

# **Tecnologie del Linguaggio Naturale**

## **Parte 3 – prof. Di Caro**

**Sgaramella Davide**

**Mazzone Giuseppe**

**Appunti aggiornati a Dicembre 2020**

## Significato del significato

Il significato può essere espresso in diversi modi:

- **Lessico:** insieme degli elementi linguistici a disposizione. Es: dizionario-vocabolario. È l'insieme degli elementi che in un sistema linguistico danno forma a diversi significati.
- **Sintassi:** indica come gli elementi di un dizionario possono essere connessi tra di loro per costruire significati. È la struttura della frase.
- **Semantica:** è l'interpretazione di una struttura lessico-sintattica. La semantica studia il significato delle parole.
- **Pragmatica:** prende in considerazione non solo il contenuto semantico, ma anche informazioni non esplicitate che riguardano il discorso o un background psicologico, cognitivo, contestuale. È la capacità di mettere in relazione le parole che usiamo con il contesto comunicativo in cui ci troviamo nel momento in cui le utilizziamo. Ad esempio: l'espressione ti muovi come un elefante per indicare che una persona si muove in maniera goffa.
- **Ambiguità:** concetto fondamentale in NLP. Si può trovare sia nel lessico che nella sintassi. L'ambiguità è un livello che maschera il vero significato alle macchine.
- **Polisemia/Omonimia.**
- **Comunicazione:** il linguaggio nasce come strumento per condividere significati presenti nella nostra mente. Serve quindi a comunicare.
- **Convenzione:** la comunicazione è un concetto ampio, sia verbale che non verbale. Nel caso verbale si usa la convenzione di trasferire un contenuto semantico attraverso dei simboli (lettere)
- **Granularità:** la struttura del linguaggio può essere più o meno profonda. L'approccio per studiarlo è una vista su questa struttura. Il contenuto semantico cambia se ci concentriamo su una parola, una frase o un testo.
- **Soggettività:** tale livello dipende dalla persona, in base ad essa e agli altri elementi sottostanti può essere più o meno complesso.
- **Cultura:** una parola (convenzione) cambia a seconda della cultura a cui si fa riferimento.
- **Senso comune:** è legato alla cultura. È il significato che diamo alle convenzioni, in maniera condivisa all'interno della cerchia culturale, viene usato per esprimere concetti diversi.
- **Esperienza personale:** dinamica che cambia il significato del significato.
- **Similarità:** è il meccanismo che ci permette di adeguarci a situazioni non previste e non note e ci permette di capire il significato di qualcosa di nuovo. Risulta fondamentale nel NLP perché i calcolatori devono avere a che fare con situazioni non note a priori. Il cervello lavora spesso con questo concetto (cosa abbiamo sentito dire, cosa abbiamo visto).

Tramite queste definizioni abbiamo creato una **ontologia** di base, ossia concetti in cui il significato hanno un valore condiviso. Non ci interessa il significato singolo, ma quello condiviso tra i concetti. Un'ontologia è una rappresentazione formale, condivisa ed esplicita di una concettualizzazione di un dominio di interesse.

## Teorie su “Word Meaning”

Per quanto riguarda il significato delle parole esistono diverse teorie:

1. **Basate su primitive:** per rappresentare il significato di una parola dobbiamo frammentarlo in piccoli contenuti semantici di natura atomica. Si frammentano significati complessivi attraverso significati primitivi. Ad esempio, per comprendere il significato di scrivania dobbiamo prima capire il concetto di tavolo e ancora prima il concetto di struttura piana con fondamenta. Il significato è dato dall'insieme di tali primitive.
2. **Basate su relazioni:** il significato che sta dietro una parola nasce dalla relazione con altre parole. Una parola di per sé non ha significato se non è impiegata in un contesto fatto di altre parole. Parliamo di una sorta di contestualizzazione. qui entra in gioco il concetto di **Logic Forms** e di inferenza: a determinate forme sintattiche posso associare poi un significato inferenziale, da questo deriva quest'altro. la "torta di mele" implica che la mela sia il gusto della torta
3. **Basate su composizioni:** si compongono significati attraverso composizioni lessico-sintattiche. Non solo una parola prende significato quando inserita in un contesto lessicale, ma la composizionalità stessa delle parole produce significati nuovi e le parole stesse prendono un significato individuale. In pratica è la composizione tra parole che dà il significato. Esempio: vino rosso. Rosso di per sé ha già un significato, tuttavia associato al vino il suo significato cambia.

## Triangolo semiotico

Un modello del significato molto utile in questo ambito è quello del triangolo semiotico. Esso rappresenta un modello del significato in cui qualsiasi concetto è rappresentato attraverso 3 componenti:



- Un **concetto** che è la rappresentazione che abbiamo in testa. Ad esempio, se parliamo di gatti il concetto corrisponde all'idea di gatto che abbiamo immagazzinato nel tempo all'interno della nostra mente.
- Una **rappresentazione** che si basa sulle convenzioni e che ci permette di comunicare il concetto. Al concetto viene associato un simbolo, cioè la parola che rappresenta il concetto. Ad esempio, la parola gatto all'interno di una conversazione ci permette di comunicare il concetto che abbiamo in testa.
- Un **referente**, ossia l'elemento reale nel mondo. Ad esempio, il gatto vero e proprio.

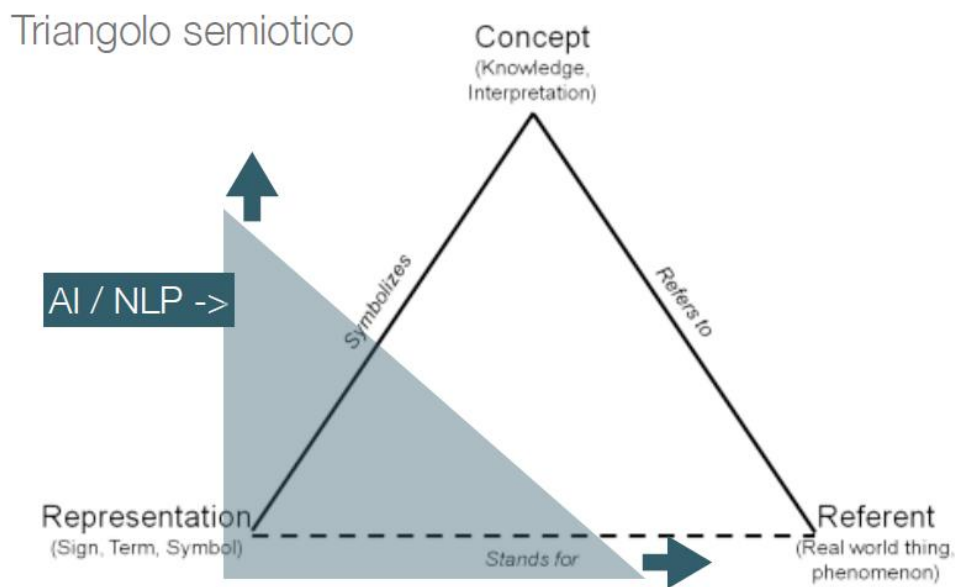
Altri aspetti interessanti del triangolo sono i seguenti:

- Per comunicare abbiamo bisogno della **rappresentazione**, ma non del **referente**.
- La rappresentazione è l'unico punto da cui può partire un algoritmo di IA/NLP.  
L'idea è che un algoritmo debba partire dalla rappresentazione per cercare verso:
  - il **concetto** – dando il significato della parola - ad esempio attraverso qualche rappresentazione semantica
  - il **referente** – ad esempio attraverso delle immagini

L'AI parte da un livello simbolico per arrivare ad una rappresentazione astratta il più possibile vicina a quella mentale della persona.

(Da rappresentazione a concetto, ma non sempre è così in quanto ci sono approcci ibridi: si cerca di prendere info da immagini e testo contemporaneamente, in questo modo si va sia verso il livello concettuale sia verso quello referenziale)

Le persone lessicalizzano i loro pensieri in modo da poterli comunicare attraverso la convenzione del linguaggio naturale (Da concetto a rappresentazione).



## Multilingual word meaning

Un ulteriore aspetto da considerare è il multilingual word meaning. Le lingue hanno molte cose in comune, ma anche differenze a livello sintattico ma anche semantico. 3 aspetti fondamentali sono:

- L'abbattimento delle barriere linguistiche è visto come una **sfida**, ma anche come una grande **opportunità**. Attraverso il confronto di testi in lingue diverse è possibile comprendere sfumature di significato di altre lingue, inoltre informazioni estratte su più lingue hanno un valore più significativo. L'integrazione dei risultati estratti da testi in diverse lingue può permettermi di individuare una semantica un po' più indipendente dalla lingua, in questo modo è anche possibile aiutare la singola estrazione sulla singola lingua. Ci sono diversi scenari: correttezza del parsing it 100%, parsing eng 90% corretto e 10% di errori. Con un parsing posso migliorare l'altro, cioè in quello italiano posso trovare eventuali errori già trovati in quello inglese, mentre in quello inglese posso trovare significati che non erano stati individuati.
- Ci sono delle **lingue rare** dove l'assenza di un dizionario rende tutto più complesso. Ad esempio, il Norvegese viene usato per alcune comunicazioni militari, citazione "La cosa".
- Abbiamo il problema dei "**concetti comunque diversi**" secondo cui alcuni concetti tradotti esprimeranno significati leggermente diversi indipendentemente dalla traduzione. Non esistono delle traduzioni perfette tra 2 lingue, questo dipende

ci sono delle differenze culturali

anche dal territorio in cui è parlata una certa lingua, fenomeno amplificato anche dalla diversa concettualizzazione di alcuni concetti (senso di giustizia ...). Alcune parole non possono essere proprio lessicalizzate. Ad esempio, parole intraducibili in da IT-EN sono: “boh”, “spaghetтата”, “abbiocco”.

