Miniere di conoscenza

INFORMATION MINING

Master in Marketing e Servuce Management

Simona Balbi

26 marzo 2014

Miniere di conoscenza

Grazie alle nuove tecnologie per raccogliere ed archiviare i dati, la quantità di informazione disponibile in formato elettronico aumenta in maniera incontrollabile

Basti pensare a che cos c'è memorizzato nel nostro cellulare (senza pensare all'elaborazione di questi dati) e possiamo immaginare la quantità di dati che, ad esempio, un'azienda possiede "inconsapevolmente"

Di solito, infatti, nessuno analizza questi dati!!!

(se non a fini amministrativi)

Le unità delle basi di dati elettroniche sono sempre superate dalla realtà: i *terabyte* (10^{12}) appartengono alla storia**,** gli *yottabyte* (10^{24}) *non sono più* sufficienti, i **petabyte sono oggi necessari, ed anche oltre (come i $ di Zio Paperone)**

Siamo sopraffatti dai dati e possiamo soltanto sospettare che al loro interno ci siano informazioni preziose, cioè CONOSCENZA (Knowledge)

Knowledge Discovery in Databases (KDD)

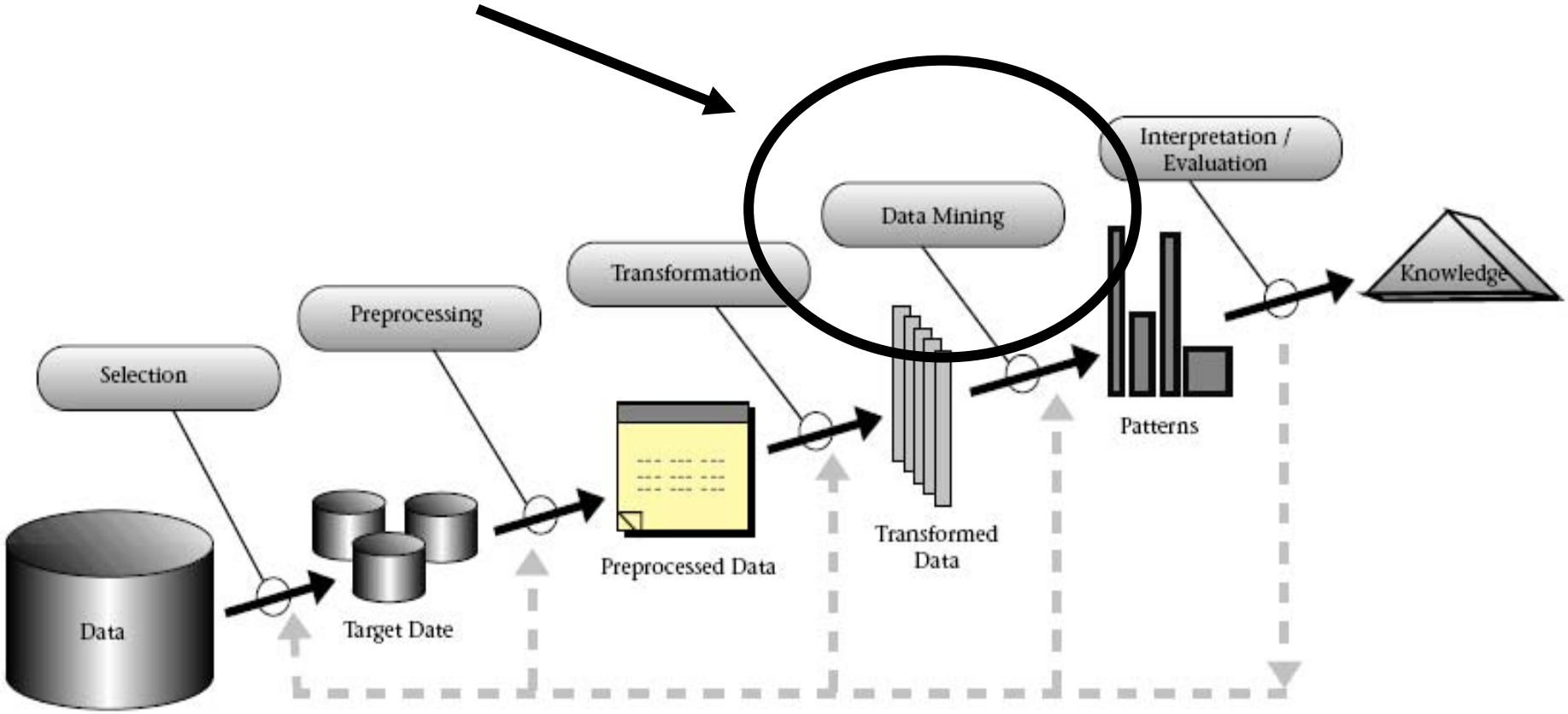
Trasformare i dati (**troppo voluminosi**) di un **livello inferiore**, in **forme più compatte** (ad esempio brevi sintesi), **più astratte** (ad esempio un'approssimazione descrittiva, oppure un modello del processo di generazione dei dati), o **più utili** (ad esempio, un modello predittivo per stimare il valore di casi futuri).

Al centro del processo è l'applicazione di specifici metodi di Data Mining (DM) per l'identificazione e l'estrazione di **pattern**

*N.B. La parola **pattern** (genreralmente **found in data**) è ricorrente nella letteratura del DM, seguendo Fayyad et al., lo possiamo intendere come "**components of models, for example, a particular rule in a classification model or a linear component in a regression model**"*

Knowledge Discovery in Databases (KDD)

KDD ha l'obiettivo di sviluppare metodi e tecniche per dare senso ai dati



(Fayyad, Piatetsky-Shapiro, Smyth, AI Magazine, 1996)

Le fasi in dettaglio 1.

Le fasi in dettaglio 1.

1. Capire il contesto applicativo e l'**obiettivo** conoscitivo del committente
2. Creare il ***target data set***: selezionare un insieme di dati, oppure concentrarsi su un sottoinsieme di variabili o di individui (campione)
   pensiamo di voler studiare l'associazioni di prodotti acquistati dai clienti di un supermercato, avendo a disposizione la scheda fedeltà, associata ad un breve questionario descrittivo, ad esempio, del titolo di studio del cliente. Questa informazione potrebbe essere inutile, in questa analisi, mentre invece potrebbe essere rilevante per una specifica campagna promozionale, ad esempio di un abbonamento ad una rivista aziendale. Spesso, può non essere conveniente o necessario analizzare l'intera base di dati e può convenire esplorarlo preventivamente e giungere alla scelta di un'analisi su base campionaria.
3. **Pulire** i dati e **pre-trattarli**: eliminare il rumore, se necessario, raccogliere le informazioni sul modello di generazione dei dati, decidere le strategie per gestire i **dati mancanti**, o tener conto, ad esempio, di incoerenze e di eventuali errori. Possono essere necessarie codifiche o standardizzazioni. Può essere necessario rendere coerenti dati provenienti da fonti differenti (necessità dei *metadati*)

Riduzione dei dati e proiezione:

individuare utili caratteristiche per rappresentare i dati in relazione dell'obiettivo

Abbinare gli obiettivi di KDD (fase 1) ad uno specifico metodo di **DM** (ad es. sintesi, classificazione, regressione, clustering, …)

L'analisi esplorativa e la **selezione del modello** e delle ipotesi: scegliere i metodi e gli algoritmi di DM al fine di trovare i pattern nei dati. Si tratta di decidere modelli e parametri appropriati (ad esempio alla natura dei dati, quindi scelte di codifica: discretizzare i dati?) e abbinare i metodi di DM con i criteri generali definiti nel processo di KDD (ad es. c'è una necessità di predizione, o di comprendere le relazioni?)

