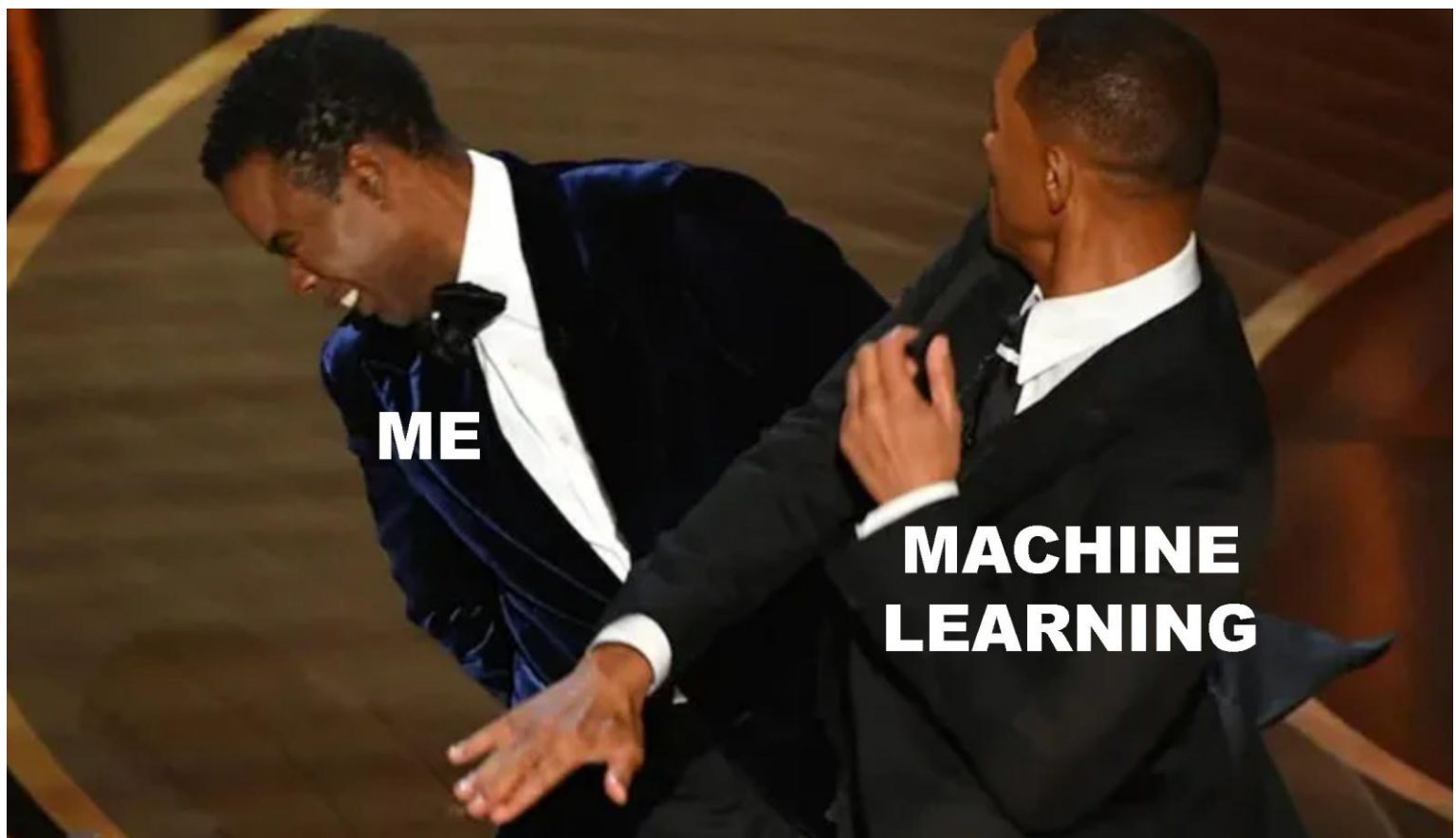
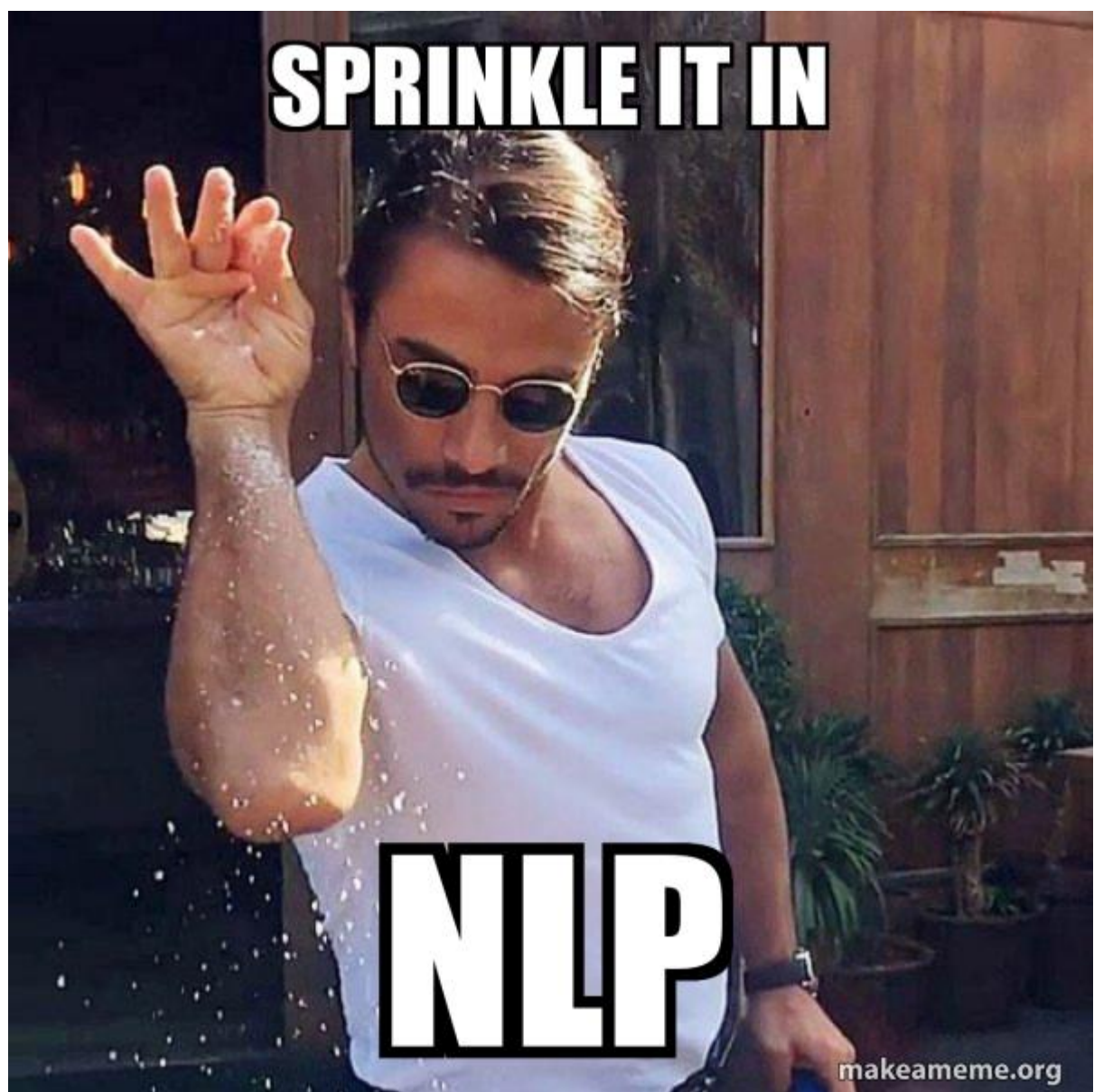


