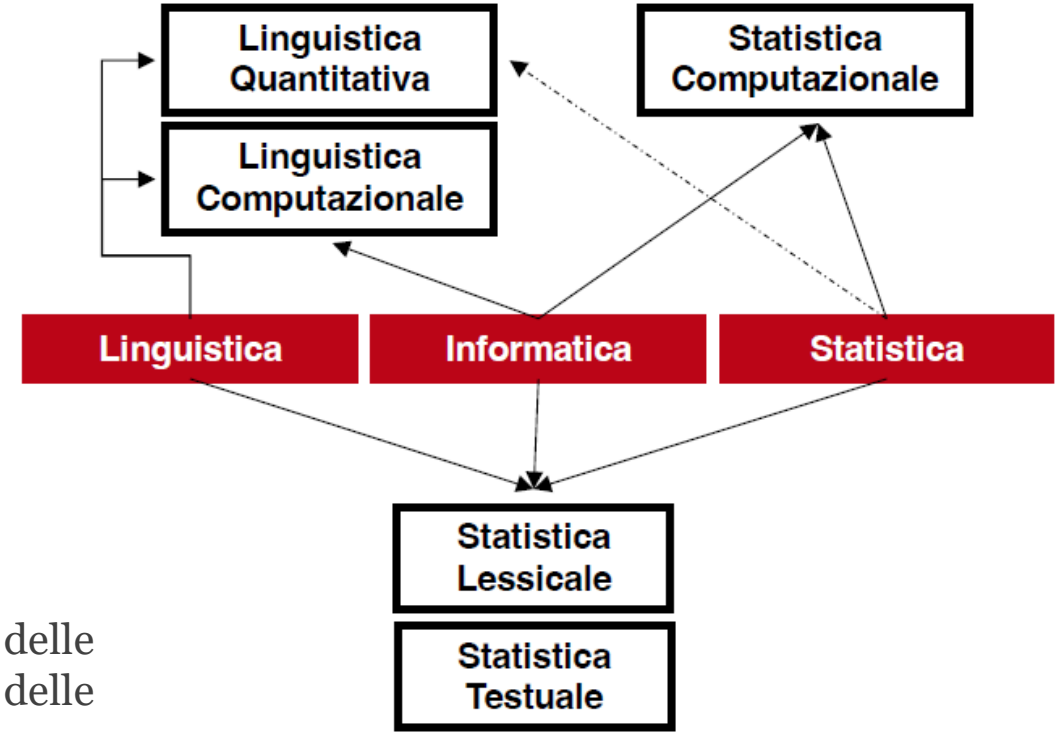
ANALISI TESTUALE

L'ANALISI TESTUALE

L'Analisi testuale è un campo di ricerca consolidato, che si occupa di analizzare in maniera automatica informazioni testuali.

È un settore di ricerca in continua evoluzione, risultato dell'intreccio della ricerca congiunta di tre ambiti disciplinari:

la linguistica, la statistica e l'informatica.



Un'analisi testuale prevede la lettura automatica dei testi, la rilevazione delle caratteristiche più rilevanti in maniera automatica e l'elaborazione delle stesse…
anche se il lavoro del ricercatore è fondamentale in ogni fase della procedura

LINGUAGGIO E LINGUA

Il **linguaggio naturale** è la facoltà - esclusiva del genere umano - di esprimere sentimenti e sensazioni, riflessioni, giudizi, di narrare fatti o descrivere aspetti della realtà in cui si vive

La **lingua** è una particolare forma di linguaggio usata da un gruppo di persone ai fini della comunicazione: è il modo concreto e storicamente determinato in cui si manifesta la facoltà del linguaggio negli esseri umani

LINGUAGGIO E LINGUA

Il **linguaggio naturale** è la facoltà - esclusiva del genere umano - di esprimere sentimenti e sensazioni, riflessioni, giudizi, di narrare fatti o descrivere aspetti della realtà in cui si vive

La **lingua** è una particolare forma di linguaggio usata da un gruppo di persone ai fini della comunicazione: è il modo concreto e storicamente determinato in cui si manifesta la facoltà del linguaggio negli esseri umani

La lingua può essere vista come un sistema di segni convenzionali, cioè un codice: la comunicazione tra individui avviene attraverso processi di codifica e decodifica

Una lingua (nella sua forma scritta) consiste di un alfabeto, una grammatica e una sintassi

(1) **ALFABETO**: è l'insieme dei simboli (grafi) utilizzati dal linguaggio

(2) **GRAMMATICA**: è l'insieme di regole che definiscono come i simboli dell'alfabeto interagiscono per formare le parole

(3) **SINTASSI**: è l'insieme di regole che definiscono come le parole sono unite per costruire le proposizioni e le frasi

La scienza che si occupa dello studio del linguaggio e delle lingue è la **linguistica**

LINGUAGGIO E LINGUA

Internet ha profondamente modificato il nostro rapporto con le fonti di informazione, rispetto ai media tradizionali…

Due sono gli elementi fondamentali:

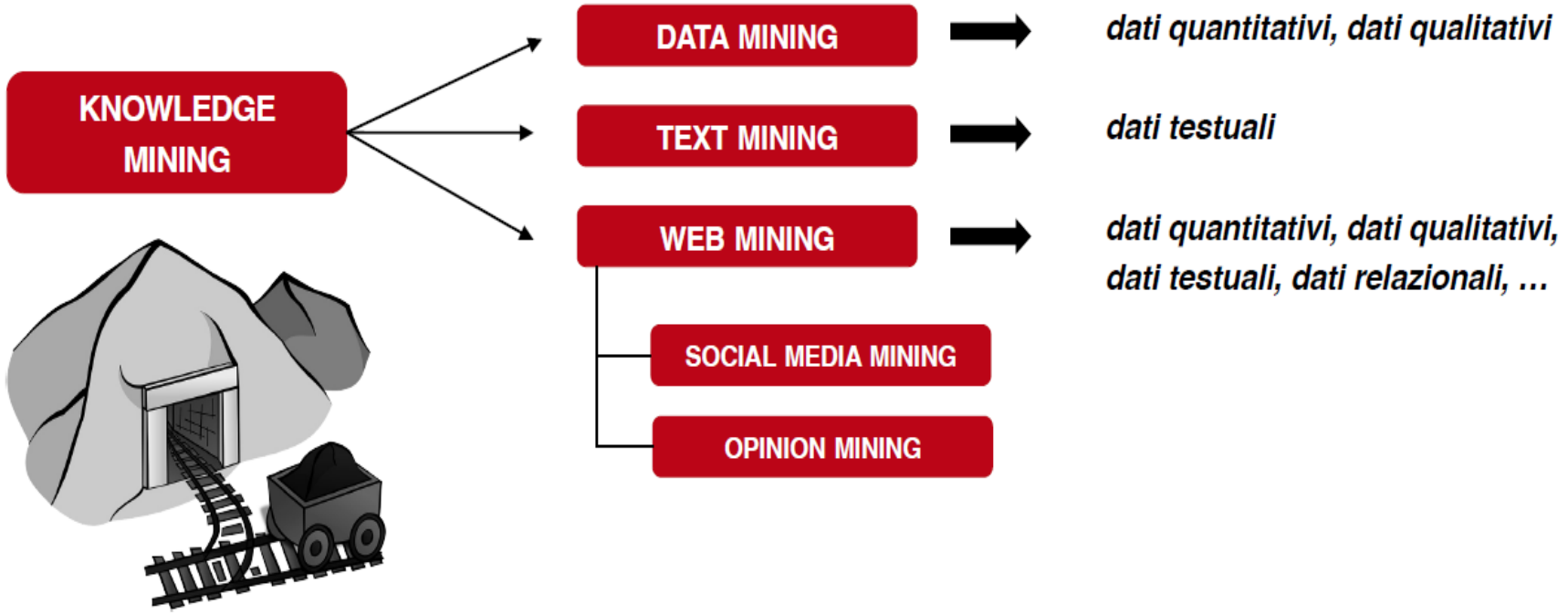
- vi è una maggiore **diffusione** dell'informazione (da un punto di vista socio-demografico, culturale, geografico)
- vi è una maggiore **diversificazione** del contenuto informativo, in relazione ai diversi bisogni conoscitivi degli utenti

La crescente mole di dati disponibili immediatamente su **supporto digitale**, spesso in forma documentaria, rende allo stesso tempo necessario e possibile il ricorso a strategie sempre più complesse per **l'estrazione**, **l'analisi** e **l'organizzazione** della conoscenza, finalizzate alla soddisfazione di uno specifico bisogno conoscitivo