Le fasi in dettaglio 3.

E' il DM: ricercare i pattern di interesse in una particolare forma di rappresentazione o un insieme di rappresentazioni (regole di classificazione, alberi, regressione, analisi dei gruppi). Il committente può essere determinante per la qualità del processo nella chiara definizione degli obiettivi

Interpretare i risultati ottenuti

E' possibile che a questo punto sia necessario tornare **iterativamente** a uno qualsiasi dei passi precedenti Questo passo può produrre la visualizzazione dei risultati

Utilizzare la conoscenza direttamente, incorporandola in un altro sistema per azioni ulteriori o semplicemente per documentare e presentare i risultati alle parti interessate. Il processo include il controllo e la soluzione di potenziali conflitti con precedenti convincimenti

Il KDD è un processo ITERATIVO e INTERATTIVO

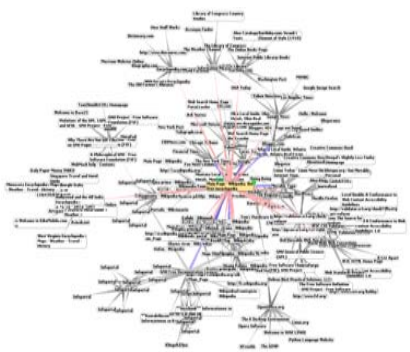
Abbondanza di dati, ma … povertà di conoscenza

Le basi di dati elettroniche sono sempre più grandi

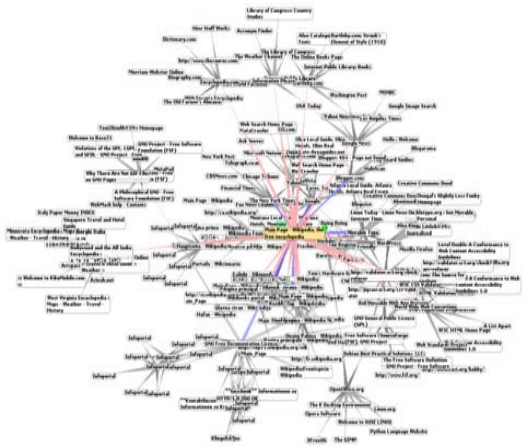
Si ragiona in termini di *terabyte* (10^{12}), ma anche, ad es. *yottabyte* (10^{24})

Siamo sommersi dai dati e sospettiamo che contengano nuova conoscenza

Questo grazie alle nuove tecnologie di **raccolta** e **memorizzazione** dei dati e soprattutto grazie al



WEB



Cos'è il Data Mining (DM): alcune definizioni in letteratura

- ✓ **DM is the process of *discovering meaningful* new correlations, *patterns*, and *trends* by sifting through *large amounts of data* stored in repositories**

- ✓ **DM is the *exploration* and *analysis*, by *automatic* and *semiautomatic* means, of *large quantitaties of data* in order to *discover meaningful patterns* and *rules***

- ✓ **DM is the *non trivial* process of identifying *valid*, *novel potentially useful* and ultimately *understandable patterns* in *data***

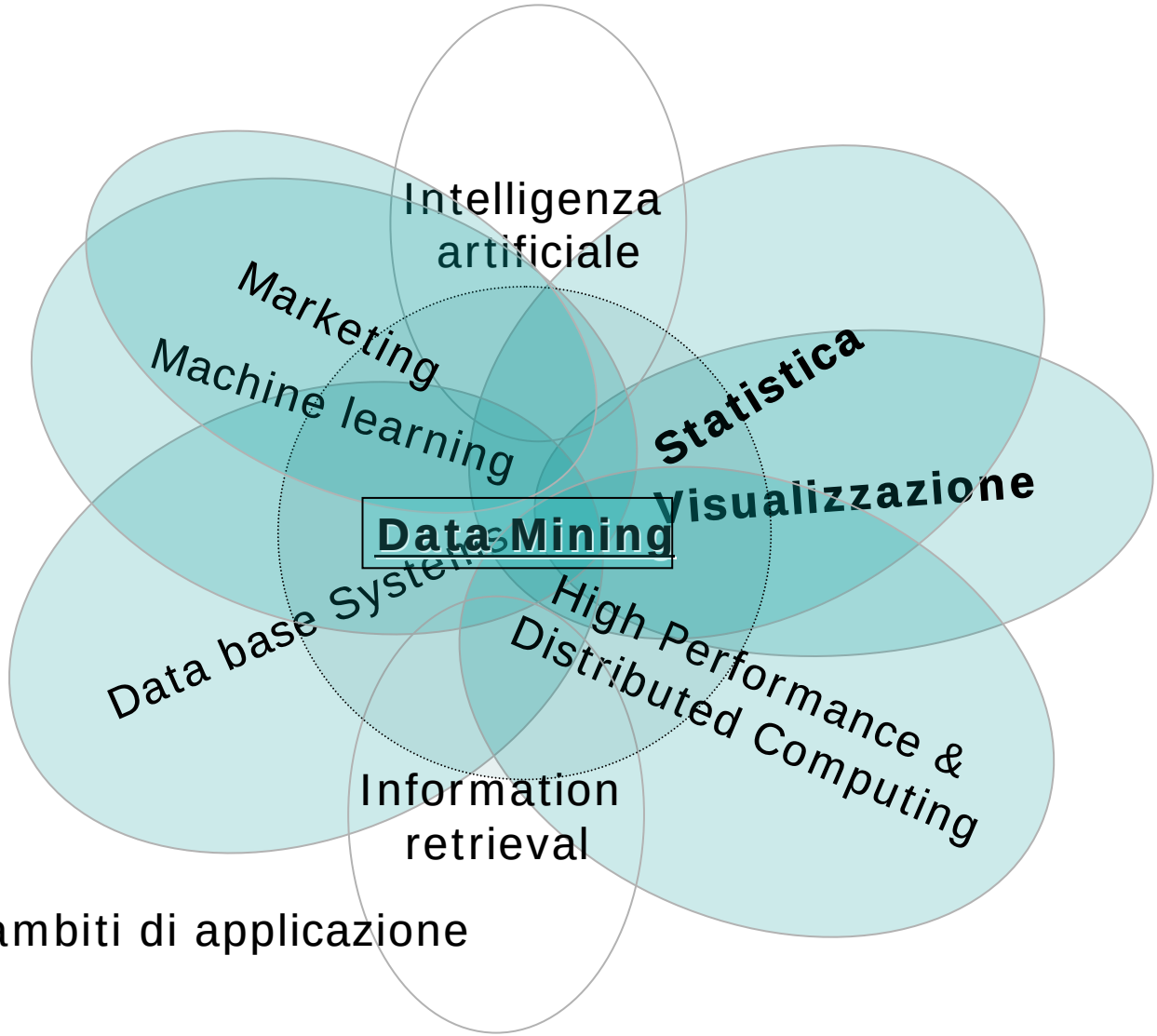
Cos'è il Data Mining

DM è la ricerca semi-automatica di strutture, associazioni, anomalie e cambiamenti, all'interno di grandi insiemi di dati