## Granularità del Significato

L'analisi semantica è l'attività di assegnazione di un significato e può essere effettuata in diversi modi a seconda del livello di granularità utilizzata. Ogni task di NLP ne ha una associata:



- **Word:** complessità già elevata, word sense disambiguation
- **Chunk:** multiword expression
- **Sentence:** si utilizza il question answering
- **Discourse:** chatbot, botta e risposta tra i partecipanti
- **Document:** summarization, generare un estratto che contenga il più possibile della semantica del primo. Creazione di un riassunto
- **Documents collection:** topic modelling, data una collezione di documenti, trovare i temi della collezione

## Word Sense Disambiguation

Una parola può avere tanti significati. In base al contesto disambighiamo le varie parole. Per essere effettuata dobbiamo avere a che fare con alcuni problemi:

- **Specificità:** granularità dei sensi troppo fine, ci sono troppi sensi per una parola. Alcuni di questi sensi potrebbero essere accorpati, in quanto molto simili tra loro. La sparsità della disambiguazione porta a problemi nelle fasi successive;
- **Copertura:** alcuni percorsi della tassonomia di WordNet non sono completi come altri, ad es i neologismi potrebbero non essere presenti;
- **Soggettività:** varia a seconda della frase.

Uno strumento che svolge WSD è **WordNet**.

## Word Sense Induction

È una variante della WSD in cui non abbiamo un dizionario (tipo WordNet) che contiene tutti i sensi di una parola.

Il senso, inteso come la categoria a cui appartiene una parola, viene dedotto cercando all'interno di una grande quantità di testi. In pratica viene fatto clustering sull'utilizzo di una determinata parola all'interno dei testi e in questo modo vengono dedotti i sensi legati a quella parola. Tale valutazione risulta più complessa e si basa sul metodo della **pseudo-word**.

Tale metodo si basa sulla clusterizzazione e si divide in 3 fasi:

1. **Merging:** concatenazione di parole (random);
2. **Clustering:** applicazione del meccanismo WSI per l'identificazione di clusters (sensi identificati automaticamente), quindi vogliamo capire a cosa ci riferiamo (frutto o animale) quando compare "canebanana". Occorre trovare i sensi della parola esempio, il WSI deve trovarne 2, uno per cane e uno per banana;
3. **Cluster-to-class evaluation:** confronta i cluster ottenuti automaticamente con le parole originali delle pseudo-word.

Se i concetti vengono separati nettamente, il procedimento è andato a buon fine. Se il procedimento non va a buon fine troviamo dei concetti mescolati tra di loro. Il metodo delle pseudo-word permette di valutare l'efficacia di un sistema WSI.

### WSD vs WSI

<b>Inventory:</b> intendiamo un dizionario come WordNet, per un termine esisteranno x sensi con degli esempi di utilizzo	<b>No Inventory:</b> approccio non supervisionato
<b>Human-Based</b>	<b>Data-Based:</b> si usa un corpus da cui estrae il senso
<b>Grammar-Based</b>	<b>Usage_Based</b>
<b>Semplice ma criticabile:</b> posso calcolare un'accuratezza e controllare manualmente quali sono i sensi corretti	Più <b>complicato</b> perché una volta trovato il senso non si può confrontare con una inventory

## Semantica lessicale

È un sottocampo della semantica che studia cosa denotano le parole di una lingua. Le parole possono denotare cose nel mondo o concetti.

Ad oggi esistono molti **dizionari elettronici**, come WordNet e BabelNet.

Esistono anche delle risorse linguistico-cognitive dette **property norms**. Tali risorse descrivono delle proprietà che vengono in mente quando si descrive un concetto e derivano da questionari a cui veniva chiesto alle persone di descrivere proprietà/caratteristiche di alcune parole che trovavano scritte. Sono feature semantiche basate su **percezione** e **immediatezza**. Ad esempio, pensando a coccodrillo le prime parole che ci vengono in mente sono: pericoloso, grosso ecc.

Alcune caratteristiche sono:

- Mischiano attributi e valori;
- Contengono variabilità linguistiche → sparsità;
- Creazione costosa;
- Informazioni CSK ad alto valore semantico in termini di similarità percepita;
- Mancano info che non hanno a che vedere con la percezione. Se l'esperimento venisse effettuato su concetti astratti produrrebbe risultati poco utilizzabili.

**Table 1** A subset of the preprocessed features for the concept *turtle* (only a sample of uncollapsed features is shown)

PF	Relation	Feature	Participant list
23	has	a shell	p15 17 18 24 28 30 39 45 48 50 52 55 56 58 59 60 61 63 64 88 132 133 135
18	does	swim	p15 18 28 30 45 48 52 55 56 58 59 60 62 88 113 131 131 133
16	does	lay	p15 17 24 28 39 55 56 59 59 60 62 87 88 113 132 133
14	does	live	p18 24 30 45 45 52 52 52 55 59 60 62 64 133
10	is	a reptile	p18 45 50 56 58 60 64 113 132 134

Questo è un esempio di risultati di un questionario: si va da info più rilevanti a quelle meno rilevanti.

Abbiamo anche delle risorse di **common-sense** che rappresentano fatti che sono condivisibili tra determinate persone (basate appunto sul senso comune). Ad esempio, ci si aspetta che tutti sappiano che lo zucchero è dolce. Una risorsa utilizzabile è ConceptNet.


Infine, abbiamo i **visual attributes** che sono informazioni semantiche relative alla conoscenza visuale/percettiva. Una risorsa utilizzabile è ImageNet, ovvero il fratello di WordNet per le immagini.


Descrivere un concetto è difficile. Dare una definizione univoca è impossibile. Se il concetto è astratto è ancora più difficile.

Inoltre, per valutare determinate risorse, ci si può basare su alcune feature:

- Potere espressivo;
- Scalabilità computazionale;
- Sorgente;
- Ambiguità, soggettività.

## Cenni teoria pre-esercitazione 2

 **Onomasiologic search:** problema di passare da qualcosa che abbiamo in mente concettualmente a cui non sappiamo dare un nome o non ce lo ricordiamo;

 **Tip-of-the-tongue problem:** derivato dal punto precedente;

**Genus-differentia mechanism:** per descrivere un concetto le cose fondamentali sono 2:

- **Genus:** per descrivere un concetto dobbiamo inserirlo all'interno di una tassonomia, circoscrivere un raggio d'azione per quel concetto. Es: dico che il mango è un frutto;
- **Differentia:** tutto quello che caratterizza quel concetto in maniera differenziale da tutti gli altri concetti; la cosa che più lo caratterizza

**Circularity:** stesso problema definito da Mazzei: evitare di definire un termine con un termine che sto cercando di descrivere. Può essere indiretta, difficile da identificare.



## Costruzione del significato

Vediamo come si costruisce il significato utilizzando più termini insieme secondo 2 importanti teorie.

### Teoria di Pustejovsky

Pustejovsky ha creato una teoria sulla semantica che permette di descrivere il significato che possiamo attribuire alle frasi (composizioni lessico-sintattiche). Nella costruzione del significato, secondo lui, la semantica interviene a diversi livelli. Prevede un approccio strutturato su cui è sviluppata tale teoria:

- **Argument Structure:** si vede una parola come una funzione con arietà specificata. Ad esempio, il verbo può avere diversi argomenti. C'è una corrispondenza /interfaccia tra una parola e la struttura sintattica che la circonda. La struttura degli argomenti di una parola può essere vista come una piccola specificazione della sua semantica. Tale specificazione, presa da sola, è inadeguata a catturare tutto il significato di una parola, ma risulta necessaria per comprendere il significato. Tale definizione è basata a partire da quella di Grimshaw;
- **Event structure:** esprime il legame dell'evento con il significato. Gli eventi vengono distinti in 3 classi:
  1. **Stato:** Mary is sick oppure The door is closed. In questo caso non c'è alcun riferimento temporale sull'inizio o la fine di un evento. Indica uno stato in cui si trova Mary;
  2. **Processo:** Mary walked. Il verbo walk indica un'attività infinita. La frase denota un processo;
  3. **Transizione:** John closed the door. Indica una transizione da uno stato in cui la porta è aperta ad uno stato in cui è chiusa. Tale definizione è basata in parte su quella di Vendler.
- **Qualia structure:** gli attributi essenziali di un oggetto sono definiti tramite entità lessicali. Questa struttura è importante perché ogni qualia descrive quattro aspetti del significato di una parola, ognuno identificato come ruolo:
  - a. **Constitutive role:** relazione tra un oggetto e le parti che lo costituiscono in modo concreto. Quindi materiali, pesi, parti;
  - b. **Formal role:** caratteristiche peculiari che distinguono un oggetto all'interno di un vasto dominio. Quindi orientamento, dimensione, colore, forma, posizione. Es: un particolare Fragole → colore rosso;
  - c. **Telic role:** definiscono lo scopo e la funzione di un oggetto. Più legato ai verbi dell'oggetto;
  - d. **Agentive role:** legato a chi crea/genera/interagisce con l'oggetto.
- **Inheritance Structure:** indica come la parola è relazione con gli altri concetti. Costruendo una tassonomia possiamo capire meglio il significato di una parola.

Teoria difficilmente applicabile a causa dell'ambiguità, ma fornisce un'ottica di semantica profonda del significato.

## Teoria di Hanks

Più semplice da implementare rispetto a quella di Pustejovsky. Il verbo è la **radice del significato**. Non esistono espressioni senza verbo. Ad ogni verbo viene associata una **valenza** che indica il numero di argomenti necessari per il verbo. In base al numero di argomenti che un verbo richiede, in alcuni casi possiamo differenziarne il significato.

Esempio

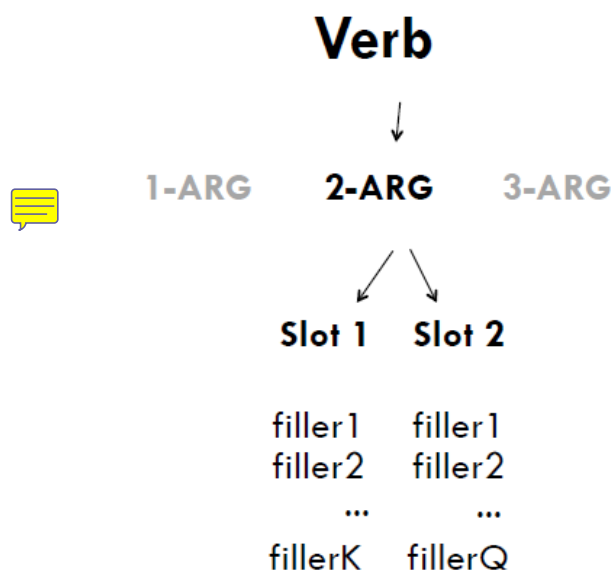
“He showered and dressed quickly” Si è fatto la doccia e si è vestito in fretta

“He showered her with kisses” L'ha inondata di baci

Shower può significare “lavare” o “inondare” e sono distinti dal numero di argomenti. Nel primo caso shower è intransitivo e ha valenza 1, in quanto non necessita di un oggetto né di una proposizione/complemento. Nel secondo caso shower è transitivo ed ha valenza 3 in quanto ha un oggetto e un complemento richiesti.