ESTRAZIONE E GESTIONE DELLA CONOSCENZA

La ricerca di conoscenza in banche dati di notevoli dimensioni è messa in atto con strategie d'analisi spesso nate in ambito informatico, ma il problema è affrontato sempre da un punto di vista statistico

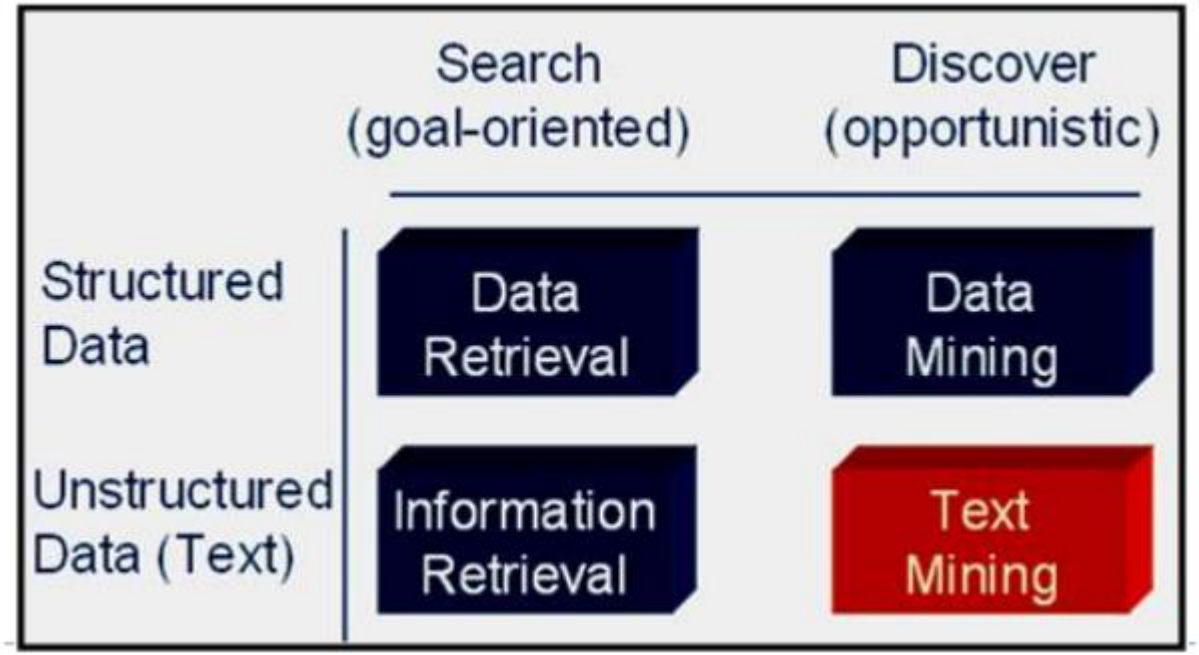
Negli ultimi anni la forbice tra l'analisi di dati strutturati e non strutturati si è ampliata a tal punto che il **Text Mining** è ritenuto un ambito di ricerca nettamente distinguibili dal **Data Mining**



DATA MINING E TEX MINING

Data Mining: estrazione di informazione da dati strutturati

Text Mining: estrazione di informazione da databases testuali non strutturati



TEXT MINING E TEXT ANALYTICS

Text Mining: esplorazione e "scavo" in un giacimento di materiali testuali (corpus) per recupero ed estrazione di informazioni

Text Analytics: applicazione di algoritmi di analisi ai testi strutturati prodotti dal processo di Text Mining

- Steps of Text Mining:
  - Information Retrieval
  - Create text corpus
  - Data Preparation and Cleaning
  - Segmentation
  - Tokenization
  - Stop-word, numbers and punctuation removal
  - Stemming
  - Convert to lowercase
  - POS tagging
  - Term-Document matrix

- Steps of Text Analytics:
  - Modeling (e.g., inferential models, predictive models or prescriptive models)
  - Training and evaluation of models
  - Application of these Models
  - Visualizing the Models

Once Term Document matrix is prepared

LE FONTI DEI DATI TESTUALI

Nell'era dell'ICT, i testi sono comunemente già disponibili in formato elettronico. A seconda del tipo di fenomeno che si vuole studiare possiamo considerare:

FONTI PRIMARIE

Testi da indagini:
- interviste libere
- interviste semi-strutturate
- interviste strutturate
- focus group

FONTI SECONDARIE

Testi da altri contesti:
- testi letterari
- articoli scientifici
- articoli giornalistici
- email
- pagine web
- social media
- altri documenti…

LE FONTI DEI DATI TESTUALI

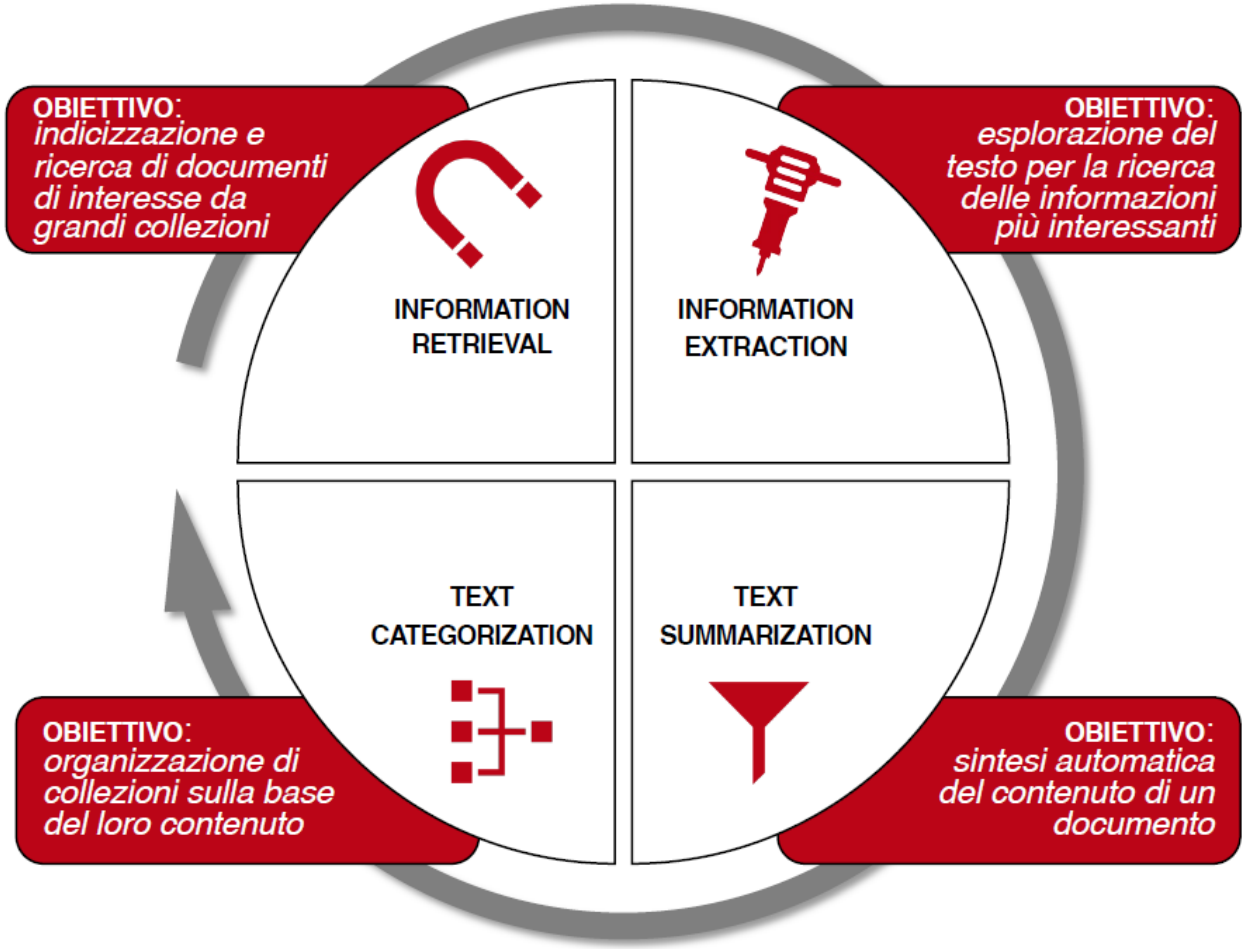
Il Text Mining riguarda la ricerca di pattern in un insieme di documenti scritti in linguaggio naturale

Può essere definito come il processo di analisi di testi per estrarre informazioni per scopi particolari.

