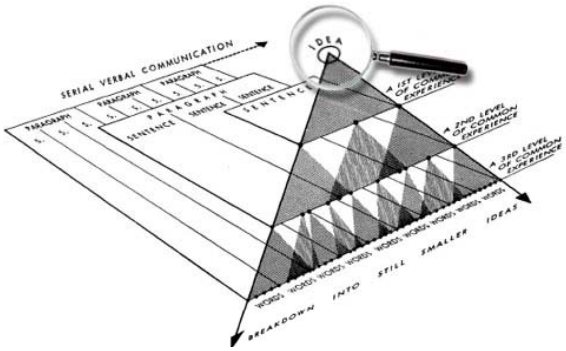
L'Analisi dei Dati Testuali



uno strumento per leggere tra le righe

Michelangelo Misuraca – università della calabria
dipartimento di economia e statistica

INTRODUZIONE

L'Analisi del Testo riguarda lo studio diretto di fonti di natura linguistica

I campi di applicazione sono numerosi

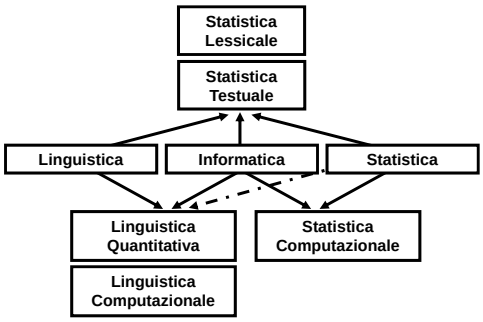
- ☑ Testi letterari
- ☑ Testi brevi (domande aperte, messaggi pubblicitari)
- ☑ Indagini qualitative in ambito psico-sociologico
- ☑ Ricerche terminologiche
- ☑ Costruzione di lessici di frequenza
- ☑ Analisi dei Forum sul Web

Simili studi sono oggi praticabili soprattutto grazie alle capacità dei computer di codificare e riconoscere i caratteri di un qualsiasi linguaggio

LO STUDIO DEL LINGUAGGIO

Il **linguaggio naturale** è la facoltà, esclusiva del genere umano, di esprimere sensazioni e sentimenti, riflessioni, giudizi; di narrare fatti o situazioni; di descrivere aspetti della realtà mediante un *medium* che sia espressione di un determinato livello comunicativo

Le interconnessioni disciplinari nello studio del linguaggio



L'*età dell'informazione* è definita da un costante e diffuso processo di produzione di "fonti testuali"

L'analisi delle informazioni testuali è per sua natura estremamente interdisciplinare: ogni livello comunicativo è approfondito da domini differenti per scopi diversi

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

L'APPROCCIO QUALITATIVO ALL'ANALISI DEL TESTO

L'analisi qualitativa dei testi è uno strumento di ricerca spesso utilizzato in campo sociologico e psicologico per lo studio del comportamento e della psiche umana

Tale analisi si esplicita nell'utilizzo di un metodo positivista come l'**Analisi del Contenuto**

Nella definizione data da K. Krippendorf, può essere definita come un metodo per inferire i significati contenuti in un testo rispetto al contesto in cui lo stesso è stato prodotto

E' una tecnica sistematica e replicabile con cui classificare il contenuto di un testo mediante l'utilizzo di categorie definite. E' possibile seguire due approcci diversi, la codifica **a priori** e la codifica **emergente**:

- nella codifica a priori la definizione delle categorie è fondata su un impianto teorico prestabilito
- nella codifica emergente è necessario un esame preliminare dei testi per individuare le categorie di contenuto (più ricercatori)

I sistemi di codifica più utilizzati sono il **KWIC** e il **KWOC**

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

MARE O MINIERA DI INFORMAZIONI?

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

Due elementi da sottolineare:

- una maggiore **diffusione** dell'informazione (da un punto di vista socio-demografico, culturale, geografico)
- 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 informativo

L'ANALISI TESTUALE: UN PROBLEMA DI ESTRAZIONE DELLA CONOSCENZA

La ricerca della conoscenza in database di notevoli dimensioni (K.D.D.) è messa in atto con un'insieme di strategie nate in ambito informatico, ma il problema è affrontato sempre più anche in ambito statistico

Negli ultimi anni la forbice tra l'analisi di dati strutturati e non strutturati si è ampliata a tal punto che **Data Mining** e **Text Mining** sono ritenuti ambiti di ricerca nettamente distinguibili

Il Text Mining ha come obiettivo l'estrazione di conoscenza a partire da grandi raccolte di fonti testuali (d'ora in avanti *documenti*)

IL BISOGNO INFORMATIVO

Le tecniche di **Mining** aiutano l'utente a soddisfare il suo bisogno informativo attraverso una serie di passaggi formali predefiniti:

- **Ricerca**
- **Browsing (Navigazione)**
- **Visualizzazione**
- **Altri task: Overview dell'intera collezione, Zoom, Filtering…**

LA STATISTICA E L'ANALISI DEL TESTO

Con l'affermarsi e il diffondersi di strumenti informatici adeguati, sia hardware che software, è stato possibile sviluppare delle tecniche d'analisi della lingua sempre più sofisticate

Gli studi sul linguaggio naturale intrapresi da linguisti, sociologi e psicologi, sono stati affiancati dal lavoro che informatici e statistici, partendo spesso da problematiche e prospettive diverse, hanno effettuato sui dati testuali

Gli approcci che si basano su metodologie statistiche fanno riferimento a strumenti di tipo quantitativo per trattare le unità linguistiche contenute in una raccolta di testi

E' in particolare alla scuola francese di **Analyse des Donneés** che va il merito di aver determinato un notevole salto di qualità nell'analisi dei dati testuali e aver prodotto le prime proposte metodologiche compatibili con quelle di taglio informatico

UN PO' DI STORIA…

fine anni '50

Il Centro Studi del Vocabolario della Lingua Francese di Beçanson porta a termine una classificazione delle opere di Corneille e la loro trasposizione su supporto informatico. La disponibilità di questa risorsa incoraggia C. Muller a sfruttarla per effettuare le prime analisi di tipo lessicometrico con l'ausilio di strumenti statistici (Statistica Lessicale). La logica implicita è che il testo analizzato può essere visto come un esemplare rappresentativo della lingua: dallo studio di una base di dati testuali è quindi possibile inferire alla lingua stessa alcuni risultati d'indagine

anni '60

J.P. Benzécri si interessa ai metodi di Analisi dei Dati non come strumento di ricerca in campo psicologico (ambito in cui tali strumenti erano nati e che inizialmente ha dato luogo agli sviluppi più numerosi), ma per l'applicazione degli stessi allo studio della lingua, ponendo le basi alla Analisi dei Dati Linguistici. L'idea portante è quella di aprire le porte ad una nuova linguistica, superando le tesi di N. Chomsky secondo cui non potevano esistere procedure sistematiche per determinare le strutture linguistiche a partire da un insieme di dati come una raccolta di testi

