Management delle informazioni e gestione della conoscenza AA 2021-22

Text Mining

Roberto Boselli
roberto.boselli@unimib.it

Babele di concetti

Facciamo chiarezza

Analisi automatica dei testi (AAT): si serve di statistica, informatica e linguistica computazionale

- Analisi lessicale: linguaggio o vocabolario
- Analisi testuale: discorso

La linguistica computazionale si occupa dell'analisi ed elaborazione del linguaggio naturale attraverso l'uso di metodologie informatiche

La linguistica computazionale si concentra sullo sviluppo di formalismi descrittivi del funzionamento del linguaggio naturale, tali che si possano trasformare in programmi eseguibili dai computer

Linguistica Computazionale

La Linguistica Computazionale è lo studio di sistemi informatici per la comprensione e la produzione di linguaggio naturale. (R. Grishman, ``Computational Linguistics - An Introduction", 1986)

La Linguistica Computazionale si occupa dello sviluppo di una teoria computazionale del linguaggio, sfruttando le nozioni di algoritmi e strutture dati provenienti dall'Informatica. (J. Allen, ``Natural Language Understanding", 1994)

Elaborazione del Parlato e del Linguaggio (Speech and Language Processing)

L'idea di dare ai computer la capacità di elaborare il linguaggio naturale è molto "antica":

- V. Bush 1945: Memex
- A. Turing 1950: "*Can machines think?*"

Esempi dalla Science Fiction:

In "*2001- A Space Odyssey*" (Kubrick, 1968) HAL 9000 è un agente artificiale che può dialogare con gli uomini utilizzando il linguaggio umano

In "*Star Wars*" (Lucas, 1977) C-3PO è un robot protocollare cioè si occupa delle comunicazioni tra umani e robot

Elaborazione del Parlato e del Linguaggio

Quali conoscenze linguistiche dovrebbero possedere tali agenti?

- articolare e decodificare i suoni di una lingua
  - fonetica articolatoria e acustica, fonologia, prosodia, ecc.
- conoscere le parole di una lingua, la loro struttura e la loro organizzazione
  - lessico e morfologia
- comporre le parole in espressioni linguistiche complesse (sintagmi, frasi, ecc.)
  - sintassi
- assegnare significati alle espressioni linguistiche semplici e complesse
  - semantica (lessicale e composizionale)
- usare le frasi nei contesti, situazioni e modi appropriati agli scopi comunicativi
  - pragmatica

Analisi computazionale dei dati linguistici

La linguistica computazionale permette di affrontare ricerche attraverso:

- metodi e strumenti informatici per la rappresentazione e gestione di grandi quantità di dati linguistici
  - digitalizzazione di testi
  - trascrizioni del parlato
  - dati linguistici raccolti sul campo
  - ecc.
- ricerche ed esplorazioni avanzate del testo
- metodi matematici e statistici per analizzare i dati linguistici ed elaborare modelli del linguaggio

Natural Language Processing

- Il Natural Language Processing (NLP) o Trattamento Automatico del Linguaggio (TAL) cerca di dotare il computer di conoscenze linguistiche allo scopo di:
  - progettare programmi e sistemi informatici che assistano l'uomo in "compiti linguistici"
    - traduzione
    - gestione dei documenti e della conoscenza, ecc.
  - sviluppare sistemi informatici che usano il linguaggio naturale per:
    - interagire con essere umani in maniera "naturale"
    - estrarre automaticamente informazioni da testi o da altri media
    - estendere dinamicamente la propria competenza linguistica

Natural Language Processing

Alcune applicazioni

- Correttori ortografici, grammaticali, ecc.
- Recupero “intelligente” di documenti
  - Information Retrieval
- Riconoscimento automatico del parlato
  - Automatic Speech Recognition (ASR)
- Sintesi automatica della voce
  - Text-To-Speech (TTS)
- Estrazione automatica di informazione da testi
  - Information Extraction (IE)
- Interrogare documenti attraverso domande in linguaggio naturale
  - Question Answering (QA)
- Traduzione (semi)-automatica di testi
  - Machine translation
- Interazione (conversazione) uomo-macchina multimodale
  - Agenti conversazionali complessi

Text Mining e Text Analytics

Intelligenza artificiale, grazie alla linguistica computazionale e al NLP, ha dotato il computer di capacità avanzate di elaborare il linguaggio e decodificarne

i messaggi, ciò ha permesso lo sviluppo di:

- 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 TM

Text Mining & Text Analytics

- 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

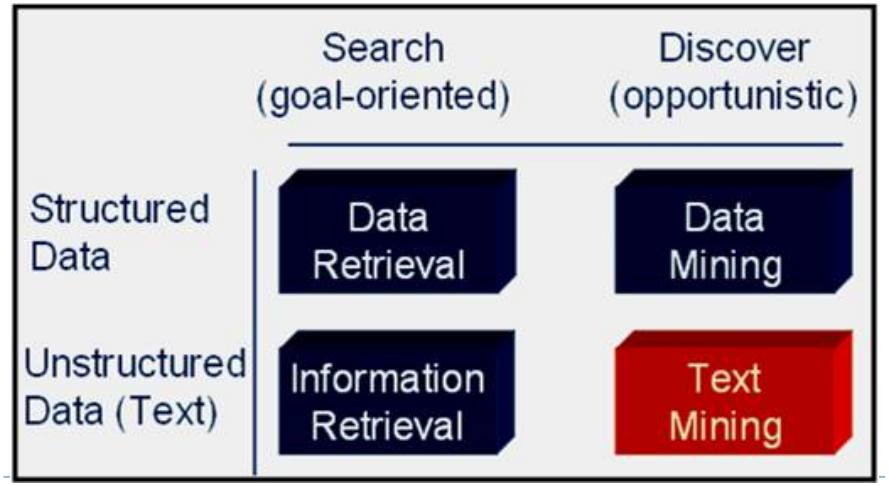
- Steps of Text Analytics:
  - Modeling (e.g., inferential models, predictive models or prescriptive models)
  - Training and evaluation of models
  - Application of these Models
  - Visualizing the Models

Once Term Document matrix is prepared

Data vs. Text Mining

Il processo di KDD si divide in

- Data Mining: estrazione di informazione da dati strutturati
  - Ambiti: Statistica, Machine learning, Database
- Text Mining o KDT (Knowledge Discovery in Texts): estrazione di informazione da databases testuali non strutturati
  - Ambiti: Info Retrieval, Info Extraction, NLP



