Lezione 2.1

Text Mining e applicazioni

Luigi Di Caro

Text Mining e applicazioni

- Panoramica
  - Il testo secondo la statistica
  - Frequenze e co-occorrenze
    - tag clouds, tag flakes
- Document-level Text Mining
  - Clustering, Categorization/Classification, Segmentation, Summarization, Information Retrieval, Browsing, Orienteering, Concept enrichment
- Lab

Il testo secondo la statistica

- **Linguistica Computazionale (o Natural Language Processing):** studio di formalismi descrittivi del funzionamento del linguaggio naturale, che permettano di essere trasformati in programmi eseguibili dai computer.

  approccio top-down

- **Statistica:** disciplina che studia qualitativamente e quantitativamente particolari fenomeni

  approccio bottom-up

Il testo secondo la statistica

approccio top-down

Studio delle regole che governano l'espressione linguistica

Estrazione delle regole attraverso l'evidenza statistica

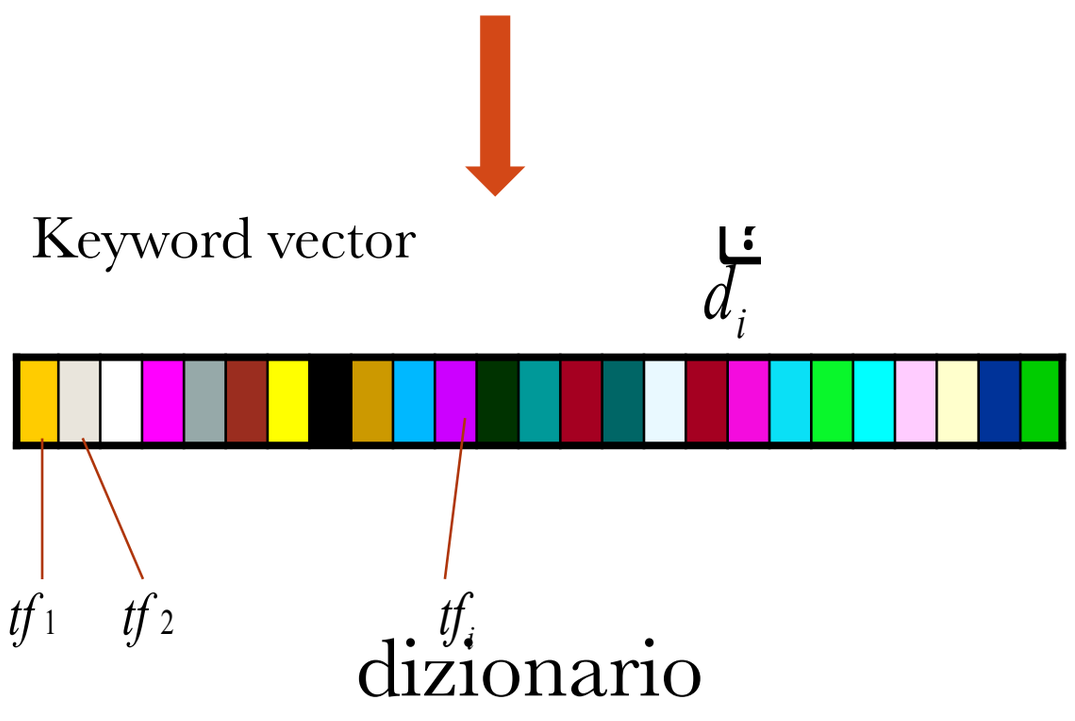
approccio bottom-up

Il testo secondo la statistica

- Le parole sono token (sequenze di caratteri contigui)
- Un testo è un insieme di token, con una certa frequenza

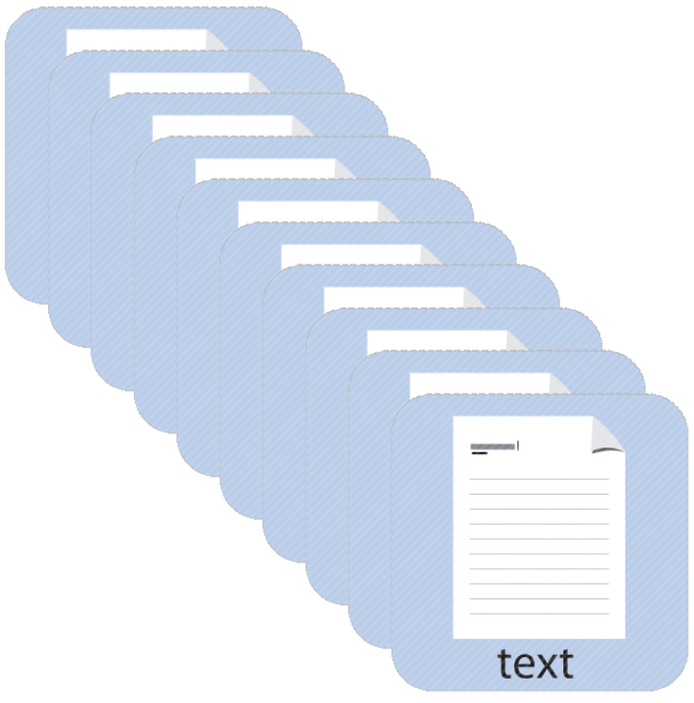
Rappresentazione Vettoriale (Vector Space Model)

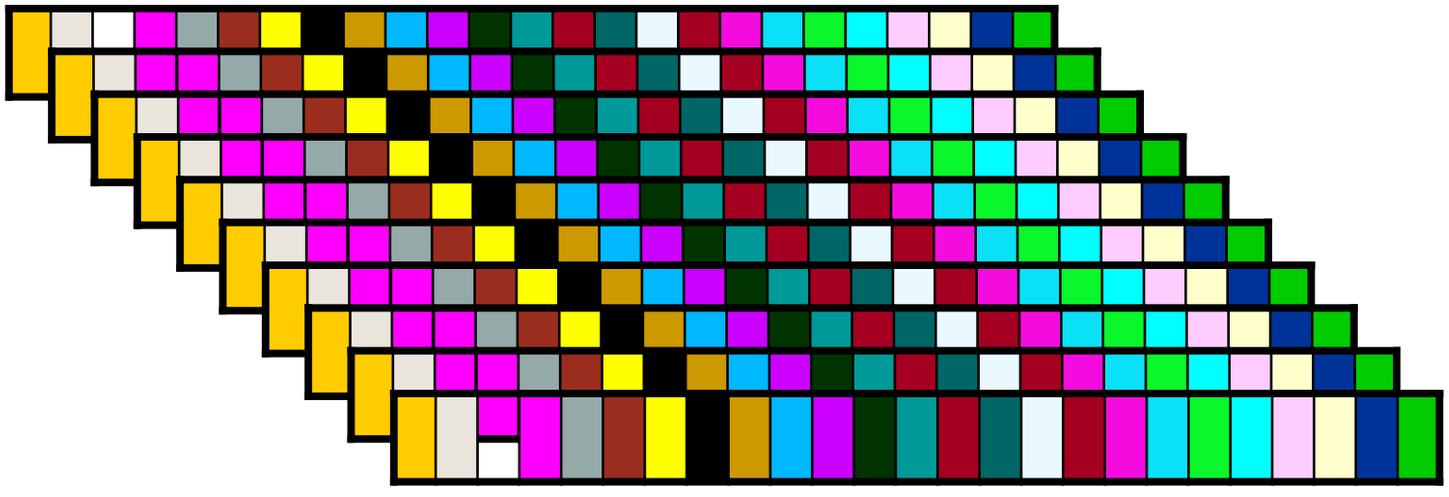
Modello algebrico di rappresentazione introdotto da Salton (1975)



Il testo secondo la statistica

Rappresentazione Vettoriale
(Vector Space Model)





Corpus →Matrice numerica

Il testo secondo la statistica

Rappresentazione Vettoriale
(Vector Space Model)

Perché?