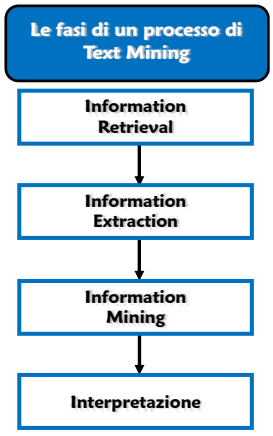
anni '80

Con le prime proposte metodologiche di L. Lebart e di A. Salem si delinea nei suoi tratti fondamentali l'impianto teorico della Statistica Testuale, che a differenza della Statistica Lessicale pone una maggiore attenzione alla testualità della base di dati analizzata. La tendenza attuale è quella di una Statistica Lessico/Testuale che utilizza un approccio "integrato", intervenendo *a priori* sul testo oggetto d'analisi e considerando a supporto delle meta-informazioni di carattere linguistico

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

ALLA SCOPERTA DELLA MATERIA INVISIBILE DELL'UNIVERSO SOCIALE

- ☑ Il sempre maggior sviluppo delle tecniche proprie della *Analisi Multidimensionale dei Dati* su basi di dati testuali consente oggi un utilizzo più proficuo dell'informazione non strutturata
- ☑ E' possibile infatti considerare l'uso dei dati testuali non soltanto come *informazione esterna* o *meta-informazione* nell'analisi e nell'interpretazione dei fenomeni sociali e economici, ma pensare ad una **Statistica Testuale**
- ☑ In tale ottica qualsiasi collezione di documenti scritti in linguaggio naturale può essere analizzata da un punto di vista statistico allo scopo scoprire ed estrarre *conoscenza*



L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

I FONDAMENTI DELLA STATISTICA TESTUALE

Volendo adottare per l'analisi dell'informazione testuale una terminologia nota, risulta necessario individuare correttamente le diverse entità in gioco

Collettivo	Insieme di elementi omogenei rispetto ad una o più caratteristiche
Dato Statistico	Osservazione di una caratteristica su una unità del collettivo
Distribuzione	Insieme dei diversi modi di presentarsi di una caratteristica nel collettivo esaminato

Il modo in cui l'informazione è rappresentata dipende dalle unità d'analisi e dalle regole del linguaggio naturale ritenute significative per il loro riconoscimento e combinazione

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

Corpus

Il collettivo analizzato è rappresentato da una raccolta coerente di materiale testuale, detta *corpus*, omogenea sotto un qualche punto di vista oggetto d'interesse

Questa definizione di *corpus* è applicabile alle fonti testuali più disparate: nel corso degli anni i campi applicativi sono stati numerosi, dalle analisi sulle domande aperte contenute nei questionari ai discorsi politici, le annate di stampa periodica, i messaggi pubblicitari, per arrivare al linguaggio utilizzato in Internet

Potenzialmente è possibile applicare le metodologie proprie della Statistica Testuale a qualsiasi ambito o disciplina che preveda l'utilizzo di un linguaggio più o meno specifico

Occorrenza

Il dato statistico rilevato è il numero di volte in cui una unità lessicale (detta occorrenza) si presenta nella raccolta in esame

Non è scontato attribuire alle occorrenze di una data parola il significato statistico di frequenza, in particolare se il *corpus* considerato non è sufficientemente ampio. Per poter effettuare dei confronti tra *corpora* di ampiezza diversa è conveniente ricorrere alle *occorrenze normalizzate*, frequenze relative ottenute dividendo le occorrenze di ogni parola per una quantità data, variabile in relazione alla dimensione del corpus (in genere 10000, 100000 o 1000000)

Vocabolario

La distribuzione statistica delle parole all'interno del corpus, ossia il vocabolario, è ottenuta misurando per ogni parola il n° di volte che si presenta nella raccolta analizzata. L'ampiezza del vocabolario V è definita dal numero di parole presenti nel testo: $V = V_1 + \ldots + V_k + \ldots + V_{fmax}$ dove V_1 è il n° di parole che si presentano una volta sola nel testo (*hapax*), V_k il n° di parole che si presentano k volte, e V_{fmax} la frequenza della parola con il maggior n° di occorrenze nel vocabolario

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

IL FENOMENO LINGUISTICO COME OGGETTO D'INDAGINE

La lingua è lo strumento di comunicazione maggiormente utilizzato e costituito da un complesso sistema di segni organizzati in struttura

Esistono vari livelli di descrizione delle strutture linguistiche, e ognuno di questi livelli è approfondito da una disciplina specifica della linguistica

Le varie teorie concordano sull'esistenza di quattro domini principali:
- ✓ la *fonologia*, che descrive come i suoni di una lingua si organizzano in sistema
- ✓ la *lessicologia*, che si occupa del senso delle parole, descrive cioè la composizione del lessico di una lingua
- ✓ la *morfologia*, che si occupa delle parole indipendentemente dal significato
- ✓ la *sintassi*, che tratta delle parole nella frase: ordine delle parole, accordo, funzioni delle parole nella frase

Le ultime due discipline, la morfologia e la sintassi, vengono fatte rientrare nel più ampio contenitore della **grammatica** di una lingua, e insieme alla lessicologia sono di fatto le più interessanti in un'ottica di trattamento statistico "multidimensionale"

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

ALCUNE ANNOTAZIONI SUL CONCETTO DI PAROLA

Ogni parola può essere considerata in relazione (1) al suo significato o (2) al ruolo che riveste nell'articolazione della lingua

⇨ Il lessema è l'unità di base del lessico e può essere una radice (*cant-* in *canto*, *cantare*, *cantante*), una parola autonoma (*figlio*, *penna*, *stella*) o una sequenza di parole fissatasi nell'uso in modo che i suoi singoli elementi non possano più essere scambiati né sostituiti con sinonimi (*per lo più*, *dopo cena*, *mulino a vento*)

⇨ Il morfema è l'unità grammaticale di base, ossia "il più piccolo elemento di un enunciato che ha significato":
(1) liberi, se possono presentarsi isolati ed avere una propria autonomia di senso
(2) legati, se non hanno autonomia e quindi non possono restare isolati

⇨ Il sintagma (dal greco "disposizione") è l'unità sintattica autonoma, e la sintassi si occupa dei modi in cui le parole si combinano mostrando connessioni di significato all'interno della frase. Una data entità sintattica può essere considerata da due punti di vista:
(a) nella sua interezza, per la funzione che ha isolatamente
(b) come parte di una unità più ampia

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

PARTI DEL DISCORSO

I linguisti solitamente raggruppano le parole proprie di una lingua in **classi** che mostrano un comportamento sintattico simile, e sovente una struttura semantica tipica

Tali classi sono comunemente indicate con il nome di *categorie grammaticali* o *categorie sintattiche*, ma con maggior precisione vengono indicate **parti del discorso** (POS)

- ✓ le **POS lessicali** (o aperte), rappresentano la classe più numerosa e sono in costante aggiornamento, poiché in esse vi è un continuo processo di acquisizione e coniazione di parole "nuove" (neologismi, barbarismi, ecc.)