Knowledge Discovery in Texts

Text preprocessing

Syntactic/Semantic text analysis

Features Generation

Bag of words

Features Selection

Simple counting

Statistics

Text/Data Mining

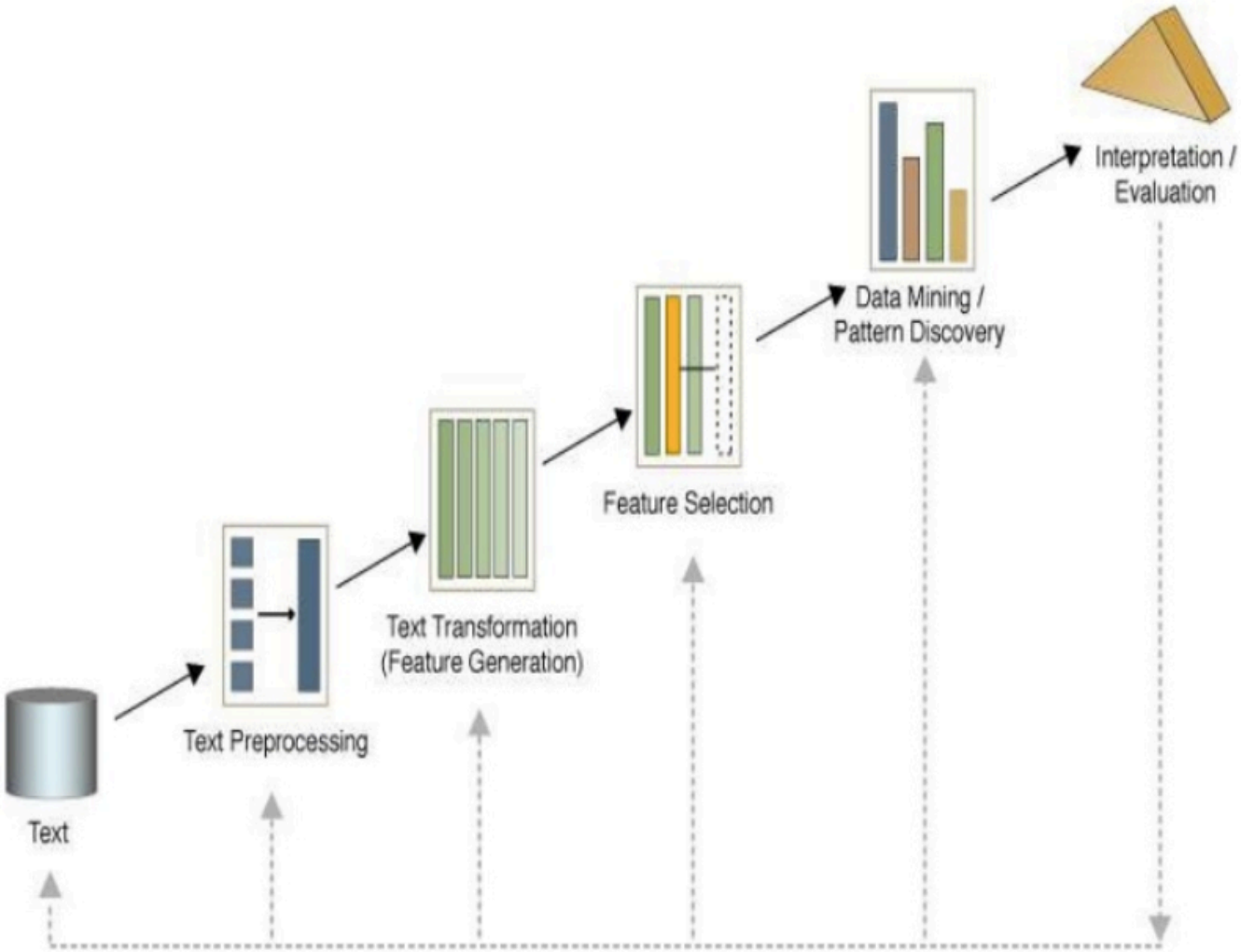
Classification- Supervised learning

Clustering- Unsupervised learning

Analyzing results

Mapping/Visualization

Result interpretation



Text Mining

Obiettivo principale: estrazione di informazione utile implicitamente contenuta in un insieme di documenti

Insieme di tecniche per:

- Analizzare archivi testuali e non
- Strutturare e classificare il contenuto
- Trovare l'informazione nascosta

"Attraverso il text mining si possono analizzare volumi immensi di informazioni e si possono identificare relazioni e strutture che altrimenti sfuggirebbero alla capacità analitica dell'essere umano"

TM, obiettivi (1)

Permette di trattare i documenti con strumenti di analisi automatica:

- riassumere e categorizzare i documenti;
- identificare la lingua in cui sono scritti;
- estrarre concetti chiave, nomi propri e frasi con più parole (n-grammi), contandone le frequenze;
- classificare un documento in funzione della rilevanza rispetto a uno specifico argomento

TM, obiettivi (2)

Aumenta la capacità di recupero automatico di documenti (e.g., Web-crawling):

- estrarre dati in vari formati;
- collegare le informazioni tra di loro in relazioni spaziali o temporali;
- scoprire legami o catene di informazioni legate fra di loro;
- raggruppare documenti in funzione del loro contenuto;
- effettuare analisi incrociate e permettere l'utilizzo congiunto di modelli statistici

TM, approccio multidisciplinare

- “Più complicato” del Data Mining perché lavora con testi non strutturati
- È un campo multidisciplinare, che impiega:
  - Data Mining
  - Information Retrieval (la raccolta di informazioni),
  - Information Extraction (l’estrazione di informazioni),
  - Analisi testuale,
  - Clustering,
  - Tecniche di visualizzazione,
  - Tecniche di trattamento dei database,
  - Machine learning,
  - …

Fonti di dati

- Siti web (social media e altro)
- Banche dati online (e.g., pubblicazioni, brevetti o articoli scientifici)
- Sorgenti informative private
- E-mail
- Opinion surveys
- Newsletters, newsgroups, mailing lists ecc.

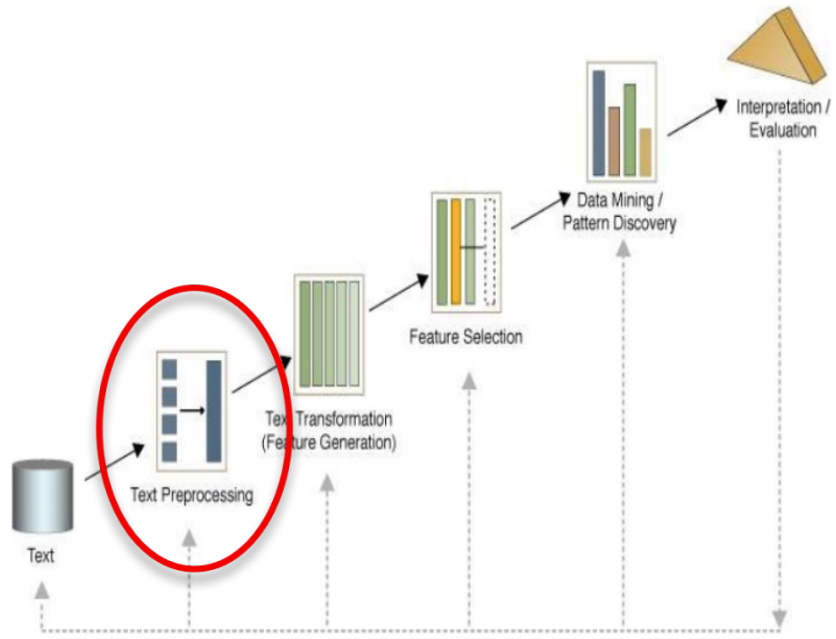
Terminologia

- **Corpus**: collezione di testi
- **Testo**: raccolta di frammenti
- **Frammento**: insieme di parole
- **Occorrenza** (Token): ogni apparizione di una parola nel testo; la frequenza di una parola in un testo è data dal numero delle sue occorrenze
- **N-Gramma**: serie di n elementi da studiare in sequenza, dove n<4/5, per identificare importanti elementi strutturali del testo
- **Documento**: termine con il quale ci si riferisce genericamente all'unità di testo indicizzato nel sistema e disponibile per l'analisi