Introduction to NLP

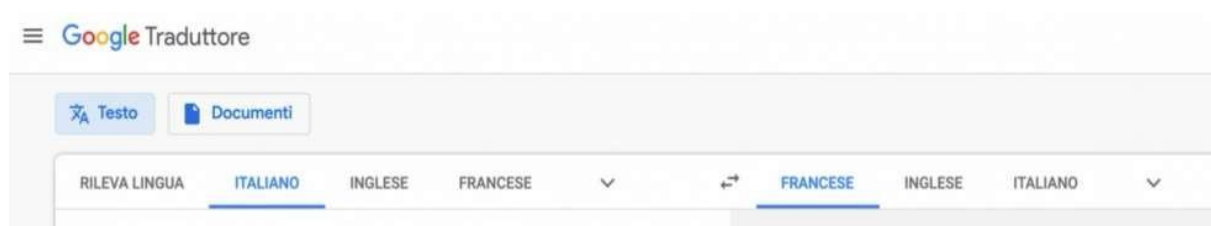
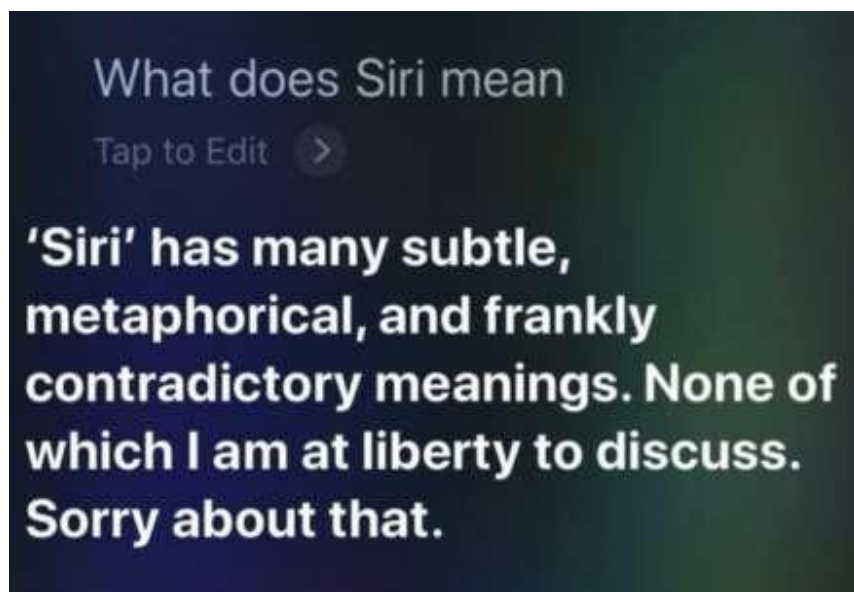