  POS LESSICALI => sostantivi, aggettivi, verbi

- ✓ le **POS funzionali** (o chiuse), con un numero di elementi limitato rispetto alle prime ma caratterizzate dal fatto di avere, all'interno di una grammatica, un ruolo ed un utilizzo definito

  POS FUNZIONALI => articoli, preposizioni, congiunzioni, pronomi, avverbi

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

PROCESSI MORFOLOGICI

La morfologia è importante per il linguaggio naturale perché la lingua è "produttiva"

In ogni testo analizzato è possibile infatti incontrare parole o forme flesse di parole non comprese nei dizionari cui si fa riferimento, parole nuove ma morfologicamente connesse a parole note, da cui è possibile inferire le diverse proprietà sintattiche e semantiche

Le categorie sono sistematicamente relazionate ai cosiddetti *processi morfologici*, quali la formazione del plurale di una parola dal singolare, del femminile dal maschile (e viceversa)

I principali processi morfologici da considerare sono:

- → FLESSIONE
- → DERIVAZIONE
- → COMPOSIZIONE

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

PROCESSI MORFOLOGICI (2)

Le **flessioni** sono modificazioni sistematiche della radice di una data parola (lessema), per mezzo di prefissi o suffissi. Non agiscono sulla categoria o sul significato della parola, ma su caratteristiche quali genere (maschile/femminile), numero (singolare/plurale) o tempo (presente/passato/futuro/...)

Il processo di **derivazione** segue un criterio per certi versi meno preciso del precedente, anche se ogni lingua ha meccanismi caratteristici differenti. Il risultato è un cambiamento più "radicale" della categoria grammaticale e spesso anche del significato e dell'uso della parola. Esempi di derivazione sono la trasformazione dei verbi in sostantivi e aggettivi, e dei sostantivi e aggettivi in avverbi

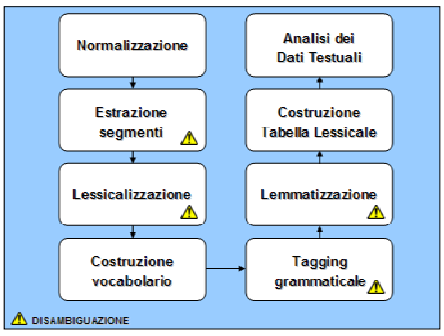
La **composizione** è la fusione di due parole distinte in una parola composta con, talvolta, significato completamente diverso da quello delle singole parole costituenti. Nella lingua italiana tale fenomeno è meno diffuso che in altre lingue, come ad esempio nell'Inglese, e necessita comunque dell'utilizzo di preposizioni e congiunzioni

In generale si definisce *gruppo nominale polirematico*, o più semplicemente "polirematica", un'espressione linguistica composta non modificabile che ha in un lessico l'autonomia di una parola singola

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

LA PREPARAZIONE DEL CORPUS

Per esprimere l'informazione contenuta nel *corpus* in modo compatto e in un formato tale da poter essere trattata statisticamente è opportuno effettuare una serie di operazioni



Criticità

- **Disambiguazione**
- **Contenuto informativo**
- **Codifica**
- **Organizzazione dei dati**
- **Obiettivo dell'analisi**

Tali operazioni, riunite sotto il nome di **pre-trattamento** del *corpus*, sono indispensabili perché i testi scritti in linguaggio naturale non possono essere trattati direttamente per mezzo di algoritmi, essendo non strutturati e con un basso livello di standardizzazione

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

LA SCELTA DELL'UNITÀ DI ANALISI

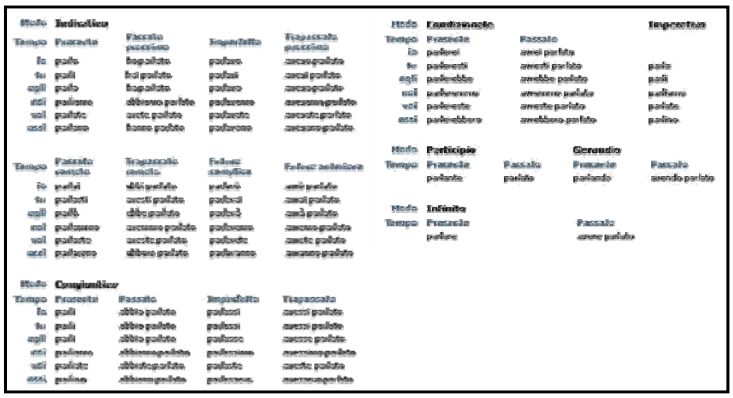
L'unità elementare del linguaggio, la **parola**, non si presta di per sé ad una definizione univoca, poiché la lingua difficilmente può essere vista in senso statistico come un "universo"

La variabilità del fenomeno "lingua" non è facilmente misurabile e comunque l'ampiezza del vocabolario è sensibilmente differente da idioma a idioma:

Verbo *parlare* in italiano

Verbo *to speak* in inglese

speak
speaks
spoke
spoken
speaking



PAROLE VUOTE E PAROLE PIENE

Le **forme principali**, altresì note come "parole piene", sono portatrici di parti sostanziali del contenuto di un *corpus*, delle sue modalità di enunciazione o di azione

Esiste un'ampia classe di forme che non hanno significato autonomo una volta estrapolate dai contesti e pertanto inutile considerare nell'ottica del trattamento statistico

Tali forme, dette **strumentali** (articoli, preposizioni, congiunzioni, pronomi), sono in genere indicate come "parole vuote" o **stop word**:
sono utili a discernere il senso generale del fenomeno analizzato, ma devono essere filtrate per semplificare l'analisi, diminuendo la presenza di rumore nella base di dati

La costruzione di un elenco di forme strumentali (**stop list**) è un problema delicato.
E' impossibile, infatti, compilare un elenco che vada bene per tutti gli scopi: non ci sono particolari problemi con le POS funzionali ma è necessario individuare di volta in volta (a seconda del contesto) quelle forme che risultano "banali", e quindi povere di contenuto informativo

FORME GRAFICHE

Una parola è, convenzionalmente, una forma grafica, ossia una sequenza di caratteri appartenenti ad un alfabeto predefinito, delimitata da due separatori (ad es. segni di interpunzione, spazi, o altri caratteri definiti ad hoc). Tale definizione, proprio perché frutto di convenzioni, risulta essere però arbitraria

L'operazione di riconoscimento all'interno del *corpus* di tutte le forme grafiche che lo compongono, conduce ad una perdita di informazione sul significato, i contesti, lo stile, e più in generale di tutti quei fenomeni generati dalla combinazione di segnali linguistici

Io non àltero mai i fatti: sono troppo altèro per farlo!
In àmbito cinematografico, il "Premio Oscar" è un riconoscimento molto ambìto
Sono molto benèfici verso gli altri, ma non ricordano mai i benefici che hanno ricevuto
Cesare ha molto intùito e perciò ha subito intuìto le intenzioni della sua ragazza
I prìncipi del Rinascimento erano affatto privi di princìpi morali
E' giunto in ufficio il ministro col suo séguito di portaborse, seguìto dalla scorta
Ho subìto un altro affronto, ma mi sono sùbito vendicato
E giunti in porto il marinaio calò ancòra una volta l'àncora della nave

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

FORME TESTUALI

Le "unità minimali di senso" possono tanto essere delle forme grafiche, quanto dei *segmenti di testo* che esprimono un contenuto autonomo. I **segmenti ripetuti** sono disposizioni di k > 2 forme che si ripetono più volte all'interno del *corpus*