Tecniche di pre-processing

Preparare i testi per successive elaborazioni (normalizzazione): pulizia dati, eliminazione parole e informazioni non utili alla classificazione ecc.

- Tokenization
- Lower Case
- Stop words
- Stemming
- Lemmatization
- Tagging



Tokenization e Lower Case

- Suddivisione del testo in singole parole (tokens), in unità linguistiche (microsegmentazione del corpus)
- 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

Tokenization algorithms

"La disastrosa campagna di Russia (1812). Il tramonto del suo dominio sull'Europa."

Word & Punctuation (divide in parole e mantiene punteggiatura)

La	disastrosa	campagna	di	Russia	(	1812	)	.	Il	tramonto	del	suo	dominio	sull'Europa	.

Whitespace (divide in parole usando spazi bianchi)

La	disastrosa	campagna	di	Russia	(1812).	Il	tramonto	del	suo	dominio	sull'Europa.

Sentence (divide in frasi usando punto)

La disastrosa campagna di Russia (1812).	Il tramonto del suo dominio sull'Europa.

Regexp: \w+ (divide in parole usando spazi bianchi senza punteggiatura)

La	disastrosa	campagna	di	Russia	1812	Il	tramonto	del	suo	dominio	sull	Europa

Regexp: \w{4,} (divide in parole lunghe almeno 4 caratteri)

disastrosa	campagna	Russia	1812	tramonto	dominio	sull	Europa

Lowercase, esempio

Raw	Lowercased
Canada CanadA CANADA	canada
TOMCAT Tomcat toMcat	tomcat

Stop Words

- Eliminare tutte quelle parole tipiche della lingua che risultano ridondanti e non utili all'estrazione del significato del testo
- Esistono degli elenchi di parole classificate come "stop words list" per ciascuna lingua: articoli, preposizioni, congiunzioni, verbi ausiliari e modali, ecc.
  - Esempio: "il" "e" "che" "avere" "essere" "non"
- Per alcuni obiettivi di analisi però la conservazione di stop words può risultare determinante (e.g., sentiment)

Stopword removal, esempio

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

Stemming

- **Operazione opzionale** per ridurre le parole alla loro radice (*stem* in inglese), detto anche tema
- Nella lingua italiana ci sono regole grammaticali che comportano la presenza di suffissi (*gliele, glieli, glielo ecc.*) e di coniugazioni (*ammo, ando, are ecc.*) rispettivamente per nomi e verbi, che spesso risultano di poco interesse al fine dell'analisi
  - Esempio: "proponeva" diventa "propon"

Stemming (Porter algorithm), esempio

	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

	original_word	stemmed_word
0	trouble	troubl
1	troubled	troubl
2	troubles	troubl
3	troublemsome	troublemsom

Lemmatization

- **Alternativa** allo stemming, è il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta appunto lemma
  - riduzione dei verbi nella loro forma base, per esempio “camminerò” , “cammina” vengono trasformati in “camminare”
- **Problemi:** una parola potrebbe essere riconducibile a più di un lemma, es. ‘giochi’ può essere considerato sia il plurale di ‘gioco’ sia la seconda persona singolare del presente indicativo del verbo ‘giocare’

Lemmatization (WordNet), esempio

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Generare n-grammi

- Tecnica categorizzazione per n-grammi:
  - tenere in considerazione non solo la singola parola (unigram) ma anche quella che la precede immediatamente (bigram), le due prima (trigram) fino a tutte (n-gram)

Questa tecnica è molto utile nel caso di categorizzazione grammaticale (POS). E' noto che nella lingua italiana, per esempio, se si ha un articolo, questo sarà seguito da un nome. Quindi nell'analisi del nome, utilizzando la tecnica del bigram, sarà più facile assegnarlo alla categoria corretta

N-Grams, esempio

"La disastrosa campagna di Russia (1812). Il tramonto del suo dominio sull'Europa."

Bi-gram (+ lowercase + tokenization)

la disastrosa	disastrosa campagna	campagna di	di russia	russia 1812	1812 il	il tramonto
tramonto del	del suo	suo dominio	dominio sull	sull europa		

Normalizzazione

- Il processo di trasformazione di un testo in una forma canonica (standard), e.g., le parole "gooood" e "gud" possono essere trasformate in ”good", la sua forma canonica.
  - Un altro esempio è la mappatura di parole quasi identiche come "stopwords", "stop-words" e "stop words" alla sola "stopwords”
- La normalizzazione del testo è importante per testi come commenti sui social media, messaggi di testo e commenti ai post di blog, in cui prevalgono abbreviazioni, errori di ortografia e uso di parole fuori dal vocabolario (oov)
- Dipende molto dal contesto e non c’è un algoritmo o un metodo standard, si usano metodi statistici, spelling-correction, o dizionari, spesso sistemi a regole o regex (espressioni regolari)

Normalizzazione, esempio

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

Informazioni testuali

- **Statistiche**: il conteggio e la frequenza delle parole, la lunghezza media della frasi, esempi per quantificare il vocabolario e l'uso dei termini di un dato corpus
- **Sintattiche**: la correzione grammaticale, la suddivisione delle sentenze e l'etichettatura, o *labeling*, sintagmatica delle parti del discorso, NER
- **Semantiche**:
  - identificare gli argomenti di un testo, **Text Classification**;
  - la classificazione della tipologia di documento, **Document Clustering**;
  - la classificazione dei topic trattati, **Topic Modeling**;
  - rilevazione se un'e-mail contenga spam, **Spam Detection**;
  - identificazione e valutazione di reviews dei clienti, **Sentiment Analysis**

Rappresentazione dei documenti

Una volta processato e pulito il testo dobbiamo rappresentarlo con un modello

Due approcci:

- **Bag of words**, la sequenza delle parole nel documento è irrilevante (come fossero estratte da un'urna), si applicano metodi di calcolo dei pesi (e.g., TF-IDF), il testo diventa un vettore di termini pesati
- **Collocazioni** = sequenze ordinate di stringhe

Dipende dall'obiettivo:

- Information Retrieval / Text Categorization -> **BoW**
- Natural Language Processing / POS / NER -> **Collocazioni**

Bag of words

- Le singole parole sono analizzate e rappresentate atomicamente come unità singole in **Bag-of-Words** (BOW)
- Il 'bagaglio' di parole, o 'set', è una rappresentazione sparsa delle parole e della loro presenza, indipendentemente dall'ordine sintattico in cui appaiono in una data frase

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	1	0	0	0	0	0	1	0	1

Esempi

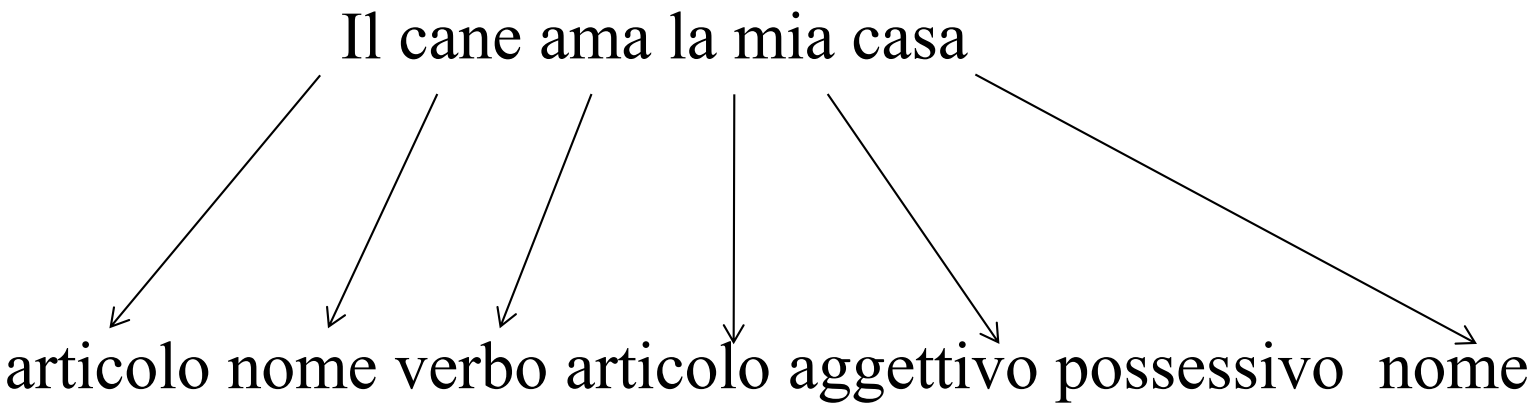
Raw Text	Processed	Steps	Task	How Pipeline Suits Task
She sells seashells by the seashore.	['she','sell','seashell','seashore']	tokenization, lemmatization, stop word removal, punctuation removal	topic modeling	We only care about high level, thematic, and semantically heavy words
John is capable.	John/PROPN is/VERB capable/ADJ ./PUNCT	tokenization, part of speech tagging	named entity recognition	We care about every word, but want to indicate the role each word plays to build a list of NER candidates
Who won? I didn't check the scores.	[u'who', u'win'], [u'i',u'do',u'not',u'check',u'score']	tokenization, lemmatization, sentence segmentation, punctuation removal, string encoding	sentiment analysis	We need all words, including negations since they can negate positive statements, but don't care about tense or word form

Natural Language Processing

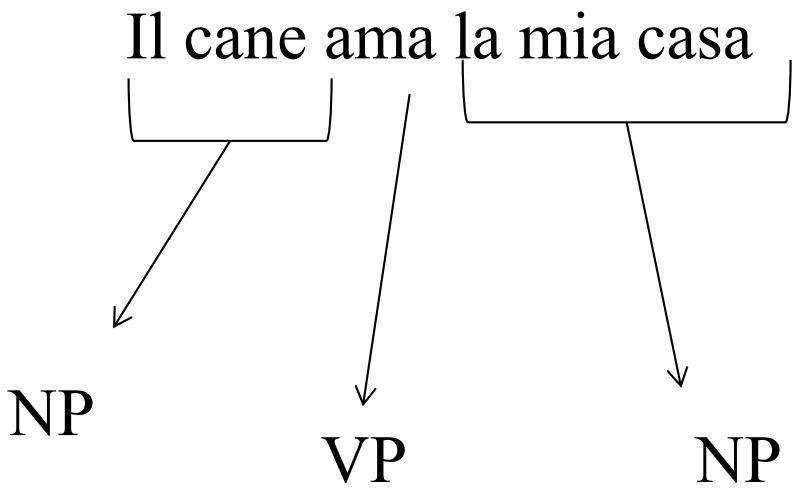
- Combinazione di regole linguistico - grammaticali della lingua oggetto di studio e di algoritmi informatici
- Permette di estrarre informazioni strutturate da testi scritti in linguaggio umano, quindi non strutturati
- Usa:
  - Algoritmi di apprendimento automatico, supervisionato e non supervisionato
  - Part-of-Speech Tagging, assegna un tag grammaticale ad ogni parola del testo sulla base della sua posizione

