto ed il grado di subordinazione o sovraordinazione fra termini appartenenti allo stesso albero gerarchico
  - il termine sovraordinato rappresenta una classe o un tutto
  - il termine subordinato rappresenta un suo elemento o parte

Relazione gerarchica (II)

- La sigla che individua i sovraordinati è BT (Broader Term), alla quale può utilmente essere aggiunta una cifra indicante la distanza in "gradini" gerarchici fra i due termini legati
  - Geometria iperbolica
    BT1 Geometria non euclidea
    BT2 Geometria
    BT3 Matematica
- La sigla identificante il rapporto inverso, cioè i subordinati del termine dato, è NT (Narrower Term)
  - Geometria
    NT1 Geometria euclidea
    NT1 Geometria non euclidea
    NT2 Geometria iperbolica
    NT2 Geometria ellittica

Relazione gerarchica (III)

- Rientrano nella categoria delle relazioni gerarchiche tre sottospecie di relazioni:
  - la relazione generica o relazione genere-specie
    - distinta eventualmente dalle sigle BTG e NTG
  - la relazione partitiva o relazione parte-tutto
    - distinta eventualmente dalle sigle BTP e NTP
  - la relazione esemplificativa o classe-istanza o specie-esempio

Rel. gerarchica generica

- identifica il legame che intercorre fra una classe o categoria ed i suoi elementi, membri o specie
  - È la tipica relazione delle classificazioni zoologiche o botaniche

Rel. gerarchica partitiva

Lo standard ISO elenca quattro sottocasi

1. Sistemi e organi del corpo
   - Sistema circolatorio
     NT1 Sistema vascolare
     NT2 Arterie
     NT2 Vene
2. Luoghi geografici
   - Canada
     NT1 Manitoba
     NT2 Winnipeg
3. Discipline e campi di studio
   - Scienze
     NT1 Chimica
     NT1 Biologia
     NT2 Botanica
4. Strutture sociali gerarchizzate
   - Corpi d'armata
     NT1 Divisioni
     NT2 Reggimenti

Rel. gerarchica esemplificativa

- Identifica il legame che intercorre fra una classe o categoria generale di cose o avvenimenti, espressa da un nome comune, ed un suo individuo, rappresentato da un nome proprio, e costituente una “classe di uno”
  - Regioni montuose <classe>
    <rel. esemplificativa>
    NT1 Alpi <individuo>
    <rel. partitiva>
    NT2 Alpi Graie <individuo>

Relazione associativa (I)

- Detta anche relazione “residuale”
  - in grado di collegare coppie di termini che non rientrano né nella casistica della relazione sinonimica, né in quella della relazione gerarchica
- È reciproca
- Indicata in ambedue i sensi con la sigla RT (Related Term)
- Ci sono due tipi di termini suscettibili di intrattenere rapporti associativi:
  1. quelli appartenenti alla stessa categoria
  2. quelli appartenenti a categorie diverse

Relazione associativa (II)

- Fra i termini appartenenti alla stessa categoria distinguiamo fra

  a) termini che hanno lo stesso termine sovraordinato, ed i cui significati hanno una zona di sovrapposizione, e che quindi, anche se hanno una definizione che li distingue esattamente, potrebbero essere adoperati dagli utenti in maniera non rigorosa (quasi intercambiabile)

  - `Barche`
    `BT Veicoli`
    `RT Navi`
  - `Navi`
    `BT Veicoli`
    `RT Barche`

Relazione associativa (III)

- Fra i termini appartenenti alla stessa categoria distinguiamo fra
  - termini che rappresentano concetti legati da una relazione di tipo “familiare” o di tipo “derivato” (un concetto che deriva dall’altro)