- Rappresentando i testi come vettori numerici, è possibile effettuare operazioni matematiche per il confronto
  - una tra tutte: Cosine Similarity (Misura coseno)

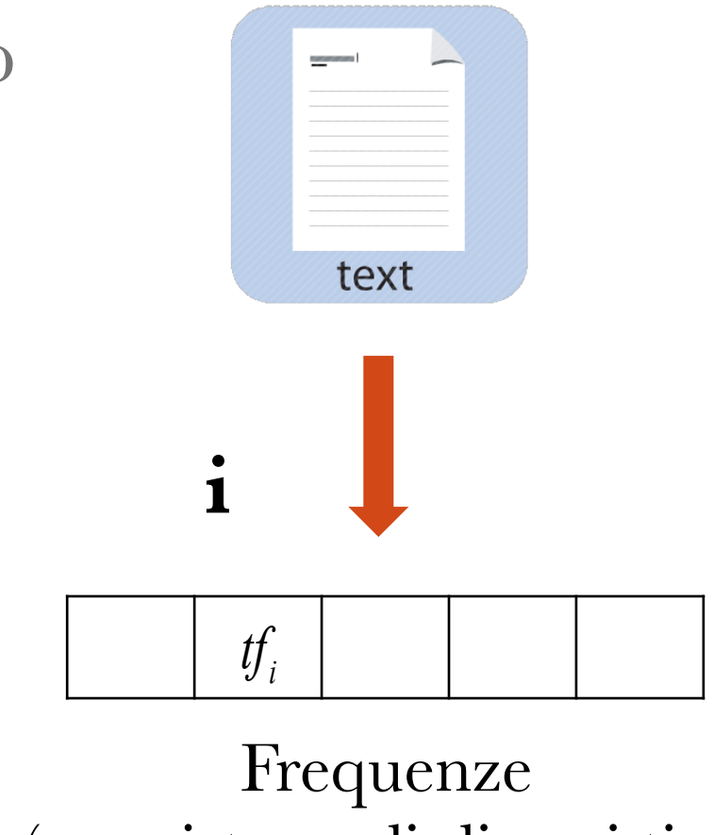
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Il testo secondo la statistica

- I metodi statistici applicati ai documenti di testo si concentrano essenzialmente su due tipi di informazioni statistiche:
  - **Frequenza** di una parola in un testo
  - Co-occorrenza di due parole in un testo

Tipo di frequenza più usato

TF IDF
Term Frequency
Inverse Document Frequency



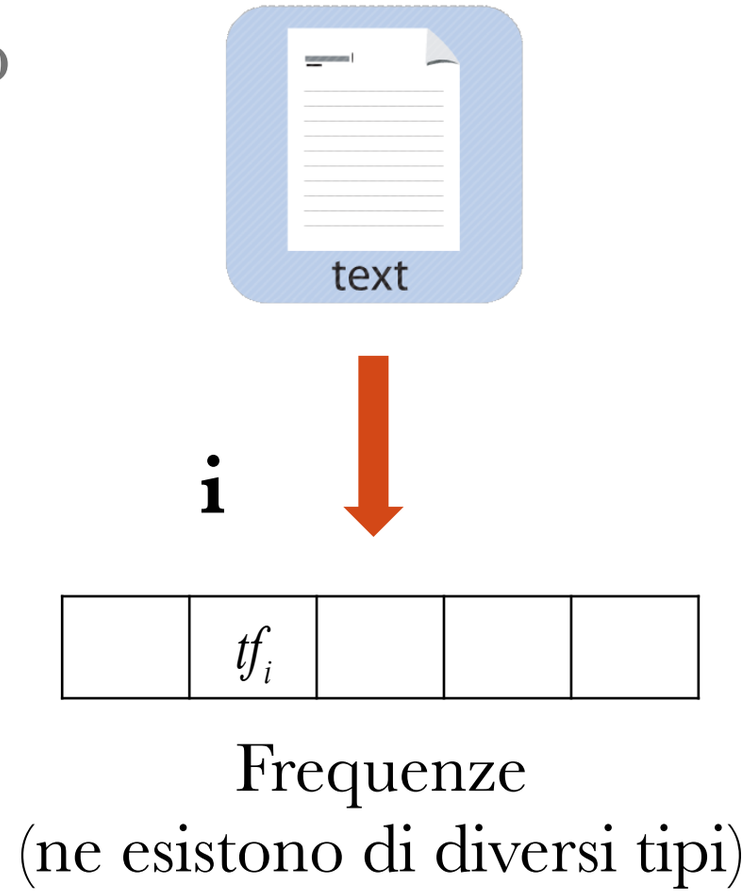
Frequenze
(ne esistono di diversi tipi)

Il testo secondo la statistica

- I metodi statistici applicati ai documenti di testo si concentrano essenzialmente su due tipi di informazioni statistiche:
  - **Frequenza** di una parola in un testo
  - Co-occorrenza di due parole in un testo

TF IDF
Term Frequency
Inverse Document Frequency -> log(nd/ndt)

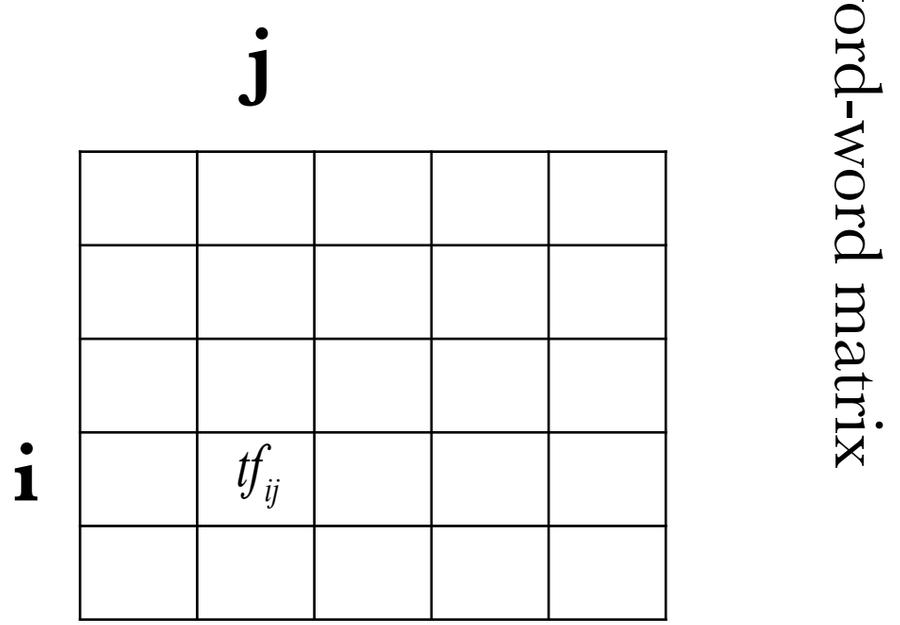
Caso 1: articolo (“the”) —> valore IDF=0
Caso 2: “cat" —> valore IDF positivo



Frequenze
(ne esistono di diversi tipi)

Il testo secondo la statistica

- I metodi statistici applicati ai documenti di testo si concentrano essenzialmente su due tipi di informazioni statistiche:
  - Frequenza di una parola in un testo
  - **Co-occorrenza** di due parole in un testo

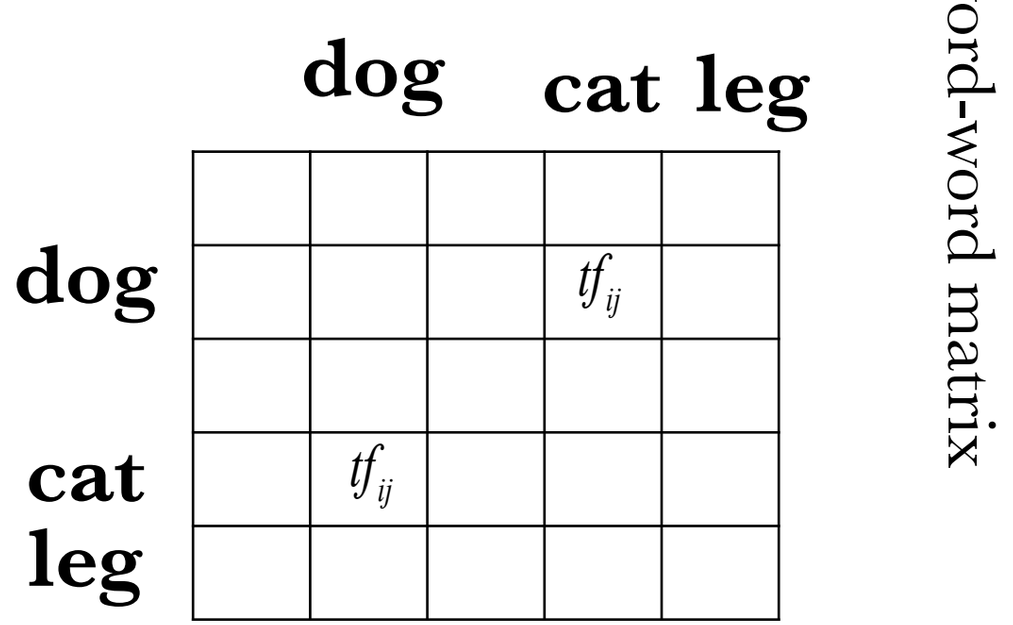


Co-occorrenze
(ne esistono di diversi tipi)

Il testo secondo la statistica

- I metodi statistici applicati ai documenti di testo si concentrano essenzialmente su due tipi di informazioni statistiche:
  - Frequenza di una parola in un testo
  - **Co-occorrenza** di due parole in un testo

Wikipedia page - "Pets"