Parole chiavi:

- ***grandi*** *insiemi di dati*
- ***scoperta*** *di strutture (NON verifica)*
- ***data driven*** *(NON hypothesis driven)*
- ***machine-driven*** *(NON human-driven)*

DM ha origine in numerose discipline



Ma anche altri ambiti di applicazione

…

NOTA BENE

Ciò che distingue il DM da un'analisi statistica non è tanto (o soltanto) la quantità dei dati che vengono analizzati o la particolarità delle tecniche impiegate: è l'integrazione fra le conoscenze sulla base di dati, la metodologia di analisi e la conoscenza del business

Il DM non è il mero utilizzo di un algoritmo informatico o di una tecnica statistica ma è un processo di business intelligence, volto all'utilizzo di quanto fornito dalla tecnologia dell'informazione e dalla metodologia statistica come supporto per le decisioni aziendali

Fare DM significa tradurre le esigenze di *business* in una problematica da analizzare, nel reperimento del *database* necessario per l'analisi, nell'applicazione della tecnica statistica, implementate in un algoritmo informatico, al fine di produrre dei risultati rilevanti per prendere una decisione strategica, che a sua volta comporterà una nuova esigenza di misurazione e, quindi, in nuova opportunità di *business*, facendo partire quello che è stato chiamato il

"circolo virtuoso della conoscenza del DM"

(Berry, Linoff, 1997)

Obiettivo: trovare oggetti che soddisfino condizioni chiaramente specificate mediante una espressione regolare o di algebra relazionale, ma anche strumenti tipici di interrogazione di banche dati

Esempi di risultati di un processo di *data retrieval*:

Risposte puntuali ad una richiesta

Quanti prodotti abbiamo venduto questo mese?

Produzione di raccolte

Quali sono i nominativi dei nostri clienti con ordinativi superiori a 10000€?

Supporto alla elaborazione di studi

N.B. *Legge di Mooers: "Un sistema di reperimento delle informazioni tenderà a non essere usato quando trovare le informazioni è "more painful and troublesome" ("più noioso e doloroso") che non trovarle"*

Data Retrieval 2.

Il data retrieval è uno strumento per interrogare banche dati, mediante query (interrogazione)

Il sistema cerca, all interno della banca dati, tutti i casi che soddisfano le condizioni poste nella query vale a dire tutti i record che presentano le caratteristiche richieste, fornendo successivamente la risposta

L'individuazione di associazioni nascoste può quindi solo procedere per tentativi, mentre l'uso di strumenti di data retrieval consente di avere risposte precise a qualsiasi domanda specifica, il data mining risponde a domande più generiche

Si tratta quindi di un approccio verificativo e non esplorativo, come nel data mining

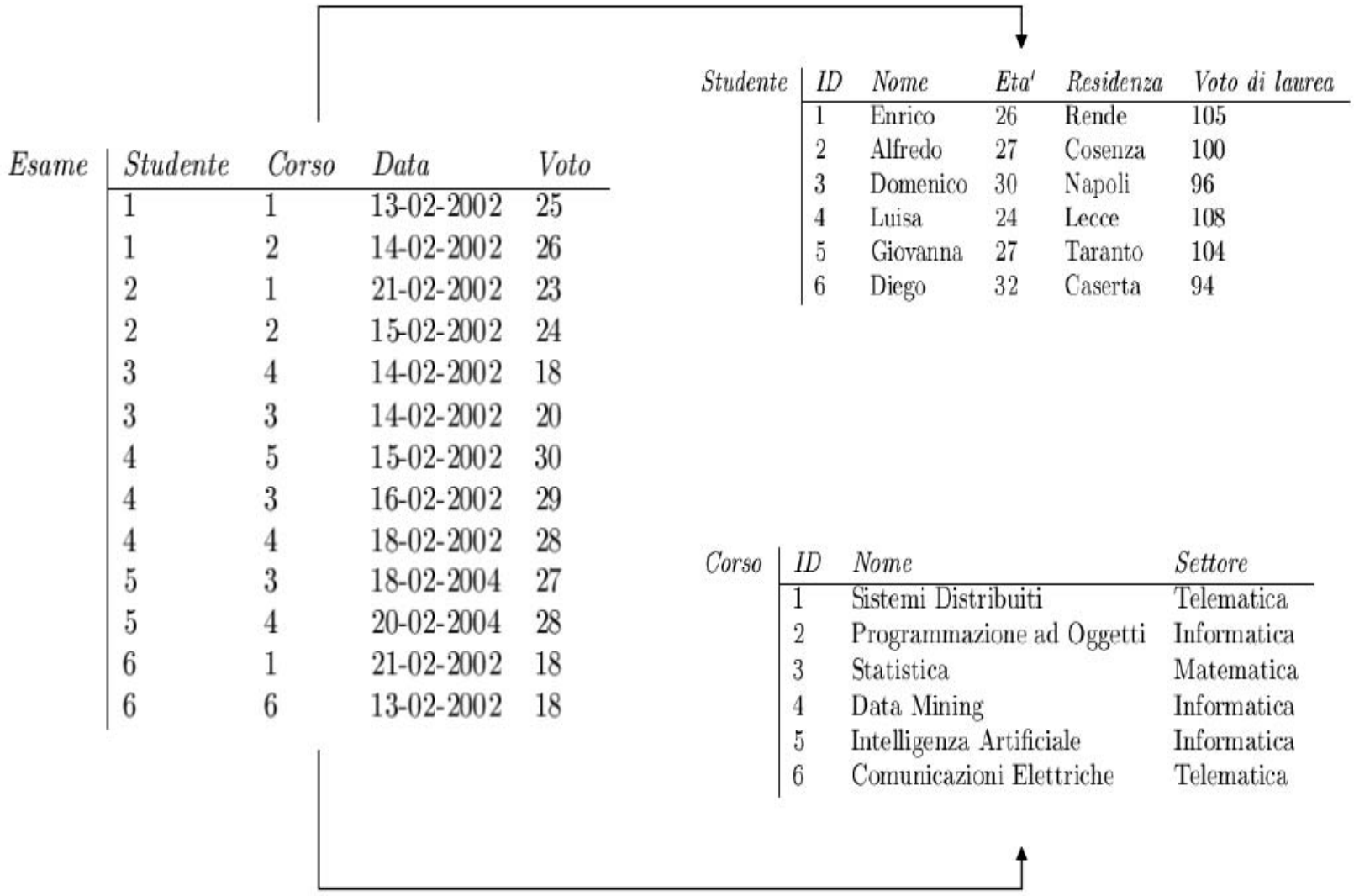
Structured Query Language - SQL

SQL è un linguaggio progettato per gestire e recuperare i dati in un sistema di gestione di basi di dati relazionali (DBMS)