Part of speech

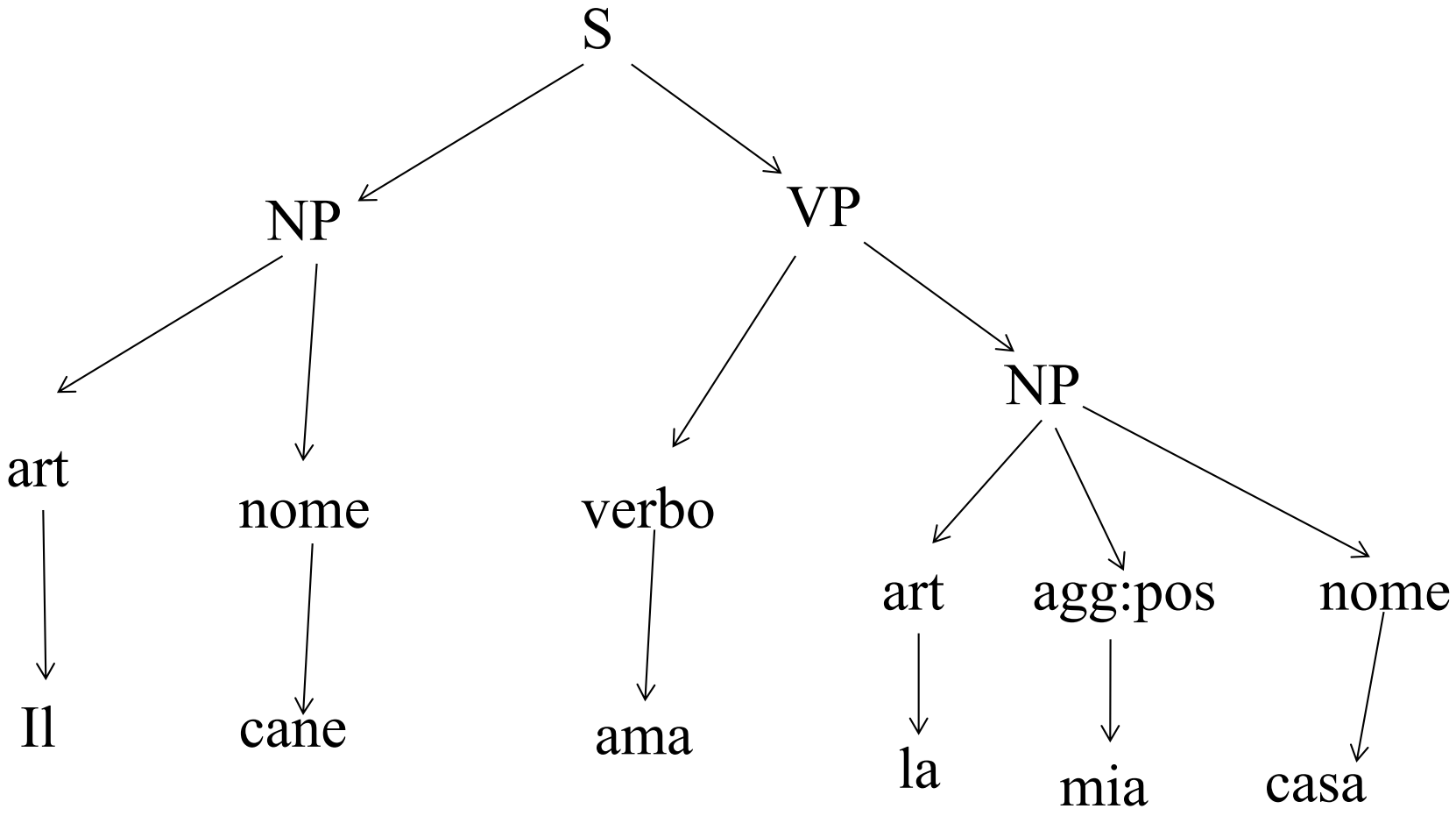
POS: Part-of-speech, la classificazione dei termini in categorie grammaticali (nomi, avverbi, verbi, ecc.)



POS: Shallow parsing



POS: Full parsing



Riconoscimento di nomi

- La fase successiva del processo identifica i vari tipi di nomi propri ed altre forme speciali, come dati e cifre
- I nomi propri appaiono frequentemente in molti tipi di testi e la loro identificazione e classificazione semplifica le successive fasi di elaborazione
- I nomi vengono identificati tramite un set di pattern (espressioni regolari) espresse nei termini del part-of-speech, delle caratteristiche sintattiche e delle caratteristiche ortografiche (ad es. l'iniziale maiuscola)
- NE: Named Entities, identificazione di nomi di entità nei testi, e loro classificazione in un insieme di categorie di interesse predefinite, per esempio persone, luoghi, organizzazioni…

  (es. Dandelion API)

Topic extraction

- La *topic extraction* è una sezione dell'Information Extraction che si occupa di estrarre gli elementi con alto potere informativo da un testo non strutturato, ad esempio:
  - Entità: nomi di persone, organizzazioni o luoghi
  - Concetti
  - Date
  - Espressioni monetarie
  - Citazioni
  - URLs
  - Numeri di telefono
- Utilizza una combinazione di tecniche di Natural Language Processing che permettono di ottenere analisi morfologiche, sintattiche e semantiche del testo e di usarle per identificare gli elementi significativi e classificarli grazie a dei **dizionari** precompilati che contengono le varie classi

Information Retrieval (IR)

Localizzare e recuperare documenti che possono essere considerati rilevanti alla luce degli obiettivi prefissati

- Seleziona un sottoinsieme rilevante di documenti da un insieme più grande e tenta di rappresentare tutto il contenuto informativo di una parte delle informazioni contenute nel testo
- Il termine IR fa riferimento all'attività di ricerca di documenti attraverso delle parole chiavi o composizioni logiche delle stesse (query), le quali a loro volta sono utilizzate per indicizzare i documenti (search engines)

Tecniche di Information Retrieval

Alcuni strumenti essenziali per l'analisi del testo:

- Modello di Spazio Vettoriale
- TF-IDF (Term Frequency - Inverse Document Frequency)
- Coseno
- Document Clustering
- Text Categorization

Modello di Spazio Vettoriale (VSM)

- I documenti sono formalizzati come un insieme di termini che possono essere pesati e manipolati
- VSM rappresenta i documenti in uno spazio *n*-dimensionale, ad ogni documento corrisponde un vettore di *n* componenti
- I documenti vengono rappresentati come vettori di termini **d=(t1 ,t2 ,…,tm)** in cui ogni $\mathbf{t_i}$ è un valore non-negativo indicante l'occorrenza (singola o multipla) del termine i-esimo nel documento **d**
  - ogni termine della collezione di documenti corrisponde a una dimensione dello spazio

Pesi dei termini

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

  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 document 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

Tre modi di assegnare pesi

1. *pesi binari*: relativi alla sola presenza del token:

$$w_{dt} = \begin{cases} 1 \; se \; t \; compare \; in \; d \\ 0 \; altrimenti \end{cases}$$

2. *term frequency*: relativi alla frequenza del token nel documento:

$$w_{dt} = tf(d,t)$$

3. *term frequency normalizzata*: si normalizza la term frequency moltiplicandola per l'inverse document frequency

$$w_{dt} = tf(d,t) \cdot idf(d,t) = tf\_idf(d,t)$$

TF-IDF

- E’ una funzione utilizzata in IR 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 di questo comportamento è di dare più importanza ai termini che compaiono nel documento, ma che in generale sono poco frequenti

TF-IDF, esempio

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

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

- 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

Matrice termini-documenti

- La descrizione dei documenti come vettori di termini permette la costruzione di una matrice, il valore nelle celle rappresenta l'occorrenza/freq di quel termine nel documento o il suo peso tf-idf

	TERMINI								
Tweet	ambiente	amnistia	andare	anno	anticorruzione	antipolitica	approvare	aspettare	assolutamente
1	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0

Similarità

- Il modello spazio vettoriale permette di individuare i documenti simili tra loro
- è necessario definire una misura di similarità tra documenti o del suo concetto inverso, cioè una misura di distanza
- La similarità *S* (o distanza *dist*) tra due documenti d_1 e d_2 può essere calcolata in diversi modi

Prodotto interno

- Qualora si adottassero i pesi binari, la similarità tra i due documenti sarebbe pari al prodotto interno dei due vettori che li rappresentano:

$$S(d_1, d_2) = \sum_{k=1}^{m} w(d_1, t_k) \cdot w(d_2, t_k)$$

- in questo modo la similarità è data dal numero di token in comune nei due documenti

Distanza euclidea

- Per definire la similarità tra due vettori-documento si può utilizzare una misura di distanza, in particolare la distanza euclidea in $\mathbb{R}^m$:

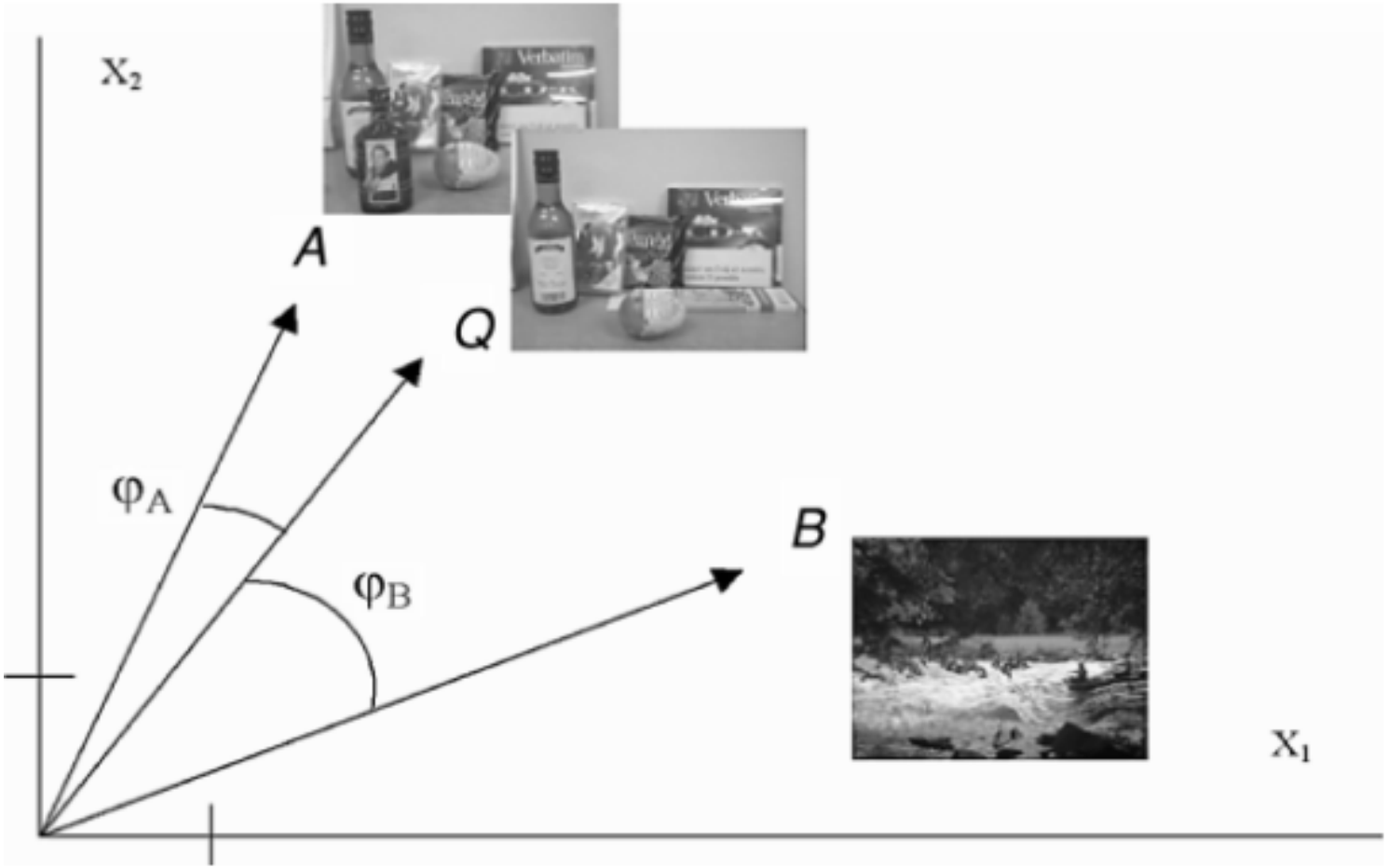
$$dist(d_1, d_2) = \sqrt{\sum_{k=1}^{m} |w(d_1, t_k) - w(d_2, t_k)|}^2$$

Coseno

- La similarità del coseno è la misura maggiormente utilizzata per calcolare la vicinanza tra due testi
- Dati due vettori A e B la similarità del coseno è definita come il coseno dell'angolo che si forma fra gli elementi che definiscono i punti nello spazio multidimensionale. Dove Ai e Bi sono gli i-esimi elementi dei vettori

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

Esempio grafico, spazio tra vettori



cos 90°= 0
cos 0° = 1

Jaccard

- **Indice di Jaccard:** Misura la similarità tramite il rapporto tra la dimensione dell'intersezione e la dimensione dell'unione dei due documenti, considerati non più come vettori ma come insiemi:

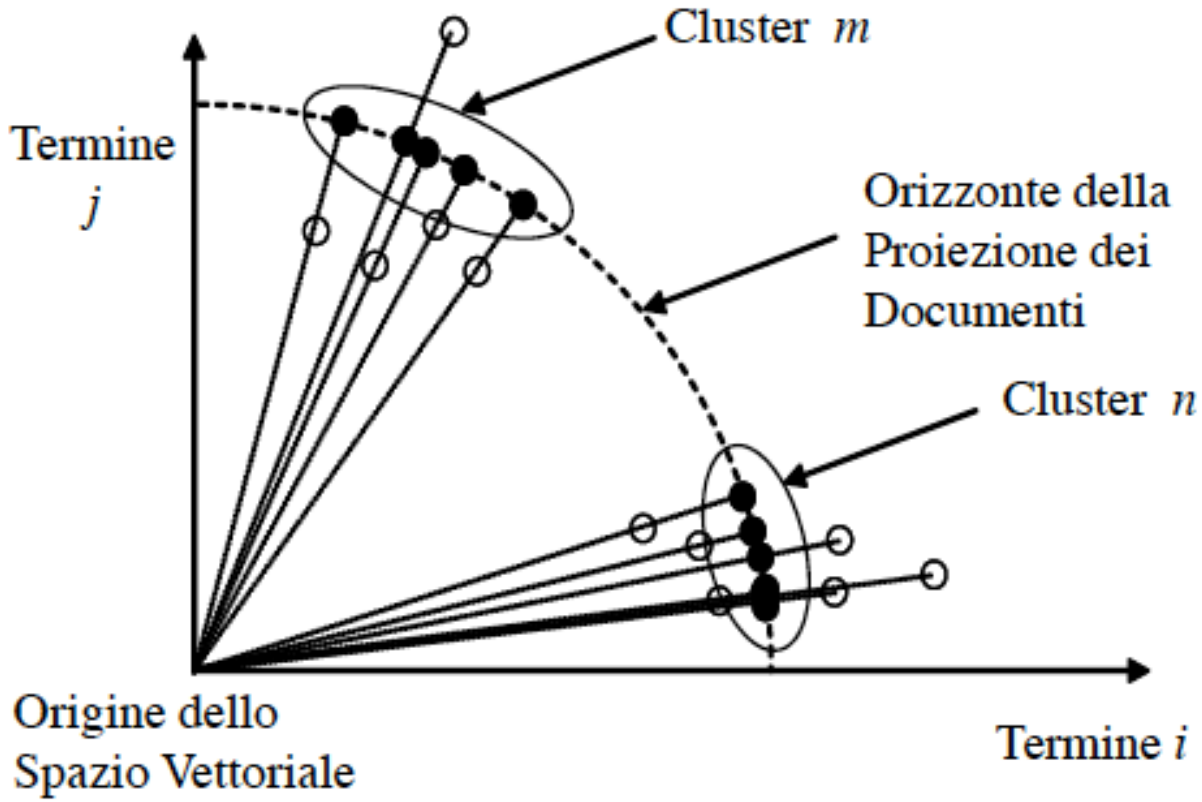
$$J(d_1, d_2) = \frac{|d_1 \cap d_2|}{|d_1 \cup d_2|}$$