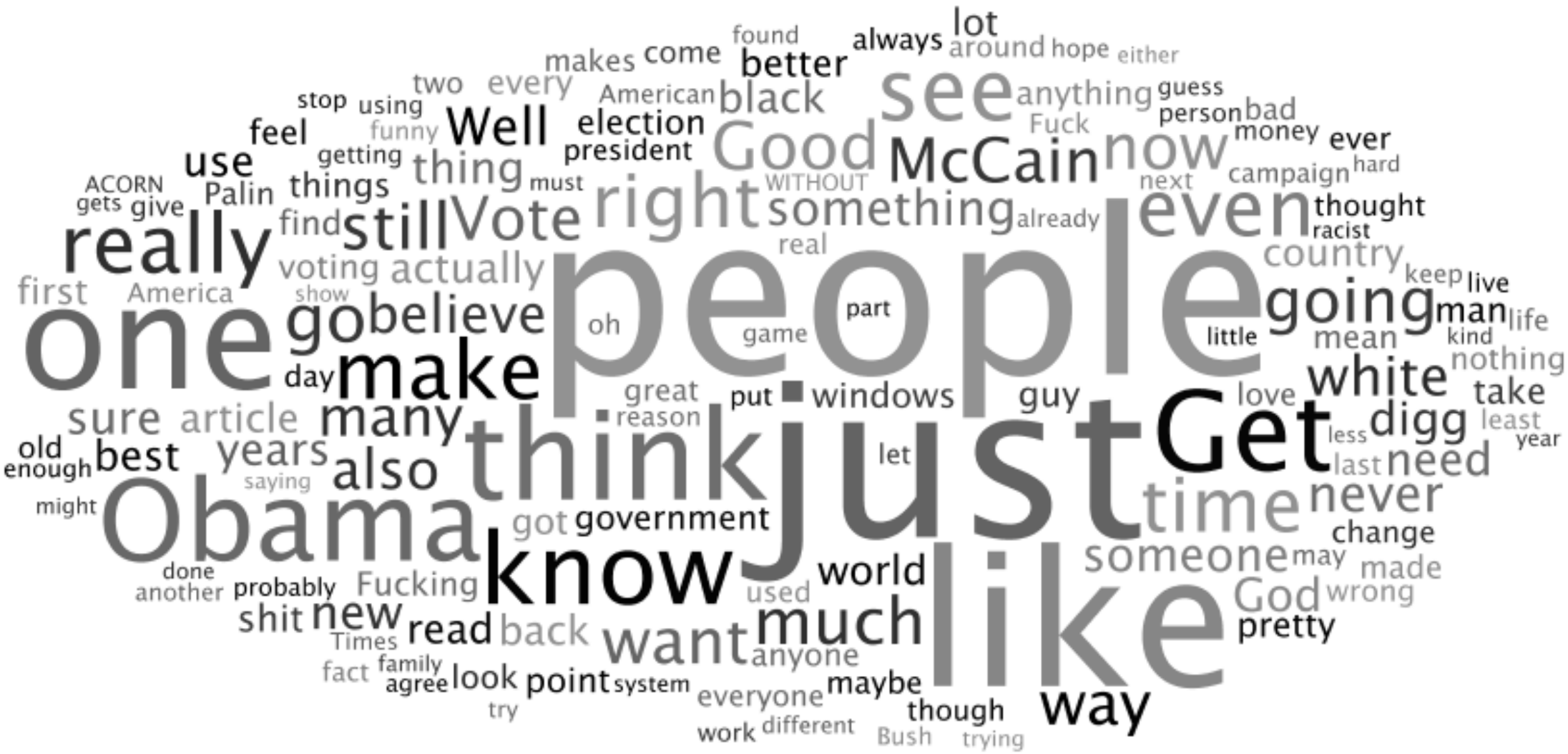
Co-occorrenze
(ne esistono di diversi tipi)

Il testo secondo la statistica

- I metodi statistici applicati ai documenti di testo si concentrano essenzialmente su due tipi di informazioni statistiche:
  - **Frequenza** di una parola in un testo
    - Indica l'importanza, la rilevanza, la "dominance", la significatività di un termine nel testo
  - **Co-occorrenza** di due parole in un testo
    - Indica la similarità tra due parole (sempre da intendersi come similarità statistico-semantica), assumendo che due parole con simile significato siano presenti negli stessi contesti
      - **Contesto?** Il contesto è un concetto che può essere associato alla vicinanza fisica delle parole nei testi, e viene spesso definito come l'insieme di parole che due termini hanno nel loro "intorno".