✓*Estrazione di informazioni*

✓*Recupero delle informazioni*

Il Text Mining comprende diverse tecniche utili a rispondere a differenti esigenze di ricerca e spesso queste tecniche sono combinate tra loro per poter costruire strategie analitiche complesse e integrate



AMBITI APPLICATIVI DEL TEXT MINING

Gli ambiti di applicazione delle tecniche di Text Mining sono molteplici e in continua evoluzione

MEDICINA
- ✓ *Analisi anamnesi paziente*
- ✓ *Analisi automatica referti*
- ✓ *Prevenzione epidemie*

LETTERATURA
- ✓ *Analisi testi*
- ✓ *Ricerche stilometriche*
- ✓ *Attribuzione autore*

AMMINISTRAZIONE & FINANZA
- ✓ *Analisi note bilanci*
- ✓ *Analisi/previsione mercati*

MARKETING
- ✓ *Customer interest*
- ✓ *Customer care*
- ✓ *Brand image*

ISTRUZIONE
- ✓ *Scoring automatico*
- ✓ *Plagiarism detection*
- ✓ *Analisi commenti studenti*

GIUSTIZIA & SICUREZZA
- ✓ *Analisi reti criminali*
- ✓ *Indagini da documenti*
- ✓ *Indicizzazione sentenze*

SCIENZE POLITICHE
- ✓ *Analisi linguaggio politico*
- ✓ *Valutazione intenzioni voto*

TURISMO
- ✓ *Analisi bisogni turisti*
- ✓ *Recommender system*

INFORMAZIONE
- ✓ *Textual data journalism*
- ✓ *Indicizzazione articoli*

SOCIAL MEDIA
- ✓ *Determinazione topic*
- ✓ *Individuazione hater*
- ✓ *Valutazione fake news*

GESTIONE POSTA
- ✓ *Ricerca spam*
- ✓ *Indirizzamento automatico*

...

IL TESTO

Qualsiasi fonte di natura scritta che contiene un insieme di concetti che siano leggibili (e comprensibili) da parte di un individuo può essere considerata un **testo**

Possiamo considerare testi brevi (formati da una o più frasi) e testi lunghi (formati da più paragrafi)

Nel Text Mining spesso un testo si definisce genericamente **DOCUMENTO**

IL TESTO

Qualsiasi fonte di natura scritta che contiene un insieme di concetti che siano leggibili (e comprensibili) da parte di un individuo può essere considerata un **testo**

Possiamo considerare testi brevi (formati da una o più frasi) e testi lunghi (formati da più paragrafi)

Nel Text Mining spesso un testo si definisce genericamente **DOCUMENTO**

TESTO GREZZO

Sai, la gente è strana
Prima si odia e poi si ama
Cambia idea improvvisamente
Prima la verità poi mentirà lui
Senza serietà, come fosse niente
Sai, la gente è matta
Forse è troppo insoddisfatta
Lei segue il mondo ciecamente
E quando la moda cambia
Lei pure cambia
Continuamente, scioccamente
Tu, tu che sei diverso
Almeno tu nell'universo
Un punto sei, che non ruota mai intorno a me
Un sole che splende per me soltanto
Come un diamante in mezzo al cuore
…

META-DATI

Anno: 1989 **Genere:** pop
Interprete: donna

IL TESTO

La parola è intesa come componente di base di un qualsiasi testo scritto

Nel Text Mining spesso una parola si definisce genericamente **TERMINE**

In un testo in formato elettronico si utilizzano set di caratteri alfanumerici e non alfanumerici

Ogni parola nel testo è rappresentata come una **stringa**, cioè una sequenza di caratteri (solitamente alfabetici) delimitata da caratteri separatori: lo spazio, i segni di interpunzione, altri specifici (es. @#)

IL CORPUS

Un insieme di testi confrontabili tra di loro e appartenenti ad uno stesso contesto si chiama **CORPUS**

Il corpus può essere costituito da

1. un **unico testo**
2. alcuni testi (sub-testi o **parti**, da due fino a qualche decina)
3. centinaia o migliaia di micro-testi (**frammenti**, risposte, messaggi, titoli)

Un corpus **rappresentativo** è un insieme di testi raccolti e selezionati in modo tale da riflettere accuratamente le caratteristiche linguistiche di una determinata lingua, dialetto, o varietà linguistica specifica in un certo contesto o periodo di tempo.

L'obiettivo principale di un corpus rappresentativo è quello di fornire una base di dati affidabile per l'analisi e la ricerca linguistica.

In generale, per un'analisi robusta e affidabile, un corpus dovrebbe idealmente avere almeno qualche milione di parole.

Tuttavia, per studi più specifici e limitati, **corpora** più piccoli possono ancora essere utili, a patto che siano **ben costruiti** e **rappresentativi** del fenomeno linguistico in esame.

TERMINOLOGIA

Forma grafica (parola o type): è una sequenza di caratteri (bytes) di un alfabeto predefinito

Occorrenza (Token): ogni parola che appare o ricorre in un testo.

Frequenza di una parola in un testo è data dal numero delle sue occorrenze

Segmento è una sequenza di parole adiacenti

Frammento di testo: insieme di parole definenti un contesto d'interesse

Testo: raccolta di frammenti

Documento: termine con il quale ci si riferisce genericamente all'unità di testo disponibile per l'analisi

Corpus: insieme di testi confrontabili tra di loro e appartenenti ad uno stesso contesto

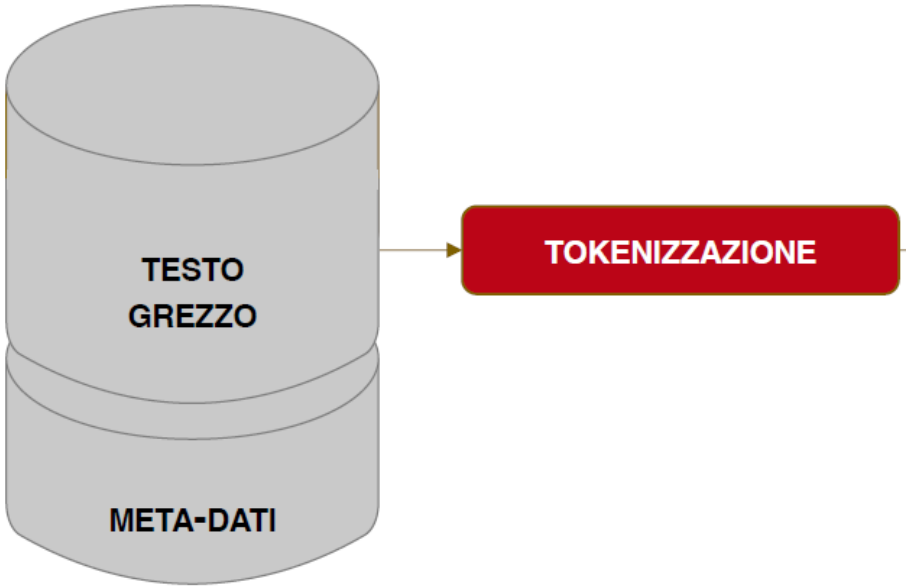
Vocabolario (V): insieme di parole diverse del Corpus. Può essere espresso o in forme grafiche (le parole così come compaiono nel corpus) o per lemmi (cioè le forme presenti nei dizionari)

Dimensione (lunghezza del corpus – N): il numero totale di occorrenze del corpus

ACQUISIZIONE DEL TESTO

Attraverso il **parsing**, il testo è acquisito e convertito in formato leggibile dalla macchina è una vera e propria scansione del testo che associa ad ogni stringa un codice numeri

In questa fase, ogni stringa-parola viene identificata come **token**: per questo motivo la fase di parsing del testo è anche detta di **tokenizzazione**



Il Tokenizer deve conoscere le caratteristiche della lingua dei testi da trattare:

- Rimozione punteggiatura, tranne alcuni casi (virgola nei numeri, barra nelle date)
- Rimozione spazi bianchi multipli

Rendere le parole uniformi dal punto di vista del formato del carattere

- Convertire le maiuscole in minuscole o viceversa

ACQUISIZIONE DEL TESTO

Il testo è decomposto nelle sue componenti costitutive in accordo ad una particolare codifica nota come **Bag-of-Words** (borsa di parole)

Il testo è visto come un multi-set delle sue parole, cioè un insieme che consente la ripetizione delle entità che contiene (molteplicità). Caratteristiche peculiari di tale codifica sono che non si tiene conto dell'ordine delle parole e non si tiene conto del loro ruolo grammaticale e sintattico nel testo

Fratelli d' Italia
1 2 3
L'Italia s'è desta,
4 3 5 6 7
Dell'elmo di Scipio
8 9 10 11
S'è cinta la testa.
12 6 13 14 15
…