Il Natural Language Processing

Il Natural Language Processing è la branca dell'informatica che si occupa di insegnare ai computer ad analizzare, comprendere ed (eventualmente) utilizzare il linguaggio naturale





Brevissima Storia del Natural Language Processing

FASE 1: Linguistica Computazionale

FASE 2: Machine Learning

FASE 3: Deep Learning e Big Data

Linguistica Computazionale

E' l'approccio al Natural Language Processing più tradizionale, che si basa sull'utilizzo di **regole linguistiche**.

Machine Learning

E' il campo dell'**Intelligenza Artificiale** che vuole dare ai computer la capacità di apprendere in modo autonomo dai dati.

Deep Learning

E' il settore del Machine Learning che sfruttando modelli ispirati al funzionamento delle **reti neurali** permette di apprendere relazioni non lineari su grandi quantità di dati

1) **REGULAR EXPRESSION**

Cosa sono le espressioni regolari ?

Un'espressione regolare (regular expression - regex) è un'insieme di caratteri che definisce un pattern di ricerca.

ES. Ricerca degli indirizzi email

co93@gmail.com

rt@nigmatica.it

PATTERN

Una serie di caratteri minuscoli e/o numeri,
seguiti da una chiocciola,
seguita da una serie di caratteri minuscoli,
più un punto seguito da 2 o 3 caratteri.

ES. Ricerca degli indirizzi email

PATTERN

Una serie di caratteri minuscoli e/o numeri, seguiti da una chiocciola, seguita da una serie di caratteri minuscoli, più un punto seguito da 2 o 3 caratteri.



REGEX

`[a-z0-9]+@[a-z]+\.[a-z]{2,3}`

2) Tokenizzazione

Cosa è la Tokenizzazione ?

Nel Natural Language Processing la **tokenizzazione** è il processo che ci permette di dividere la frase nei suoi componenti, chiamati **Tokens**

Testo: Oggi è una splendida giornata per studiare

Tokens

Oggi	è	una	splendida	giornata	per	studiare
------	---	-----	-----------	----------	-----	----------

Testo: Cos'è questa fretta? Facciamolo un'altra volta

Tokens

Cos è questa fretta ? Facciamolo un'altra volta

Una strategia molto utilizzata consiste nel dividere per ogni carattere non alfanumerico...

Testo: Io ho studiato al Politecnico di Torino

Tokens

Io ho studiato al Politecnico di Torino

bisogna anche riconoscere quali parole costituiscono una singola entità

Serve un'intero set di regole

I tokenizzatori sono specifici per un linguaggio



per ogni lingua è necessario un set di regole

Cosa sono le Stop Words ?

Nel Natural Language Processing le **Stop Words** sono quella parole di uso comune che non portano nessuna informazione utile al testo

Esempi tipici di Stop Words

CONGIUNZIONI

(e, anche, pure, quindi, dunque o, ma, altrimenti....)

AVVERBI

(ora, prima, sempre, no, non, forse, dove, quando, come, perché...)

PREPOSIZIONI

(di, a, da, in, con, su, per, tra, fra ...)

PRONOMI

(io, me, tu, lui, lei, noi, voi, loro ...)



Rimozione delle Stop Words

Dato che le stop words sono parole superflue,
è buona norma rimuoverle dal testo.

Oggi è una splendida giornata per studiare

Oggi è una splendida giornata per studiare

 = STOP WORD

Cosa è lo Stemming ?

Nel Natural Language Processing lo **Stemming** è un processo che permette di ridurre le parole alla loro forma base, chiamata appunto **stem** (radice). Una parola è composta da una **radice** e da una **desinenza**,

DESINENZA
Studiare
RADICE

La **radice** è fissa, mentre la **desinenza** può variare

DESINENZA
Studio
RADICE

Giocare a bowling con i **miei** amici mi **diverte** molto

E' un processo semplice che consiste nel troncare la parte finale della parola in base ad un set di regole

Gioc	a	bowling	con	i	mi	amic	mi	divert	molto
------	---	---------	-----	---	----	------	----	--------	-------

Perché usare lo Stemming ?

Lo stemming ci permette di ridurre, anche di molto, il numero di parole uniche nel nostro testo.

4) La Lemmatizzazione

Cosa è la Lemmatizzazione ?

Nel Natural Language Processing la **Lemmatizzazione** è un processo che permette di ridurre le parole dalla loro forma flessa alla loro forma canonica, detta appunto **lemma**



I verbi vengono ridotti all'infinito e le parole alla prima persona maschile singolare

Stemming VS Lemmatizzazione

Lo scopo dello stemming e della lemmatizzazione è lo stesso, quest'ultima è una tecnica più sofisticata che porta a migliori risultati, quindi andrebbe favorita.