Si tratta di un linguaggio di programmazione interattivo. Le interrogazioni (query) hanno la forma di comandi, e consente all'utente di specificare la descrizione dei risultati desiderati

Es. SELECT * FROM books WHERE price > 100.00 ORDER BY title;

che consente di ottenere la lista dei titoi dei libri più cari contenuti nella base di dati su cui si è fatta l'interrogazione. In dettalgio, la query identifica nella tabella *books* tutte le righe nelle quali la colonna *price* contiene un valore superiore a 100,00. Il risultato è ordinato alfabeticamente, rispetto al campo *title*. L'asterisco (*) indica che tutte le colonne della tabella *books* per le quali è valida la richiesta devono essere incluse nell'insieme rei risultati



Esame

Studente	*Corso*	*Data*	*Voto*
1	1	13-02-2002	25
1	2	14-02-2002	26
2	1	21-02-2002	23
2	2	15-02-2002	24
3	4	14-02-2002	18
3	3	14-02-2002	20
4	5	15-02-2002	30
4	3	16-02-2002	29
4	4	18-02-2002	28
5	3	18-02-2004	27
5	4	20-02-2004	28
6	1	21-02-2002	18
6	6	13-02-2002	18

Studente

ID	*Nome*	*Eta'*	*Residenza*	*Voto di laurea*
1	Enrico	26	Rende	105
2	Alfredo	27	Cosenza	100
3	Domenico	30	Napoli	96
4	Luisa	24	Lecce	108
5	Giovanna	27	Taranto	104
6	Diego	32	Caserta	94

Corso

ID	*Nome*	*Settore*
1	Sistemi Distribuiti	Telematica
2	Programmazione ad Oggetti	Informatica
3	Statistica	Matematica
4	Data Mining	Informatica
5	Intelligenza Artificiale	Informatica
6	Comunicazioni Elettriche	Telematica

Es. tratto da G.Manco **Tecniche Di Data Mining per l'Analisi dei Dati**

Qualche semplice "query"

- Che voti hanno avuto gli studenti calabresi
- Che media hanno gli studenti del corso di DM
- Chi ha avuto il voto migliore

Dalle "query" al supporto alle decisioni

- Quanti studenti che hanno ottenuto un voto di laurea superiore a 100 hanno avuto un voto alto (>27) agli esami sia di informatica che di Statistica ?
- Qual è l'andamento temporale della media dei voti in Informatica, rispetto alla media in Statistica ?

Quanti studenti che hanno ottenuto un voto di laurea superiore a 100 hanno avuto un voto alto (>27) agli esami sia di informatica che di Statistica ?

ID	Nome	Voto Statistica	Voto Informatica	Voto LAUREA
1	Enrico	27	26	105
2	Alfredo	20	24	100
3	Domenico	19	23	96
4	Luisa	30	30	108
5	Giovanna	27	27	104
6	Diego	23	21	94

- **Quali esami vengono sostenuti insieme di solito?**

Studente	*ID*	*Nome*	*Eta'*	*Residenza*	*Voto di laurea*
	1	Enrico	26	Rende	105
	2	Alfredo	27	Cosenza	100

Esame	*Studente*	*Corso*	*Data*	*Voto*
	5	3	18-02-2004	27
	5	4	20-02-2004	28
	6	1	21-02-2002	18
	6	6	13-02-2002	18

ID	*Sist. Distr.*	*Progr. a Oggetti*	*Statistica*	*Data Mining*	*Int. Artificiale*	*Com. El.*
1	Si	Si	No	No	No	No
2	Si	Si	No	No	No	No
3	No	No	Si	Si	No	No
4	No	No	Si	Si	Si	No
5	No	No	Si	Si	No	No
6	Si	No	No	No	No	Si

1	Sistemi Distribuiti	Telematica
2	Programmazione ad Oggetti	Informatica
3	Statistica	Matematica
4	Data Mining	Informatica
5	Intelligenza Artificiale	Informatica
6	Comunicazioni Elettriche	Telematica

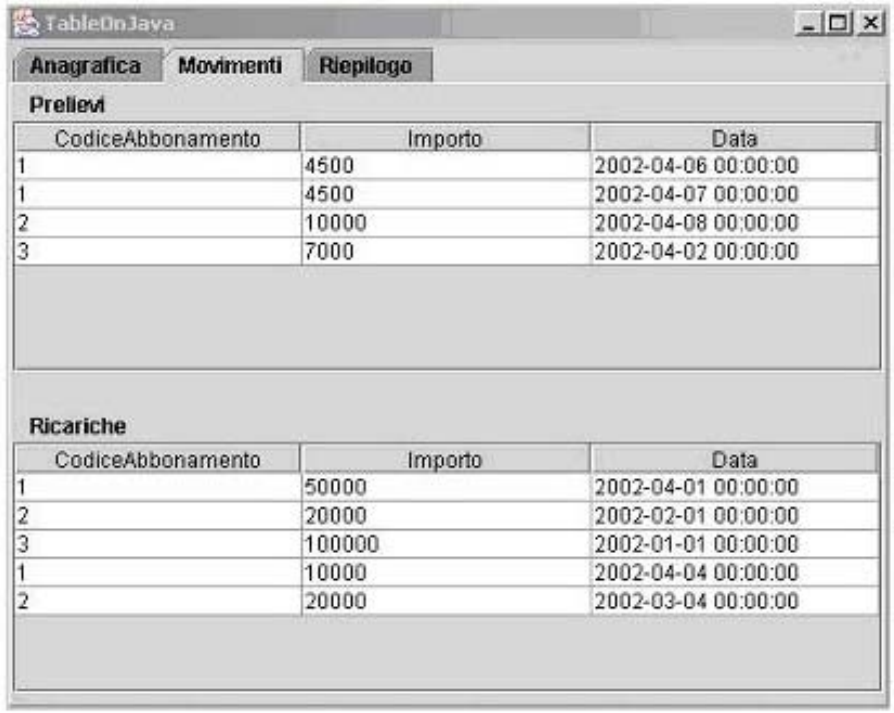
Es. tratto da G.Manco **Tecniche Di Data Mining per l'Analisi dei Dati**

Miniere

INFORMATION MINING

- DATA MINING → *dati strutturati*
- TEXT MINING → *dati non strutturati*
- WEB MINING → *dati strutturati e non strutturati*

Dati Strutturati



Dati Strutturati

Esame

Studente	*Corso*	*Data*	*Voto*
1	1	13-02-2002	25
1	2	14-02-2002	26
2	1	21-02-2002	23
2	2	15-02-2002	24
3	4	14-02-2002	18
3	3	14-02-2002	20
4	5	15-02-2002	30
4	3	16-02-2002	29
4	4	18-02-2002	28
5	3	18-02-2004	27
5	4	20-02-2004	28
6	1	21-02-2002	18
6	6	13-02-2002	18