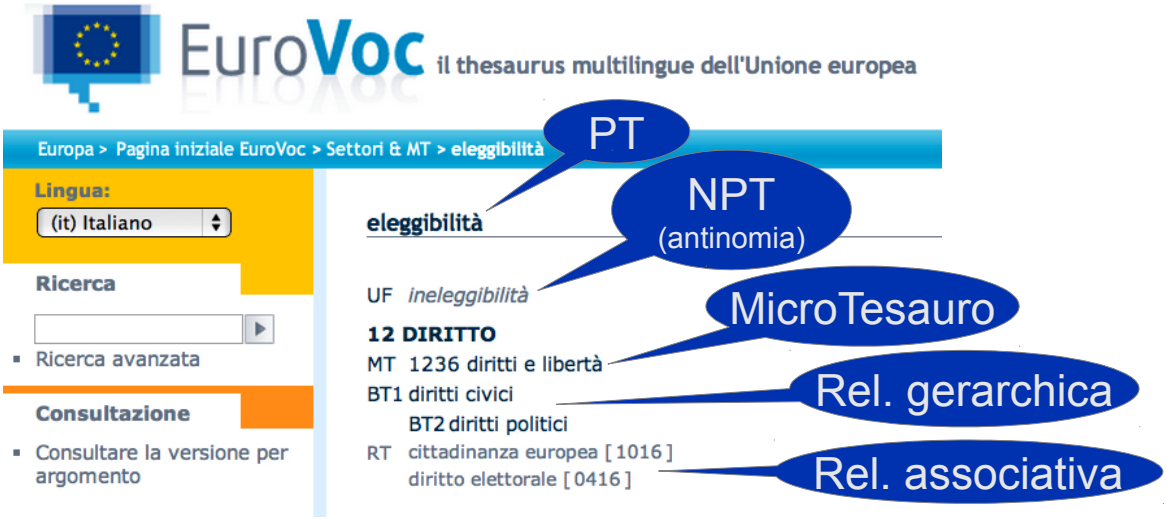
- Equini
  - NT1 Asini
  - NT1 Cavalli
  - NT1 Muli
- Asini
  - BT1 Equini
  - RT Cavalli
  - RT Muli
- Cavalli
  - BT1 Equini
  - RT Asini
  - RT Muli
- Muli
  - BT1 Equini
  - RT Asini
  - RT Cavalli

Relazione associativa (IV)

- Fra i termini appartenenti a categorie diverse si configurano diverse tipologie di rapporti che possono motivare una relazione associativa
  1. una disciplina e il suo oggetto di studio (zoologia e animali)
  2. un processo od operazione e il suo agente o strumento (termometro e misurazione della temperatura)
  3. una azione e il suo prodotto (scrittura e documenti)
  4. una azione e chi o cosa la subisce (potatura e piante; pesca e pesci)
  5. oggetti e fenomeni e loro proprietà (magneti e magnetismo)
  6. concetti e loro origini (Tedeschi e Germania)
  7. concetti legati da rapporti causali (inquinamento e sostanze inquinanti)
  8. una cosa e il suo antidoto (piante ed erbicidi)
  9. un concetto e la sua unità di misura (frequenza e hertz)
  10. locuzioni sincategorematiche (cioè termini composti) e loro nomi sottocategoriali (piante fossili e piante)

Tesauri: esempio

- Thesaurus pubblico accessibile online: EuroVoc
  - multilingue e pluridisciplinare
  - comprende la terminologia dei settori d'attività dell'UE
  - http://eurovoc.europa.eu/



Tesauri: conclusione

- I tesauri sono utilizzati per allargare, rilassare la ricerca effettuata dall'utente anche ai termini semanticamente simili a quelli ricercati
- Ovviamente i documenti contenenti i termini affini anziché quelli ricercati hanno una rilevanza diversa
  - Da quelli contenenti i termini dell'utente
  - A seconda della relazione che li lega ai termini dell'utente
- Ogni SRI nel suo algoritmo di ranking (ordinamento dei documenti recuperati secondo la rilevanza) tiene conto di ciò in modo diverso