230 token
100 type

Type	Occ.
Italia	10
la	9
è	7
a	6
L	6
alla	5
chiamò	5
stringiamci	5
coorte	5
…	…

Il vocabolario può essere visto come una
distribuzione di frequenza

COSTRUZIONE DEL VOCABOLARIO

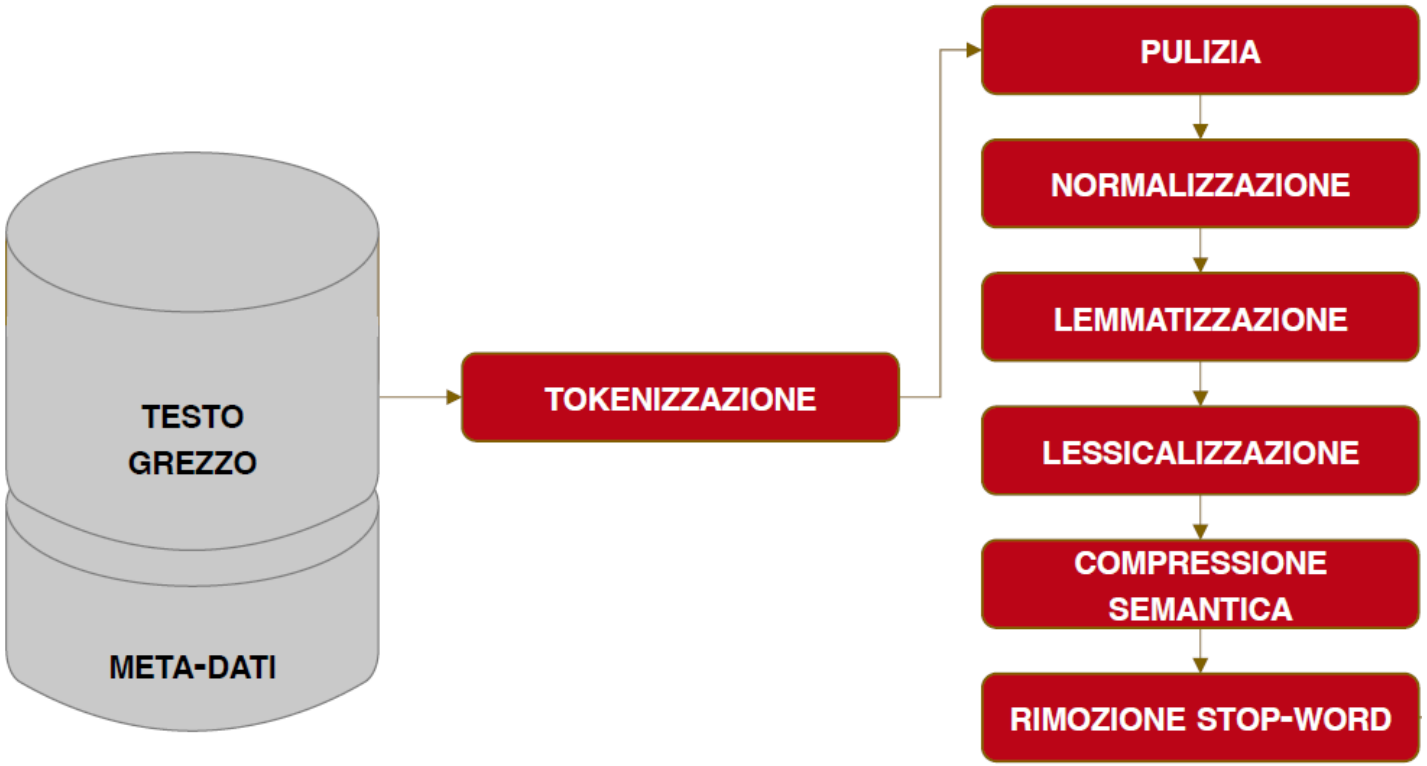
A seguito del parsing, i token identici (stringhe cui corrisponde un uguale codice numerico) sono impilati e conservati in un particolare elenco, chiamato **vocabolario**

Il vocabolario contiene la lista di stringhe distinte riconosciute nel testo, note come **type**, insieme al corrispondente numero totale di occorrenze, cioè quante volte il token compare nel testo

PRE-TRATTAMENTO DEI TESTI

Per ottenere dei **dati testuali** (strutturati) è necessario intervenire sulla collezione di testi oggetto di studio attraverso una sequenza di procedure articolate

Il cosiddetto **pre-trattamento** serve per ottenere una base di dati idonea ad essere analizzata da un punto di vista statistico



PULIZIA E NORMALIZZAZIONE DEL TESTO

Attraverso le **fasi di pulizia** e **normalizzazione** si agisce sul testo per eliminare possibile «rumore» e migliorare la qualità dei dati per le successive strategie di analisi

I caratteri e le stringhe non alfabetici sono solitamente eliminati dal testo (separatori, simboli speciali, numeri, emoticon, ...), poiché non sono portatori di informazioni interessanti (salvo casi particolari)

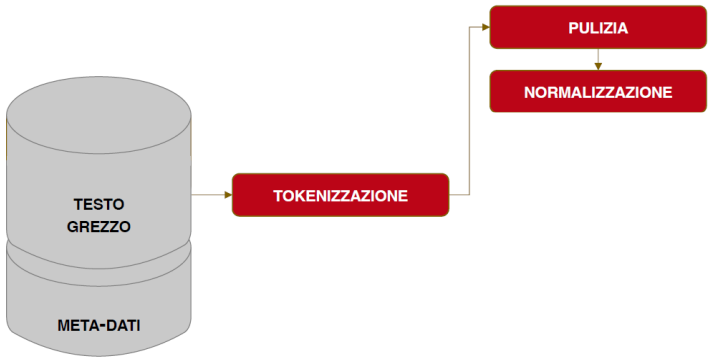
La normalizzazione serve per standardizzare il testo, in modo da evitare possibili sdoppiamenti

- riduzione da maiuscolo a minuscolo (case folding)
  [Es. Casa/casa; Quando/quando; Stato/stato; Rosa/rosa]

- standardizzazione varianti ortografiche
  [Es. socioeconomico ↔ socio-economico; U.E. ↔ UE]

- espansione degli acronimi, delle abbreviazioni e delle contrazioni
  [Es. s.p.a. → società per azioni; dott. → dottore; un' → una]

- Ricostruzione degli accenti
  [Es. caffe' → caffè]



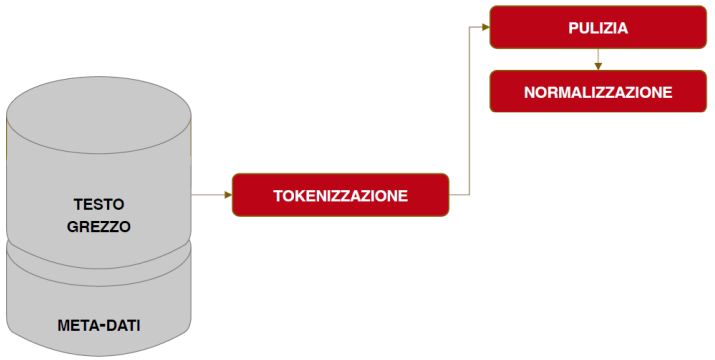
PULIZIA E NORMALIZZAZIONE DEL TESTO

Attraverso le **fasi di pulizia** e **normalizzazione** si agisce sul testo per eliminare possibile «rumore» e migliorare la qualità dei dati per le successive strategie di analisi

I caratteri e le stringhe non alfabetici sono solitamente eliminati dal testo (separatori, simboli speciali, numeri, emoticon, ...), poiché non sono portatori di informazioni interessanti (salvo casi particolari)

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

Ogni fase di trasformazione del testo deve comunque essere sempre effettuata con grande attenzione



LE PARTI DEL DISCORSO

Le parole possono essere raggruppate in classi che mostrano un comportamento sintattico simile e, spesso, una struttura semantica tipica.