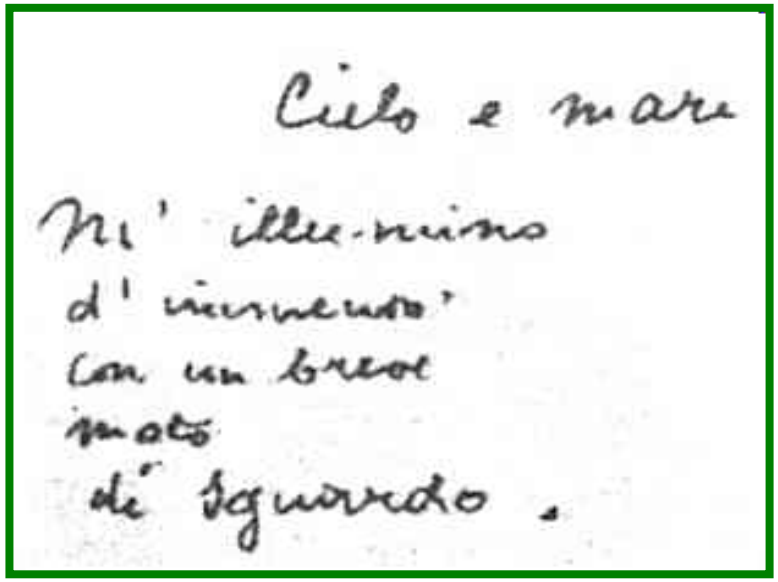
Studente

ID	*Nome*	*Eta'*	*Residenza*	*Voto di laurea*
1	Enrico	26	Rende	105
2	Alfredo	27	Cosenza	100
3	Domenico	30	Napoli	96
4	Luisa	24	Lecce	108
5	Giovanna	27	Taranto	104
6	Diego	32	Caserta	94

Corso

ID	*Nome*	*Settore*
1	Sistemi Distribuiti	Telematica
2	Programmazione ad Oggetti	Informatica
3	Statistica	Matematica
4	Data Mining	Informatica
5	Intelligenza Artificiale	Informatica
6	Comunicazioni Elettriche	Telematica

Dati NON Strutturati



Linguaggio Naturale:

Difficile da modellare e da analizzare

Text Mining

Molte le definizione in letteratura:

- ✓ **The non trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data**

  (as in Frawley, Piatetsky-Shapiro, Matheus - 1992)

- ✓ **The exploration and analysis of textual (natural-language) data by automatic and semi-automatic means to discover new knowledge**

  (as in Auvil, Searsmith - 2003)

Cosa significa informazione "prima sconosciuta" ?

➡ **informazione che l'autore esprime nel testo**

➡ **informazione che lo stesso scrivente può ignorare**

Cosa fa il Text Mining

Sotto l'etichetta Text Mining (TM), si indica il processo finalizzato alla scoperta di informazione interessante da documenti non strutturati, grazie all'utilizzo di strumenti statistici, dell'intelligenza artificiale, dell'informatica, della linguistica

- ✓ **Categprizzazione dei testi**
- ✓ **Sintesi e *abstract***
- ✓ **Identificare andamenti**
- ✓ **Trovare relazioni di associazione e di dipendenza non note**
- ✓ **Costruire strumenti di Decision Making**
- ✓ **Visualizzazione delle relzioni fra documenti, parole documenti e parole**

Divide et impera…

TM utilizza strumenti di analisi sviluppati in diversi ambienti linguistica, statistica, informatica, non necessariamente con lo stesso obiettivo

Information Retrieval

Trovare i documenti che contengono l'informazione utile relativaall'obiettivo di conoscenza specifico

Information Extraction

-Strutturazione della base dei dati
- Selezione dei dati e loro organizzazione

Information Mining

Espòorazione dei documenti e visualizzazione del contebuto attraverso tecnica statisiche

Knowledge Management

Categorizzazione e sviluppo di strumenti di supporto alle decisioni

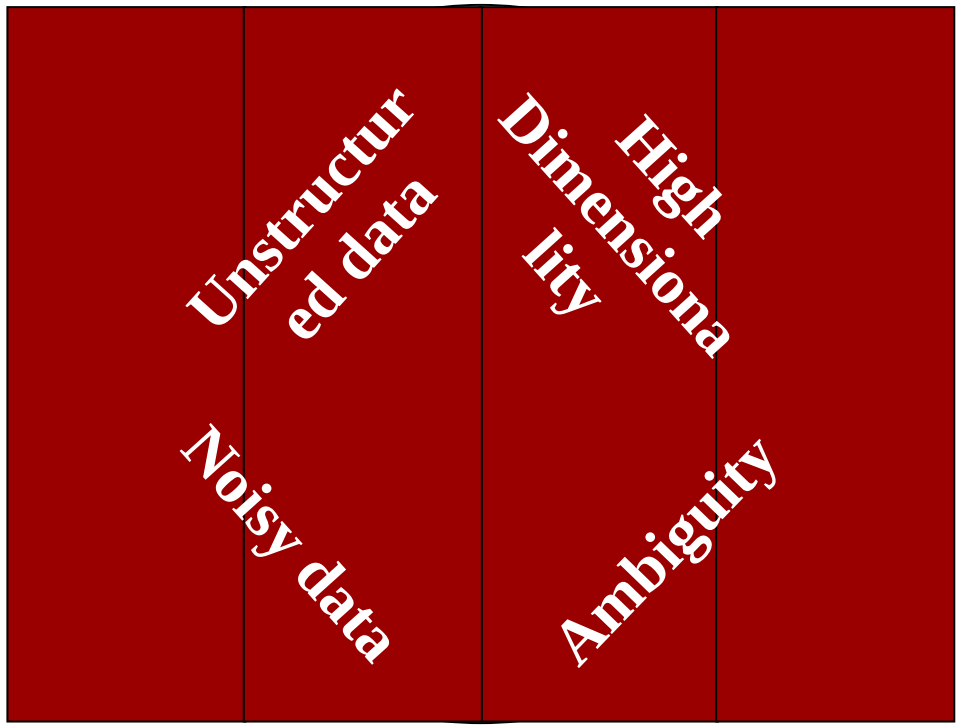
Nuovi strumenti per nuovi dati

I dati testuali sono sempre più importanti e disponibili Nel KDD

I dati testuali sono dati non convenzionali

Da un punto di vista statistico è necessario:

- ✓ **il PRETRATTAMENTO**
- ✓ **la riduzione di dimensionalità**
- ✓ **la riduzione della variabilità**
- ✓ **la disambiguazione**



Il Pretrattamento : il Natural Language Processing (NLP)

NLP ha l'obiettivo di analizzare dei testi (da un punto di vista grammaticale, sintattico, …), così da "comprenderne" il contenuto

NLP nasce nell'ambito della Intelligenza Artificiale e il suo obiettivo iniziale (negli anni Cinquanta del Novecento) era quello di predisporre de itraduttori automatici (con esiti all'inizio piuttosto discutibili)

Oggi, sono disponibili analizzatori sintattici e altri strumenti estremamente sofisticati, inclusi buoni traduttori automatici (soprattutto per la lingua inglese)