- La **distanza di jaccard**, invece, misura la dissimilarità tra i due documenti. Si può calcolare in due modi equivalenti:

$$J_{dist} = 1 - J(d_1, d_2) = \frac{|d_1 \cup d_2| - |d_1 \cap d_2|}{|d_1 \cup d_2|}$$

Document Clustering

- È reso possibile dal calcolo delle somiglianze fra documenti nel Modello di Spazio Vettoriale



Cluster analysis

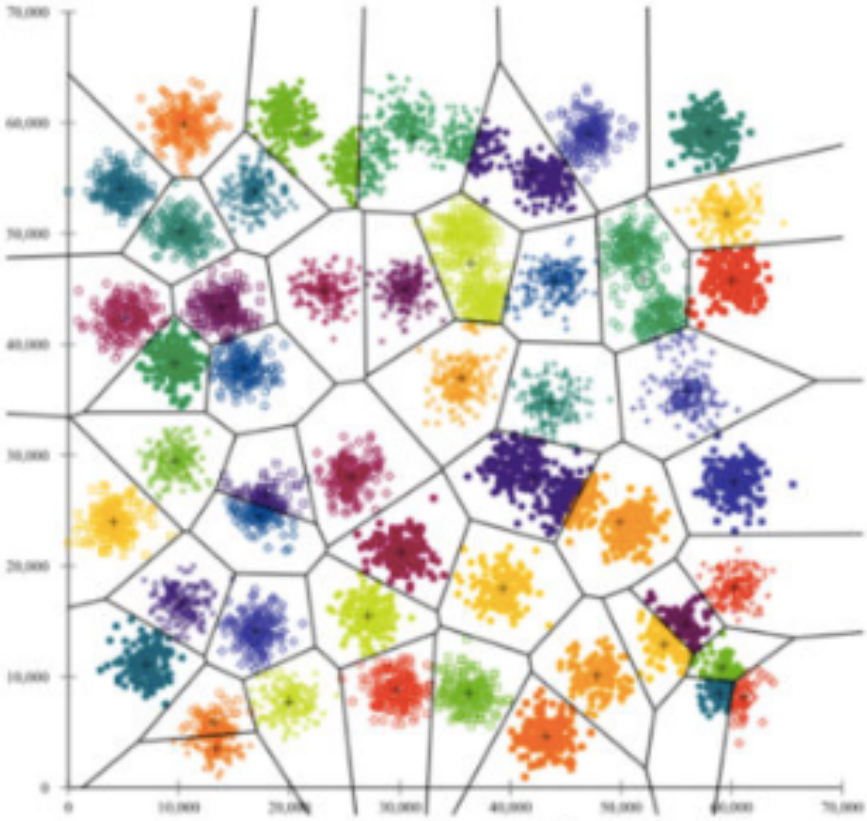
- Insieme di tecniche statistiche il cui obiettivo è costituito dall'individuare raggruppamenti di oggetti che abbiano due caratteristiche complementari:

  A) al loro interno, la massima somiglianza tra gli elementi che li costituiscono (gli oggetti appartenenti a ciascun cluster) = varianza interna (within cluster variance)

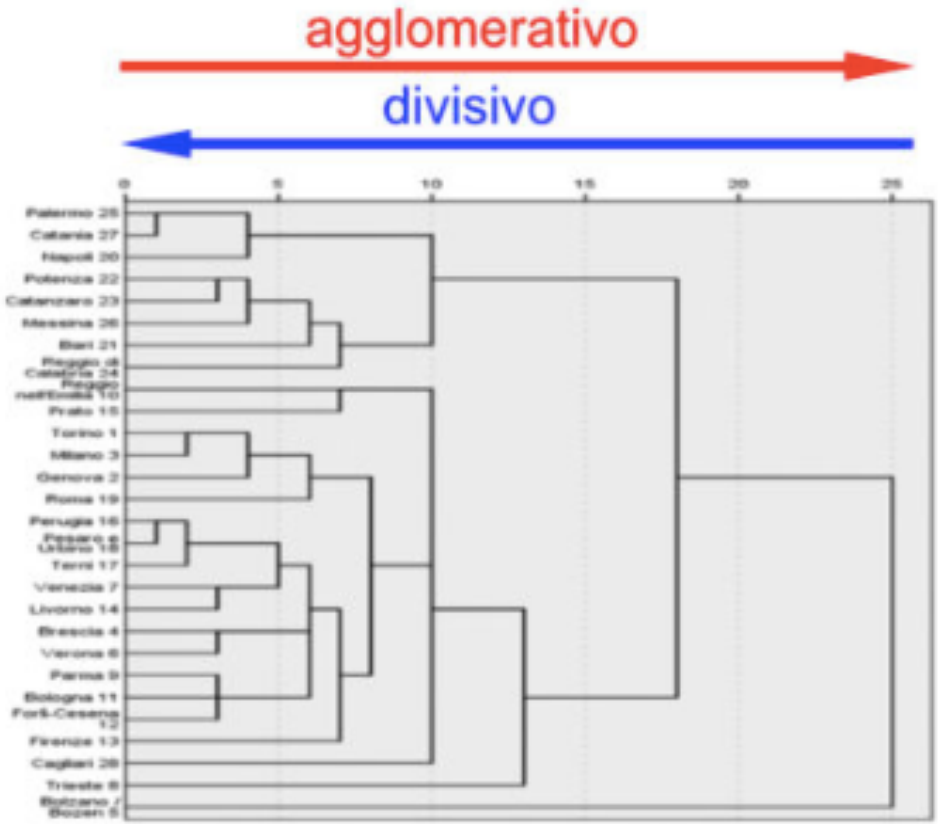
  B) tra di loro, la massima differenza = varianza esterna (between cluster variance)

- I metodi della Cluster Analysis vengono distinti in due tipi:
  - Hierarchical methods, i cui algoritmi ricostruiscono l'intera gerarchia degli oggetti in analisi (il cosiddetto "albero")
  - Partitioning methods, i cui algoritmi prevedono che l'utilizzatore abbia preventivamente definito il numero di cluster in cui l'insieme degli oggetti in analisi va diviso