Le classi sono indicate con il nome di **parti del discorso** (**POS**)

POS LESSICALI → sostantivi, aggettivi, verbi, avverbi

I *sostantivi* identificano nel discorso qualcosa o qualcuno di cui si può parlare o a cui si può pensare
- Gli *aggettivi* rappresentano le caratteristiche che qualificano le entità rappresentate dai sostantivi
- I *verbi* descrivono una azione effettuata da una entità o uno stato posseduto da questa
- Gli *avverbi* indicano le caratteristiche che qualificano azioni e stati (come gli aggettivi per le entità)

POS FUNZIONALI → articoli, preposizioni, congiunzioni, pronomi

Le POS funzionali, con un numero di elementi limitato rispetto alle prime, servono a creare il contesto

STEMMING E LEMMATIZZAZIONE

Per ridurre l'eterogeneità linguistica a livello morfologico si può intervenire in due diversi modi

STEMMING → le forme flesse sono private delle desinenze e ricondotte al loro tema (**stem** in inglese)

ES.: pesce, pescare, pescato, pescatore → pesc-
Si possono utilizzare le POS per limitare possibili errori nel processo di riduzione

LEMMATIZZAZIONE → le forme flesse sono ricondotte al loro **lemma** (forma canonica)

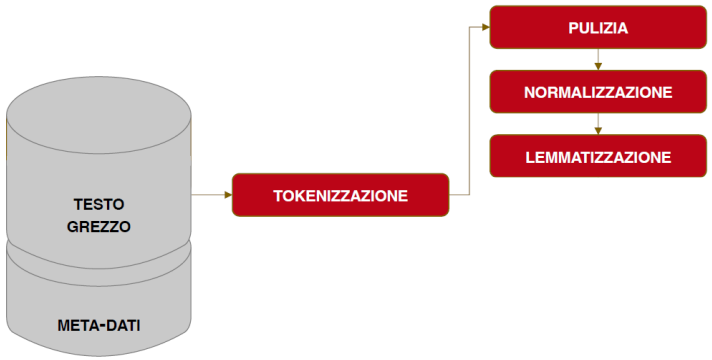
Esempio:

pescavo, pescò, peschiamo → pescare [VERBO]
studenti, studentessa → studente [SOSTANTIVO]

Si devono usare sempre le POS per effettuare il processo di riduzione

Lo Stemming la e Lemmatizzazione 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



SEGMENTAZIONE O LESSICALIZZAZIONE

La **Segmentazione** consiste nell'individuazione delle parole multiple all'interno di un testo.

La segmentazione è utile per trovare le locuzioni poliremiatiche, locuzioni quotidiane e i nomi composti.
Con la segmentazione si individuano **segmenti del testo**, ossia combinazioni di parole successive.

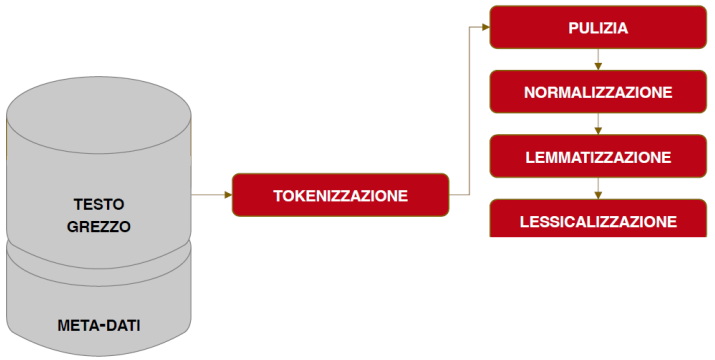
Esempio di parole composte:
"search engine marketing"
"search engine optimization"

Le due parole hanno i primi due termini uguali (search engine) ma il terzo termine associa un significato diverso all'intero gruppo.

Analizzando le singole parole, l'una dopo l'altra, non potremmo accorgerci di questa accezione semantica. Per trovare il vero significato delle parole dobbiamo analizzarle come un segmento di termini.

Allo stesso modo ci sono alcune locuzioni quotidiane e modi di dire che hanno un significato diverso da quello letterale.

Esempio:
"a tutta velocità", "tornare alla carica", "indietro tutta", "a buon mercato"



SEGMENTAZIONE O LESSICALIZZAZIONE

Nei software di text mining sono già presenti delle liste di default con le sequenze di termini più comuni del linguaggio.

Tuttavia, per avere un risultato migliore è consigliabile aggiungere alle liste di default anche delle liste di segmenti specifici della materia o dell'argomento trattato nel testo.

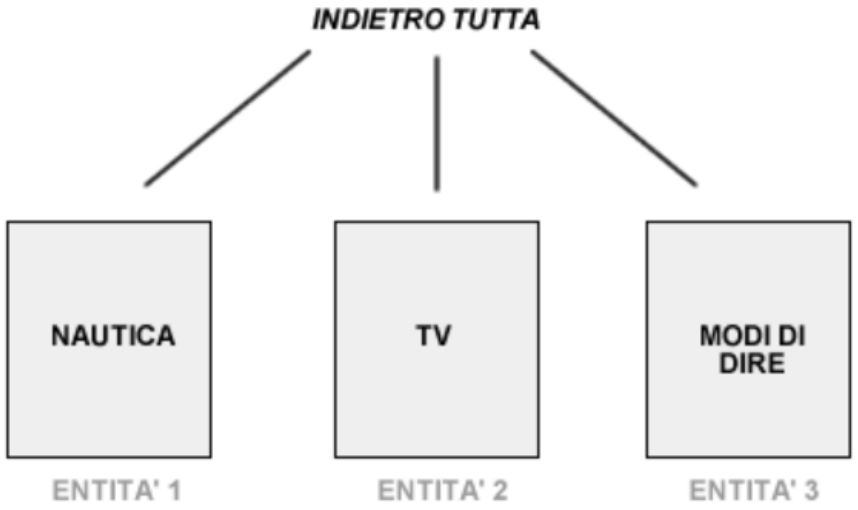
In molti casi l'interpretazione corretta del significato dei segmenti lessicali varia radicalmente a seconda del contesto in cui si trovano.
In pratica, uno stesso segmento letterale può essere associato a entità semantiche differenti a seconda del contesto in cui si trova

SEGMENTAZIONE O LESSICALIZZAZIONE

Nei software di text mining sono già presenti delle liste di default con le sequenze di termini più comuni del linguaggio.

Tuttavia, per avere un risultato migliore è consigliabile aggiungere alle liste di default anche delle liste di segmenti specifici della materia o dell'argomento trattato nel testo.

In molti casi l'interpretazione corretta del significato dei segmenti lessicali varia radicalmente a seconda del contesto in cui si trovano.
In pratica, uno stesso segmento letterale può essere associato a entità semantiche differenti a seconda del contesto in cui si trova



Ad esempio "indietro tutta" è una manovra ben precisa in ambito nautico mentre significa "fermarsi e ricominciare tutto da capo" nel linguaggio comune, oppure un noto programma tv degli anni '80 in ambito televisivo.

Qual è il significato corretto del segmento?
Per capirlo dobbiamo **analizzare il contesto**, ossia le altre parole della frase, del testo o del sito (**co-occorrenze**)

LA COMPRESSIONE SEMANTICA

La **compressione semantica** consente di compattare il vocabolario, sostituendo le parole con basso valore semantico. Preliminarmente, è utile effettuare una categorizzazione semantica per evitare casi di ambiguità ed evitare possibili effetti distorsivi. La compressione può essere effettuata tramite liste, facendo riferimento ai **campi semantici** di ogni parola appartenente al vocabolario

CAMPI SEMANTICI GERARCHICI

Si fa riferimento ad una parola gerarchicamente superiore

iponimi → iperonimo
pioppo, quercia → albero

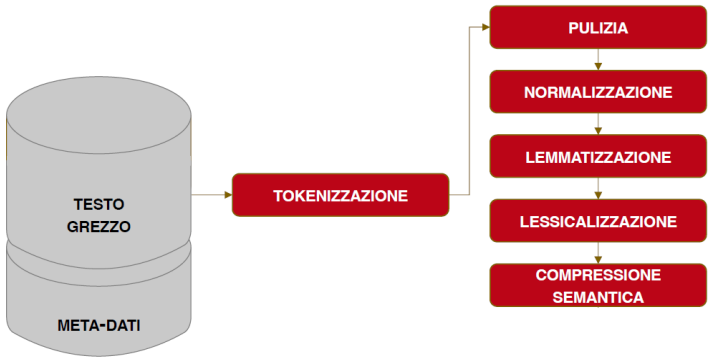
meronimi → olonimo
corteccia, tronco → albero

CAMPI SEMANTICI EQUIVALENTI

Si fa riferimento ai sinonimi, scegliendo il più frequente

ragazza, fanciulla → ragazza

ruscello, torrente → torrente



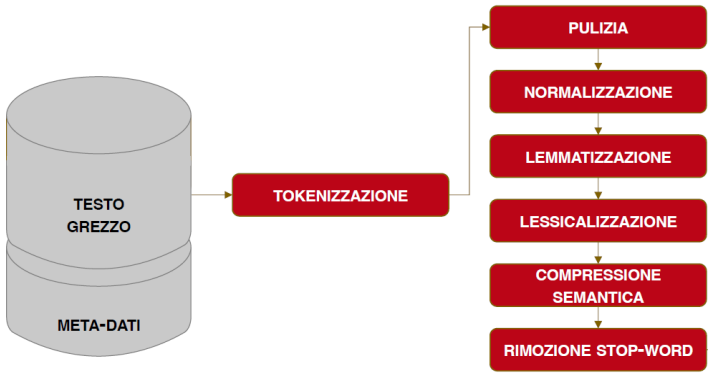
RIMOZIONE DELLE STOP WORDS

Le parole lessicali sono anche note come **parole piene**, portatrici di parti sostanziali del contenuto di un testo, nelle sue modalità di enunciazione o azione