- How does spaCy compare to NLTK?

SPACY

- Over 400 times faster
- State-of-the-art accuracy
- Tokenizer maintains alignment
- Powerful, concise API
- Integrated word vectors
- English only (at present)

NLTK

- Slow
- Low accuracy
- Tokens do not align to original string
- Models return lists of strings
- No word vector support
- Multiple languages

Cosa è il Part of Speech Tagging (POS) ?

Nel Natural Language Processing il **Part of Speech Tagging (POS)** è il processo di identificazione della **parte del discorso (part of speech)** di ogni parola di un testo.

Esempi di POS Tag

Tipo	Tag	Esempi
Aggettivi	ADJ	nuovo, buono, alto
Preposizione	ADP	su, di, a, con, da
Avverbi	ADV	presto, adesso, davvero
Congiunzioni	CONJ	e, o, ma, mentre, se
Articoli determinativi	DET	il, la, tutti, alcuni
Nomi	NOUN	casa, cane, uomo
Numeri	NUM	10, cento, 13:30
Particella	PRT	ti, ci, vi
Pronomi	PRON	Lei, esso, io, egli
Verbi	VERB	ha, era, potrebbe
Punteggiatura	.	., ! ? : ;
Altro	X	lol, xke, cmq

io	non	sono	bello	ma	sono	intelligente
PRON	ADV	VERB	ADJ	CONJ	VERB	ADJ

Perchè il POS Tagging ?

Il POS Tagging è un'operazione di basso livello utilizzata per eseguire operazione di più alto livello.

Lemmatizzazione

Cosa è la Named Entity Recognition (NER) ?

Nel Natural Language Processing la **Named Entity Recognition (NER)** è il processo di identificazione della classe di appartenenza di una parola all'interno di un documento di testo.

Esempi di classi	Esempi di classi	Esempi di classi
<div>PERSONE</div> <ul style="list-style-type: none"> • Elon Musk • Stephen Hawking 	<div>ORGANIZZAZIONI</div> <ul style="list-style-type: none"> • Google • Amazon 	<div>QUANTITA'</div> <ul style="list-style-type: none"> • 100 • quindici • un milione

Esempio di NER

PERSONA
Jeff Bezos, il fondatore di ORGANIZZAZIONE Amazon, ha acquistato
il ORGANIZZAZIONE Washington Post per DENARO 250 milioni di dollari nell'DATA Ottobre 2013.

Cosa è la Sentiment Analysis ?

Nel Natural Language Processing la **Sentiment Analysis**
è il processo di identificazione dell'emozione, positiva o negativa,
espressa in un contenuto testuale

Esempio di Sentiment Analysis

Mi è davvero **piaciuto** questo corso, il contenuto è **ben** organizzato, in particolare mi è **piaciuto** Tutto **fantastico**, davvero **consigliato**.

PIACIUTO = 0.6

FANTASTICO = 0.95

BEN = 0.4

CONSIGLIATO = 0.75

$2 * \text{PIACIUTO} + \text{BEN} + \text{FANTASTICO} + \text{CONSIGLIATO} = \text{SENTIMENT}$

$$2 * 0.6 + 0.4 + 0.95 + 0.75 = 3.2$$

Il corso è inutile, poco dettagliato e gli esempi sono banali.
Sconsigliato.

Il corso è **nutile**, **poco** dettagliato e gli esempi sono **banali**.
Sconsigliato

INUTILE = - 0.9

BANALI = - 0.6

POCO = - 0.3

SCONSIGLIATO = - 0.75

$$-0.9 - 0.3 - 0.6 - 0.75 = -2.55$$

INUTILE = 0.9

BANALI = 0.6

POCO = 0.3

SCONSIGLIATO = 0.75

Il corso mi è **piaciuto**, ma alcune parti sono **confuse**.