TM ha acquisito questi strumenti per perseguire I suoi obiettivi

Sul Web, è possibile trovare, anche gratuitamente, il *software* necessario per eseguire molte di queste operazioni (e.g. TreeTagger[*], o gli strumenti linguistici di Google)

[*] TreeTagger è un pacchetto per l'annotazione grammaticale di testi e per la loro lemmatizzazione.
E' stato sviluppato da Helmut Schmid all'interno del TC project presso l'Institute for Computational Linguistics dell'Università di Stoccarda. TreeTagger è disponibile per moltissime lingue, prime fra tutte, il tedesco, l'inglese, il francese, l'italiano, l'olandese, lo spagnolo, il bulgaro, il russo, il greco, il portoghese, il cinese, il francese antico e il latino. E' possibile adattarlo a qualsiasi lingua, a condizione che si disponga di un lessico e di un *corpus* di apprendimento opportunamente *"taggato"* a mano

NLP è alla base della maggior parte dei processi di TM

L'obiettivo principale di ricerca sul NLP è analizzare (grammaticalmente, identificandone le singole POS e il loro ruolo) e comprendere il linguaggio

Nasce all'interno dell'Intelligenza Artificiale, si sviluppa, negli anni '50 del Novecento, con le prime applicazioni di traduttori automatici e con scarsissimo successo.

Ad oggi, pur disponendo di analizzatori sintattici e di altri strumenti sofisticati (compresi efficienti *traduttori*), l'obiettivo è ancora lontano

Resta il fatto che molti strumenti sviluppati per l'NLP sono oggi diffusamente utilizzati nel TM

Strumenti di NLP entrati nel TM (1)

Tokenization (token=segno):

suddivisione del testo in unità (token) *elementari (parole, numeri, date, segni di punteggiatura, ecc.), di regola delimitate da spazi*

Stemming (stem=ramo):

estrazione della radice di una parola, rimovendo affissi e desinenze (es. ridendo, ridere, rideva *a* ride- *, così come* riso *e* risata*)*

Lemmatization (lemma)

*identificazione della voce del vocabolario della lingua (*lemma*) a partire da una parola con desinenza*

N.B. Stemming e *Lemmatization* differiscono, perché quest'ultimo deve anche risolvere problemi di ambiguità poiché una forma flessa può provenire da più di un lemma:

canti cantare/canto

botte botte/botta

Strumenti di NLP entrati nel TM (2)

Finding Collocation (Term Extraction):

dove per collocation *si intende un'espressione che consiste di due o più parole, che corrispondono ad un uso linguistico convenzionale:* strumenti di distruzione di massa *La loro caratteristica è che il loro significato non può essere ricavato dal significato dei singoli termini che la compongono*

Finding N-grams:

dove per n-gram *si intende una sequenza generica di n parole (bi-grammi; tri-grammi; tetra-grammi), non legati ad un particolare uso idiomatico*

Anaphora resolution:

dove anafora *è la «relazione tra un'espressione linguistica ed un'altra che la precede»:*

«E' venuto Luigi?» «Sì, lo ho incontrato al bar»; lo è un'anafora per Luigi. E' un compito molto difficile (anche per un essere umano) e anche i migliori software per il trattamento di testi non riescono il più delle volte a risolvere un'anafora

Word Sense Disambiguation:

consiste nel determinare quale significato abbia una parola "ambigua" nel contesto dell'analisi. La risoluzione del problema può avvenire o avvalendosi di dizionari *o utilizzando metodi di apprendimento basati su testi-campione, o richiedendo l'intervento esterno del ricercatore*

Parole ambigue

Esistono diverse situazioni di ambiguità:

Polisemia: un lemma cui corrispondono più significati

Farfalla: è un insetto, ma anche un elemento del motore che prende il nome dall'insetto

Omografia: la stessa sequenza di caratteri è comune a due lemmi:

Fine sostantivo maschile = obiettivo

Fine sostantivo femminile = termine

Fine aggettivo = elegante; sottile

N.B. Quando i problemi non sono di natura semantica, ma sintattica ci si trova di fronte a problemi di *part-of-speech tagging* (es. canto)

Una strategia di TM

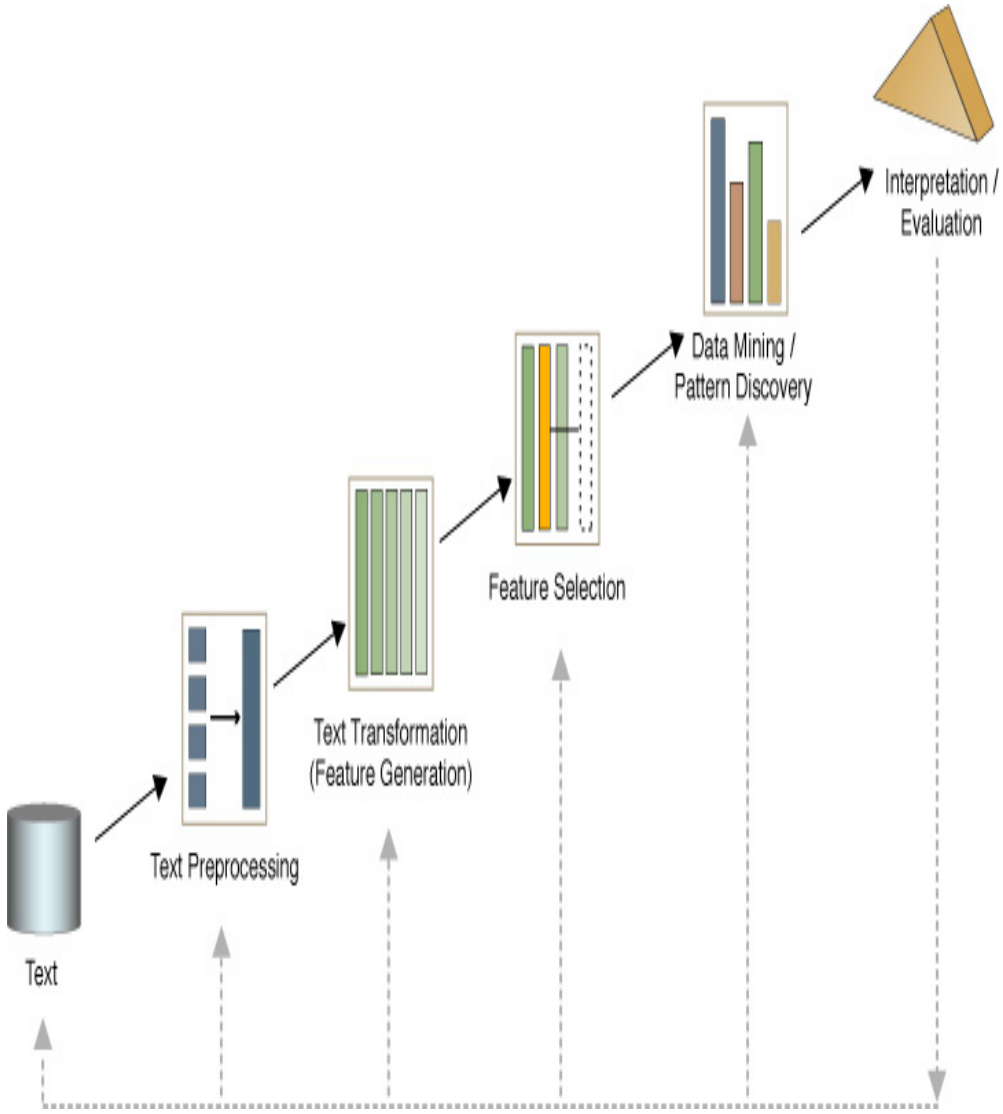
Pretrattamento
Parsing /Normalizzazione/Pulizia