Tali sequenze possono essere vuote o incomplete, formate cioè solo da parole grammaticali o da parti di sintagmi, oppure caratteristiche, se costituiscono unità di senso indipendenti

In un'ottica di Statistica Testuale è opportuno considerare come unità elementare di analisi la **forma testuale**, una componente significativa minima del discorso che non può essere ulteriormente decomponibile, sia essa semplice, composta o complessa

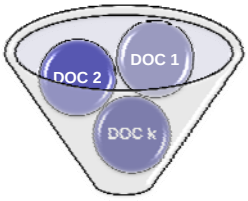
Quando parliamo di forme testuali consideriamo allora contemporaneamente delle forme grafiche, dei lemmi, delle unità minimali (segmenti, poliformi, polirematiche)

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

LA CODIFICA DELL'INFORMAZIONE TESTUALE

Lo schema maggiormente utilizzato per codificare *corpora* testuali in linguaggio naturale è il cosiddetto **Bag-of-Words** (BOW). Tale codifica consente di trasformare ogni documento (o frammento di testo) contenuto nel *corpus* così da strutturare i dati e poterli sottoporre a trattamento statistico

Ogni documento D_j è visto come un vettore nello spazio delle forme del vocabolario:



$$D_k = (w_{1k}, w_{2k}, \dots, w_{ik}, \dots, w_{lk})$$

Ogni termine $w_{i,j}$ è il peso della i-esima forma nel j-esimo documento. E' possibile considerare, a seconda del tipo di analisi effettuata, differenti schemi di ponderazione

IL "PESO" DELLE PAROLE (1)

Nell'analizzare i *corpora* da un punto di vista statistico è rilevante il tipo di codifica che si utilizza per "numerizzare" i documenti. Si ricorre innanzi tutto alla loro trasformazione in "vettori", attraverso la codifica Bag-of-Words

$$D_j = (w_{1j}, w_{2j}, \dots, w_{ij}, \dots, w_{pj})$$

dove il valore assunto da w_{ij} rappresenta l'importanza della *i* - esima forma nel *j* - esimo documento, assegnato in base ad un sistema di pesi che esprime il contenuto informativo di ogni forma presente nei documenti

① **Pesi booleani**

$$weight_{ik} = \begin{cases} 1 & \text{se la forma } i \text{ è presente nel documento } k \\ 0 & \text{se la forma } i \text{ non è presente nel documento } k \end{cases}$$

② **Pesi basati sulla frequenza**

$weight_{ik} = f_{ik}$ ← frequenza assoluta della forma *i* nel documento *k*

③ **Pesi basati su frequenze normalizzate ∝ *Term frequency* (TF)**

$$weight_{ik} = \frac{f_{ik}}{max\ f_k}$$ ← n° di occorrenze della forma più frequente in *k*

IL "PESO" DELLE PAROLE (2)

Il tipo di codifica prescelto dipende dal tipo di analisi che si vuole effettuare sui dati

In talune strategie di trattamento del linguaggio naturale, ed in particolar modo nelle tecniche connesse all'Information Retrieval, si preferisce utilizzare dei sistemi di pesi "complessi"

Term Frequency – Inverse Document Frequency **(Salton & Buckley, 1988)**

$$w_{ij} = \frac{f_{ij}}{max\ f_j} \cdot \log \frac{N}{n_i}$$

peso locale

peso globale

N = n° di documenti nel *corpus*

n_i = n° di documenti con la forma *i*

Una delle peculiarità del TF-IDF è che non esiste uno schema ideale e molto è lasciato all'esperienza del ricercatore e alla validità degli esperimenti empirici condotti negli anni

ESEMPIO DI UTILIZZO DEL TF-IDF

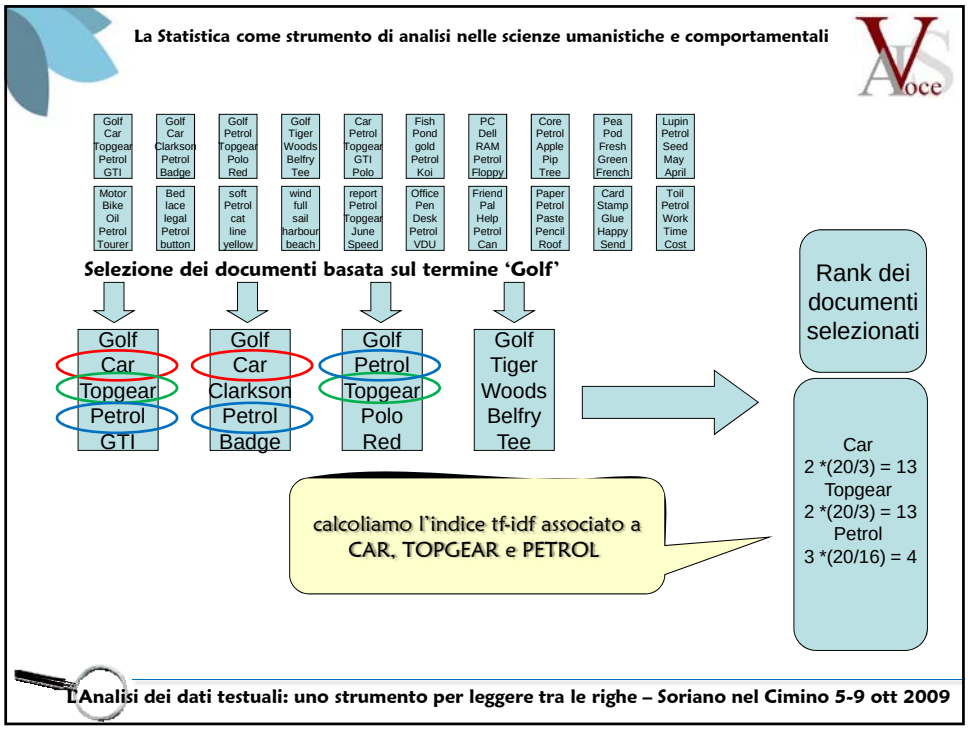
Supponiamo di voler selezionare in un dataset di 20 documenti soltanto quelli in cui si parla della GOLF GTI