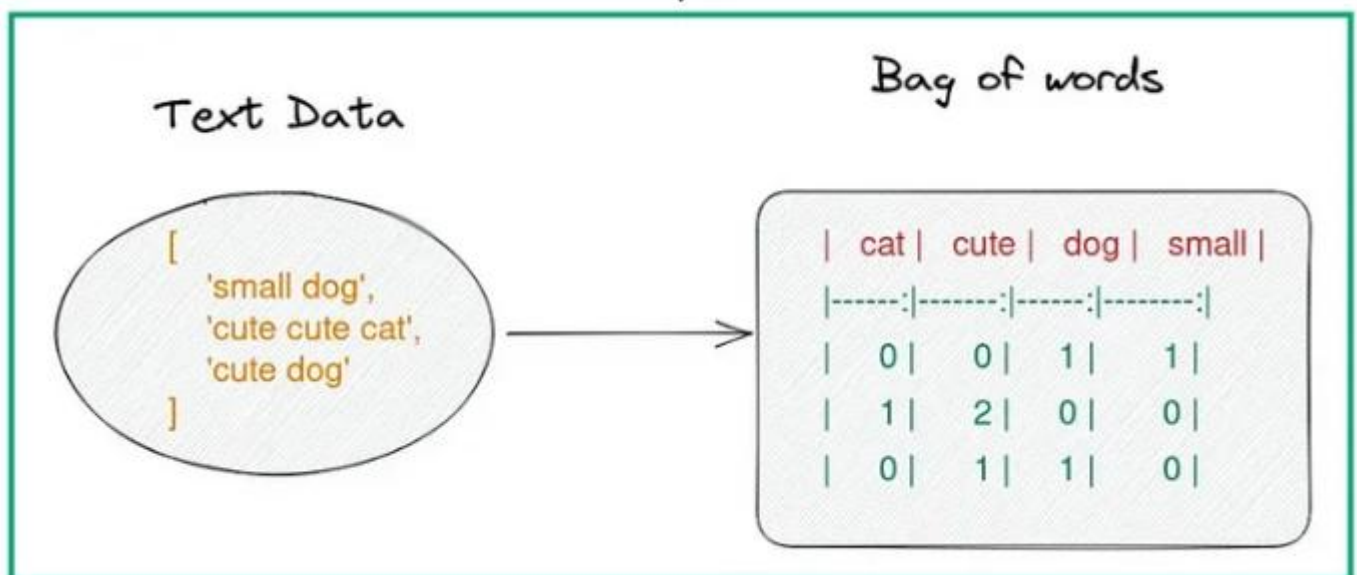
PIACIUTO = 0.6

CONFUSE = - - 0.4

Sfruttando un dizionario annotato insieme a regole lessicali
(**VADER** - Valene Aware Dictionary for Sentiment Reasoning)

Cosa è il Bag of Words ?

Il modello Bag of Words è un modello semplicistico per la rappresentazione dei testi che ci permette di trattarli come documenti di lunghezza comune all'interno dei quali la disposizione delle parole non ha importanza



Corpus di testo

- Oggi è una bellissima giornata
- La mia ragazza è bellissima
- Oggi è giornata di paga

Esempio di Bag of Words

Corpus di testo

- Oggi è una bellissima giornata
- La mia ragazza è bellissima
- Oggi è giornata di paga



VOCABOLARIO

- oggi
- è
- una
- bellissima
- giornata
- la
- mia
- ragazza
- di
- paga

Esempio di Bag of Words

VOCABOLARIO

- oggi
- bellissima
- giornata
- ragazza
- paga



RIMOZIONE DELLE STOP WORDS

VOCABOLARIO

- oggi
- è
- una
- bellissima
- giornata
- la
- mia
- ragazza
- di
- paga

Frase 1

Oggi è una bellissima giornata



Rappresentazione Bag of Words

oggi	bellissima	giornata	ragazza	paga
1	1	1	0	0

Frase 2

La mia ragazza è bellissima



Rappresentazione Bag of Words

oggi	bellissima	giornata	ragazza	paga
0	1	0	1	0

Frase 3

Oggi è giornata di paga



Rappresentazione Bag of Words

oggi	bellissima	giornata	ragazza	paga
1	0	1	0	1

Rappresentazione Bag of Words del corpus di testo

oggi	bellissima	giornata	ragazza	paga
1	1	1	0	0
0	1	0	1	0
1	0	1	0	1

Cosa è il TF*IDF ?

Il modello TF*IDF (Term Frequency * Inverse Document Frequency) è un modello per la rappresentazione delle parole nel Natural Language Processing che penalizza le parole comuni nel corpus di testo e dà più peso a quelle più rare

TERM FREQUENCY

Misura la frequenza di ogni termine in un documento

INVERSE DOCUMENT FREQUENCY

Misura l'importanza di ogni termine all'interno dell'intero corpus

DOCUMENT FREQUENCY

Quanti documenti contengono una determinata parola ?

oggi	bellissima	giornata	ragazza	paga
2	2	2	1	1

Esempio di TF*IDF

DOCUMENT FREQUENCIES

oggi	bellissima	giornata	ragazza	paga
2	2	2	1	1



$$IDF = \ln \left(\frac{\text{Numero di documenti}}{\text{Document Frequency}} \right)$$



INVERSE DOCUMENT FREQUENCIES

oggi	bellissima	giornata	ragazza	paga
0.4	0.4	0.4	1	1

TERM FREQUENCY della frase 1

Oggi è una bellissima giornata

oggi	bellissima	giornata	ragazza	paga
1/3	1/3	1/3	0/3	0/3

TERM FREQUENCY della frase 2

La mia ragazza è bellissima

oggi	bellissima	giornata	ragazza	paga
0/2	1/2	0/2	1/2	0/2

TERM FREQUENCY della frase 3

Oggi è giornata di paga

oggi	bellissima	giornata	ragazza	paga
1/3	0/3	1/3	0/3	1/3

Esempio di TF*IDF

TERM FREQUENCY

oggi	bellissima	giornata	ragazza	paga
0.3	0.3	0.3	0	0

oggi	bellissima	giornata	ragazza	paga
0	0.5	0	0.5	0

oggi	bellissima	giornata	ragazza	paga
0.3	0	0.3	0	0.3

INVERSE DOCUMENT FREQUENCY

oggi	bellissima	giornata	ragazza	paga
0.4	0.4	0.4	1	1

Frase 1

Frase 2

Frase 3

Il TF*IDF è il prodotto tra term frequency e inverse document frequency

TF*IDF

oggi	bellissima	giornata	ragazza	paga
0.12	0.12	0.12	0	0

Frase 1

oggi	bellissima	giornata	ragazza	paga
0	0.2	0	0.5	0

Frase 2

oggi	bellissima	giornata	ragazza	paga
0.12	0	0.12	0	0.3

Frase 3

Rappresentazione TF*IDF del corpus di testo

oggi	bellissima	giornata	ragazza	paga
0.12	0.12	0.12	0	0
0	0.2	0	0.5	0
0.12	0	0.12	0	0.3