Features Generation
Tokenisation

Features Selection
Lessicalizzazione/Lemmatizzazione
Codifica

Text Mining
Analisi Esplorative

Knowledge Management
Categorizzazione del testo
Strumenti di supporto alle decisioni



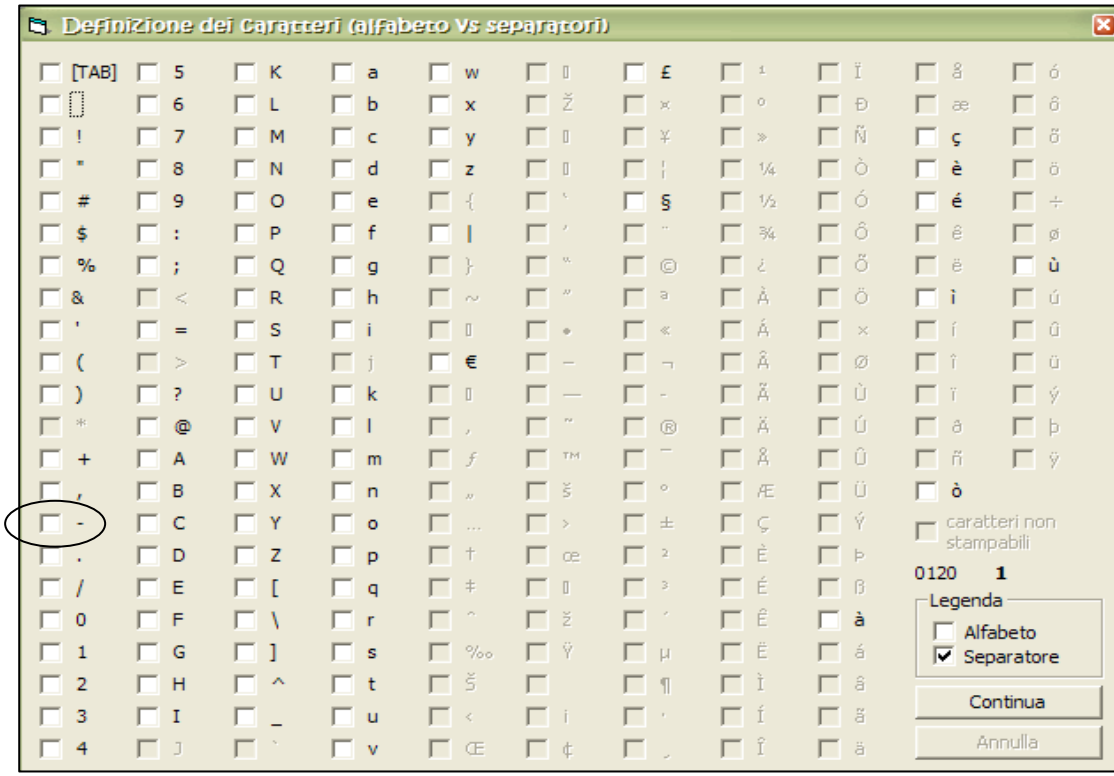
Parsing

Il testo è considerato come una sequenza di simboli appartenenti ad un codice
All'interno del codice, è necessario identificare l'insieme di caratteri che definiscono l'**alfabeto** e l'insieme complementare dei **separatori** fra parole

Questa procedura è detta **parsing**:
I documenti sono scansionati e rappresetati da una lista di forme

Generalmente i separatori sono simboli (.,:;!?), parentesi, «,»,“, -, _. E gli altri caratteri speciali (es. #£$&@§), ma possono essere differenti in casi particolari e a seconda della lingua

Taltac →



Trattino : unisce o separa ?

Normalizzazione - Pulizia

Normalizzazione

- Maiuscole/minuscole
- formati standard per date, orari, numeri, …
- accenti
- translitterazioni (se ci sono parole che provengono da alfabeti diversi)

Pulizia

- Identificazione e rimozione di *stopper words* … parole con scarso significato semantico, come *il, lo, la, i, gli, le, e, di, da, che, …*

  Le *stopper word* possono anche dipendere dal contesto, es. *teorema* e *dimostrazione* in un documento di matematica, *paziente* in articoli di medicina, ...

Tokenization

Nell'ambito del Text Mining ogni voce di un vocabolario è chiamata TOKEN
E' così possibile costruire un vocabulario attraverso la *TOKENIZATION*

Dal punto di vista del Knowledge Discovery, si vuole esplorare la raccolta di documenti in modo da far emergere l'informazione più significativa: quale è il livelli di testo da considerare? le singole parole? parole multiple? frasi?

Statistica e Informatica possono fornire validi strumenti, ma vanno considerati ke regole del linguaggio naturale, le strutture grammatiche e lessicali, I domini semantici

Unità minimali di senso

Una **parola** può convenzionalmente essere definita come una **sequenza di caratteri** appartenenti ad un determinato alfabeto e delimitate da caratteri speciali, chiamati **delimitatori**

- ✓ Lessicalizzazione
- ✓ Lemmatizzazione
- ✓ Part of Speech tagging
- ✓ Semantic tagging (Word Sense Disambiguation)

Le unità minimali di senso più frequentemente utilizzate sono:

- o forme grafiche
- o lemmi (raramente radici)
- o forme testuali
- o (quasi)- segmenti ripetuti