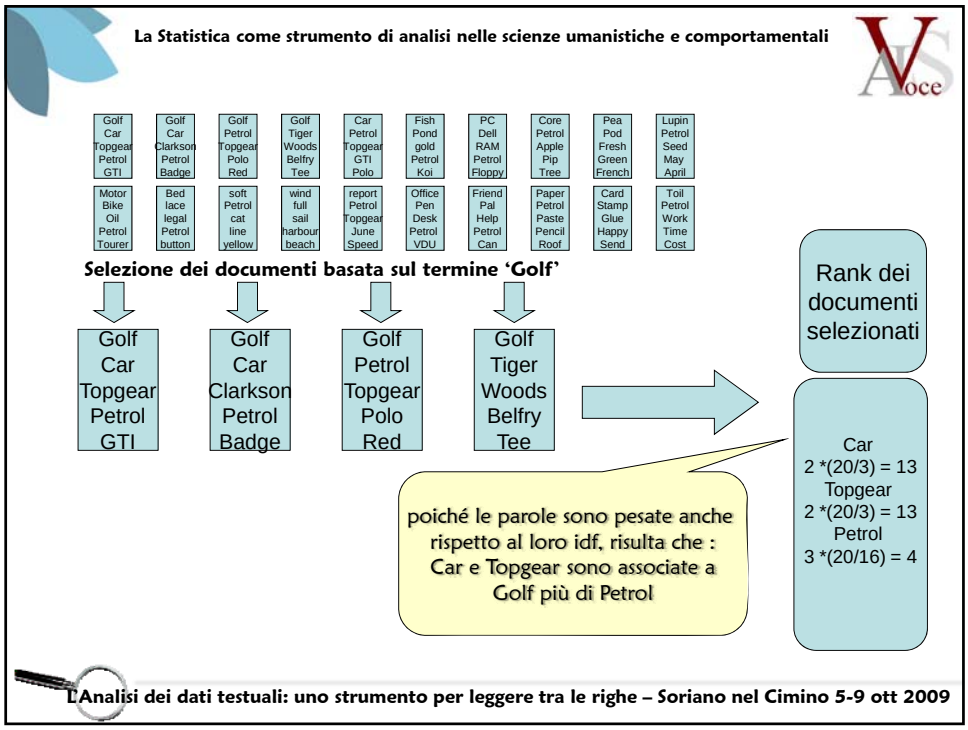
Golf Car Topgear Petrol GTI	Golf Car Clarkson Petrol Badge	Golf Petrol Topgear Polo Red	Golf Tiger Woods Belfry Tee	Car Petrol Topgear GTI Polo	Fish Pond gold Petrol Koi	PC Dell RAM Petrol Floppy	Core Petrol Apple Pip Tree	Pea Pod Fresh Green French	Lupin Petrol Seed May April
Motor Bike Oil Petrol Tourer	Bed lace legal Petrol button	soft Petrol cat line yellow	wind full sail harbour beach	report Petrol Topgear June Speed	Office Pen Desk Petrol VDU	Friend Pal Help Petrol Can	Paper Petrol Paste Pencil Roof	Card Stamp Glue Happy Send	Toil Petrol Work Time Cost

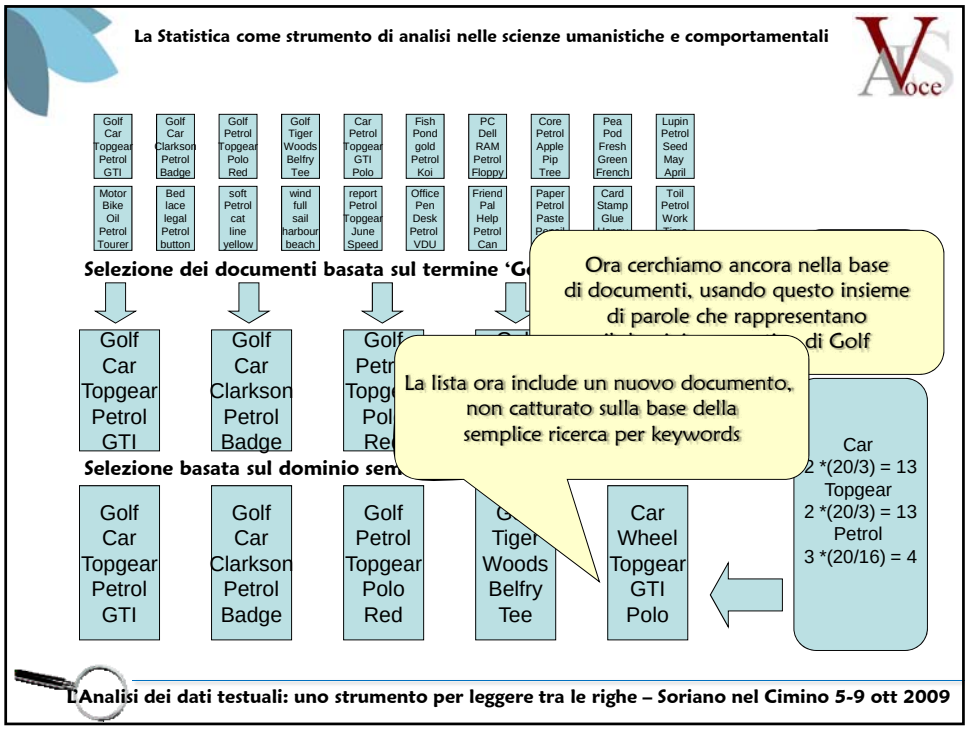
Selezione dei documenti basata sul termine 'Golf'

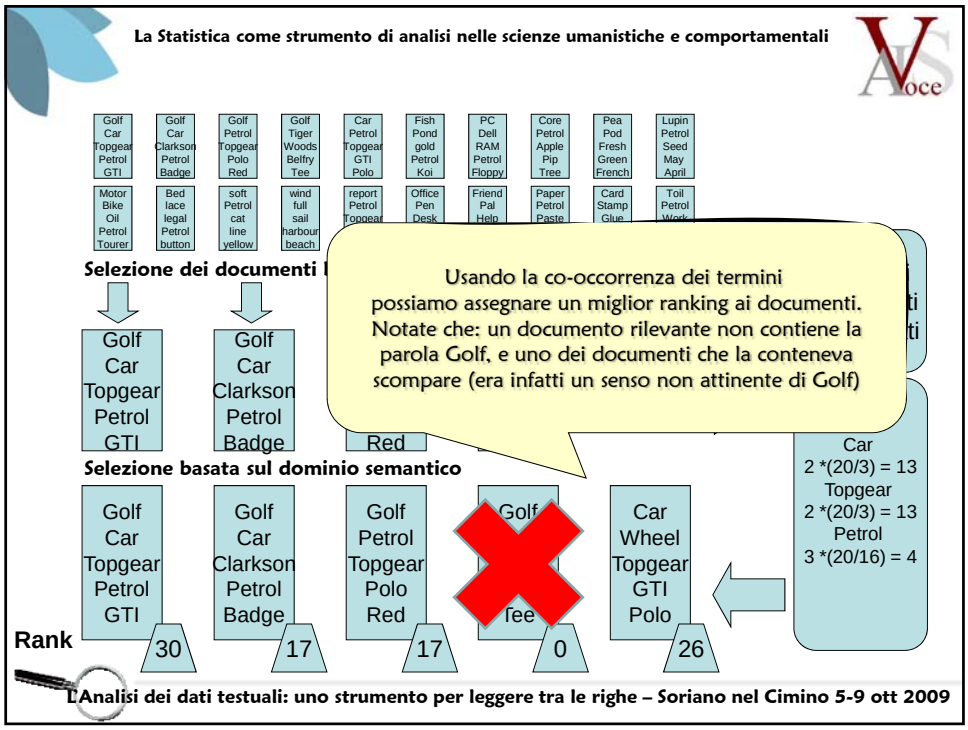
- Golf Car Topgear Petrol GTI
- Golf Car Clarkson Petrol Badge
- Golf Petrol Topgear Polo Red
- Golf Tiger Woods Belfry Tee

vediamo quali sono le parole più rilevanti associate alla parola Golf in questi 4 documenti