Le parole funzionali non hanno significato autonomo una volta estrapolate dai contesti e quindi inutili da considerare nell'ottica di un trattamento statistico → si parla di **parole vuote** o **stop-word**

La costruzione di una lista di parole vuote (stop-list) è un problema delicato: è impossibile compilare un elenco che vada bene per tutti gli scopi, inoltre ci sono parole piene che a seconda del contesto, per la loro alta occorrenza, risultano banali ai fini delle strategie di analisi

Lingua	Prima	Dopo
it	E venne l'acqua che spense il fuoco che bruciò il bastone che picchiò il cane che morse il gatto che si mangiò il topo che al mercato mio padre comprò	venne acqua spense fuoco bruciò bastone picchiò cane morse gatto mangiò topo mercato padre comprò
en	this is a text full of content and we need to clean it up	text full content need clean



RICOSTRUZIONE DEL TESTO PRE-TRATTATO

TESTO GREZZO

Sai, la gente è strana
Prima si odia e poi si ama
Cambia idea improvvisamente
Prima la verità poi mentirà lui
Senza serietà, come fosse niente
Sai, la gente è matta
Forse è troppo insoddisfatta
Lei segue il mondo ciecamente
E quando la moda cambia
Lei pure cambia
Continuamente, scioccamente
Tu, tu che sei diverso
Almeno tu nell'universo
Un punto sei, che non ruota mai intorno a me
Un sole che splende per me soltanto
Come un diamante in mezzo al cuore
…

TESTO PRE-TRATTATO

sapere gente essere_strano
prima odiare poi amare
cambiare_idea improvvisamente
prima verità poi mentire
senza_serietà essere_niente
sapere gente essere_matto
troppo insoddisfatto
seguire mondo ciecamente
quando moda cambiare
pure cambiare
continuamente, scioccamente
essere_diverso
almeno universo
punto essere ruotare mai intorno
sole splendere soltanto
diamante in_mezzo cuore
…

META-DATI

Anno: 1989
Interprete: donna
Genere: pop

ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS

Un **corpus** è l'insieme dei **testi** oggetto dell'analisi. Il nostro esempio consta di un solo testo, ma volessimo analizzare tutte le opere di Ugo Foscolo, sarebbe preferibile organizzare tutti i testi in un corpus

Il testo originale è stato suddiviso in frammenti corrispondenti alle frasi (divise da segni di interpunzione forti)

All' ombra de' cipressi e dentro l' urne
Confortate di pianto è forse il sonno
Della morte men duro? Ove più il Sole
Per me alla terra non fecondi questa
Bella d' erbe famiglia e d' animali,
E quando vaghe di lusinghe innanzi
A me non danzeran l' ore future,
Né da te, dolce amico, udrò più il verso
E la mesta armonia che lo governa,
Né più nel cor mi parlerà lo spirto
Delle vergini Muse e dell' Amore,
Unico spirto a mia vita raminga,
Qual fia ristoro a' dì perduti un sasso
Che distingua le mie dalle infinite

```
sepolcri.txt.1 :
"All' ombra de'  cipressi e dentro l' urne Confortate di pianto è forse il sonno ..."

sepolcri.txt.2 :
"Ove più il Sole Per me alla terra non fecondi questa Bella d' erbe famiglia e d'..."

sepolcri.txt.3 :
"Vero è ben, Pindemonte!"
```

ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS

Un **corpus** è l'insieme dei **testi** oggetto dell'analisi. Il nostro esempio consta di un solo testo, ma volessimo analizzare tutte le opere di Ugo Foscolo, sarebbe preferibile organizzare tutti i testi in un corpus

> *All' ombra de' cipressi e dentro l' urne*
> *Confortate di pianto è forse il sonno*
> *Della morte men duro? Ove più il Sole*
> *Per me alla terra non fecondi questa*
> *Bella d' erbe famiglia e d' animali,*
> *E quando vaghe di lusinghe innanzi*
> *A me non danzeran l' ore future,*
> *Né da te, dolce amico, udrò più il verso*
> *E la mesta armonia che lo governa,*
> *Né più nel cor mi parlerà lo spirto*
> *Delle vergini Muse e dell' Amore,*
> *Unico spirto a mia vita raminga,*
> *Qual fia ristoro a' dì perduti un sasso*
> *Che distingua le mie dalle infinite*

Il testo originale è stato suddiviso in frammenti corrispondenti alle frasi (divise da segni di interpunzione forti)

	ombra	cipresso	dentro	urna	confortare	pianto	forse	sonno	morte	meno
sepolcri.txt.1	1	1	1	1	1	1	1	1	1	1
sepolcri.txt.2	0	0	0	0	0	0	0	0	1	0
sepolcri.txt.3	0	0	0	0	0	0	0	0	0	0
sepolcri.txt.4	0	0	0	0	0	0	0	0	0	0
sepolcri.txt.5	0	0	0	0	0	0	0	0	0	0
sepolcri.txt.6	0	0	0	0	0	0	1	0	0	0
sepolcri.txt.7	1	0	0	0	0	0	0	0	0	0
sepolcri.txt.8	0	0	0	1	0	0	0	0	0	0
sepolcri.txt.9	0	0	0	0	0	0	0	0	0	0
sepolcri.txt.10	0	0	0	0	0	0	0	0	0	0
sepolcri.txt.11	0	0	0	0	0	0	0	0	0	0
sepolcri.txt.12	0	0	0	0	0	1	0	0	0	0
sepolcri.txt.13	1	0	0	1	0	0	0	0	0	0
sepolcri.txt.14	0	0	0	0	0	0	1	0	0	0
sepolcri.txt.15	1	0	0	0	0	0	1	0	0	0
sepolcri.txt.16	0	0	0	0	0	0	0	0	0	0

ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS: **UNITÀ DI ANALISI DEI TESTI: PAROLE, TOKEN E TYPES**

All' ombra de' cipressi e dentro l' urne
Confortate di pianto è forse il sonno
Della morte men duro? Ove più il Sole
Per me alla terra non fecondi questa
Bella d' erbe famiglia e d' animali,
E quando vaghe di lusinghe innanzi
A me non danzeran l' ore future,
Né da te, dolce amico, udrò più il verso
E la mesta armonia che lo governa,
Né più nel cor mi parlerà lo spirto
Delle vergini Muse e dell' Amore,
Unico spirto a mia vita raminga,
Qual fia ristoro a' dì perduti un sasso
Che distingua le mie dalle infinite

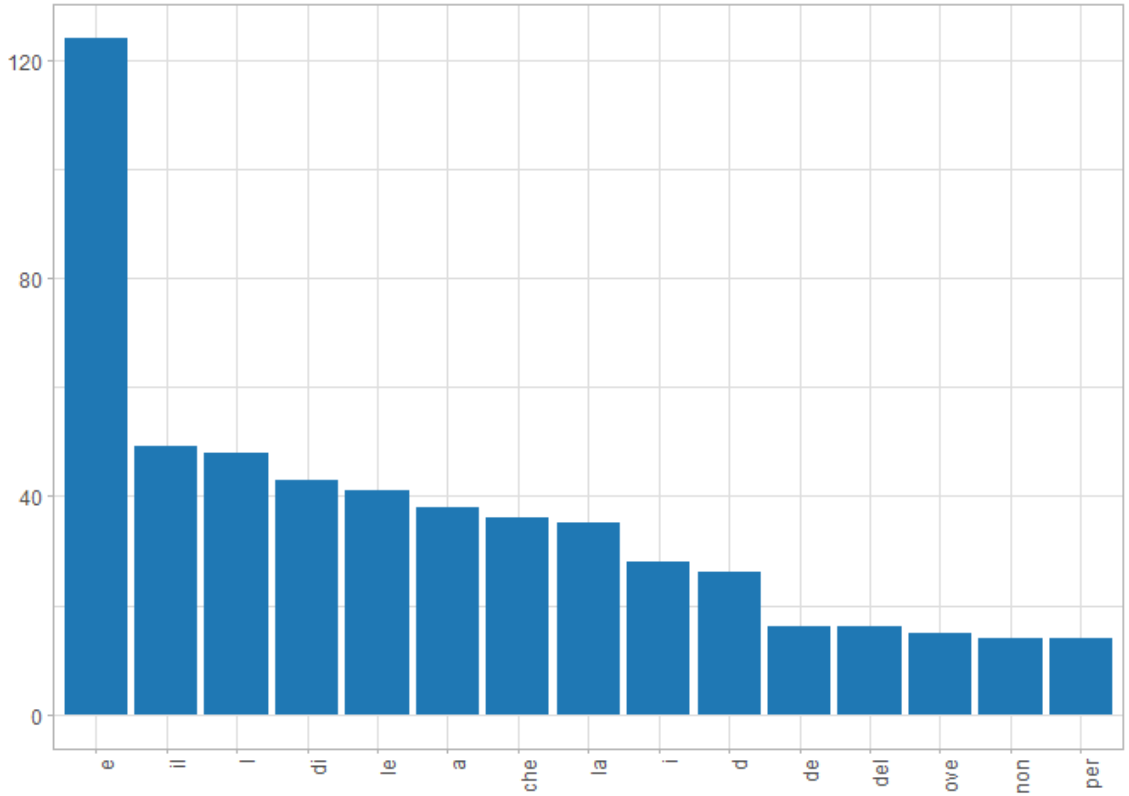
```
Caratteri: 10929
Types: 1096
Tokens: 2363
Sentences: 56
```