Una volta determinato il numero di argomenti di un verbo dobbiamo specificarli mediante un certo numero di **slot**. Ogni slot può avere un certo numero di valori che lo riempiono, detti **filler**. Ogni filler può avere associati dei **tipi semantici** che rappresentano delle generalizzazioni concettuali strutturate come una gerarchia (donna → umano → essere vivente...). Raggruppiamo i vari filler secondo alcuni types. Essi sono gruppi semantici generici o categorizzazioni/clusterizzazioni dei filler.

Es verbo mangiare, ottengo 3 semantic types ovvero vegetables, cereal, meat.



L'immagine rappresenta la struttura di Hanks con verbo, argomenti, slot e filler. La valenza in questo caso è 2.

Inoltre, secondo Hanks, **il significato è una specifica combinazione di semantic type** legati ai vari slot di una specifica valenza legata a un verbo. Hanks aggiunge piccole varianti/modifiche. Una è quella in cui alcune combinazioni possono essere unite perché sono varianti sintattiche con stesso significato.

Es: una frase attiva trasformata in passiva, il significato rimane lo stesso.

**Problemi:** dobbiamo definire con esattezza i tipi semantici (dipendono dal dominio e dalla loro applicazione) e quindi non sappiamo a quale livello gerarchico dobbiamo fermarci. Ad esempio, ci fermiamo al concetto di “persona” o introduciamo anche gli “animali” per poter raccogliere tutto sotto il concetto di “esseri viventi”? inoltre potremmo avere dei termini che si riferiscono ad un certo livello di generalizzazione a seconda del contesto.

Esempio: analizzando “lo studente va a scuola” dobbiamo capire quali proprietà attivare in relazione al tipo semantico. Per lo studente dobbiamo utilizzare un tipo semantico studente o persona? Essere vivente sarebbe troppo generico, visto che non permette di attivare abbastanza proprietà specifiche per lo studente.

Tali problemi possono essere risolti concentrandosi sulle **proprietà**. Ciò è realizzato con una terza teoria detta **affordance linguistica**. Affordance specifica la proprietà che un oggetto, quando viene descritto/si parla di esso, fornisce dei suggerimenti su come utilizzarlo, quindi fornisce indicazioni sulla possibilità d'uso; iniziamo ad attribuirgli delle affordances o inferire delle proprietà. Queste possibilità d'uso le abbiamo anche se l'oggetto non è mai stato visto.

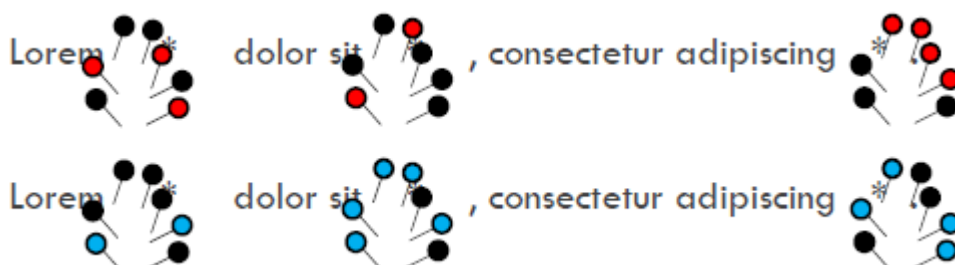
~~Il concetto di affordance può essere esteso anche alla linguistica. In presenza di parole non note possiamo ricavarne il significato dal contesto in cui compaiono. Da questo possiamo dire che il contesto crea delle proprietà semantiche, infatti il significato che attribuiamo ad una parola è dato dalle proprietà delle parole che associamo ad essa. Le proprietà possono essere di due tipi: **conosciute o assunte**.~~

~~Esistono inoltre dei pattern, i quali raccolgono istanze linguistiche o filler. All'interno di queste istanze potranno esserci delle “omissioni” di parole. Tali omissioni potranno essere coperte con alcune parole che hanno determinate caratteristiche per farlo (Img 1).~~

~~La frase iniziale sarebbe Lorem \* dolor sit \*, consectetur adipiscing \*.~~



~~Possiamo parlare di “Conceptual Instances” quando si avranno delle combinazioni di proprietà semantiche attribuibili ai lessemi/parole usate nel contesto (Img 2).~~



~~Il livello concettuale passa dal lessicale di hanks al semantico delle linguistic affordances.~~

In aggiunta, se abbiamo dei pattern completamente diversi, ma ci sono similarità (**properties overlap**) tra le proprietà delle parole, si può effettuare un **graded match** (potenziale similarità  $> 0$ ) fra i pattern anche se sono diversi. Questo è possibile grazie alla sovrapposizione delle proprietà e una combinazione simile cross pattern.

Con questo approccio il potere espressivo cresce notevolmente perché possiamo comprendere parole che non sono presenti nel corpus iniziale andando a studiarne le proprietà.

Le proprietà possono essere ottenute (op. non semplice):

- Considerando proprietà latenti derivanti dalla similarità;
- Estraeendole da risorse linguistiche come WordNet o BabelNet;
- Tramite questionari o studi cognitivi.

I dati sono presi da risorse linguistiche come Corpora (wikipedia), annotazioni manuali, ML oppure Hackatons.

In ogni caso non abbiamo bisogno di grandi quantità di dati perché non esiste un corpus "completo" e il numero di proprietà utilizzate cresce secondo il logaritmo dei termini introdotti. Tale numero dopo un po' tende a convergere.

Es: anche se aumentano tanto i termini, le proprietà aumentano lentamente. Se abbiamo visto 100 frutti avremmo trovato la maggior parte delle proprietà dei frutti. Aggiungerne altri 100 porterà ad aggiungere qualche proprietà, ma la maggior parte saranno già state trovate.

## Tipi di semantica

Saranno analizzati i seguenti argomenti:

- Text-mining;
- Semantica documentale e la sua visualizzazione;
- Semantica distribuzionale;
- Ontologie.

### **Text Mining**

È il processo per ricavare informazioni di alta qualità dal testo. I problemi di text mining vennero proposti per risolvere il problema del **question answering**. L'idea è che un utente umano possa fare delle domande a cui il calcolatore deve rispondere in base alle informazioni che trova in un testo. Le domande riguardano fatti, definizioni ecc.

L'idea è di associare ad ogni domanda diverse informazioni come:

Il punto è che per ogni domanda, è necessario una specifica analisi e delle diverse informazioni semantiche

- Parole
- Punteggiatura
- Lemmi
- Sinonimi

Facciamo una distinzione:

**Linguistica Computazionale (NLP):** studio di formalismi descrittivi del funzionamento del linguaggio naturale, che permettano di essere trasformati in programmi eseguibili dai computer → approccio top-down.

**Statistica:** disciplina che studia qualitativamente e quantitativamente particolari fenomeni → approccio bottom-up, data-driven, si parte dai dati.

Entrambi sono approcci complementari alla semantica.

Dal punto di vista statistico le parole sono dei token (sequenze di caratteri contigui) e un testo è un insieme di token che hanno una certa frequenza. Non siamo interessati ai caratteri e alla sintassi di una parola. In base a questa assunzione un documento può essere visto come un vettore di coppie (parole, #occorrenze) detto **Vector Space Model**.

Si rappresenta il testo mediante un vettore, ad ogni parola nel dizionario viene associato un indice crescente. Il testo lo posso vedere come il numero di occorrenze (**tf, term frequency**) di quelle parole nel testo. Questa rimane comunque un'approssimazione.

Se effettuiamo questo procedimento separatamente per tanti documenti otteniamo una matrice in cui ogni riga rappresenta un documento e ogni colonna un possibile vocabolo. La matrice risulta sparsa. Sulle righe avremo rappresentati i testi, sulle colonne i termini. Ogni cella darà info su quanto un termine occorre in un testo.

Questa rappresentazione permette di rappresentare i testi come vettori numerici. In questo modo possiamo misurare la similarità tra i documenti ed il loro topic (ottenibile analizzando poche parole) tramite la **cosine-similarity**:

$$sim = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Vedendo la similarità tra i testi si effettua un primo step per il ragionamento.

Paper: analisi numerica di testi → 80% significato semantico deriva dall'uso delle parole e il 20% deriva da come le parole vengono utilizzate. È una stima basata sull'inglese.

La similarità del coseno permette di memorizzare la similitudine tra due vettori calcolando il coseno tra di loro. Nel caso del confronto di testi, i due vettori contengono la frequenza dei termini, ossia il numero di occorrenze di una parola in un testo. Il k-esimo elemento di ogni vettore conterrà il numero di occorrenze della k-esima parola.

Per definizione di coseno, dati due vettori, si otterrà sempre un vettore di similitudine compreso tra -1 e +1. -1 indica una corrispondenza esatta, ma opposta (un vettore contiene l'opposto dei valori presenti nell'altro); +1 indica due vettori uguali.

Nel caso dell'analisi dei testi poiché le frequenze dei termini sono sempre valori positivi si otterranno valori che vanno da 0 a 1 dove 1 indica che le parole contenute nei due testi sono le stesse, ma non necessariamente nello stesso ordine. 0 indica che non c'è alcuna parola che appare in entrambi.

### Altri metodi statistici

I metodi statistici applicati ai documenti di testo si basano su due tipi di informazione:

- **Frequenza** di una parola in un testo;
- **Co-occorrenza** di due parole in un testo.

### Frequenza

Indica l'importanza, la rilevanza, la "dominance", la significatività di un termine nel testo.

Ci si basa su 2 informazioni statistiche: la **Term Frequency (TF)** e la **Inverse Document Frequency (IDF)**.