Tag Clouds: prima (banale) applicazione

- Indovinate il topic



Tag Clouds: prima (banale) applicazione

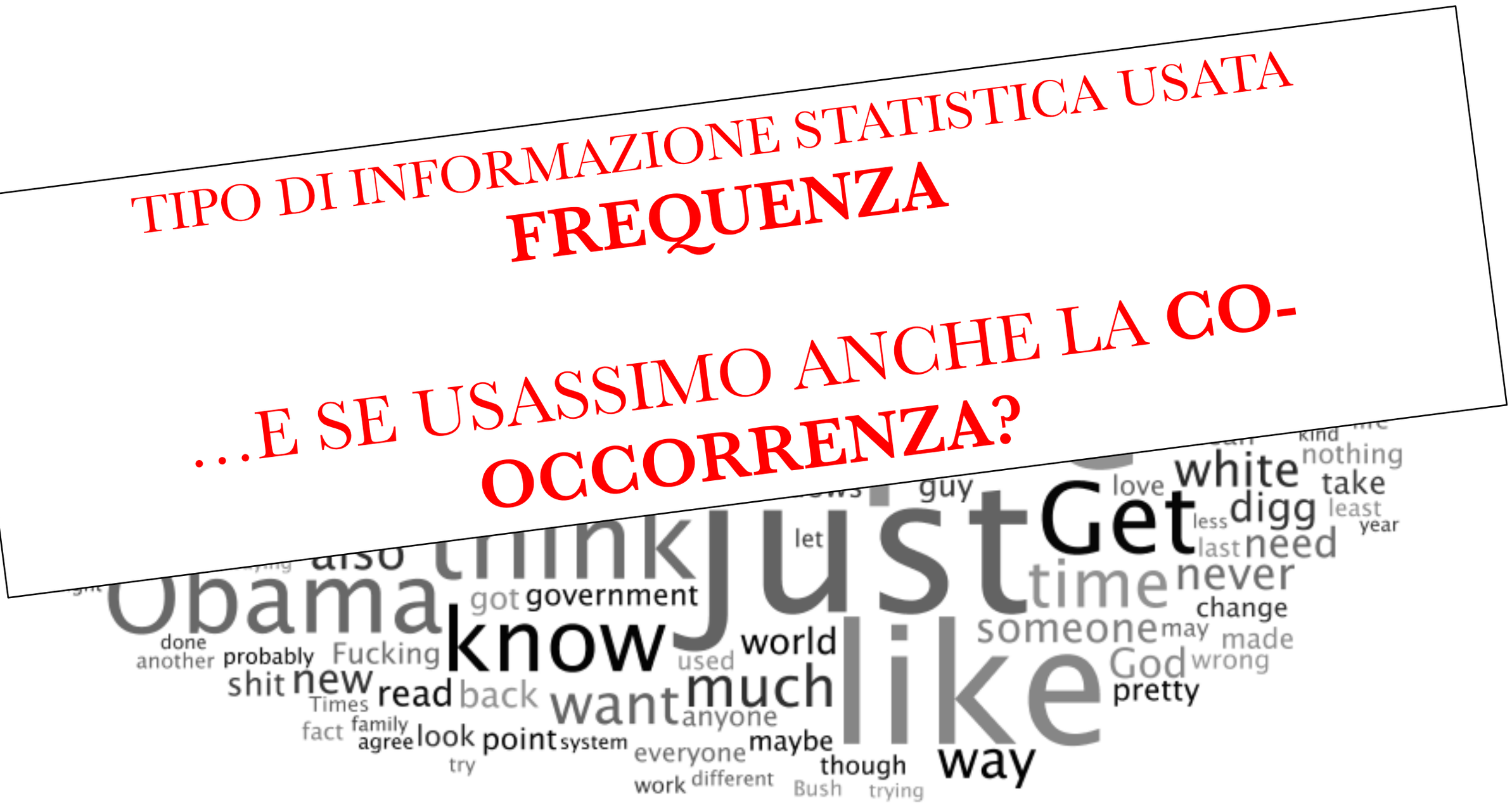
Tag Clouds: prima (banale) applicazione

TIPO DI INFORMAZIONE STATISTICA USATA

FREQUENZA

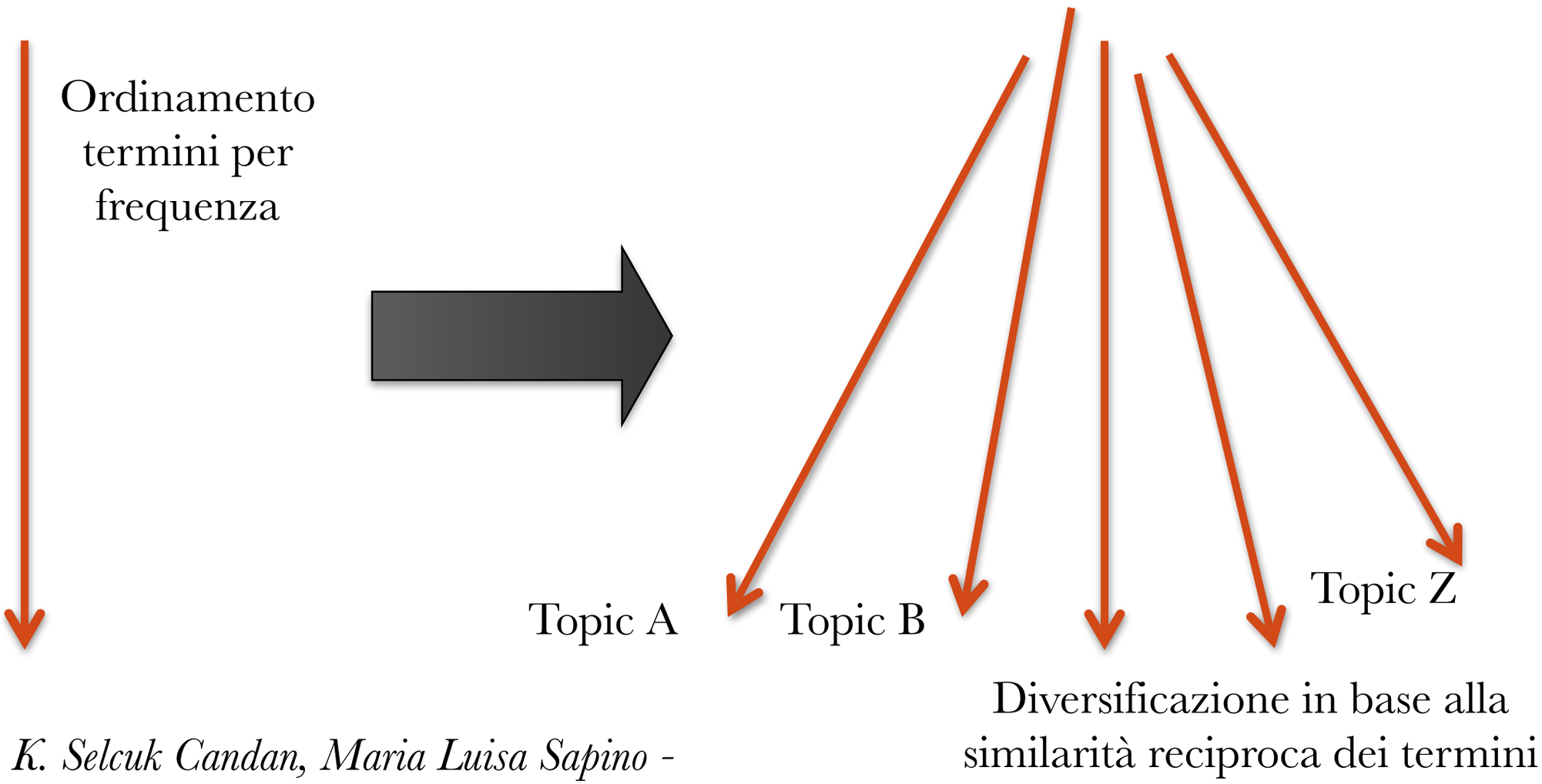
…E SE USASSIMO ANCHE LA CO-

OCCORRENZA?



Tag Flakes: seconda (meno banale) applicazione

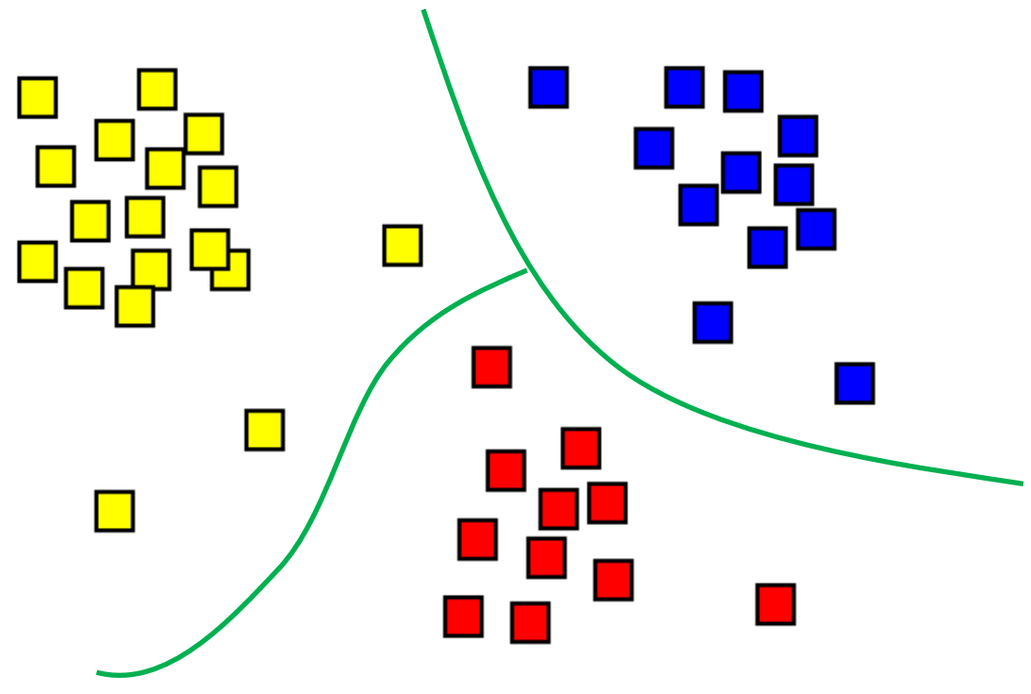
- Estrazione automatica di una gerarchia di termini



Luigi Di Caro, K. Selcuk Candan, Maria Luisa Sapino - "Navigating within News Collections using Tag-Flakes", Journal of Visual Languages and Computing, 2011.

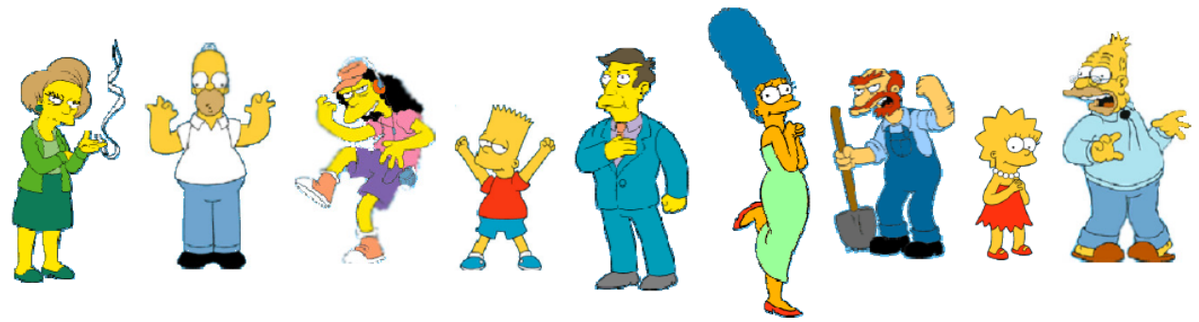
Document Clustering

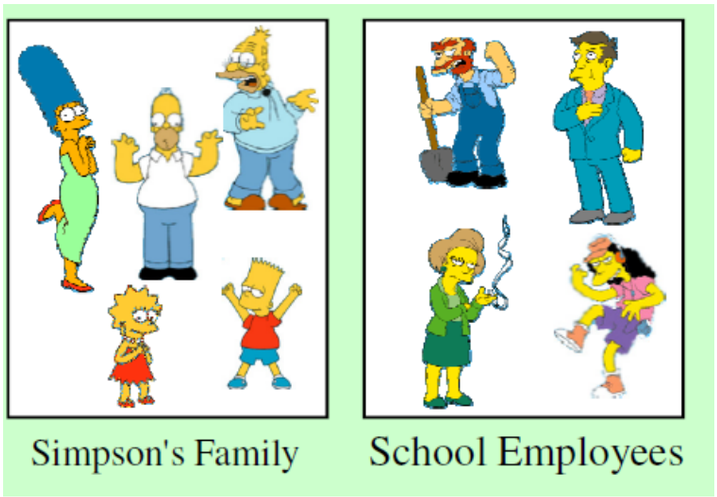
- Clustering
  - Tecniche di selezione e raggruppamento di elementi omogenei in un insieme di dati
  - Tutte le tecniche di *clustering* si basano sul concetto di distanza tra due elementi