La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali
Golf Car Topgear Petrol GTI
Golf Car Clarkson Petrol Badge
Golf Petrol Topgear Polo Red
Golf Tiger Woods Belfry Tee
Car Petrol Topgear GTI Polo
Fish Pond gold Petrol Koi
PC Dell RAM Petrol Floppy
Core Petrol Apple Pip Tree
Pea Pod Fresh Green French
Lupin Petrol Seed May April
Motor Bike Oil Petrol Tourer
Bed lace legal Petrol button
soft Petrol cat line yellow
wind full sail harbour beach
report Petrol Topgear June Speed
Office Pen Desk Petrol VDU
Friend Pal Help Petrol Can
Paper Petrol Paste Pencil Roof
Card Stamp Glue Happy Send
Toil Petrol Work Time Cost
Selezione dei documenti basata sul termine 'Golf'
Golf Car Topgear Petrol GTI
Golf Car Clarkson Petrol Badge
Golf Petrol Topgear Polo Red
Golf Tiger Woods Belfry Tee
Rank dei documenti selezionati
Car
2 *(20/3) = 13
Topgear
2 *(20/3) = 13
Petrol
3 *(20/16) = 4
calcoliamo l'indice tf-idf associato a CAR, TOPGEAR e PETROL
L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009


La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali
Golf Car Topgear Petrol GTI
Golf Car Clarkson Petrol Badge
Golf Petrol Topgear Polo Red
Golf Tiger Woods Belfry Tee
Car Petrol Topgear GTI Polo
Fish Pond gold Petrol Koi
PC Dell RAM Petrol Floppy
Core Petrol Apple Pip Tree
Pea Pod Fresh Green French
Lupin Petrol Seed May April
Motor Bike Oil Petrol Tourer
Bed lace legal Petrol button
soft Petrol cat line yellow
wind full sail harbour beach
report Petrol Topgear June Speed
Office Pen Desk Petrol VDU
Friend Pal Help Petrol Can
Paper Petrol Paste Pencil Roof
Card Stamp Glue Happy Send
Toil Petrol Work Time Cost
Selezione dei documenti basata sul termine 'Golf'
Golf Car Topgear Petrol GTI
Golf Car Clarkson Petrol Badge
Golf Petrol Topgear Polo Red
Golf Tiger Woods Belfry Tee
Rank dei documenti selezionati
Car
2 *(20/3) = 13
Topgear
2 *(20/3) = 13
Petrol
3 *(20/16) = 4
poiché le parole sono pesate anche rispetto al loro idf, risulta che : Car e Topgear sono associate a Golf più di Petrol
L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

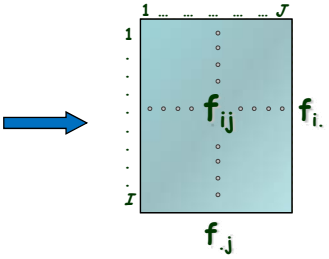

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali
Golf Car Topgear Petrol GTI
Golf Car Clarkson Petrol Badge
Golf Petrol Topgear Polo Red
Golf Tiger Woods Belfry Tee
Car Petrol Topgear GTI Polo
Fish Pond gold Petrol Koi
PC Dell RAM Petrol Floppy
Core Petrol Apple Pip Tree
Pea Pod Fresh Green French
Lupin Petrol Seed May April
Motor Bike Oil Petrol Tourer
Bed lace legal Petrol button
soft Petrol cat line yellow
wind full sail harbour beach
report Petrol Topgear June Speed
Office Pen Desk Petrol VDU
Friend Pal Help Petrol Can
Paper Petrol Paste
Card Stamp Glue
Toil Petrol Work
Selezione dei documenti basata sul termine 'G
Ora cerchiamo ancora nella base di documenti, usando questo insieme di parole che rappresentano
di Golf
La lista ora include un nuovo documento, non catturato sulla base della semplice ricerca per keywords
Selezione basata sul dominio sem
Car Wheel Topgear GTI Polo
Car 2 *(20/3) = 13 Topgear 2 *(20/3) = 13 Petrol 3 *(20/16) = 4
L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009


La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali
Golf Car Topgear Petrol GTI
Golf Car Clarkson Petrol Badge
Golf Petrol Topgear Polo Red
Golf Tiger Woods Belfry Tee
Car Petrol Topgear GTI Polo
Fish Pond gold Petrol Koi
PC Dell RAM Petrol Floppy
Core Petrol Apple Pip Tree
Pea Pod Fresh Green French
Lupin Petrol Seed May April
Motor Bike Oil Petrol Tourer
Bed lace legal Petrol button
soft Petrol cat line yellow
wind full sail harbour beach
Usando la co-occorrenza dei termini possiamo assegnare un miglior ranking ai documenti. Notate che: un documento rilevante non contiene la parola Golf, e uno dei documenti che la conteneva scompare (era infatti un senso non attinente di Golf)
Selezione basata sul dominio semantico
Car Wheel Topgear GTI Polo
Car 2 *(20/3) = 13 Topgear 2 *(20/3) = 13 Petrol 3 *(20/16) = 4
Rank
30
17
17
0
26
L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

LA TABELLA LESSICALE

Seguendo lo schema BOW, i documenti sono organizzati in una matrice **T**, chiamata **tabella lessicale**, con p righe e q colonne. Sulle righe si trovano le forme di maggior interesse selezionate dal vocabolario del *corpus* dopo aver effettuato le operazioni di pre-trattamento, mentre sulle colonne i documenti

	doc 1	doc 2	doc 3	...	doc q
forma 1	1	0	0	...	1
forma 2	0	2	1	...	0
forma 3	1	0	1	...	1
⋮	⋮	⋮	⋮		⋮
forma p	1	0	0	...	0



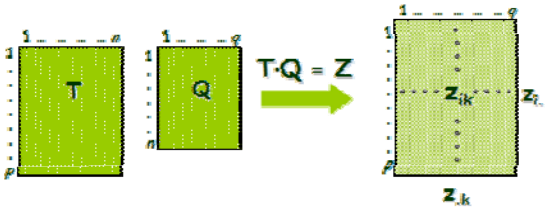
In questa matrice {forme × documenti} ogni cella $f_{i,j}$ contiene l'occorrenza della forma i-esima nel documento j-esimo. Se le occorrenze sono ponderate in accordo a qualche sistema di pesi allora la cella non contiene più la semplice frequenza ma una funzione della stessa

LE TABELLE LESSICALI AGGREGATE

Il principale vantaggio della codifica Bag-of-Words è la relativamente bassa complessità computazionale, derivante dal fatto di poter trattare i testi analizzati come dei "vettori" e quindi calcolare facilmente delle misure statistiche

Per contro, le matrici coinvolte nell'analisi sono *sparse*, ossia con molte celle nulle, poiché non è pensabile ritrovare in ogni documento tutte le forme del vocabolario considerate rilevanti ai fini della ricerca svolta

Se per ogni documento è possibile registrare delle informazioni esterne, diverse a seconda del fenomeno del quale si sta analizzando il linguaggio peculiare, è possibile ottenere delle **tabelle lessicali aggregate**