Term frequency: indica quante volte un termine compare in un documento. È definita come segue:

$$tf_{i,j} = \begin{cases} 1 + \log_2(f_{i,j}) & \text{con } f_{i,j} > 0 \\ 0 & \text{altrimenti} \end{cases}$$

$tf_{ij}$  ha una crescita logaritmica perché se un termine compare mille o duemila volte non fa tanta differenza. Indica l'importanza di un termine nel testo.



Inverse document frequency: valuta quanto un termine è specifico per un documento. Va a valutare in quanti documenti un certo termine compare. È definita come segue:

$$idf_i = \log_2\left(\frac{N}{n_i}\right)$$

N è il numero di documenti e  $n_i$  è il numero di documenti in cui il termine  $i$  appare. Questa quantità permette di escludere tutte le parole che compaiono molte volte in un documento, ma non sono specifiche del documento, come ad esempio gli articoli hanno  $tf$  alto,  $idf$  quasi 0.

Inoltre, ci dice che ci sono dei termini per cui vale la pena dare uno score molto alto per certi termini in quanto sono caratterizzanti (si trovano solo in specifici documenti) per la semantica del testo in cui si trovano.

Per valutare quanto sia interessante una parola viene utilizzato il prodotto delle due frequenze:

$$tf - idf = tf_{i,j} \cdot idf_i$$

### Co-occorrenza

La co-occorrenza di due parole in un testo indica la similarità tra due parole assumendo che due parole con significato simile siano presenti negli stessi contesti. Il **contesto** si riferisce alla vicinanza fisica delle parole nei testi e viene spesso definito come l'insieme di parole che due termini hanno nel loro "intorno".

Un altro approccio per determinare quanto due parole siano correlate in un testo consiste nel costruire una matrice di co-occorrenza che mostra per ogni parola quanto spesso un'altra parola le compare vicino. Ciò ci permette di implementare la nozione di contesto in cui una parola appare.

### Applicazioni del Text Mining

Il text mining è applicato in diversi contesti:

**Tag cloud:** le parole figurano con una dimensione proporzionale alla loro frequenza. Possiamo includere un'informazione di co-occorrenza andando a rappresentare le parole ad una distanza proporzionale alla loro co-occorrenza. È un modo per comunicare il topic.

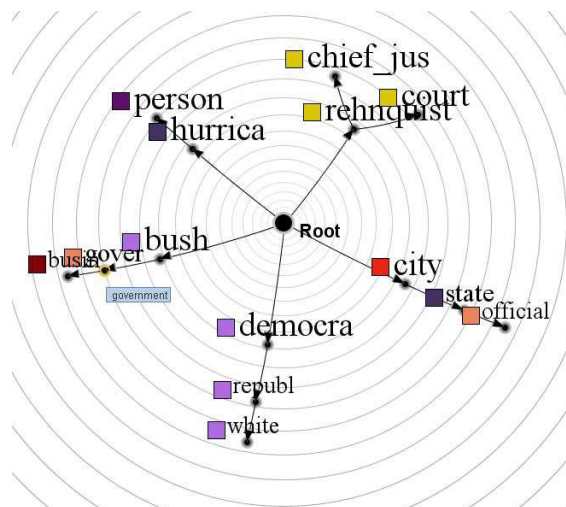
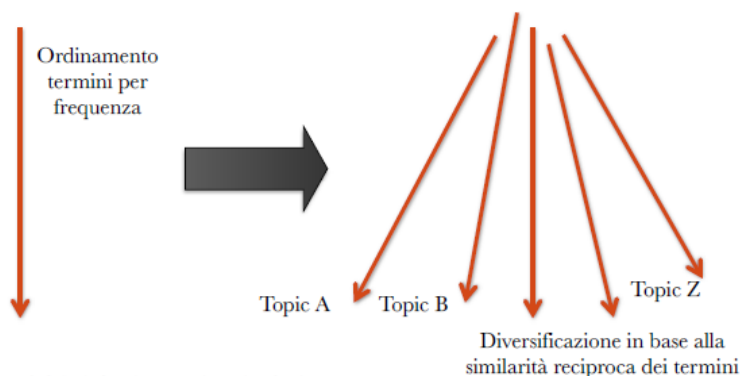


**Tag flakes:** è un particolare tipo di tag cloud in cui viene fatto un ordinamento in base alla frequenza e successivamente si separano le dimensioni su una base semantica. Ciò permette di:

- Organizzare in le tag cloud in base al significato;
- Estrarre in maniera automatica una gerarchia di termini;



- Estrarre in maniera automatica una gerarchia di termini in cui sia rappresentata la reciproca similarità. Per fare ciò serve un buon algoritmo di costruzione della gerarchia. In questo modo si individuano diversi topic che hanno al loro interno parole semanticamente correlate.



**Document clustering:** utilizzando tecniche di apprendimento non supervisionato possiamo raggruppare i documenti che hanno qualcosa in comune. Non esiste un clustering perfetto, in quanto stiamo ottimizzando una funzione. Tutto dipende da cosa vogliamo evidenziare e dalla funzione che ottimizziamo. In alcuni casi le tecniche di clustering vengono utilizzate in cascata con altre (pipeline), ad esempio potremmo effettuare il clustering e successivamente creare una tag cloud per ogni cluster. Le tecniche di clustering si basano sul concetto di distanza tra due elementi.

Ci sono metodi e misure differenti che si possono utilizzare. Ad esempio, calciatori di una squadra oppure calciatori di una certa nazionalità.

**Document categorization/classification:** l'obiettivo è associare un documento ad una certa etichetta di una tassonomia. Possiamo farlo tramite tecniche di apprendimento non supervisionato.



Una prima tecnica consiste nel prendere il vettore che rappresenta il documento e verificare quale fra tutte le etichette della tassonomia compare più volte nel documento. La tecnica è troppo semplice e fornisce risultati approssimativi.

La seconda tecnica consiste nell'andare a costruire una matrice  $n \times n$  dove  $n$  è il numero di etichette della tassonomia. Inizialmente la matrice coincide con la matrice identità  $n \times n$  (**Inizializzazione**). Gli indici dei vettori indicano i concetti, come se fosse un dizionario di co-occorrenze, 1 indica che quel vettore contiene il concetto associato, con 0 indico gli altri vettori che invece non lo contengono.

Successivamente, ogni elemento della tassonomia cede una parte di informazione ai nodi con lui comunicanti (**Propagazione**). Prendiamo i livelli della matrice a 2 a 2 in maniera contigua. Il valore di un vettore viene propagato, cioè andrò a trasferire parte del valore ad un altro vettore perché in qualche modo sono correlati. Es. il vettore 5 trasferirà parte del suo valore in quello che rappresenta America oppure World. il nuovo valore sarà normalizzato affinché la somma dei valori nel vettore dia 1.



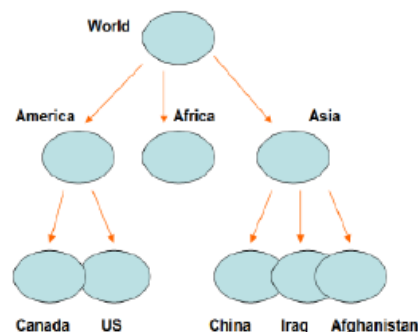
Per capire qual è il nuovo valore da inserire esiste un **fattore di propagazione** (stabilito a priori presente su ogni arco della tassonomia) che deve essere moltiplicato per l'1 del vettore 5 ad esempio.

Questa operazione avviene per un certo numero di iterazioni, ciò avviene perché così facendo si riesce a propagare nella tassonomia info dalla radice fino alle foglie e viceversa. La propagazione è bidirezionale.

Se itero troppo rischio di introdurre rumore e ottenere risultati sbagliati.

Per associare un documento alla tassonomia si riduce il vettore del documento alla stessa dimensione del numero di etichette. In questo procedimento di riduzione si mantengono le componenti del vettore documento che coincidono con le etichette della tassonomia. Infine si calcola la similarità del vettore documento ridotto con ogni vettore della tassonomia e si associa al documento l'etichetta il cui vettore presenta una maggiore similarità.

Esempio: Tassonomia geografica



Concept vectors

	world	Asia	Africa	America	Afghanistan	Iraq	China	Canada	US
$\vec{c}_w$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\vec{c}_a$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\vec{c}_f$	0.100	0.012	0.873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\vec{c}_{am}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\vec{c}_{af}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\vec{c}_{ir}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\vec{c}_{ch}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\vec{c}_{ca}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\vec{c}_{us}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806

**Document segmentation:** lo scopo è di individuare diversi elementi del discorso all'interno del documento in modo da analizzare l'evoluzione di un tema o delle relazioni. Il metodo più famoso è quello del **text tiling**. L'idea è quella di separare un test in finestre di lunghezza fissata a priori. All'interno di ogni finestra viene calcolata una **coesione intra-gruppo**, che misura quanto sono legate le parole all'interno di una finestra. Una misura potrebbe essere la co-occorrenza, ma ne esistono altre. A questo punto si cercano i **break point**, ossia quei punti in cui la coesione intra-gruppo ha un crollo. È molto probabile che in tali punti finisca una sezione del documento relativa ad un argomento e ne cominci una nuova. Questo ci permette di abbandonare la lunghezza fissa delle finestre e andare ad individuare delle finestre in base ai break-point.