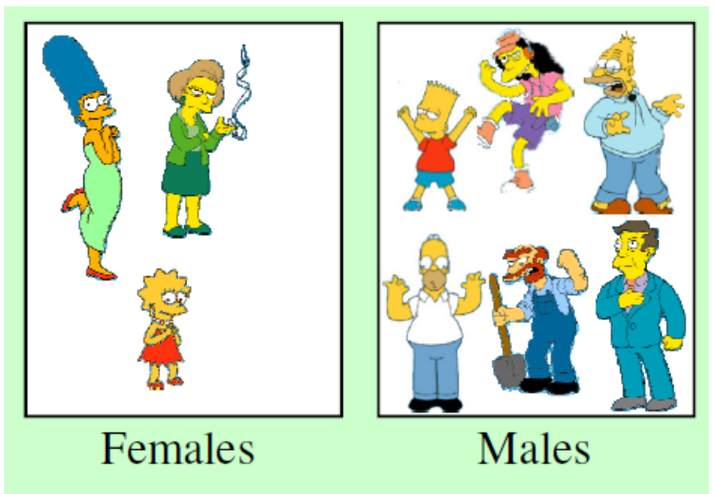


Document Clustering

- Clustering
  - I dati possono essere raggruppati in modi differenti a seconda della misura di distanza adottata

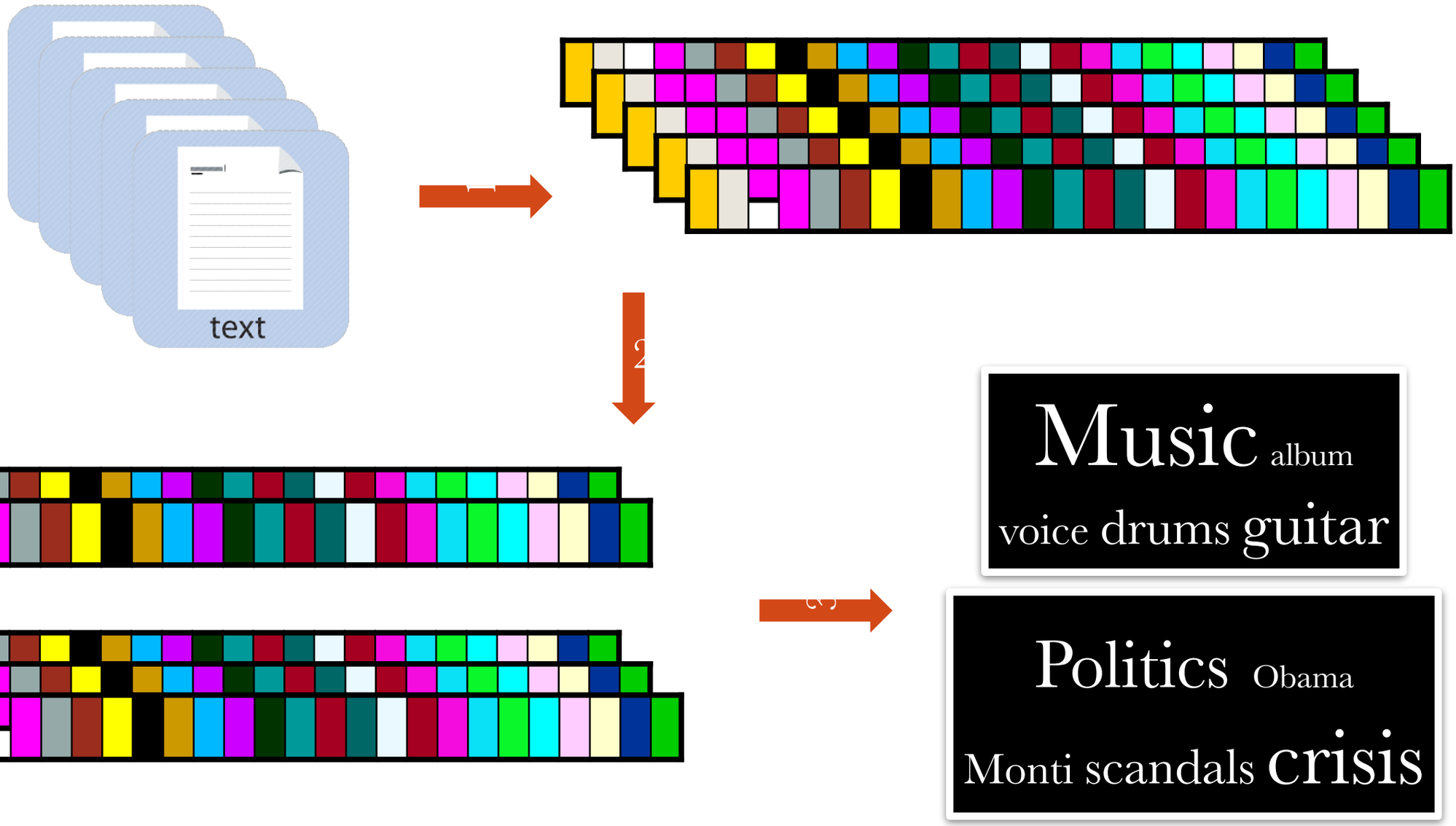






Document Clustering

- Esempio di utilizzo combinato di Document Clustering con Tag Clouds



Document Categorization / Classification

- Come associare documenti di testo ad una tassonomia?

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

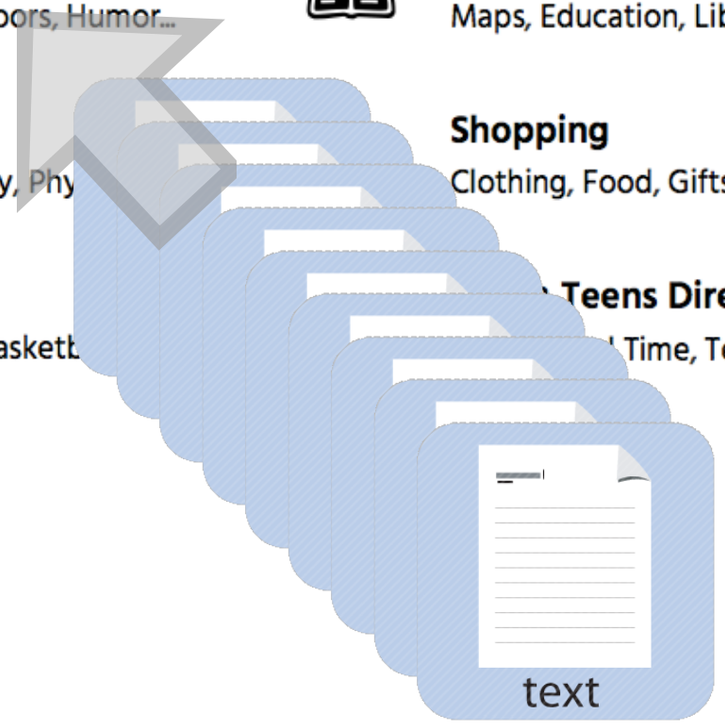
Science
Biology, Psychology, Phy

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

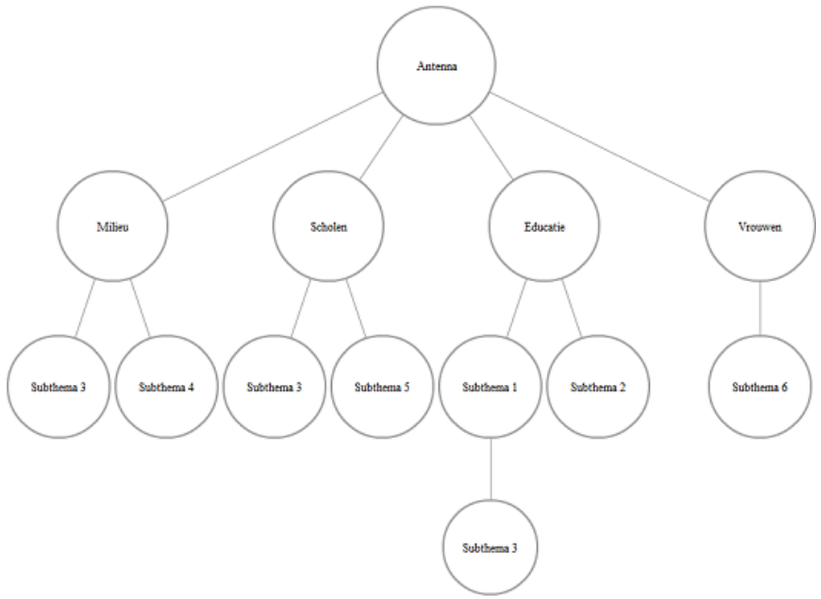
Sports
Baseball, Soccer, Basketb

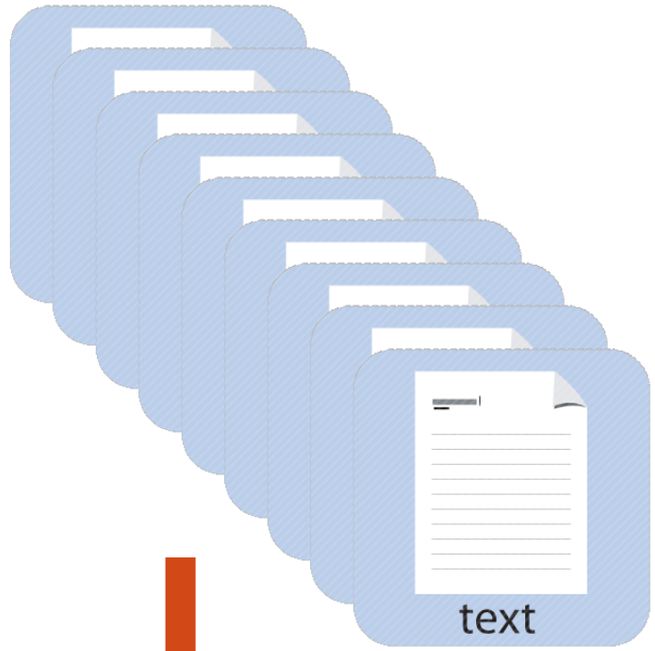
Teens Directory
Time, Teen Life...