E' possibile confrontare l'ampiezza del testo (N, 2.363) — ciò che comunemente si intende per "lunghezza" — con quella del vocabolario (V, 1.096)

ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS: **STOP WORDS**

Le 15 forme più frequenti dei Sepolcri non sono significative dal punto di visto del contenuto, essendo tutte articoli, preposizioni e congiunzioni, ovvero di **parole vuote** (**stop words**).



ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS: **NORMALIZZAZIONE**

Si è proceduto alla normalizzazione:

- l'eliminazione delle parole vuote, o stop words: parole vuote prive di un significato autonomo, come gli articoli o le congiunzioni;
- il trattamento delle maiuscole, identificando (o meno) le entità significative, ovvero i nomi propri (di persona, di cosa o di luogo);
- il trattamento delle forme flesse;
- l'individuazione e il trattamento dei poliformi, o multi-word (come "Presidente della Repubblica", o "anche se").

Il testo del primo segmento dei Sepolcri è:

> *"All' ombra de' cipressi e dentro l' urne Confortate di pianto è forse il sonno Della morte men duro?"*

I tokens, esclusa la punteggiatura, sono:

```
Tokens consisting of 1 document.
sepolcri.txt.1 :
 [1] "All"        "ombra"      "de"         "cipressi"   "e"
 [6] "dentro"     "l"          "urne"       "Confortate" "di"
[11] "pianto"     "è"          "forse"      "il"         "sonno"
[16] "Della"      "morte"      "men"        "duro"
```

Tolte le maiuscole e le parole vuote (*stopwords*), abbiamo:

```
Tokens consisting of 1 document.
sepolcri.txt.1 :
 [1] "ombra"      "cipressi"   "dentro"     "urne"       "confortate"
 [6] "pianto"     "forse"      "sonno"      "morte"      "men"
[11] "duro"
```

ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS: **LEMMATIZZAZIONE**

Si è proceduto alla lemmatizzazione eliminando le le forme flesse

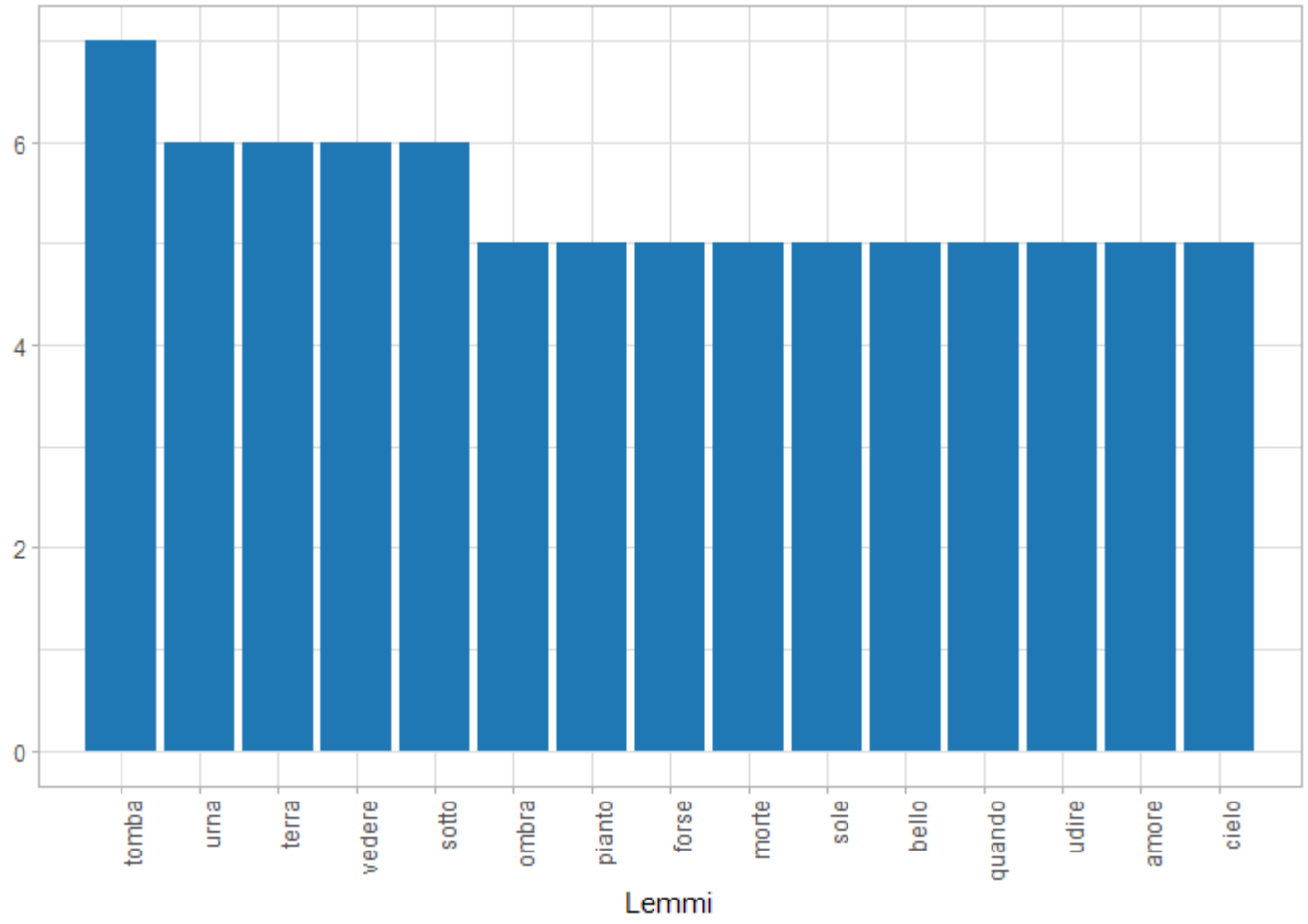
```
Tokens consisting of 1 document.
sepolcri.txt.1 :
 [1] "ombra"      "cipressi"   "dentro"     "urne"       "confortate"
 [6] "pianto"     "forse"      "sonno"      "morte"      "men"
[11] "duro"
```

```
Tokens consisting of 1 document.
sepolcri.txt.1 :
 [1] "ombra"      "cipresso"   "dentro"     "urna"       "confortare"
 [6] "pianto"     "forse"      "sonno"      "morte"      "meno"
[11] "duro"
```

ESEMPIO: I SEPOLCRI DI UGO FOSCOLO

DAL TESTO AL CORPUS: **LEMMATIZZAZIONE**

Una volta normalizzato il testo dell'intera ode, i 15 lemmi più frequenti sono quelli rappresentati nel grafico in figura, certamente più informativo di quello precedente.



DESCRIZIONE QUANTITATIVA DI UN CORPUS

Per poter descrivere un *corpus* da un punto di vista quantitativo è necessario individuare alcune grandezze caratteristiche.

Data una collezione di documenti, indichiamo con:

N → la **dimensione del *corpus***
(insieme di tutte le forme che compongono il *corpus*)

V→ **l'ampiezza del vocabolario**
(insieme di tutte le forme uniche presenti nel *corpus*)

Il numero di types rappresenta l'ampiezza del vocabolario (V)

Nella notazione tipica del Text Mining e del trattamento automatico dei testi le diverse forme del corpus sono solitamente indicate come **token**, mentre le forme appartenenti al vocabolario sono indicate come **type**

DESCRIZIONE QUANTITATIVA DI UN CORPUS

N → la **dimensione del *corpus***
(insieme di tutte le forme che compongono il *corpus*)

V→ **l'ampiezza del vocabolario**
(insieme di tutte le forme uniche presenti nel *corpus*)

Il numero di types rappresenta l'ampiezza del vocabolario (V)

In un corpus composto di 7.940 occorrenze di parole, sono stati rilevati 1.610 termini diversi, e, fra questi, ad esempio il termine «deve» appare 28 volte (i-esima classe di occorrenze).