**Riassunto automatico di documenti:** è effettuato in due modi:



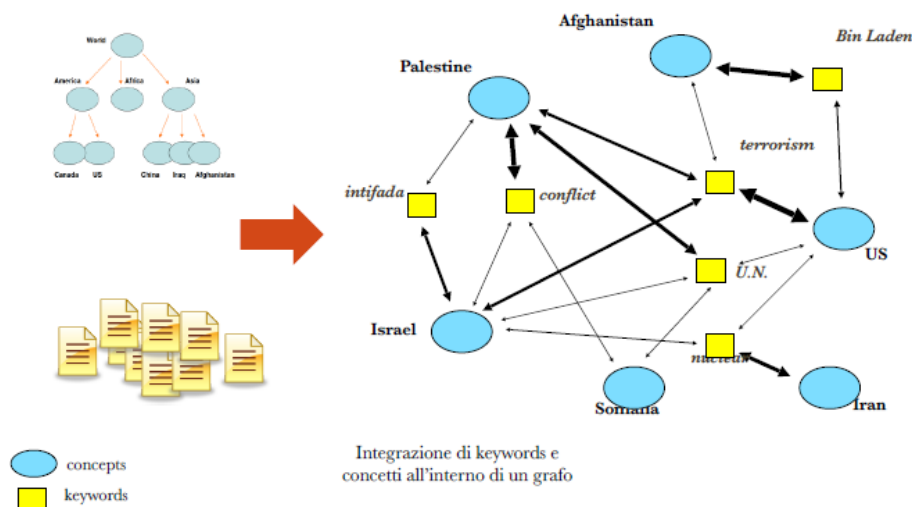
▪ **Estrattivo:** dato il testo si cerca di attribuire un valore di importanza alle frasi attraverso metodi di keyphrases, TextRank ecc. Ciò ci permette di evidenziare le frasi maggiormente significative e utilizzarle nel riassunto;



**Astrattivo:** non selezionano le frasi, ma vanno a generare nuovi documenti. Questa tecnica richiede l'utilizzo di tecniche di NLG. È più complessa. Per effettuare una valutazione del riassunto si utilizzano delle metriche dette ROGUE (Recall-Oriented Understudy for Gisting Evaluation).

**Orienteering browsing:** il problema è che non sempre le persone sanno come cercare le informazioni. Lo scopo è quello di sviluppare dei meccanismi basati sul concetto di orienteering per aiutare gli utenti a navigare tra i topic sfruttando le relazioni tra i documenti di testo. Ad esempio, evidenziando relazioni latenti tra documenti (Wikipedia), abilitando la navigazione guidata dal contesto o supportando la comprensione dei dati.

**Information retrieval:** è lo scopo del text mining. Inizialmente si cercava un match tra le parole chiave e i documenti. Nel tempo si è evoluto andando a considerare anche link, snippet ed altre informazioni. Nell'information retrieval troviamo l'idea di rete semantica. La rete semantica viene si forma quando esplicitiamo i legami fra le tre componenti fondamentali: query, documenti e concettualizzazione. Infatti, mediante una query recuperiamo dei documenti. Tali documenti sono relativi a dei concetti nella concettualizzazione. Ciò crea un collegamento implicito tra query e concettualizzazione.



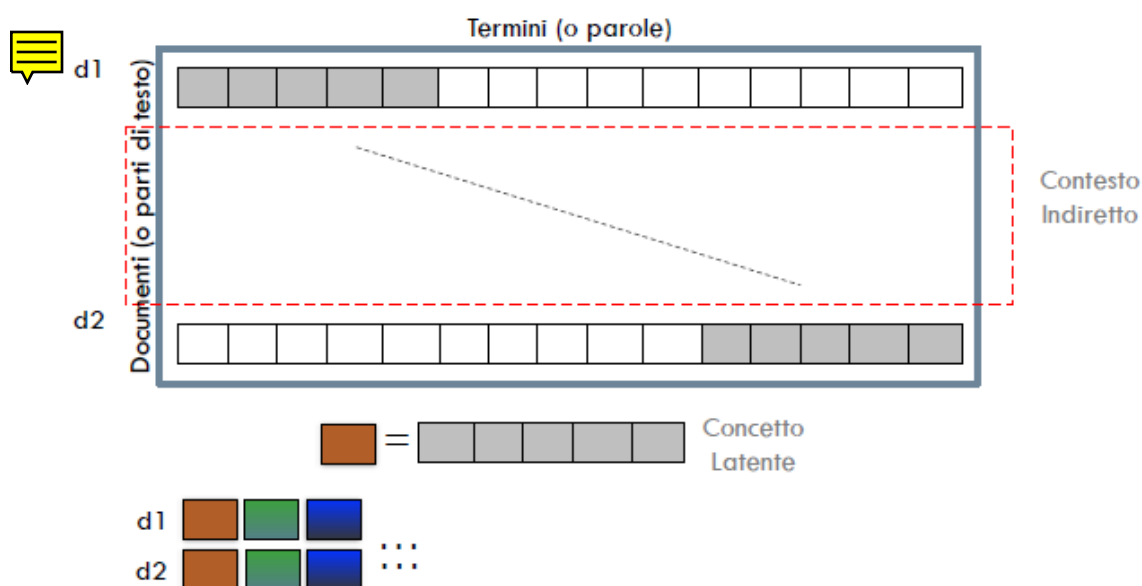
## Semantica documentale e visualizzazione

In quest'ambito, su grandi quantità di testi/documenti, possono essere utilizzati diversi modelli:

**Topic modeling:** è un modello statistico e probabilistico che analizza l'uso del linguaggio ed individua automaticamente gli argomenti in una collezione di testi (anche molto vasta). Il modello è non supervisionato perché non necessita di dati annotati. Un topic è una lista pesata di parole (in base all'importanza delle parole stesse) estratta dal documento. Definire l'etichetta è compito del lettore che interpreta la lista di parole e ne definisce l'etichetta. **Problemi:** non è sempre facile interpretare l'output del documento, una lista di parole non è chiara come un discorso o documento. I topic estratti non sempre sono utili. A volte il topic modeling richiede l'intervento umano, può anche essere semi-supervisionata. In un articolo di giornale ci sono più argomenti, a differenza di articoli scientifici che riguardano un unico topic. Inoltre, i topic sono qualcosa di più strutturato dal punto di vista della granularità.

**Latent semantic analysis (LSA):** è basato sulla decomposizione SVD (Singular Value Decomposition) che dà luogo a vettori meno sparsi. La SVD permette di approssimare una matrice numerica di partenza  $n \times m$  con una moltiplicazione tra 3 matrici a dimensionalità ridotta, la centrale è quadrata e diagonale detta dei **singular values**; essa permette di riorganizzare documenti e termini (**Term Document Matrix**, sparsa, molti valori = 0), raccoglie la ridondanza dei dati andando ad accorpare più dimensioni (**Concetti latenti**) insieme sfruttando la co-occorrenza. I valori vicini a 1 rappresentano parole molto simili, i valori vicini a 0 rappresentano parole diverse. Se facessi una semplice similarità tra d1 e d2 avrei come risultato 0 in quanto c'è un valore 0 che mi annulla il calcolo. Tuttavia, bisogna considerare il **contesto indiretto**. In d1 uso determinate parole, diverse da quelle usate in d2, ma sono semanticamente simili perché co-occorrono all'interno del contesto indiretto, esse condividono uno spazio semantico comune. d1 e d2 non sono dissimili, parlano di cose simili ma usando termini diversi.

L'operazione viene effettuata sull'intero corpus, il risultato è una serie di concetti latenti che sono ordinati per **importanza/dominance**.




I documenti vengono ridefiniti in un nuovo spazio vettoriale di  $n$  concetti latenti che riassumono  $n$  clusterizzazioni/combinazioni lineari dei termini usati nel corpus originale.

Quindi dovendo calcolare la cos similarity nel nuovo spazio vettoriale avrò sicuramente un valore  $> 0$  perché entrambi  $d1$  e  $d2$  avranno un alto valore di associazione per un concetto latente (Es. il colore marrone).

La similarità quindi non si basa più sulla semplice corrispondenza lessicale, ma va anche oltre andando a considerare il concetto latente.


**Problemi:** Il modello (non generativo) non generalizza bene su documenti non ancora visti. Inoltre, dopo la fattorizzazione potremmo avere dei valori negativi difficili da interpretare.

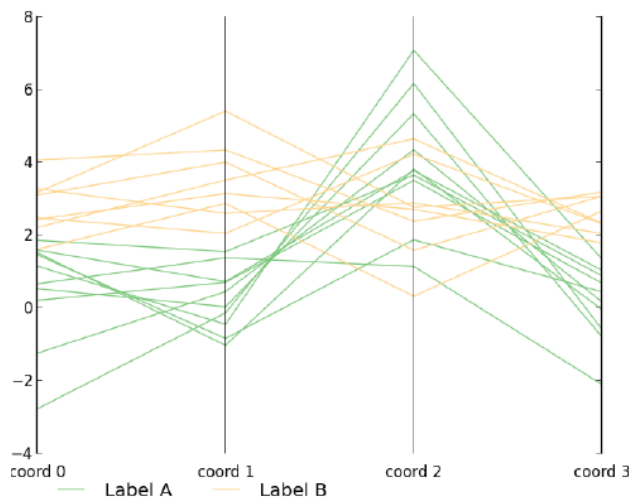
Per risolvere tali problemi sono state proposte diverse varianti:

- **Non-negative Matrix Factorization (NNMF);**
- **Latent Dirichlet Allocation (LDA):** versione probabilistica della LSA (pLSA) che sfrutta la statistica bayesiana. Si basa sull'assunzione che un documento è un mix di topics e ogni parola ha una certa probabilità di comparsa in ogni topic. Ogni documento è un misto di un piccolo numero di argomenti. La presenza di ciascuna parola è attribuibile a uno degli argomenti (non è il risultato del topic modeling, ma è un'etichetta inserita da una persona dopo aver letto il testo) del documento.
- **Dynamic Topic Modeling (DTM)** si concentra sull'evoluzione dei topic nel tempo a patto di avere un corpus annotato con valori temporali per i topic. 

## Visualizzazione

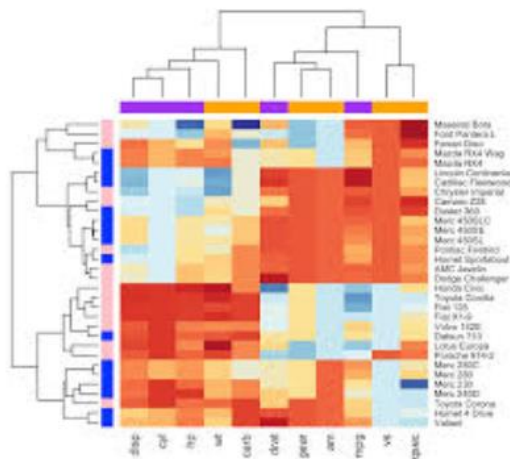
Un problema comune al text mining e alla semantica documentale è la visualizzazione. Il problema è rappresentare uno spazio  $n$ -dimensionale (quello dei termini) in uno spazio bidimensionale. Si utilizzano delle tecniche di **text visualization**:

 **Parallel coordinates:** per ognuna delle coordinate dei termini abbiamo una barra che ci indica le occorrenze di quella coordinata in un certo documento. Una coordinata corrisponde ad un termine (coord0, coord1, ...) La barra corrisponde ad un documento (labelA, labelB). La tabella sarà formata da una riga per ogni documento. Le colonne rappresentano i termini trovati nel documento.