I limiti del Bag of Words

CORPUS DI TESTO

- Un film magnifico.
- Un capolavoro da vedere con la famiglia.
- Pessimo, 90 minuti sprecati.

CORPUS DI TESTO (3 frasi)

- Un film magnifico.
- Un capolavoro da vedere con la famiglia.
- Pessimo, 90 minuti sprecati.

DIZIONARIO (13 parole)

un
film
magnifico
capolavoro
da
vedere
con
la
famiglia
pessimo
novanta
minuti
sprecati

DIZIONARIO (13 parole)

un
film
magnifico
capolavoro
da
vedere
con
la
famiglia
pessimo
novanta
minuti
sprecati

RIMOZIONE DELLE
STOP WORDS



DIZIONARIO (6)

magnifico
capolavoro
vedere
famiglia
pessimo
sprecati

- Un film magnifico.
- Un capolavoro da vedere con la famiglia.
- Pessimo, 90 minuti sprecati.

magnifico	capolavoro	vedere	famiglia	pessimo	sprecati
1	0	0	0	0	0
0	1	1	1	0	0
0	0	0	0	1	1



ONE HOT ENCODING

- | | |
|--|----------------------|
| • Un film magnifico. | • [1, 0, 0, 0, 0, 0] |
| • Un capolavoro da vedere con la famiglia. | • [0, 1, 1, 1, 0, 0] |
| • Pessimo, 90 minuti sprecati. | • [0, 0, 0, 0, 1, 1] |

PROBLEMI

1. Abbiamo perso tutte le informazioni sulla sequenza, cioè l'ordine delle parole all'interno della frase.
2. Ogni osservazione avrà un numero di features pari al numero di parole nel dizionario.

Cosa è il Word Embedding ?

Il Word Embedding è un modello di codifica del testo all'interno di uno spazio vettoriale in cui i vettori di parole semanticamente simili si trovano più vicini.

CORPUS DI TESTO

- Questo corso è brutto.
- Questo corso è pessimo.
- Questo corso è bello.
- Questo corso è stupendo.

DIZIONARIO

questo
corso
è
brutto
pessimo
bello
stupendo

DIZIONARIO

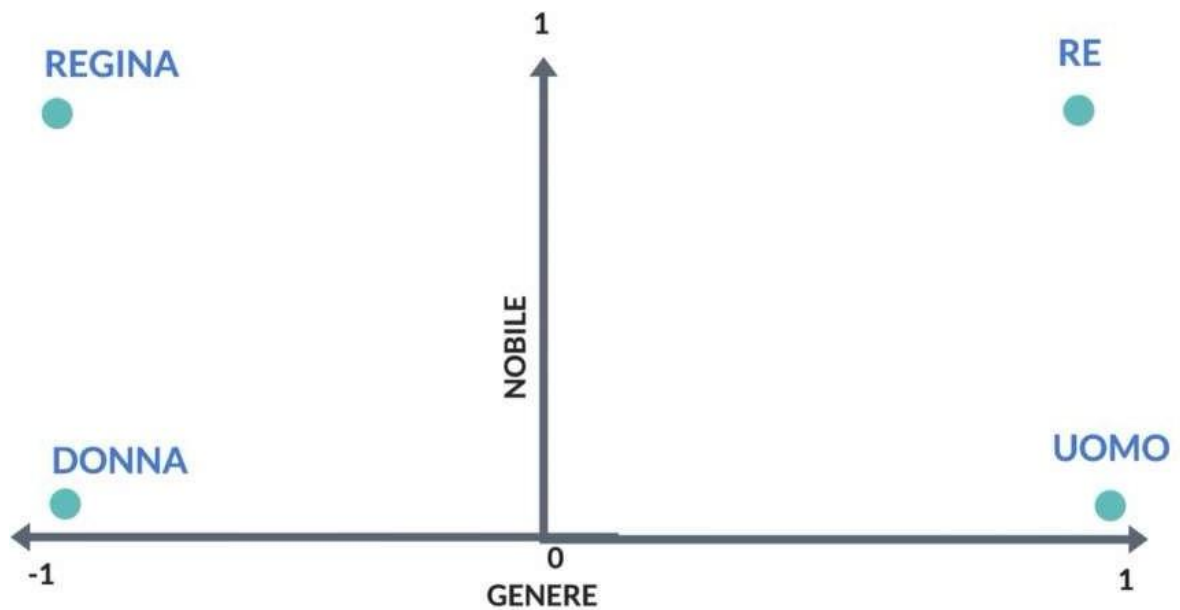
questo
corso
è
brutto
pessimo
bello
stupendo