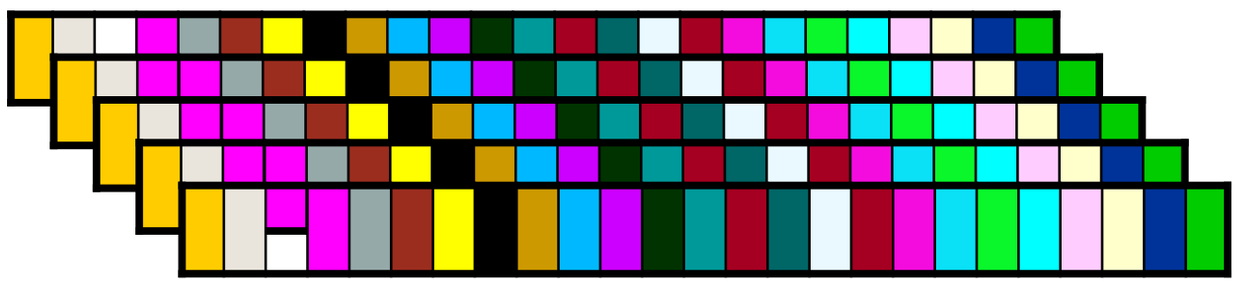


Document Categorization / Classification

- Come associare documenti di testo ad una tassonomia?

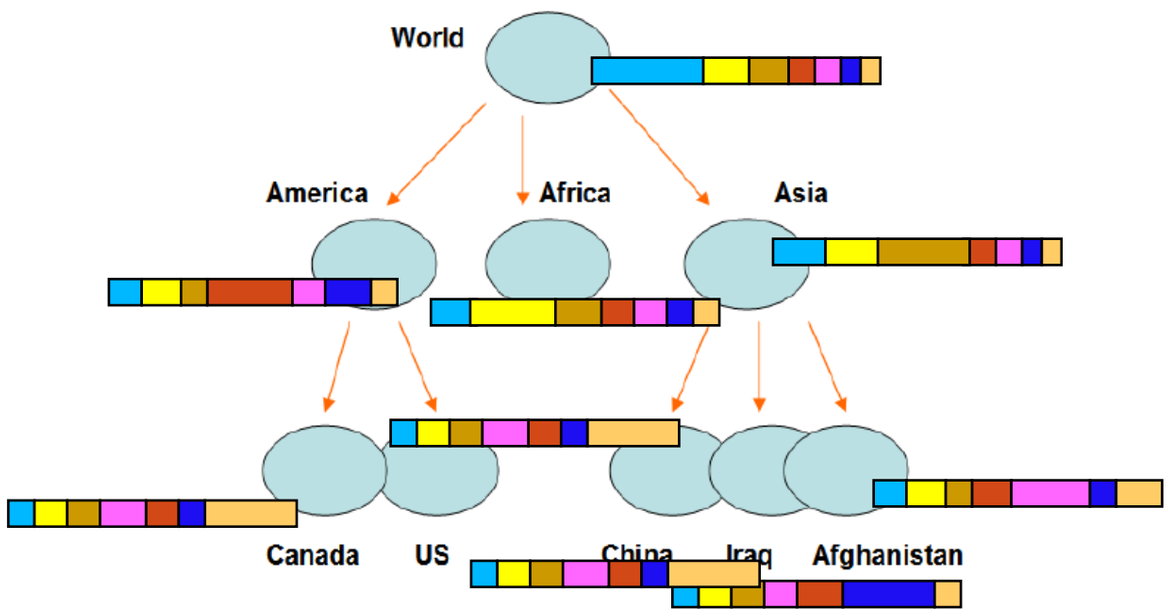






Document Categorization / Classification

- Come associare documenti di testo ad una tassonomia?



Ad ogni nodo viene associato un vettore numerico, le cui dimensioni rappresentano i nodi stessi della tassonomia

1 - Inizializzazione

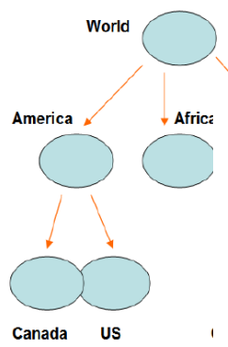
2 - Processo di propagazione

Jong Wook Kim, K. Selçuk Candan: CP/CV: concept similarity mining without frequency information from domain describing taxonomies. CIKM 2006

Document Categorization / Classification

- Come associare documenti di testo ad una tassonomia?

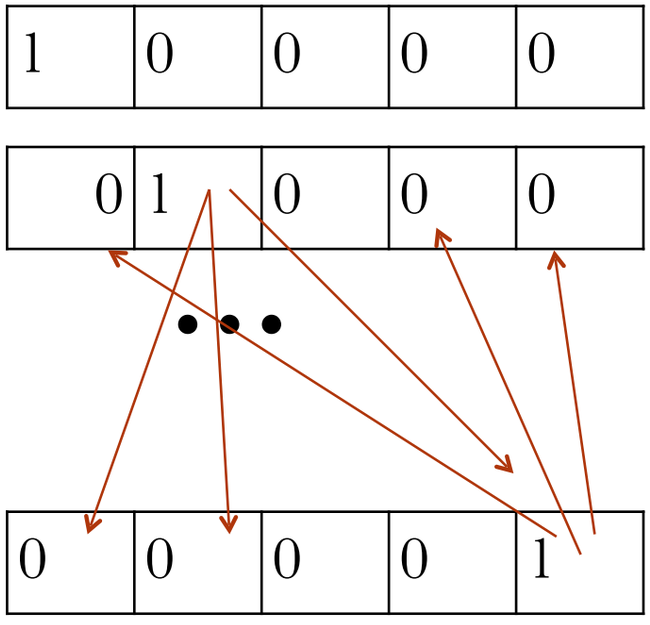
	c1	c2			c5
c1	1	0	0	0	0
c2	0	1	0	0	0
...					
c5	0	0	0	0	1

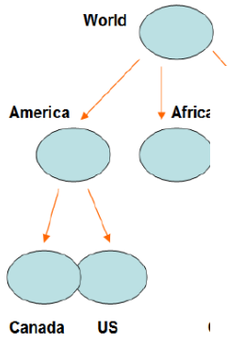


initialization

Document Categorization / Classification

- Come associare documenti di testo ad una tassonomia?