Proprietà dei testi

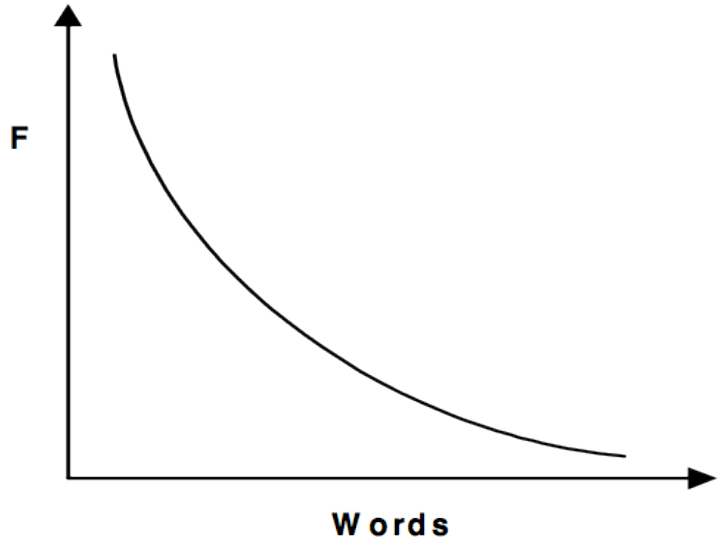
- Dall'indicizzazione e analisi dei testi si possono osservare alcune leggi empiriche
  - Leggi che non sono dimostrabilmente sempre vere
  - Leggi approssimative
    - Legge di Zipf
    - Legge di Heaps
  - Proposte a metà del 1900 da linguisti

Legge di Zipf (I)

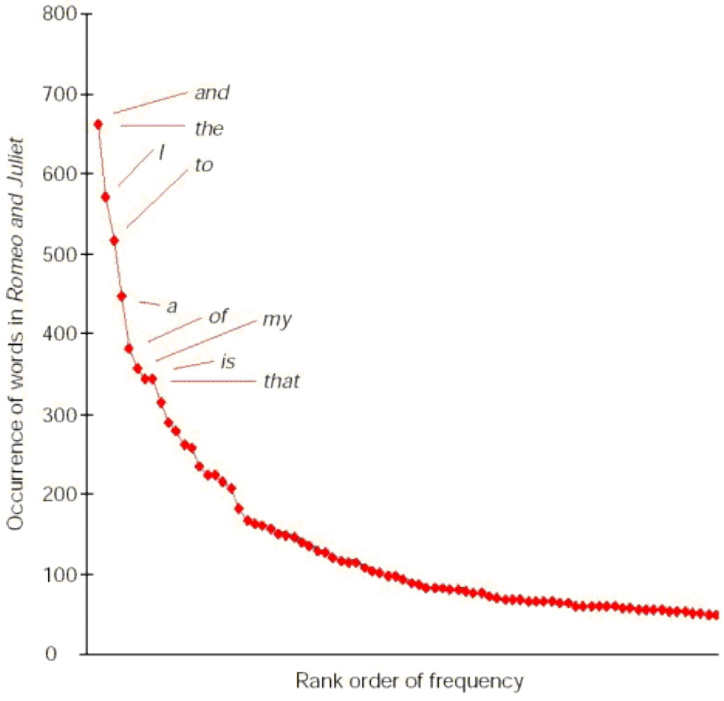
- Come sono distribuiti in un documento/collezione i diversi termini?
- Legge di Zipf
  - Ordinando i termini in un doc/collezione per frequenza decrescente si ha che
    frequenza · rango = costante
  - In altre parole, la frequenza f_i dell'*i*-esimo termine più frequente è $f_i = \frac{f_1}{i}$
    f_1 è la frequenza del termine più frequente (rango=1)
  - Il secondo termine appare la metà del primo, il terzo un terzo del primo e così via...

Legge di Zipf (II)

→ Pochi termini appaiono molto spesso
  – Congiunzioni, articoli, ecc...
→ Tanti termini appaiono pochissimo