vedremo …

****ART1
Certi segnali non ingannano. Alla ricomparsa della mendicità, della disoccupazione di massa, delle "classi pericolose" nelle periferie delle metropoli, delle mense e ostelli per i poveri è venuta ad aggiungersi in questa fine secolo, a ulteriore riprova delle conseguenze disumane della mondializzazione economica, la figura sociale del bambino-lavoratore. Già nel XIX secolo, all'epoca del liberismo trionfante e dello sviluppo industriale, l'aggravamento delle disuguaglianze e l'"inferno" operaio si manifestavano in particolare nello sfruttamento generalizzato, senza limiti d'età, del lavoro infantile. Nel suo celebre rapporto del 1840 (1), Louis Villermé descrisse le condizioni fisiche e morali dei bambini operai in Francia, con un orario di lavoro che era allora di 14 ore al giorno: "questa moltitudine di bambini, alcuni di appena sette anni, magri, sparuti, coperti di cenci, che si recano nelle fabbriche a piedi nudi, sotto la pioggia e tra il fango: pallidi e snervati, offrono uno spettacolo di miseria, sofferenza e abbattimento". La descrizione di questa realtà denunciata anche da romanzieri quali Charles Dickens, Victor Hugo, Hector Malot, Jules Vallés, Emile Zola e Edmondo De Amicis era però ben lungi dall'impressionare certi liberisti, che la consideravano come un "male necessario". E ci fu chi scrisse: "Questa miseria offre un salutare spettacolo a tutta la parte ancora sana delle classi meno fortunate, poiché è fatta per incuterle spavento ed esortarla così alle difficili virtù delle quali ha bisogno per arrivare a una condizione migliore (2)". Davanti a un tale cinismo, come non comprendere, ad esempio, la rivolta di Carlo Marx, che nel 1848, nel suo Manifesto del partito comunista, denuncerà la grande industria, "che spezza nella classe proletaria ogni legame di famiglia, e trasforma i fanciulli in semplici articoli di commercio e strumenti di lavoro"; e reclamerà "l'abolizione del lavoro infantile nelle fabbriche nella sua forma attuale (3)"? La storia ha dimostrato che la progressiva abolizione del lavoro infantile e l'istruzione obbligatoria sono state, in Europa come nell'America del Nord, le condizioni indispensabili per lo sviluppo economico e sociale. Ma si è dovuto attendere il 1990 per l'entrata in vigore della Convenzione dei Diritti dell'Infanzia, ratificata praticamente da tutti gli stati del mondo nell'ambito dell'Onu, ad eccezione degli Stati Uniti, che stabilisce un'età minima per l'ingresso nel mondo del lavoro, auspicata dall'Oil (Organizzazione internazionale del lavoro) fin dal 1973. Malgrado ciò, si valuta che i bambini lavoratori sono circa 250 milioni, a volte addirittura di età inferiore ai cinque anni. Questi dati riguardano in maggioranza i paesi poveri del Sud, ma molti bambini sono sfruttati anche in quelli del Nord. Nell'insieme dell'Unione europea ad esempio, il numero dei lavoratori non ancora quindicenni supererebbe attualmente i 2 milioni. In particolare nelle zone più violentemente colpite dalle ristrutturazioni ultraliberiste, come il Regno Unito. Ma persino nei paesi considerati "socialmente avanzati" come la Danimarca o l'Olanda, il fenomeno dei bambini che lavorano ha fatto la sua ricomparsa. "Anche in Francia, afferma un esperto del Fondo delle Nazioni unite per l'Infanzia (Unicef), varie decine di migliaia di bambini sarebbero di fatto lavoratori salariati, sotto la copertura dell'apprendistato; e il 59% degli apprendisti lavorerebbero più di 40 ore la settimana, a volte anche fino a 60 (4)". Su scala planetaria, il numero dei bambini lavoratori non cessa di aumentare. In alcuni paesi è un flagello di massa. I bambini sfruttati al di sotto dei sei anni d'età sono decine di milioni (5). In America latina lavora un bambino su 5, un Africa uno su tre; in Asia, uno su due! In queste regioni, il settore che più utilizza il lavoro dei bambini è quello agricolo. Spesso vi si pratica la servitù per indebitamento.

Società	197
Medioriente	157
Cultura	143
Conflitti	108
Globalizzazione	95
Balcani - guerre e processi di pace	60
Finanza mondiale	50
Islam e Musulmani	49
UE	49
ex URSS	48
Organismi Internazionali	47
Povertà ed esclusione	42
Terrorismo	40
Dittature	39
Storia	39
Media e giornalismo	37
Comunicazione e Internet	35
Ambiente e sviluppo sostenibile	33
Donne	27
Destra ed estrema destra	26
Immigrazione	25
Europa dell'Est	24
Nucleare	24
Lavoro e disoccupazione	23
Cinema	18
Imperialismo	18
Minoranze	18
Sanità	16

Vocabolario ordinato per frequenza

una	439	03	J	DET+NUM+
punito	433	06	J	A+V
se	428	02	J	CONG+N+P
che	423	03	J	A+CONG+N
dell	370	04	PREP	PREP
numero	369	06	N	N
ovvero	347	06	CONG	CONG
Chiunque	343	08	PRON	PRON
le	337	02	J	DET+PRON
fatto	320	05	J	A+N+V
delle	320	05	PREP	PREP
reato	318	05	N	N
La	317	02	J	DET+N+PR
alla	312	04	J	N+PREP
dalla	308	05	J	PREP+V

N	sostantivo	PREP	preposizione	FORM	forma idiom.
A	aggettivo	CONG	congiunzione	NM	nome proprio
V	verbo	PRON	pronome	DAT	data
AVV	avverbio	ESC	interiezione	NUM	numerale
DET	determinante	J	ambigua	O	stranierismo

Parole … parole … parole

La scelta dell'unità elemetnste è molto delicata: la variabilità e la complessità del fenomeno dipendono anche dalla linga (es. un verso regolare italiano / un verbo irregolare inglese)

parlare

Verbo regolare: Più di 80 forme differenti

to speak

- speak
- speaks
- spoke
- spoken
- speaking

verbo irregolare: 5 forme differenti

Modo	Indicativo			
Tempo	Presente	Passato prossimo	Imperfetto	Trapassato prossimo
io	parlo	ho parlato	parlavo	avevo parlato
tu	parli	hai parlato	parlavi	avevi parlato
egli	parla	ha parlato	parlava	aveva parlato
noi	parliamo	abbiamo parlato	parlavamo	avevamo parlato
voi	parlate	avete parlato	parlavate	avevate parlato
essi	parlano	hanno parlato	parlavano	avevano parlato

Tempo	Passato remoto	Trapassato remoto	Futuro semplice	Futuro anteriore
io	parlai	ebbi parlato	parlerò	avrò parlato
tu	parlasti	avesti parlato	parlerai	avrai parlato
egli	parlò	ebbe parlato	parlerà	avrà parlato
noi	parlammo	avemmo parlato	parleremo	avremo parlato
voi	parlaste	aveste parlato	parlerete	avrete parlato
essi	parlarono	ebbero parlato	parleranno	avranno parlato

Modo	Congiuntivo			
Tempo	Presente	Passato	Imperfetto	Trapassato
io	parli	abbia parlato	parlassi	avessi parlato
tu	parli	abbia parlato	parlassi	avessi parlato
egli	parli	abbia parlato	parlasse	avesse parlato
noi	parliamo	abbiamo parlato	parlassimo	avessimo parlato
voi	parliate	abbiate parlato	parlaste	aveste parlato
essi	parlino	abbiano parlato	parlassero	avessero parlato

Modo	Condizionale		Imperativo
Tempo	Presente	Passato	
io	parlerei	avrei parlato	
tu	parleresti	avresti parlato	parla
egli	parlerebbe	avrebbe parlato	parli
noi	parleremmo	avremmo parlato	parliamo
voi	parlereste	avreste parlato	parlate
essi	parlerebbero	avrebbero parlato	parlino

Modo	Participio		Gerundio	
Tempo	Presente	Passato	Presente	Passato
	parlante	parlato	parlando	avendo parlato

Modo	Infinito	
Tempo	Presente	Passato
	parlare	avere parlato

Garzanti Linguistica