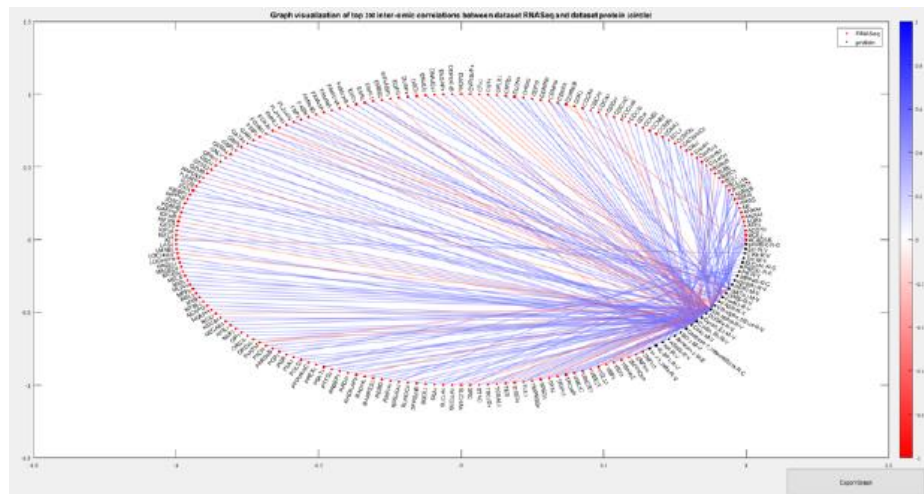


**RadViz (Radar visualization):** ogni termine rappresenta un punto di attrazione nella circonferenza (i punti sulla circonferenza sono i topic). I vari documenti si dispongono a una distanza dai topic proporzionale al numero di occorrenze di quel termine. Utile per mappare dati espressi con 5/10 dimensioni diverse. **Problema:** possiamo avere dei punti vicini che in realtà sono dissimili. Problema dovuto a coincidenze sfortunate di combinazioni di valori. Ad esempio, 2 valori per dei topic (9, 4) molto forti che portano il punto ad accentrarsi. Al centro potrebbero trovarsi dei punti con valori per il topic 1 e 8. Visualizzazione utile fino ad un certo punto. Colori associati a delle classi (Sport, economia, politica ...).

**HeatMap:** i valori contenuti nella matrice sono rappresentati da colori.



**Correlation Circle:** i termini sono disposti lungo la circonferenza e collegati fra loro mediante delle rette che ne indicano la correlazione. Utile per le matrici di co-occurrenza.



## **Semantica distribuzionale**

È una rivisitazione in chiave linguistica delle tecniche di text-mining. Di seguito un po' di storia della semantica distribuzionale:

**Harris:** nel 1954 per primo afferma che parole le quali occorrono negli stessi contesti tendono ad avere significati simili



**Firth:** similmente, nel 1957, afferma che una parola è caratterizzata da quelle che la accompagnano

**Furnas:** nel 1983 sostiene che la congiunzione di varie parole ci permette di specificare più precisamente l'oggetto del discorso

**Deerwester:** nel 1990 introduce i concetti latenti con cui afferma che esista una struttura sottostante che viene parzialmente oscurata dalla scelta "casuale" delle parole che viene fatta per comporre un discorso

**Bley:** nel 2003 introduce una visione probabilistica secondo cui il tema del documento influenza in modo probabilistico le parole utilizzate nel documento

**Turney:** sempre nel 2003 impiega per la prima volta delle coppie di parole. L'intuizione è che se una certa coppia di parole (x; y) presenta gli stessi pattern della parola (w; z) allora possiamo costruire una sorta di "proporzione" nel significato delle due coppie

La semantica distribuzionale è circondata da altre aree e assume nomi diversi a seconda del contesto in cui viene utilizzata:

- Nell'ambito dell'information retrieval viene detta vector space model;
- Nell'ambito della cognitive science viene nominata come studio dei conceptual spaces;
- Nella teoria dei grafi viene studiata mediante le matrici di adiacenza.

### **Matrici**

Fino ad ora abbiamo utilizzato solo matrici. Ciò è dovuto al fatto che le matrici sono una **approssimazione** e ciò semplifica i task, anche se porta risultati meno precisi; questo è dovuto anche al fatto che il linguaggio ha un'elevata complessità e la matrice non può contenere tutto il patrimonio informativo riguardo al linguaggio. Inoltre, metodi esatti non catturerebbero sfumature dell'uso del linguaggio derivanti dalla soggettività oppure ironia, sarcasmo... mentre la DS funziona anche in questi casi, giustificandone il suo utilizzo.

Con le matrici infatti perdiamo l'ordinamento delle parole, avendo solo le frequenze.

Esistono 2 correnti di pensiero/rappresentazioni tra cui le matrici si collocano:

- **Rappresentazione puramente simbolica:** semplice da utilizzare, ma povera di informazione;
- **Rappresentazione associazionistica/connessionistica:** proviene da studi di psicologia in cui si afferma che la rappresentazione del significato nella nostra mente è funzione di tutto il nostro vissuto, quindi dovremmo sempre tenerne conto. Specifica che il significato è una cosa non tangibile, non rappresentabile se non attraverso la connessione di una miriade di elementi la stessa miriade di elementi.

Questa rappresentazione è carica di informazione, ma difficile da implementare. Gli oggetti sono rappresentati mediante delle **quality** (features). Essa inoltre, è in linea con la **prototype theory**: quando associamo degli elementi ad un concetto con determinate caratteristiche, possiamo comunque far riferimento all'elemento in generale che sarà il **prototype** (es. gatto) a cui associamo però un'**immagine mentale** condivisa.

Es. ci sono vari tipi di gatto, ma bene o male l'immagine condivisa del gatto è sempre la stessa.

Le matrici facilitano la condivisione della conoscenza. L'idea infatti è di andare a porre sulle righe le entità basilari e sulle colonne le qualità. Possiamo applicare delle tecniche di tassellazione di Voronoi per dividere lo spazio rappresentato dalle matrici in regioni. In questo modo il significato viene rappresentato tramite una regione di spazio vettoriale costruito in base alle entità e alle qualità.

Visto che utilizziamo le matrici possiamo utilizzare delle tecniche note per questo tipo di dato: misure di similarità, strategie di pesatura - in modo da capire quali pesi mettere all'interno delle celle della matrice – trasformazioni matriciali, tecniche di clustering.

Per ottenere risultati migliori si applicano delle operazioni di pre-processing:

- **Normalizzazione:** procedimento lessicale, sintattico o morfologico che consiste di:
  - Tokenizzazione, stemming e lemmatizzazione.
- **Denormalizzazione:** si aumenta il carico dimensionale ma su aspetti semantici e non aspetti morfologici/lessicali come nella normalizzazione. È un procedimento semantico attuato tramite alcune tecniche:
  - **Named entities** con cui cerchiamo di estrarre nomi dal testo;
  - **Semantic roles** con cui andiamo a determinare i ruoli all'interno della frase. I ruoli solitamente sono **agent**, **patient** e **means**.

Turney nel 2010 fornisce una classificazione delle matrici nell'ambito della semantica distribuzionale. Secondo lui ne esistono di tre tipi:

**Term document matrix:** su ogni riga abbiamo un documento e su ogni colonna un termine. Su queste matrici possiamo eseguire le operazioni di: similarità, clustering, classificazione, segmentazione e parzialmente question answering.


**Term content matrix:** generalizzano le term document matrix. Su ogni riga c'è il contesto e su ogni colonna un termine. Il **contesto** non deve essere necessariamente un documento, ma può essere una frase, un paragrafo o una dipendenza sintattica. Le operazioni supportate sono: similarità, costruzione di cluster, classificazione di parole, generazione automatica di un thesaurus, disambiguazione, etichettatura semantica dei ruoli.





**Pair-pattern matrix:** proposte da Turney. Su ogni riga abbiamo coppie di parole e su ogni colonna un pattern. I pattern possono essere relazioni. Ad esempio, X causa Y; X è risolto da Y...

Non si utilizzano triple di parole per due motivi:

- Costo computazionale eccessivo
- Inutile perché otterremmo matrici troppo sparse 


Su questa matrice possiamo misurare similarità relazionali e similarità di pattern, con queste ultime effettuiamo un clustering sui pattern.


L'utilizzo di questa matrice permette di cogliere una maggiore profondità semantica perché permette di rispondere ad una **domanda** del tipo: individua tutte le X tali che X causa il cancro (effettua una sorta di proporzione tra significati). Non cerco una risposta/documento in una collezione, ma cerco una relazione, quindi si possono effettuare domande come quella specificata. Posso cercare quello che voglio a condizione che sia la causa (o altre relazioni) di qualcos'altro. Si arriva a un livello semantico più profondo.

## Similarità

La semantica distribuzionale viene detta anche semantica della similarità. Quine negli anni 50 afferma che la similarità è molto importante perché ci permette di mettere ordine mediante categorizzazione e ci fornisce delle percezioni. Infatti, gli eventi futuri non saranno uguali a quelli passati, ma in molti casi cercheremo delle similarità con gli eventi passati per cercare di capire come comportarci con le nuove situazioni.

## Definizioni di similarità

**Semantic similarity:** concettualmente si ha quando si parla di sinonimi o quasi-sinonimi, parole che in certi contesti hanno lo stesso significato o quasi lo stesso significato. Spesso malinterpretata e usata. 

**Semantic relatedness:** riguarda dei concetti che condividono delle proprietà. Il significato di questa quantità è generico, spesso inutilizzabile. Automobile e motore (meronimia) sono legati da una semantic relatedness. 