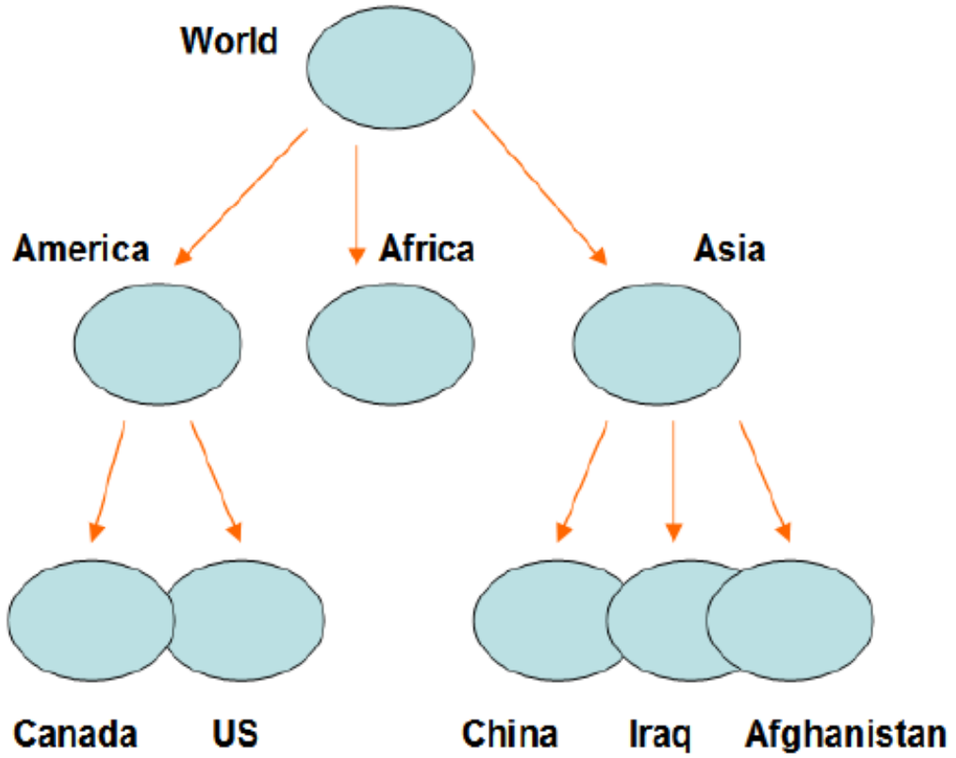




Processo di propagazione

Document Categorization / Classification

Esempio: Tassonomia geografica

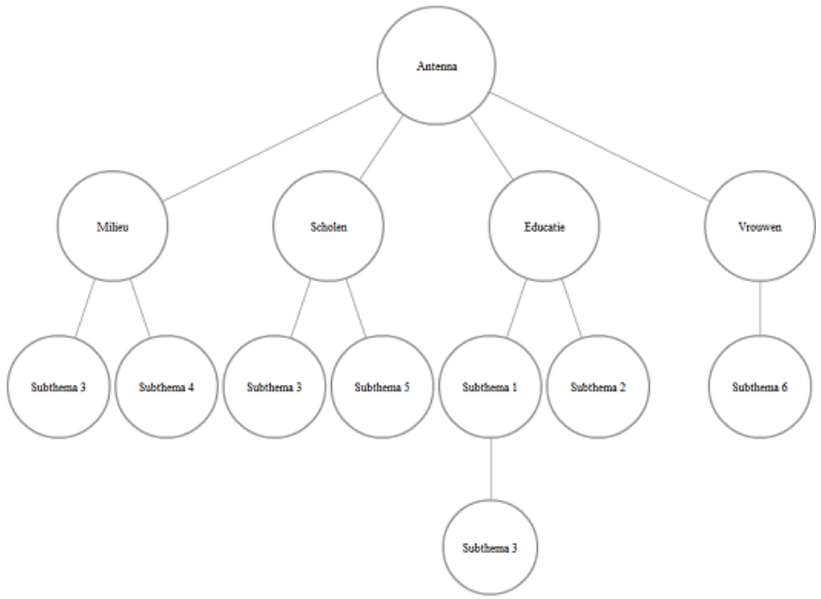


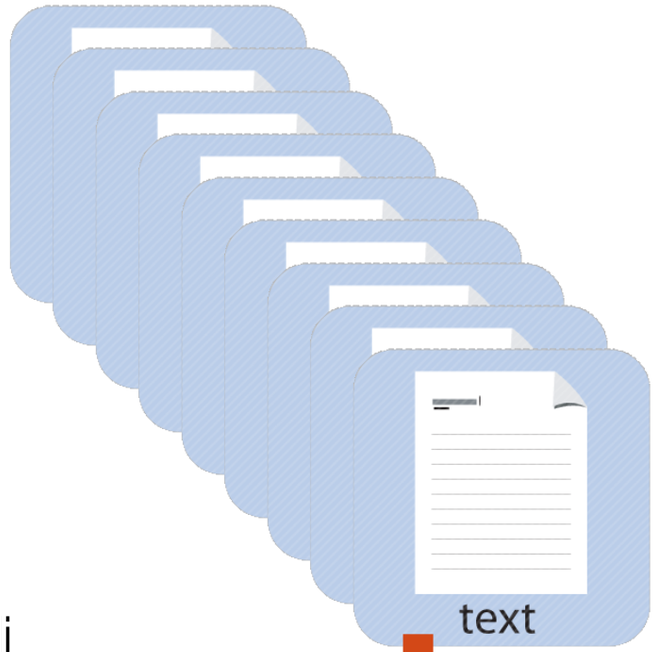
Concept vectors

	world	Asia	Africa	America	Afghanistan	Iraq	China	Canada	US
$\vec{cv}_{world}$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\vec{cv}_{Asia}$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\vec{cv}_{Africa}$	0.100	0.012	0.873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\vec{cv}_{America}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\vec{cv}_{Afghanistan}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\vec{cv}_{Iraq}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\vec{cv}_{China}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\vec{cv}_{Canada}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\vec{cv}_{US}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806

Document Categorization / Classification

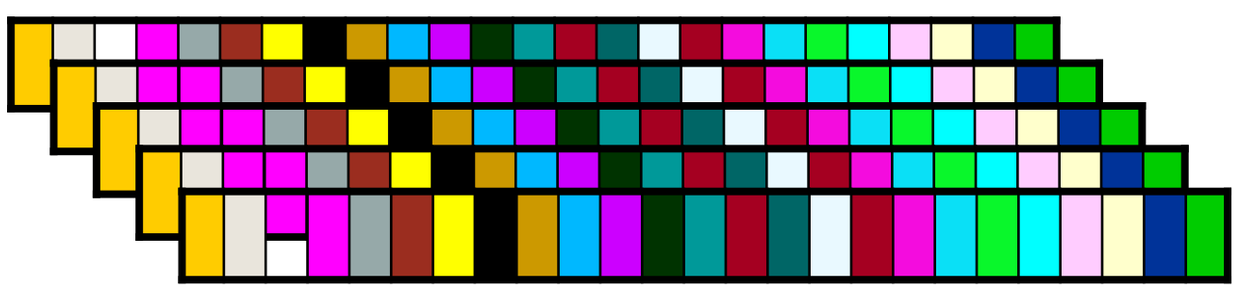
- Come associare documenti di testo ad una tassonomia?





Fusione dei due spazi vettoriali
E calcolo similarità per classificazione (unsupervised)

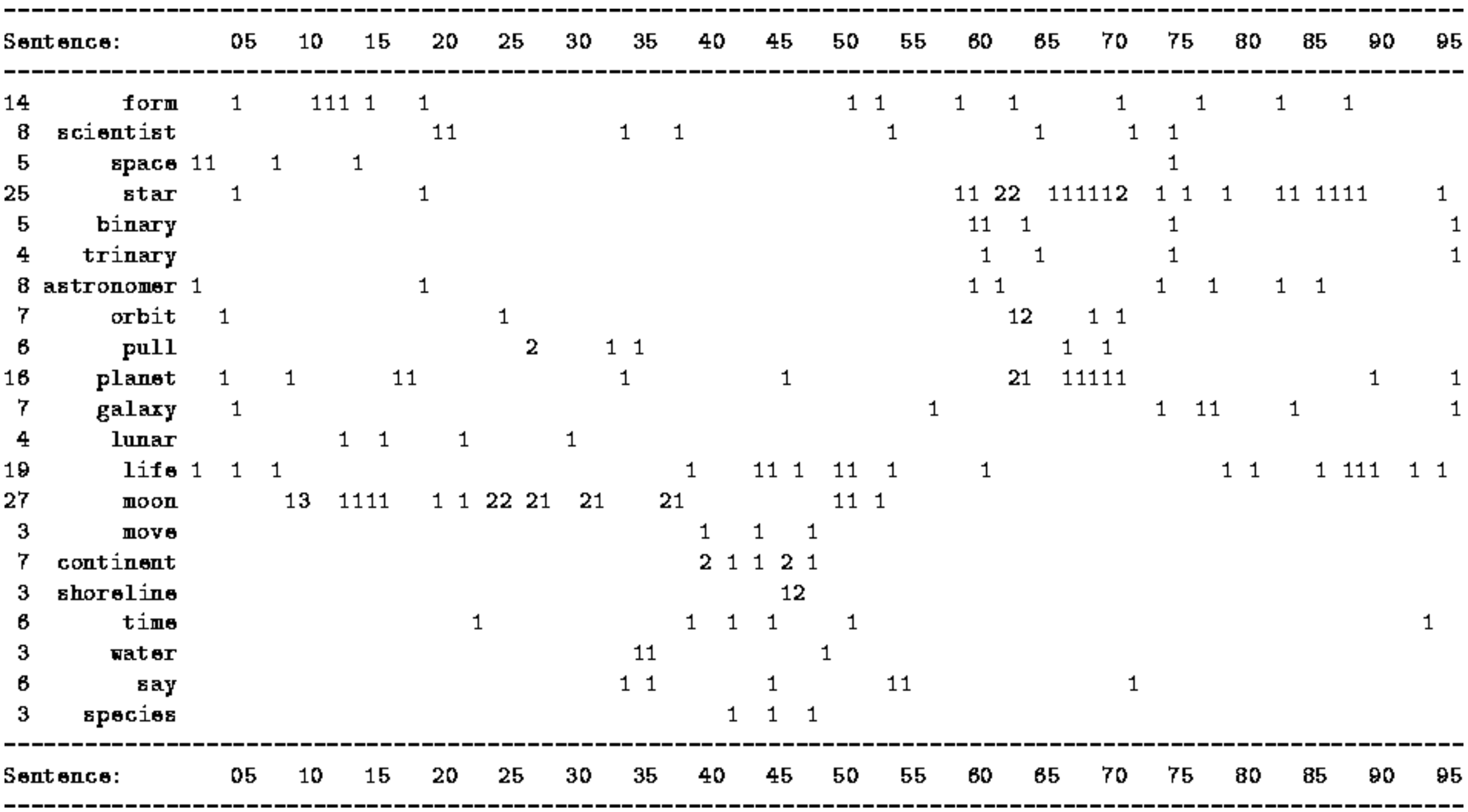
	world	Asia	Africa	America	Afghanistan	Iraq	China	Canada	US
$\vec{cv}_{world}$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\vec{cv}_{Asia}$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\vec{cv}_{Africa}$	0.100	0.012	0.873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\vec{cv}_{America}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\vec{cv}_{Afghanistan}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\vec{cv}_{Iraq}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\vec{cv}_{China}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\vec{cv}_{Canada}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\vec{cv}_{US}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806