ONE HOT ENCODING

[1,0,0,0,0,0,0]
[0,1,0,0,0,0,0]
[0,0,1,0,0,0,0]
[0,0,0,1,0,0,0]
[0,0,0,0,1,0,0]
[0,0,0,0,0,1,0]
[0,0,0,0,0,0,1]

	UOMO	DONNA	RE	REGINA
GENERE	1	-1	0.95	-0.95
NOBILE	0.02	0.03	0.98	0.97



ONE HOT ENCODING

[1,0,0,0,0,0,0]
 [0,1,0,0,0,0,0]
 [0,0,1,0,0,0,0]
 [0,0,0,1,0,0,0]
 [0,0,0,0,1,0,0]
 [0,0,0,0,0,1,0]
 [0,0,0,0,0,0,1]

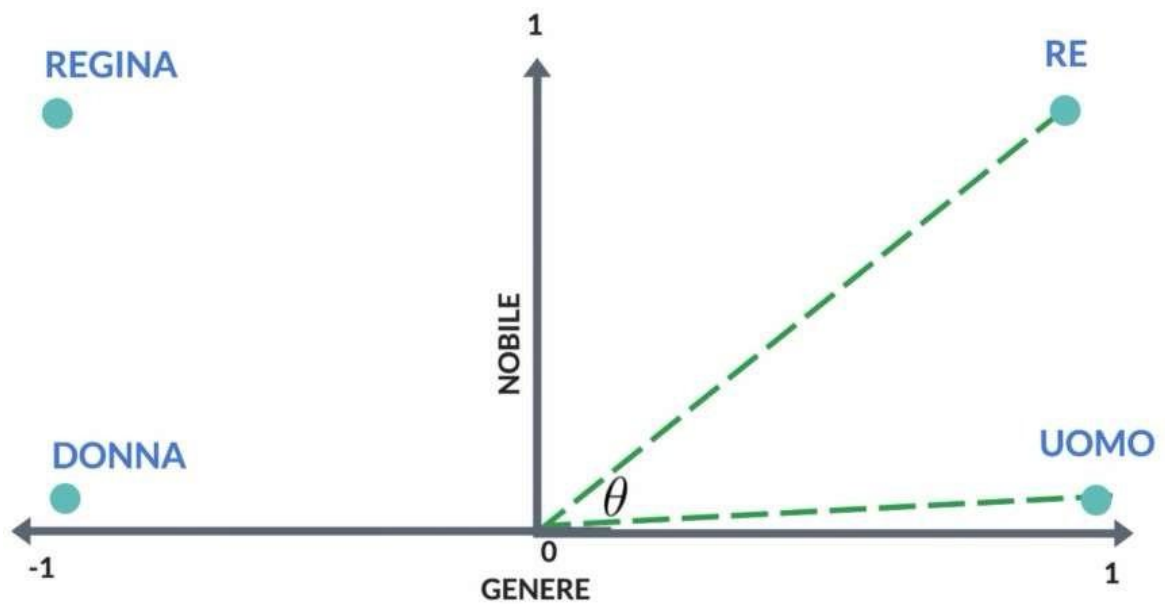


Word Embedding

UN ESEMPIO

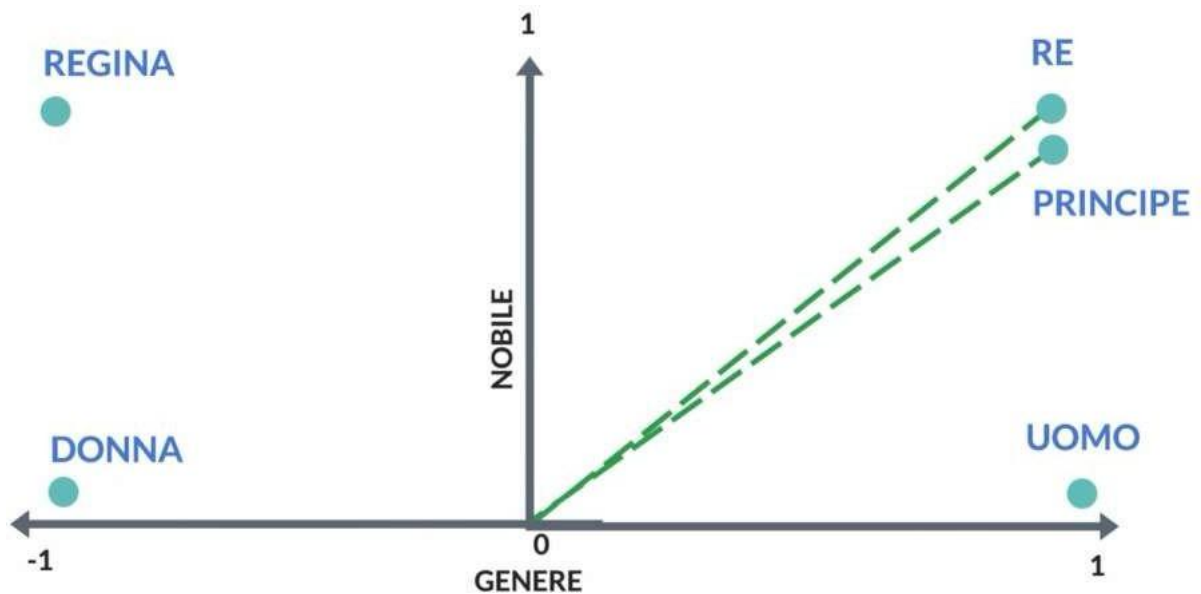
Ci permette di passare da una rappresentazione **sparsa** ad una **densa** in cui parole simili hanno rappresentazioni (vettori) simili