**Attributional similarity:** simile della semantic relatedness, raramente usata.

**Taxonomical similarity:** riguarda concetti che condividono degli iperonimi (concetti più generali). Risulta più misurabile

**Relational similarity:** studia la similarità tra coppie di parole. Ad esempio, dog:bark – cat:meow

**Semantic association:** analizza le parole che co-occorrono più frequentemente. Simile alla relatedness, ma orientata alla corpus analysis. Ad esempio – culla, bambino.

## Problemi

Le rappresentazioni matriciali non tengono conto dell'ordine delle parole nel testo. La quantità di informazione persa da un algoritmo di IR in assenza di un ordine delle parole è stimata attorno al 20%.



Per mitigare il problema abbiamo diversi modi:

- Pair-pattern matrix sensibili all'ordine delle parole;
- Si utilizzano di ordinamento che forniscono informazioni aggiuntive sull'ordine.

Spesso le matrici sono orientate al significato di singole parole. Ciò porta ad utilizzare delle combinazioni lineari dei vettori (+, \*) per cercare di lavorare sulla semantica compositiva.

Tecniche di **arricchimento matriciale** aggiungono informazioni alla matrice mediante delle risorse esterne (semantic knowledge) esistenti.

Un'ultima direzione in cui la ricerca si è orientata è quella della **filaments of meaning in wordpsace** (significato filamentare). Secondo questa idea, la semantica delle parole non è contenuta in spazi vettoriali di dimensione elevata, ma in piccole porzioni di tale spazio. Il significato di una parola dipende quindi da una rappresentazione ristretta.

## Ontology learning

L'apprendimento di ontologie consiste nell'estrazione semi-automatica di concetti e relazioni rilevanti a partire da una collezione di documenti o altri insiemi di dati al fine di creare un'ontologia.

Per risolvere il problema dell'apprendimento di un'ontologia solitamente si utilizzano strumenti automatici. Cimiano afferma che l'apprendimento di ontologie assomiglia al **Reverse Engineering**. Data una conoscenza di un certo dominio con la sua rappresentazione e la sua scrittura, vogliamo tornare alla concettualizzazione di partenza. C'è una similarità con il triangolo semiotico.

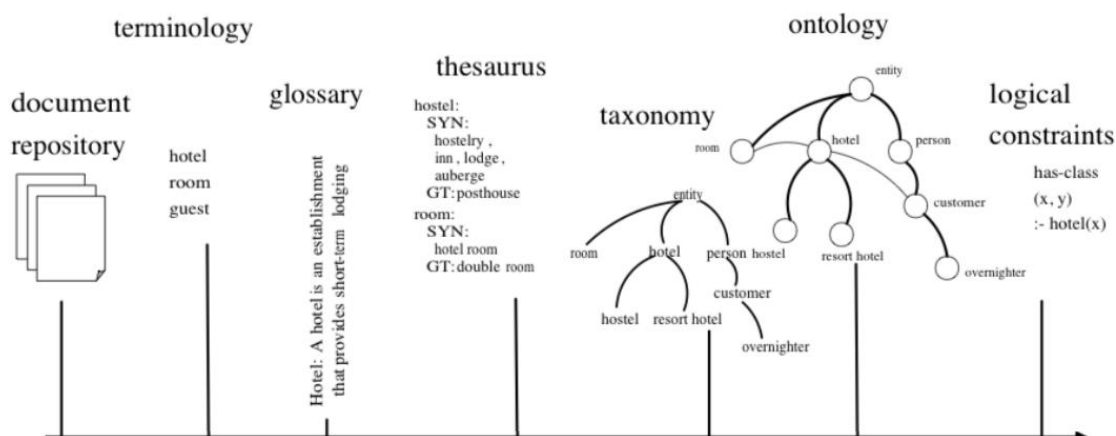
Tuttavia, occorre anche elencare i problemi relativi al RE:

- World Knowledge non codificata, quindi non è uguale per tutti, infatti potremmo avere visioni del mondo molto diverse;
- Domain Knowledge non usata completamente, parzialità e incompletezza dei dati di partenza su cui fare RE. Si utilizza una conoscenza comune.

L'apprendimento di ontologie si differenzia da altre tecniche:

- **Ontology population:** ontologia esistente, associare istanze a concetti e relazioni all'interno dell'ontologia;
- **Ontology annotation:** simile al precedente, si parte da un'ontologia già esistente e da una base documentale per taggare il testo con delle informazioni concettuali;
- **Ontology enrichment:** data un'ontologia e una base documentale si cercano istanze degli elementi dell'ontologia per ottenere più informazioni sull'ontologia stessa. Se ad esempio un termine si presenta con una frequenza elevata si va a ristrutturare l'ontologia sia a livello di concetti (aggiungendo nodi) che a livello di relazioni (aggiungendo archi). Ristrutturando l'ontologia possiamo trovare dei concetti non visti in precedenza.

## Tipi/Livelli di formalizzazione




Visto che le ontologie riguardano la formalizzazione della conoscenza, potremmo chiederci: ma quanto sono davvero formali? Nell'immagine vengono mostrate una serie di rappresentazioni dei dati in cui il grado di formalizzazione cresce da sinistra verso destra.

La minor formalizzazione possibile è quella del testo non strutturato (livello documentale), a cui segue quella della terminologia in cui abbiamo una serie di termini espressi per dominio di interesse e il glossario (termine → definizione), a cui segue la rappresentazione a thesaurus che fornisce relazioni tra le parole. Ancora più formale è la tassonomia che fornisce una gerarchizzazione dei concetti. A questo punto abbiamo proprio l'ontologia che fornisce una gerarchia ben strutturata che include diversi tipi di relazioni, regole e assiomi. Sono presenti anche delle logical constraints, cioè regole che permettono di fare inferenze, reasoning (livello più complesso).

## Task dell'Ontology Learning

Ce ne sono diversi:

- **Term Extraction:** trovare nomi per concetti e relazioni;
- **Synonym Extraction:** si va a trovare un thesaurus;
- **Concept Extraction:** *intensionale*, estrarre definizioni o aspetti definizionali legati ai concetti, *estensionale*, creare le istanze possibili di un determinato concetto;
- **Relation Extraction:** trovare relazioni tra concetti esistenti; 
- **Population:** si usano 2 tecniche:
  - *name entity recognition*: lega una parola che rappresenta un luogo ad esempio Stati Uniti si lega al concetto nominato nell'ontologia di US si riconosce che è un concetto geografico e lo si lega a un preciso concetto nell'ontologia;
  - *Information extraction*;
- **Concept Hierarchies Induction:** si parte dai concetti preesistenti e si vuole indurre il rapporto tassonomico tra di essi.

Ci concentriamo sull'induzione gerarchica dei concetti, che può essere realizzata tramite 3 paradigmi:

- **Pattern lessico sintattici:** definiamo dei pattern che abbiamo trovato nel testo che indicano una relazione semantica. I pattern sono appresi in maniera automatica partendo da un training set con migliaia di frasi, ognuna delle quali esprime una sola relazione semantica;
- **Distributional hypothesis:** è un approccio di semantica distribuzionale in cui se prendiamo due coppie di parole per cui gli elementi di ogni coppia sono a distanza simile tra di loro, allora è possibile che le due coppie siano in relazione allo stesso modo;
- **Nozione di sussunzione:** è un insieme di tecniche che permettono di costruire delle gerarchie.

Per portare a termine questi tre task abbiamo tre metodologie: NLP, FCA, ML.

## NLP (Natural Language Processing)

Basato sul ruolo dei **PoS** in quanto indicativi per la semantica. Ne esistono diversi (Nome → Sostantivo/Aggettivo, Articolo, Pronome, Verbo, Avverbio).

Successivamente si eseguono operazioni di **preprocessing** (Tokenizzazione, Stemming/lemmatizzazione, PoS tagging, ossia l'analisi grammaticale. In questo modo riconosciamo la categoria lessicale di ogni parola nel contesto in cui è usata. Named entity

recognition). Questo passaggio permette la normalizzazione del testo, aiuta ontology learning.

Dopo il preprocessing abbiamo una fase di **analisi sintattica** con un sistema a regole basato su alberi di parsing. Uso regole sulle dipendenze sintattiche: se ho un determinato verbo, prendo il complemento oggetto del verbo e lo associo a un concetto nell'ontologia regole basate su alberi a dipendenze.

Successivamente si effettua un'**analisi semantica** tramite: vector space model, Distribuzioni di probabilità, tecniche di rilevanza dei termini come tf, tf-idf

Infine, possiamo utilizzare **risorse linguistiche** esterne a supporto come WordNet, BabelNet e FrameNet.


## FCA (Formal Concept Analysis)

Sono delle matrici, abbiamo oggetti generici sulle righe descritti da attributi generici sulle colonne. La matrice di incidenza che lega oggetti e attributi è il concetto di formal content.

Abbiamo tre elementi fondamentali:

- Gli oggetti che equivalgono ai concetti;
- Gli attributi associati agli oggetti come le features venivano associate ai concetti;
- L'incidenza che rappresenta il fatto che un certo oggetto possieda o meno un attributo.

La relazione tra oggetti e attributi è espressa tramite la matrice di context analysis. In ogni riga abbiamo un oggetto e in ogni colonna un attributo.