Quindi:

N= 7.940

V= 1.610

i=28

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Dopo aver effettuato il parsing del corpus è possibile costruire il vocabolario

$$V = V_1 + V_2 + \ldots + V_i + \ldots V_{f_{max}} = \sum_{i=1}^{f_{max}} V_i$$

In generale V_i rappresenta il numero di forme che hanno nell'intero corpus un numero di occorrenze pari a i.

$$N = V_1 + 2V_2 + \ldots + iV_i + \ldots f_{max}V_{f_{max}} = \sum_{i=1}^{f_{max}} iV_i$$

Il primo elemento **V1** è il numero di **hapax** (hapax legomenon= detto una volta sola) del corpus, cioè tutte le forme presenti soltanto una volta

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Dopo aver effettuato il parsing del corpus è possibile costruire il vocabolario

$$V = V_1 + V_2 + \ldots + V_i + \ldots V_{f_{max}} = \sum_{i=1}^{f_{max}} V_i$$

In generale V_i rappresenta il numero di forme che hanno nell'intero corpus un numero di occorrenze pari a i.

$$N = V_1 + 2V_2 + \ldots + iV_i + \ldots f_{max}V_{f_{max}} = \sum_{i=1}^{f_{max}} iV_i$$

Il primo elemento **V1** è il numero di **hapax** (hapax legomenon= detto una volta sola) del corpus, cioè tutte le forme presenti soltanto una volta

Frequency:	1	2	3	4	5	...	F_{max}
Count:	V_1	V_2	V_3	V_4	V_5	...	$V_{F_{max}}$

Dove **V1** rappresenta l'insieme di parole che appaiono una sola volta, **V2** quelle che ricorrono due volte,... **Vfmax** il numero di parole con il maggior numero di occorrenze nel vocabolario

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Estensione lessicale $\frac{V}{N} x100$ *Numero di forme distinte rispetto alla estensione del corpus*

Il ***type/token ratio*** fornisce indicazioni riguardo alla possibilità di analisi statistica del *corpus*: per valori superiori al 20% il *corpus* non è sufficientemente esteso per poter cogliere la ricchezza del linguaggio da analizzare

Ricercatezza nel linguaggio $\frac{V_1}{V} x100$ *Numero di hapax rispetto al numero di forme distinte*

La percentuale di *hapax* nel vocabolario indica quanto il linguaggio utilizzato nel *corpus* è «ricercato»: per valori superiori al 50% il *corpus* è costituito da troppe forme originali e quindi non è trattabile statisticamente

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Dato un token t, si definisce la **document frequency** (**df**) del token come:

df(t) = # documenti contenenti il token t

Dato un documento d e un token t, la **term frequency** (**tf**) indica il numero di volte in cui il token t è contenuto nel documento d:

tf(d,t) = # occorrenze token t nel documento d

Dato un token t, si può calcolare la sua rilevanza tramite la **inverse documenti frequency** (**idf**):

$$idf\ (t) = log\frac{n}{df(t)}$$

Se un token è presente in tutti o quasi gli n documenti, allora sarà poco rilevante in termini di informazione

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Inoltre si utilizza la funzione **TF-IDF** per misurare l'importanza di un termine rispetto ad un documento e/o ad una collezione di documenti

Tale funzione aumenta proporzionalmente al numero di volte che il termine è contenuto nel documento, ma cresce in maniera inversamente proporzionale con la frequenza del termine nella collezione.

L'idea alla base è di **dare più importanza ai termini che compaiono nel documento, ma che in generale sono poco frequenti**

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Esempio di calcolo TF-IDF

Un documento ha i seguenti termini con le seguenti frequenze:

innovation (3), document (2), work (1)

Assumiamo una collezione di 10.000 docs in cui le frequenze globali dei termini sono:

innovation (50), document (1300), work (250)

Ne segue:

innovation: tf= 3/3=1; idf= log(10000/50) = 5,3; tf*idf=5,3

document: tf= 2/3=0,6; idf= log(10000/1300) = 2,0; tf*idf=1,3

work: tf= 1/3=0,3; idf= log(10000/250) = 3,7; tf*idf=1,2

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

DESCRIZIONE QUANTITATIVA DI UN CORPUS

Si può osservare che le parole di un vocabolario hanno sempre una distribuzione, in termini di occorrenze, ben definita. La sua forma è nota dalla **legge di Zipf**. Essa stabilisce che le parole di un vocabolario si distribuiscono in maniera tale che la frequenza (*f*) e il rango (*r*) di una parola sono inversamente proporzionali, ovvero:

$$f * r = c$$

dove *c* è un valore costante.

La **frequenza** di una parola in un testo è data dal numero delle sue occorrenze

Il **rango** di una parola è la posizione occupata da quella parola in una lista di tutte le parole

In un famoso studio Zipf ha esaminato l'Ulysse di Joyce (vocabolario composto da 260.000 occorrenze):

r	f	c
10	2653	26530
100	265	26500
1000	26	26000
10000	2	20000

LE DIVERSE TIPOLOGIE DI MATRICI

Dal punto di vista statistico può essere analizzato mediante 3 tipi di matrici di dati:

1. Matrice frammenti x forme
2. Matrice forme x testi
3. Matrice forme x forme

LE DIVERSE TIPOLOGIE DI MATRICI

Dal punto di vista statistico può essere analizzato mediante 3 tipi di matrici di dati:

1. **Matrice frammenti x forme**
2. Matrice forme x testi
3. Matrice forme x forme

In una matrice **frammenti x forme**, i frammenti di testo sono sulle righe e in colonna sono poste le forme selezionate per lo studio, le quali devono essere considerate come variabili relative a ciascun frammento.
Nel caso di una indagine con domande aperte, le righe ella matrice sono le risposte degli intervistati, mentre le colonne sono le unità lessicali selezionate dal vocabolario del corpus.

A questa matrice può essere associata una serie di variabili aggiuntive (sesso, classe d'età, titolo di studio) che possono essere utilizzate come variabili illustrative.

Frammenti	**Parti del Corpus**						**Metadati**	
	Casa	Famiglia	….	i	….	Zoom	Sesso	Titolo studio
1	1	1	….	0	….	1	M	Scuola primaria
2	0	1	….	0	….	0	M	Liceo
….	…..	…..	…..	…..	…..	…..	….	
I	0	1	….	1	….	1	F	Laurea
…..	…..	…..	…..	…..	…..	…..		
n	1	1	…..	0	…..	1	F	Dottorato

LE DIVERSE TIPOLOGIE DI MATRICI

Dal punto di vista statistico può essere analizzato mediante 3 tipi di matrici di dati:

1. Matrice frammenti x forme
2. **Matrice forme x testi**
3. Matrice forme x forme

Una matrice **forme x testi** presenta sulle righe il vocabolario selezionato per l'analisi e in colonna le parti nelle quali il corpus è suddiviso.

L'informazione statistica interna alla matrice è la frequenza (numero di occorrenze assolute) con cui una forma testuale ricorre in ciascun testo.
Questa matrice non è più sparsa come nel caso della matrice frammenti x forme e permette di lavorare sui contenuti più agevolmente.
Questo tipo di tabella è utilizzata nell'analisi delle corrispondenze lessicali ACL per la costruzione di mappe semantiche

Forme testuali	**Parti del Corpus**					
	1	2	….	i	….	P
Casa	20	56	….	55	….	65
Famiglia	21		….	76	….	64
….	….	….	….	….	….	….
J	….	….	….	23	….	14
….	….	….	….	….	….	….
zoom	50	12	….	54	….	33

LE DIVERSE TIPOLOGIE DI MATRICI

Dal punto di vista statistico può essere analizzato mediante 3 tipi di matrici di dati:

1. Matrice frammenti x forme
2. Matrice forme x testi
3. **Matrice forme x forme**

Una matrice **forme x forme** presenta sia in righe che in colonna le forme selezionate per lo studio. Questo tipo di matrice registra il tipo di co-occorrenza tra le forme all'interno dei testi

	Casa	Auto	Possedere	Vita	Denaro
Casa	1				
Auto	0	1			
Possedere	0	1	1		
Vita	1	1	0	1	
Denaro	1	0	0	1	1

ANALISI TESTUALE

Si ringrazia il Prof. Michelangelo Misuraca per aver condiviso il suo materiale didattico dal quale è stato estratto parte delle slide presentate per il text mining