Gerarchici vs. Partizionali



clustering partizionale



clustering gerarchico
(*dendroide*)

Silhouette

- Misura che confronta la distanza di ogni elemento dal centroide del gruppo a cui viene assegnato con quella dai centroidi degli altri cluster
- La silhouette varia tra -1 e 1, un valore alto indica che l'oggetto è molto legato al gruppo di appartenenza, un valore negativo mostra la non corretta classificazione dell'elemento che quindi risulterà essere più vicino al centroide di un altro cluster

$$s(i) = \frac{d_k - d_i}{\max\{d_i, d_k\}}$$

Clustering nel TM

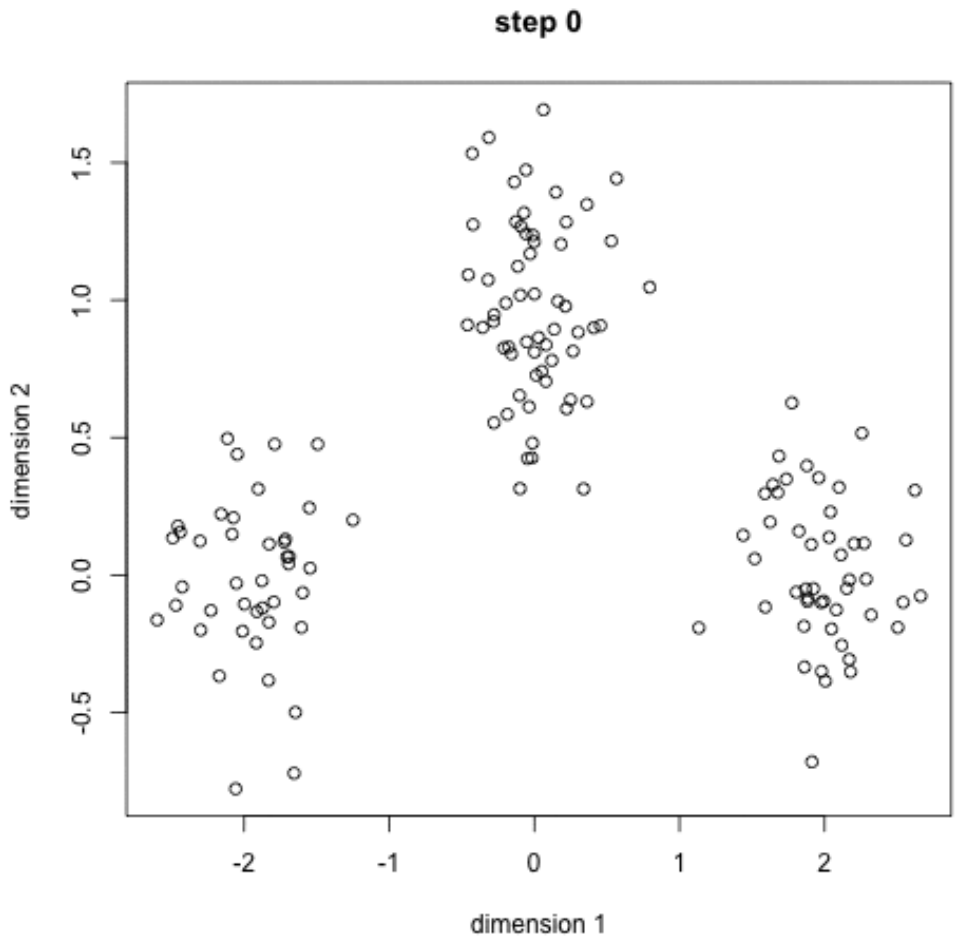
- Spesso è complicato stabilire a priori delle categorie in cui dividere i documenti, per questo risulta utile la **cluster analysis** che divida i documenti in gruppi omogenei
- Dopo l'analisi sarà necessario indagare il contenuto dei gruppi per individuare i principali argomenti trattati dai documenti che vi appartengono, per far ciò si utilizza il **centroide** del gruppo
- Nel text mining si utilizzano le tecniche di clustering in modo tale da non dover analizzare milioni di righe di testo ma solo **le parole chiave contenute nei centroidi** dei vari cluster individuati

K-means

- Tra gli algoritmi di clustering è il più utilizzato nel TM:
  1. si scelgono in modo casuale *k* punti che saranno i *k* centroidi dei cluster
  2. si calcola per ogni elemento la distanza dai *k* centroidi e lo si assegna al cluster corrispondente al centroide più vicino
  3. si ricalcolano i centroidi dei *k* cluster ottenuti al punto 2
  4. SE non si è operato nessuno spostamento di unità nei cluster ➔ FINE
  5. ALTRIMENTI tornare al punto 2

L'algoritmo essenzialmente consiste nei punti 2 e 3 che vengono ripetuti in loop finché i centroidi dei cluster non cambiano più

Algoritmo k-means



K-means, finalità

- L'obiettivo delle K-means è massimizzare la similarità entro i gruppi e minimizzarla tra i gruppi
- L'algoritmo risulta quindi essere un **problema di ottimizzazione** in cui la funzione obiettivo è la somma delle similarità delle unità con il centroide del loro gruppo
- Si possono ottenere risultati diversi in base alla scelta dei k centroidi iniziali
- Conviene quindi eseguire l'algoritmo più volte inizializzando diversamente i centroidi e poi scegliere la clusterizzazione migliore

Text Categorization

- Applicazione in molti campi all'interno del Text Mining: e-mail classification, spam filtering, opinion mining …
- Attività che ha l'obiettivo di classificare testi digitali in linguaggio naturale assegnando in maniera automatica collezioni di documenti ad una o più classi appartenenti ad un insieme, set di classi, predefinito
- Utilizza algoritmi di apprendimento per definire un modello di classificazione

Approcci per la classificazione

- Machine learning

  Tipologie di apprendimento

  - Supervisionato (training e test set)
  - Debolmente supervisionato
  - Non supervisionato
- Basati su regole
  - Individuazione e manutenzione delle regole
  - Gestione conflitti

Topic Modeling

- Tra le tecniche di classificazione non basate sulla cluster analysis esiste la branca dei Topic Models:
- **Latent Dirichlet Allocation (LDA),** ha l'obiettivo di spiegare la correlazione fra le parole chiave e topic (argomenti) simili fra loro
- Si presume che un documento sia una miscela di un piccolo numero di argomenti e che l'utilizzo di ogni parola è attribuibile a uno dei temi del documento
- Un contenuto viene frazionato in frasi più corte e semplici e si presuppone che ogni frase sia correlata semanticamente all'altra, ciò significa che se la prima frase parla di mele (argomento: FRUTTA), difficilmente o quanto meno improbabile la frase successiva parlerà di elefanti (argomento: ANIMALI)