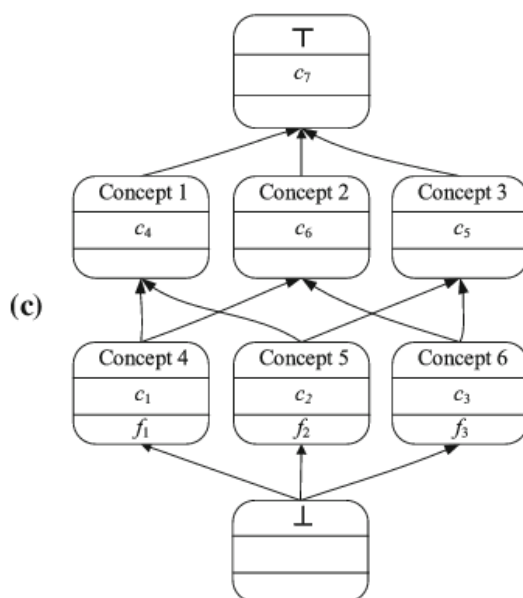


	$c_1$	$c_2$	$c_3$	$c_4$	$c_5$	$c_6$	$c_7$
$f_1$	x			x		x	x
$f_2$		x		x	x		x
$f_3$			x		x	x	x

(a)

T	$\{f_1, f_2, f_3\}, \{c_7\}$
Concept 1	$\{f_1, f_2\}, \{c_4, c_7\}$
Concept 2	$\{f_1, f_3\}, \{c_6, c_7\}$
Concept 3	$\{f_2, f_3\}, \{c_5, c_7\}$
Concept 4	$\{f_1\}, \{c_1, c_4, c_6, c_7\}$
Concept 5	$\{f_2\}, \{c_2, c_4, c_5, c_7\}$
Concept 6	$\{f_3\}, \{c_3, c_5, c_6, c_7\}$
$\perp$	$\{\emptyset\}, \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$

(b)



(a) data un feature, posso vedere l'enumerazione dei concetti che hanno quella feature e viceversa. es di formal context, i concetti possono esseri termini, le features possono essere il contesto/documenti/patern sintattici in cui si trovano oppure i concetti latenti dopo applicazione di LSA.

(b) Stato successivo di a.

- (c) detto lattice, da questo si può costruire una tassonomia. Ad esempio: data una formal context, esistono dei concetti che hanno più proprietà di altri, sono più generici. Meno proprietà sono più specifici. Attraverso quindi un rapporto di inclusione/esclusione/sovrapposizione delle feature possiamo capire quali sono i concetti più generali e quelli più specifici.

## ML (Machine Learning)

Possiamo utilizzare approcci **supervisionati** come classificatori bayesiani, alberi di decisioni, SVM o reti neurali, approcci **non supervisionati** come il clustering, che può essere divisivo o agglomerativo.

- **Divisivo**: è un approccio top down in cui tutti gli elementi si trovano inizialmente in un singolo cluster che viene suddiviso ricorsivamente in sotto-cluster;
- **Agglomerativo**: è un approccio bottom-up in cui si parte dall'inserimento di ciascun elemento in un cluster differente e successivamente si accorpano i cluster a coppie.

## Open information extraction

Rappresenta un insieme di tecniche per estrarre in maniera efficiente grandi quantità di informazioni da corpora di grandi dimensioni. Lo scopo è di estrarre triple della forma



(arg1, verbal phrase, arg2) e creare un sistema che estragga queste triple su cui è possibile effettuare un'analisi più strutturata piuttosto che una semplice analisi sul testo. Analisi un po' più avanzate rispetto a quelle su testo libero.

Queste triple sono delle proposizioni che descrivono dei fatti, ad esempio "Dante wrote the Divine Comedy" → (Dante, wrote, Divine Comedy).

La rappresentazione è molto semplice, quindi potremmo avere dei dati rumorosi, tuttavia è il necessario compromesso per avere alta efficienza per risolvere task come question answering su larga scala.

Un **vantaggio** di questo approccio è che diventa facile apporre dei vincoli logici per richiedere dati più specifici. Ad esempio, richiedere che la data di nascita di una persona sia successiva ad un certo anno.

Il procedimento segue due fasi:

- Una fase di PoS tagging (e/o dependency parsing);
- Una fase relativa all'estrazione degli elementi delle triple che può essere eseguita in base a WSD, dipendenze sintattiche e altro ancora;
- Una fase di filtering: In base alle triple estratte, si cerca di rimuovere triplette rare, oppure effettuare clustering e revisione delle estrazioni.

Ci sono diversi **problemi** nell'ambito dell'OIE:

- Non esiste un approccio rigoroso e unico, ciò potrebbe portare a tralasciare alcune parti;
- Esistono molti metodi di estrazioni diversi da loro difficili da comparare. Non esiste un gold standard a cui affidarsi;
- È stato dimostrato che le triple non funzionano bene in contesti reali.

## NLP e social media

Ci concentriamo sui tweet. I dati sono testi di 140 (ora 280) caratteri spesso incompleti e dipendenti dal contesto, spesso non corretti dal punto di vista grammaticale, con abbreviazioni, link, emoticon, hashtag e altri elementi. L'essere umano è in grado di cogliere il significato, ma un parser sintattico no.

I task da portare a termine sono diversi:

- Bisogna individuare delle strutture dedicate che permettano di cogliere al meglio il significato dei tweets;
- Bisogna capire cosa fare con gli elementi che sono notizie. Tale ambito è detto **news analytics** e misura diversi attributi qualitativi e quantitativi di un testo, come ad esempio sentimento, rilevanza, novità ecc.
- Abbiamo i task di opinion **mining** e **sentiment analysis**. L'output di un processo di opinion mining è un insieme di opinioni più o meno oggettive. Nel caso della sentiment analysis gli output vengono classificati come positivi/negativi
- Abbiamo inoltre l'operazione di **scraping** che va a recuperare le informazioni da twitter (o altri siti) per effettuare delle analisi. Lo scraping in pratica è la collezione di dati da social media e altri siti web.

I principali ambiti di ricerca ad oggi riguardano lo scraping, la pulizia dei dati, la protezione dei dati, il processamento dell'informazione e la visualizzazione dei dati:

- **Scraping**: determinare il valore commerciale esociale dei dati;
- **Pulizia dei dati**: come pulire i dati twitter;
- **Processamento dell'informazione**: come trattare dati multilingue, errori, il fatto che ogni social network possieda un proprio bacino di utenti diverso per età, grado medio di istruzione ecc. (FB vs IG vs LIN).

## Sensori sociali

L'analisi approfondita dei dati provenienti dai social media è importante perché posso rappresentano dei **sensori sociali**. Ciò può essere fatto andando a comprendere cosa sia un **termine emergente**. In base al termine possiamo costruire un **topic**, cioè un insieme di parole minimale che esplicita e disambigua l'argomento della discussione.

**Termine emergente**: un argomento di discussione è definito emergente in un dato intervallo di tempo se è estensivamente trattato in quel dato intervallo e molto poco nei precedenti.

La differenza tra termini emergenti e termini caldi è la seguente:

**Termini emergenti**: sono quelli che in un breve periodo di tempo presentano uno scostamento forte dal numero di volte che vengono utilizzati mediamente. Ad esempio, un piccolo di utilizzi della parola "terremoto" indica che probabilmente si è verificato un terremoto. Se un termine presenta un picco e quel picco è periodico, il termine non è emergente. Ad esempio, la parola "morning" presenta un picco nelle prime ore del giorno, tuttavia questo picco si presenta ogni giorno, quindi la parola non è emergente. Se consideriamo il concetto di storia ogni keyword può essere emergente in funzione del numero di intervalli di tempo che consideriamo.

**Termini caldi:** pure essendo utilizzati tanto, lo sono sempre, quindi non sono così interessanti. Ad esempio, la parola “computer”, pur essendo utilizzata molto, non indica un evento particolare, proprio perché utilizzata spesso.

Ad ogni parola di un tweet  $tw_j$  si associano una serie di pesi:

$$\vec{tw}_j = \{w_{j,1}, w_{j,2}, \dots, w_{j,v}\}$$

I pesi sono calcolati come:

$$w_{j,x} = 0.5 + 0.5 \cdot \frac{tf_{j,x}}{tf_j^{max}}$$

È possibile definire un grafo di utenti grazie alle connessioni tra di loro esistenti (followers, following). Il grado di influenza è detto page rank.

Nutrimento

Un altro aspetto da considerare in twitter è l'impatto delle informazioni sulla base di chi le pubblica. È normale che i tweet di Obama avranno un impatto maggiore dei tweet di un utente comune. Per tale motivo di considera associa un valore di autorità ali autori sulla base del numero di connessioni che hanno con gli altri.

Per comprendere l'importanza di una parola possiamo pensare che ogni parola sia un organismo vivente: con nutrimento abbondante – inteso come tweet che riportano la parola – il suo ciclo di vita è prolungato, altrimenti la parola muove in quanto il nutrimento diventa insufficiente.

Il nutrimento è calcolato considerando il concetto di autorità:

$$nutr_k^t = \sum_{tw_j \in TW_k^t} w_{k,j} \cdot auth(user(tw_j))$$

Una volta individuati i termini emergenti possiamo andare a determinare i **topic correlati** andando ad analizzare l'intervallo in cui i termini sono stati riportati sulla rete, e analizziamo la relazione semantica esistente tra tutti i termini emergenti. Quindi ci interessiamo alla **co-occorrenza fra i termini** e calcoliamo anche la frequenza dei singoli termini.

Un **topic** può essere visto come un insieme minimali di termini correlati, ad esempio Obama, Insurance, Health.

Un'applicazione di quanto detto finora sta nell'estrazione automatica del **contenuto affettivo dei documenti testuali**. Al momento la maggior parte dei sistemi riesce a catturare soltanto una differenza a grana grossa, come sentimenti positivi/negativi. È difficile determinare il tipo di emozione associato ad un tweet (tristezza, rabbia, ecc.).

Considerando un tweet positivo/negativo non abbiamo molta informazione perché un testo può contenere diversi indicatori affettivi e può avere più oggetti sotto valutazione. Si utilizzano tecniche di **aspect based analysis** in cui cerchiamo di catturare delle relazioni precise tra ciò che viene valutato e la connotazione emotiva associata. Ciò viene fatto

utilizzando degli oggetti da valutare che vengono inseriti all'interno di una ontologia fornita e popolata tramite ontology population.

Per effettuare queste operazioni possiamo utilizzare diversi strumenti:

- **Opinion mining Markup Language:** è un formalismo XML per l'annotazione "sentiment" sui testi. Ci permette di specificare l'ontologia e l'attribuzione dei sentimenti. Mediante dei tag possiamo distinguere maggiormente tra elementi soggettivi (positivi/negativi), oggettivi (osservazione), domande, suggerimenti.
- **SentiVis:** è uno strumento che mediante un'analisi della struttura sintattica (parsing a dipendenze) propaga i valori di sentiment all'interno della struttura sintattica. Se ad esempio un nodo interno dell'albero presenta un certo valore (0.7) che indica un sentimento positivo, tale valore può essere propagato alle foglie comportando una differente visualizzazione dei risultati.