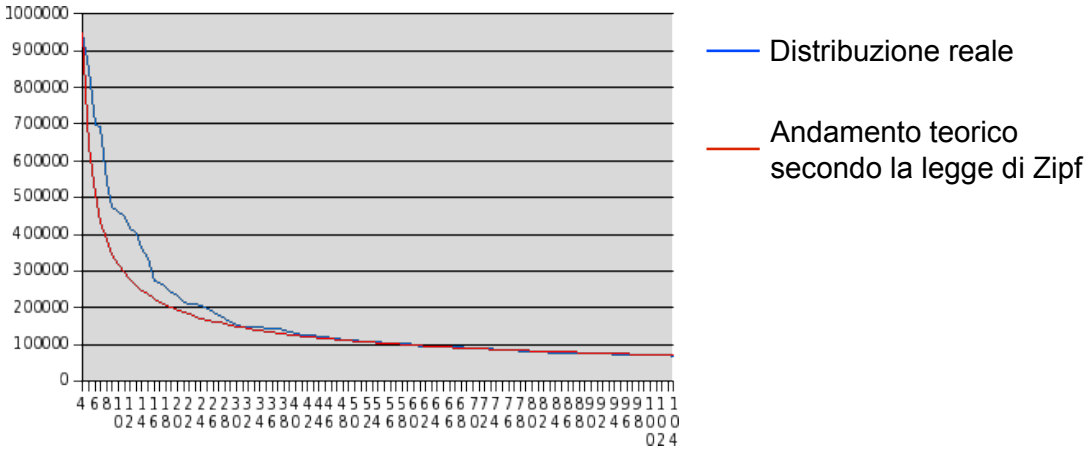


Legge di Zipf: esempio (I)



Legge di Zipf: esempio (II)

- Si applica anche ai termini usati dagli utenti nelle ricerche
- AOL Query Database: collezione di ~20M di ricerche web eseguite da ~650k utenti in tre mesi (marzo-maggio 2006)



- Principio del minimo sforzo:
  - Si tende a minimizzare lo sforzo di trovare parole nuove per esprimere concetti ricorrenti

Legge di Zipf: utilizzo

- Le parole troppo o troppo poco frequenti non sono utili nell'indicizzazione
  - Troppo → presenti in tutti i documenti quindi non aiutano a discriminare tra i documenti
  - Troppo poco → troppi termini usati raramente, “ingrassano” l'indice senza molti effetti positivi
- La legge di Zipf può essere usata come guida per eliminare le stopword e trovare i termini utili all'indicizzazione da includere nel vocabolario

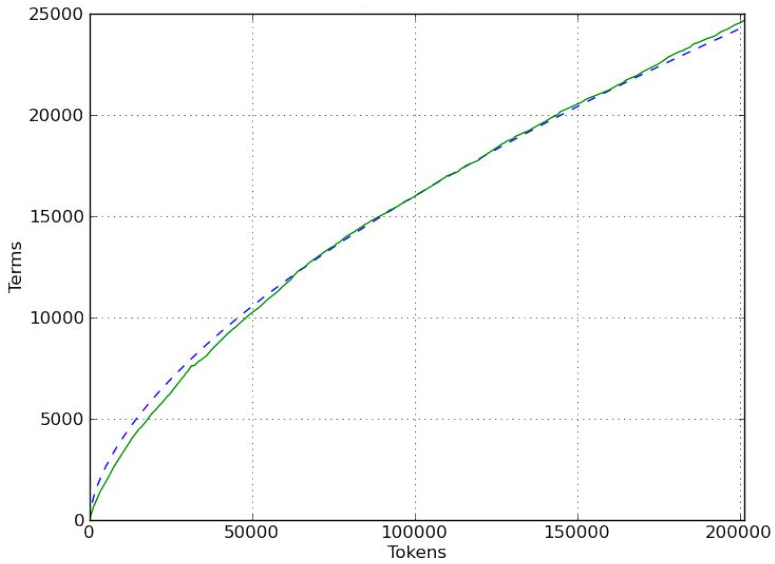
Legge di Heaps (I)

- E qual'è il rapporto tra il numero di termini distinti e il numero di documenti nella collezione??
- Dato una collezione con *n* parole la dimensione V del vocabolario d'indicizzazione è data da:

$$V = Kn^{\beta}$$

- Dove *K* e *β* sono parametri liberi
  - Dipendono dal testo (e quindi dalla lingua)
  - In inglese tipicamente $10<K<100$ e $0.4<\beta<0.6$

Legge di Heaps (II)



- Asse orizzontale: numero di parole nella collezione
- Asse verticale: numero di termini nel vocabolario
- Attenzione ai diversi ordini di grandezza sui due assi!!!

Legge di Heaps (III)

- Man mano che il vocabolario cresce, ogni documento avrà sempre meno termini “nuovi” non ancora inseriti nel vocabolario e avrà invece sempre più termini già “visti” e “conosciuti”
- Ogni documento aggiunge sempre termini al vocabolario rispetto al documento analizzato precedentemente
- La curva cresce inizialmente velocemente per poi rallentare sempre più
  - Diventa orizzontale quando tutti i termini della lingua sono stati inseriti nel vocabolario
- I documenti possono essere moltissimi/infiniti, ma i termini invece sono sempre in numero limitato