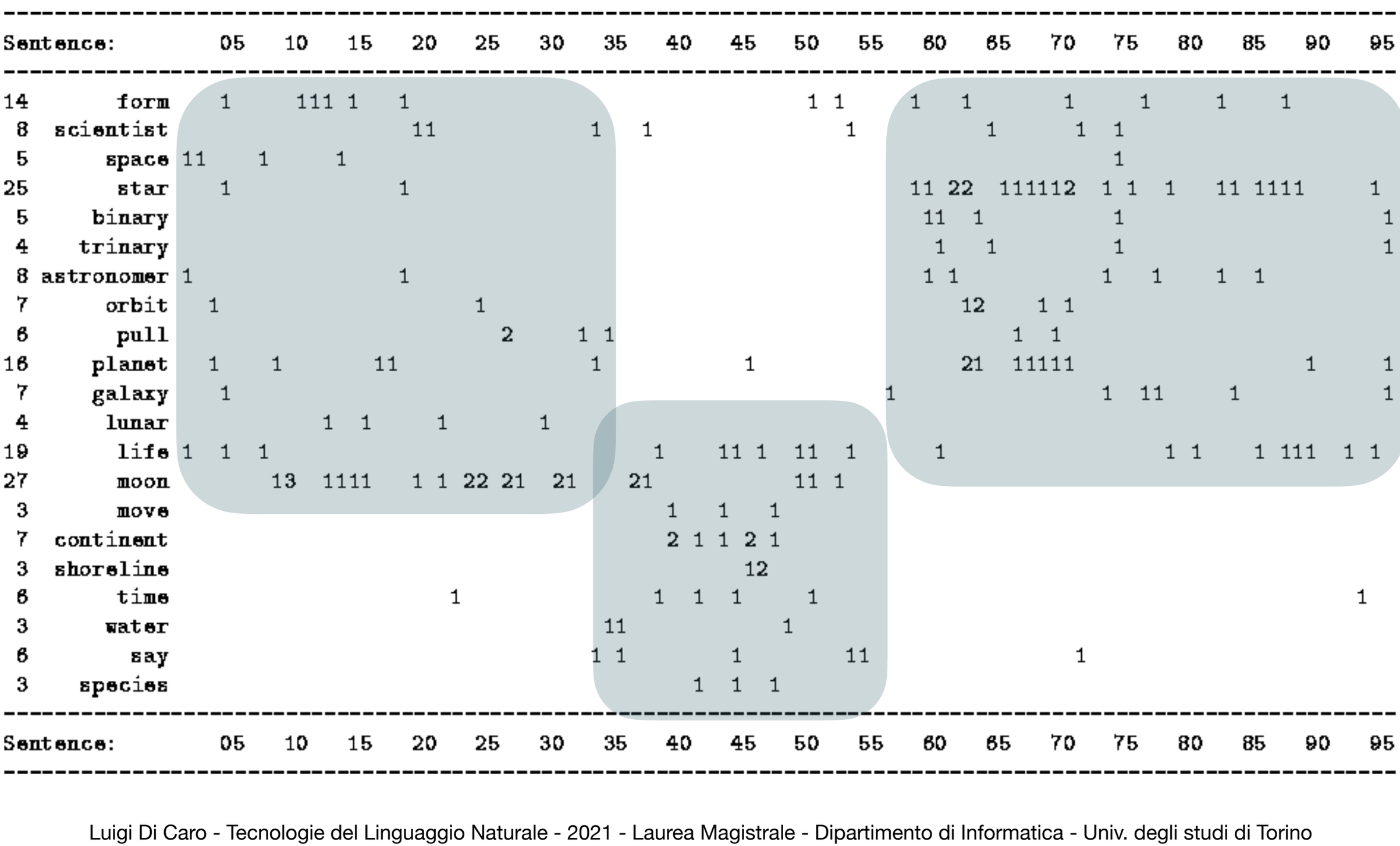
Document segmentation

- Individuazione di elementi diversi del discorso
  - evoluzione di temi, entità, relazioni durante la lettura del testo
  - metodi automatici per la segmentazione
    - basati su word count, co-occurrence, etc.
    - uno dei più famosi: *text tiling*

Document segmentation



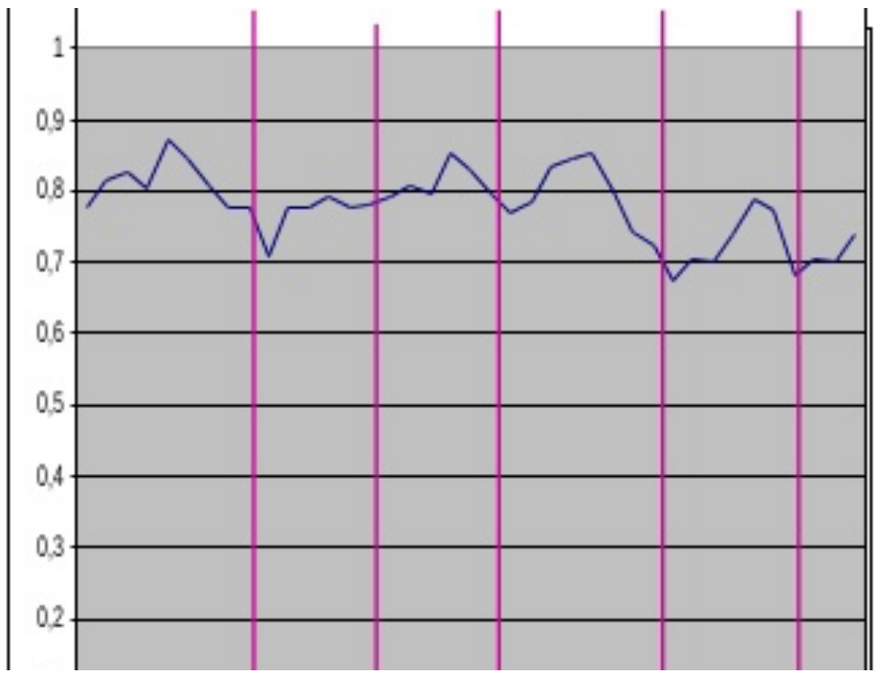
Document segmentation



Document segmentation

- *text tiling*
  - *separazione del testo in finestre di lunghezza fissa*
  - *calcolo della coesione intra-gruppo*
  - *ricerca di parti di testo a bassa coesione circondate da parti di testo ad alta coesione (break point)*
- *riadattamento finestre rispetto al break point più vicino*

TextTiling: Segmenting text into multi-paragraph subtopic passages. MA Hearst - Computational linguistics, 1997



Document summarization

- riduzione del testo, mantenimento massimale della semantica contenuta

- metodi
  - estrattivi (keyphrases, TextRank, etc.)
  - astrattivi (implica NLG)

- valutazione: ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Orienteering & Browsing

- Problema di partenza: a volte le persone non sanno come cercare le informazioni
  - I sistemi possono essere basati sul concetto di "*orienteering*", ovvero una serie di meccanismi che aiutano gli utenti a navigare attraverso i topic sfruttando le relazioni tra i documenti di testo.
  - *Evidenziare* relazioni latenti tra i documenti
  - *Abilitare* la navigazione guidata dal contesto
  - *Supportare* la comprensione dei dati
    - —> **orientamento**

O'Day,V.& Jeffries, R. (1993). Orienteering in an information landscape: how information seekers get from here to there. In Proc. CHI 1993, 438-445.

Information Retrieval

- Recupero di un documento di interesse
  - Usando query basate su keyword
    - o basate su concetti (si veda categorization)
  - Inizialmente basata su modello booleano (match diretto tra query e contenuti)
  - Sviluppi
    - Navigazione aumentata attraverso links, snippets, etc.
    - Integrazione di immagini, video, mappe
    - Modelli avanzati di interazione (ad es. chatbot, visualization)

Laboratorio - Segmentation

- Implementare un semplice algoritmo su Text segmentation

- Usare come test un input di k paragrafi presi da differenti temi (ad es. pagine Wikipedia)

- Il vostro sistema è in grado di trovare i giusti “tagli”?