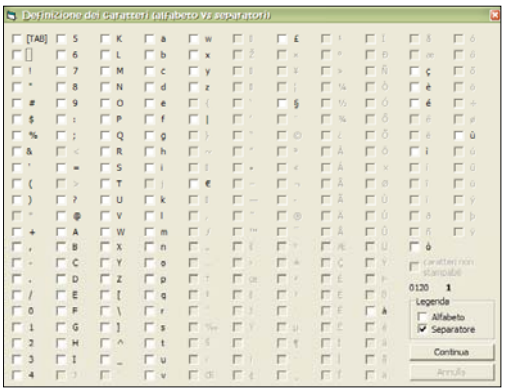


ACQUISIZIONE DEI TESTI

Il testo è considerato come una successione di simboli appartenenti ad un codice, in cui è necessario definire ed identificare un insieme di caratteri che definiscono l'**alfabeto** e un insieme complementare di caratteri che agiscono da **separatori** delle "parole"

L'individuazione dei due diversi set di caratteri avviene per mezzo di una procedura detta di **parsing**, in cui i documenti oggetto d'analisi vengono scansionati e ricondotti ad un elenco di forme grafiche

Generalmente sono considerati come separatori la punteggiatura (.,:;!?), le parentesi, le virgolette, i trattini e i caratteri speciali (es. #£$&@§), ma in alcuni casi si rende necessario definirli *ad hoc*



COSTRUZIONE DEL VOCABOLARIO

Ogni vocabolario è una rappresentazione concreta del discorso di un parlante o di un autore (secondo che si tratti di linguaggio parlato o scritto): è un fatto attualizzato e "individuale", un'espressione delle <parole> nel significato *saussuriano* del termine

Il lessico, in quanto insieme virtuale di segni linguistici, costituisce invece quello stock mentale di radici lessicali (lessemi), esistente nella memoria collettiva di una comunità o di un individuo, da cui possono essere estratte le parole di ogni potenziale discorso

E' possibile "ordinare" le forme del vocabolario in base a criteri diversi:

- ☑ **lessicometrico**: ordinamento decrescente del n° di occorrenze, sono utilizzati ad esempio per costruire dei lessici di frequenza
- ☑ **lessicografico**: ordinamento alfabetico, sono utilizzati quando ad es. si vogliono ricondurre le forme ai lessemi o ai lemmi

Ordinamento per occorrenze

Forma Grafica	Occorrenze
di	8
un	5
o	3
una	2
parole	2
del	2
discorso	2
è	2
fatto	1
attualizzato	1
…	1

Ordinamento alfabetico

Forma Grafica	Occorrenze
attualizzato	1
autore	1
che	1
collettiva	1
comunità	1
concreta	1
costituisce	1
cui	1
da	1
del	2
…	…

NORMALIZZAZIONE

Attraverso la fase di normalizzazione si agisce sui caratteri non separatori per eliminare alcune delle possibili fonti di *sdoppiamento del dato*

Uno dei problemi più comuni e di non facile trattazione è quello della composizione, ossia la costruzione di forme derivate o composte a partire dalle forme semplici, utilizzando il segno "-" (in Inglese *hyphen*). Nell'Italiano coesistono spesso grafie diverse di una stessa forma, evenienza sempre più diffusa dai linguaggi specialistici come quello giornalistico

Le normalizzazioni "basate su liste" consentono ad esempio di ridurre il tasso delle unità lessicali ambigue, ricorrendo ad una *etichettatura* di forme e/o sequenze di forme la cui specificità andrebbe perduta nelle fasi successive di trattamento

I principali obiettivi:

- ⇨ cristallizzare alcune sequenze di discorso, come nel caso dei nomi di giornali
- ⇨ etichettare univocamente le forme e le sequenze ambigue utilizzando le differenze Maiuscolo/minuscolo
- ⇨ uniformare la grafia delle forme presenti nel corpus (toponimi o celebrità scritti in minuscolo: venezia, andreotti ecc.)

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

ESTRAZIONE DI SEGMENTI

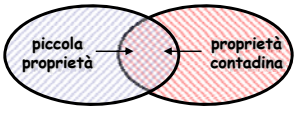
Nella *segmentazione* l'obiettivo è riconoscere e isolare all'interno dei testi analizzati le unità minimali di senso

A	B	C	D	E	F	G
Sep	fg	fg	fg	fg	Sep	fg	fg	fg	Sep

→ AB BC CD EF FG
ABC BCD EFG
ABCD

Analisi dei segmenti *con ridondanza*

piccola proprietà	13 (non contadina 6)
piccola proprietà contadina	7
proprietà contadina	11 (non piccola 4)



Lessicalizzazione dei poliformi *senza ridondanza*

piccola proprietà contadina	7
piccola proprietà	6 (è carente di informazione)
proprietà contadina	4 (è carente di informazione)

La ridondanza dei segmenti estratti (segmenti più lunghi inclusi in quelli più corti) garantisce il riconoscimento di strutture semantiche e "frasi modali", la lessicalizzazione (polirematiche, locuzioni grammaticali) diminuisce drasticamente il livello di ambiguità delle parole

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

SELEZIONE DEI SEGMENTI

A questo punto è possibile scegliere i segmenti a maggiore contenuto informativo, in modo particolare poliformi e poliremiche, e marcarli all'interno del *corpus*, per poi procedere alla costruzione di un nuovo vocabolario di forme testuali

Indice di rilevanza del Segmento
(assorbimento di forme grafiche come capacità selettiva del senso espresso dalla sequenza)

$$IS = \left(\sum_{1} \frac{f_{segm}}{f_{fgr}} \right) * P$$

L'indice si annulla quando il segmento è composto solo da *parole vuote*, e ha un max pari a L^2. L'indice consente di scartare i segmenti vuoti o irrilevanti in termini di grado d'assorbimento

Conoscendo il valore massimo dell'indice di assorbimento è possibile costruire anche un indice relativo, che varia tra 0 e 1

IL PROBLEMA DELLA DISAMBIGUAZIONE

Uno dei problemi più complessi da gestire nella preparazione della base di dati è dato dalla presenza, all'interno dei *corpora* considerati, di forme che presentano una certa "ambiguità"

Le ragioni dell'ambiguità possono essere di natura **lessicale** e **semantica**:

☑ le forme possono essere identiche a livello di rappresentazione lessicale e distinguersi unicamente nei loro contesti sintagmatici (ad es. sono forme flesse di lemmi differenti)

→ sono identificate da una medesima stringa di caratteri dell'alfabeto o dalla medesima pronuncia (**omonimia**)

☑ le forme possono essere nettamente distinte a livello lessicale ma mostrare relazioni di natura extra-linguistica, di tipo semantico o meno, perché possono riferirsi a concetti differenti

→ una stessa forma, con un unico significante, può assumere significati differenti secondo i contesti in cui è usata (**polisemia**)

OMONIMIA E POLISEMIA

Nella lingua italiana l'omonimia si manifesta sotto forma di **omografia**, e una delle maggiori difficoltà sta nel fatto che, a differenza di altri idiomi, gli accenti tonici non sono indicati

(ESEMPIO)