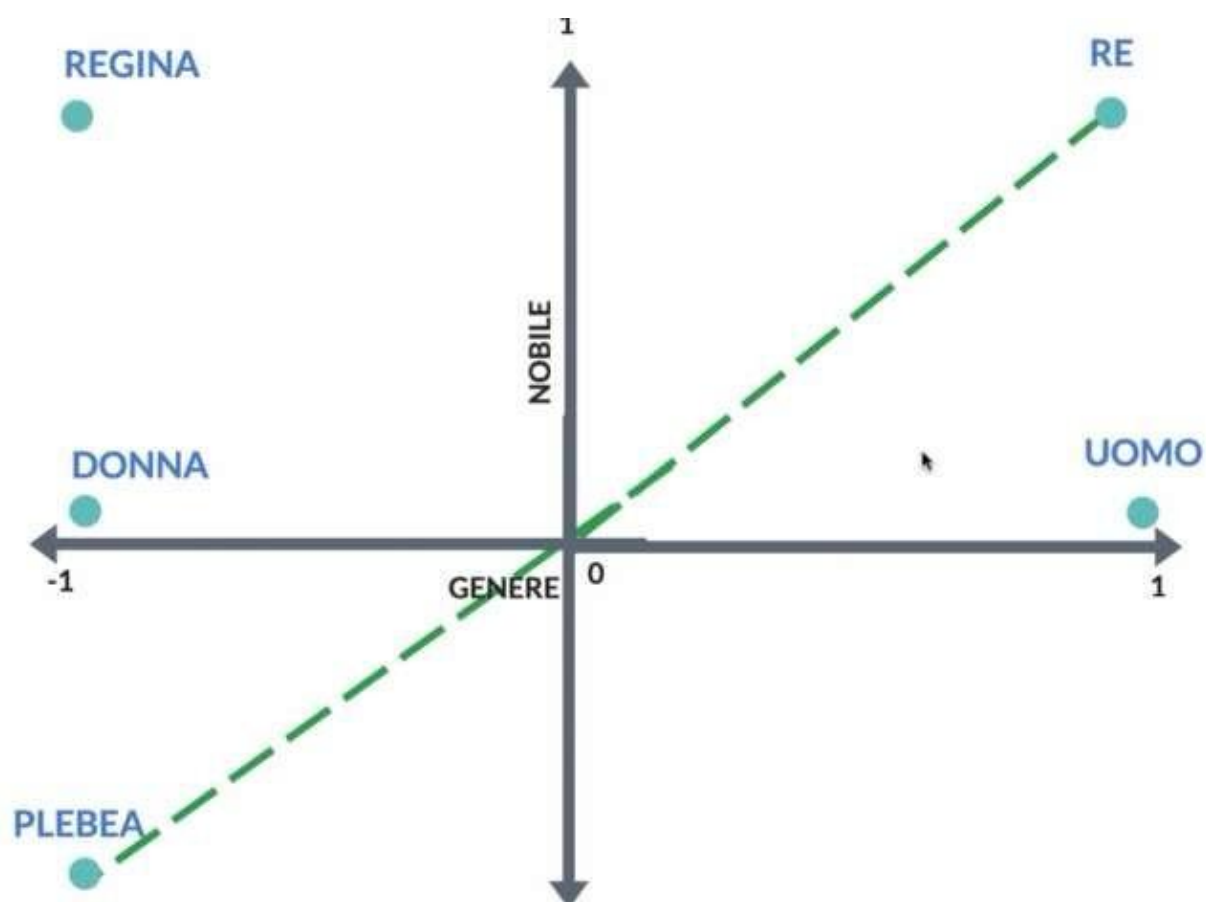
	QUESTO	FILM	E'	BRUTTO	PESSIMO	BELLO	STUPENDO
EMBEDDING 1	0	0	0	-0.5	-0.8	0.5	0.8
EMBEDDING 2	0.98	0.91	0	0.95	0.92	0.88	0.93
EMBEDDING 3	0.91	0.75	-0.15	0.23	0.21	0.23	0.19



Similarità del coseno (cosine similarity)

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





Word Embedding

VANTAGGI

- Permette di mantenere le informazioni sulla sequenza.
- Permette di ridurre la dimensionalità delle osservazioni.
- Permette di misurare somiglianza e attinenza delle parole