Fine sostantivo maschile = obiettivo
Fine sostantivo femminile = termine
Fine aggettivo = elegante, sottile

La polisemia, uno dei fenomeni più studiati dell'ambiguità lessicale, è frutto dello sviluppo nel tempo di una cultura e della lingua che la esprime:

(ESEMPIO)

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

Grazie all'esistenza della polisemia si è in grado di rappresentare i vari significati tramite un'unica forma, realizzando un'economia indispensabile per l'efficienza della stessa lingua e aumentando il potere simbolico del linguaggio

ANALISI DELLE CONDORDANZE

Se si analizzano *corpora* di grandi dimensioni è impensabile non ricorrere a metodologie di disambiguazione automatica, ma in taluni casi il costo in termini di perdita d'informazione è considerevole

1.	istica economia e commercio ed	economia	aziendale per l inserimento ne
2.	discipline tecniche e i laureati in	economia	che desiderano acquisire prof
3.	ata verso le problematiche della	economia	di impresa prevede una rotazi
4.	ormazione fisica matematica ed	economia	e commercio ad indirizzo ban
5.	canica ingegneria gestionale ed	economia	e commercio con il massimo
6.	tica o scienze dell informazione	economia	e commercio denso manufactu
7.	nica gestionale informatica ed in	economia	e commercio e a diplomati ad
8.	e diplomati geometra laureati in	economia	e commercio ecc in questo ca
9.	i marketing produzione logistica	economia	e commercio ed economia azi
10.	ste principalmente ingegneria ed	economia	e commercio ma senza alcuna
11.	progetti erp si richiede laurea in	economia	e commercio o ingegneria ges
12.	ale scienze dell informazione ed	economia	e commercio requisito fondam
13.	almente laureati in ingegneria ed	economia	e commercio senza alcuna pr
14.	ng per internet start up laurea in	economia	e commercio statistica knowle
15.	vi e formativi aziendali laurea in	economia	e commercio statistica scienz
16.	oni o dell informatica o attinenti l	economia	e la gestione d impresa viene
17.	marketing e vendite la laurea in	economia	è la più indicata per diventare
18.	enze dell informazione statistica	economia	è necessaria una conoscenza
19.	zzazione informatica ingegneria	economia	giurisprodenza informatica ric
20.	arnig center lauree più richieste	economia	ingegneria gestionale meccan
21.	emente ai laureati in ingegneria	economia	matematica fisica ed informati
22.	ere laureati in giurisprudenza in	economia	o in scienze politiche le tecno
23.	tunità per brillanti neolaureati in	economia	o ingegneria senza alcuna pre
24.	ovani laureati in giurisprudenza	economia	scienze dell informazione o di

Il primo passo, una volta individuate le possibili fonti di ambiguità, è quello di usare l'**Analisi delle Concordanze**, studio sistematico dei contesti locali

Per ogni forma ambigua, indicata come *pivot*, si considera un insieme di forme adiacenti, per migliorare la monosemia della stessa

Dopo aver selezionato i frammenti che contengono l'ambiguità, si individuano e analizzano le forme che con essa "co-occorrono" più frequentemente

LA LEMMATIZZAZIONE

Il passaggio obbligato, a questo punto, è quello del riconoscimento delle POS e del loro marcaggio attraverso il *tagging grammaticale*

Per **lemma** si intende la "forma canonica" con cui una data voce è presente nel dizionario

Il principio alla base di tale strategia è che le varianti morfologiche più comuni di una forma hanno significato simile e vengono usati in contesti simili: si parla di ***invarianti semantiche***

Isofrequenze
Le diverse flessioni di uno stesso lemma hanno spesso la stessa quantità di occorrenze: quando ciò non accade si ha un cumulo di usi secondo diverse funzioni sia grammaticali (più categorie) sia semantiche (più accezioni)

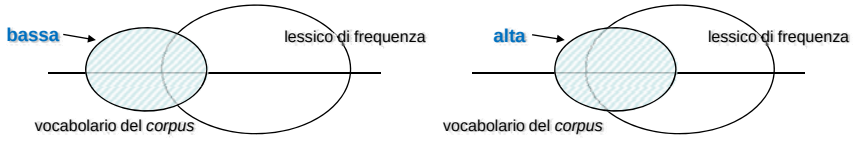
NOMI CONCRETI (scfrequenti)		NOMI ASTRATTI (non scfrequenti)				AGGETTIVI (isofrequenti)		AGGETTIVI / AVVERBI AGGETTIVI POLIFORMI (non isofrequenti)			
libro	178	sviluppo	1725	(≠ accezioni)		grave	289	stesso	1571	o stesso	503
libri	163	sviluppi	129			gravi	265	stessa	906		
zona	285	sistema	1570	(≠ accezioni)		scolastica	131	tutto	5898	dl/del/in/a/per tutto	
zone	280	sistemi	367			scolastico	137			tutto ciò/ tutto questo	
											2589
sentenza	282	materia	604	in materia	423	legittima	77	tutta	1753		
sentenze	271	materie	296			legittimo	79				
								buono	311		
festa	131	corso	1066	in corso	401	leggendario	11	buon	636		
feste	133			nel corso	306	leggendaria	8	buona	1038		
		corsi	288			leggendari	3	buone	208		
						leggendarie	3	buoni	249		

LINGUAGGIO PECULIARE

Un aiuto oggettivo al lavoro dell'analista è dato dal confronto del vocabolario relativo alla raccolta di testi considerata con **lessici di frequenza** relativi al tipo di linguaggio oggetto di studio

Un lessico di frequenza è un particolare tipo di vocabolario ottenuto da una raccolta di testi, relativi ad uno specifico fenomeno, di dimensione notevole

copertura lessicale



Per individuare le unità "peculiari" all'interno del *corpus* analizzato è possibile confrontare la lista di forme al termine del processo di pre-trattamento con il modello di riferimento offerto dal lessico di frequenza relativo

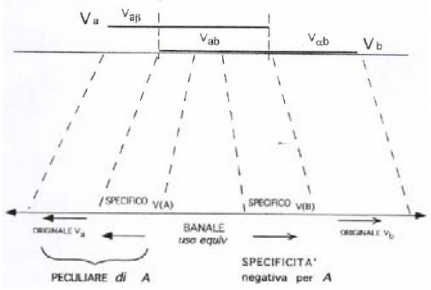
SPECIFICITA'

La specificità di ciascuna forma è valutabile attraverso lo **scarto standardizzato**

$$z_i = \frac{f_i - f_i^*}{\sqrt{f_i^*}}$$

f_i rappresenta il n° di occorrenze normalizzate dell'*i*-esima forma nel *corpus* considerato e f_i^* il corrispondente valore nel lessico di frequenza assunto come paradigma

E' la radice di una statistica Chi-quadro, considerando la frequenza della forma nel lessico come frequenza teorica: consente di valutare la *sovra / sotto-utilizzazione* delle forme rispetto alla loro frequenza nei normali contesti d'utilizzo



E' determinante, in ogni caso, la scelta soggettiva del ricercatore rispetto al tipo di forme da considerare, scelta influenzata fortemente dal grado di conoscenza del fenomeno che si sta analizzando