Natural Language Processing

Cos'è un Linguaggio Formale ?

Dato un insieme di simboli Σ detto alfabeto, un linguaggio formale è un *sottoinsieme* di tutte le possibili concatenazioni dei simboli:

$$L \subseteq \Sigma^*$$

Un linguaggio formale non è ambiguo (una concatenazione di simboli ha una interpretazione univoca) ed esprime le sue regole in maniera canonica

Cos'è il Linguaggio Naturale ?

- **Strumento di comunicazione** tra persone;
  - Fatti, idee e conoscenze (sul mondo esterno ed interiore)
  - Emozioni
  - Ordini
- E' ambiguo! (*"La vecchia porta la sbarra"*)

Obiettivi del Natural Language Processing

L' Elaborazione del Linguaggio Naturale (*Natural Language Processing, NLP*) ha come obiettivo principale:

Costruzioni di modelli e di strumenti informatici in grado di eseguire specifici task per la comunicazione *uomo – macchina*

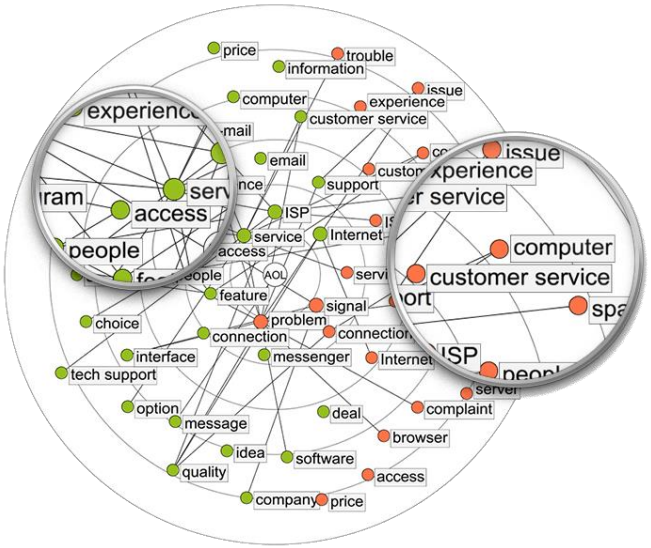
Sempre maggiore quantità di conoscenza condivisa in testi in Linguaggio Naturale *machine readable (ES: sul Web)*

Applicazioni del Natural Language Processing

- Speech recognition Given a sound clip of a person or people speaking, determine the textual representation of the speech.
- Named entity recognition (NER) Given a stream of text, determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization).
- Sentiment analysis Extract subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects.
- Topic segmentation and recognition. Given a chunk of text, separate it into segments each of which is devoted to a topic, and identify the topic of the segment.
- Spam Detection

Topic Detection

- Estrazione **AUTOMATICA** di parole **TIPICHE** all'interno di documento.
  - Text Summerization
  - Social Network Monitoring
  - ...



Tf-Idf

- La funzione di peso **Tf-Idf** (*term frequency–inverse document frequency*) è una funzione utilizzata in **Information Retrieval** per misurare **l'importanza** di un **termine** rispetto ad un **documento** e ad una **collezione di documenti**.

Tf-Idf

- La funzione è composto da due parti distinte:
  - **Tf (Term Frequency).** Misura la frequenza di una parola all'interno di un documento specifico. Quante più occorrenze della parola si presentano nel documento, tanto maggiore è il valore dell'indicatore **Tf**.
  - **Idf (Inverse Document Frequency).** Questo indicatore misura la frequenza inversa di una parola tra tutti i documenti. E' molto alto nei termini specifici, mentre è molto basso nelle parole comuni.

Tf-Idf

$$Tf_{x,y} = (N_{x,y} / N_{*,y})$$

$$Idf_x = \log (D / D_x)$$

$$Tf\text{-}Idf_{x,y} = Tf_{x,y} * Idf_x$$

x → Parola Cosiderata
y → Documento che contiene la parola x

Tf-Idf – Esempio

- Il documento **A** contiene **100 parole,** nel quale il termine *«hello»* compare **5** volte. Il fattore **Tf** per il termine «hello» è:

$$\mathbf{Tf}_{\mathbf{hello,A}} = \frac{5}{100} = 0.05$$

- Assumiamo di avere ora **1000 documenti** nella collezione e *«hello»* compare in **10** di questi:

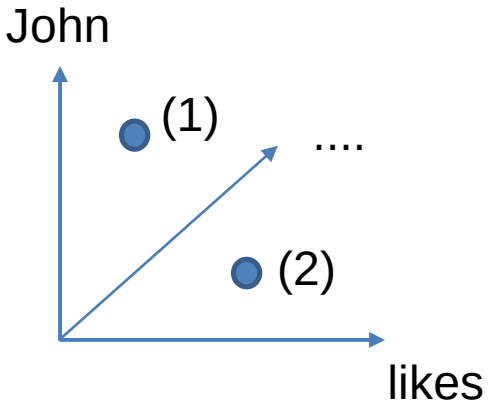
$$\mathbf{Idf}_{\mathbf{hello}} = \log\frac{1000}{10} = 2$$

- Quindi:

$$\mathbf{Tf\text{-}Idf}_{\mathbf{hello,A}} = 0{,}05 * 2 = 0.1$$

Tf-Idf – In Python?

Il modello Bag Of Words



Example:

```
(1) John likes to watch movies.
(2) John also likes to watch football games.
```

	John	likes	to	watch	movies	also	football	games	Mary	too
(1)	1	1	1	1	1	0	0	0	0	0
(2)	1	1	1	1	0	1	1	1	0	0

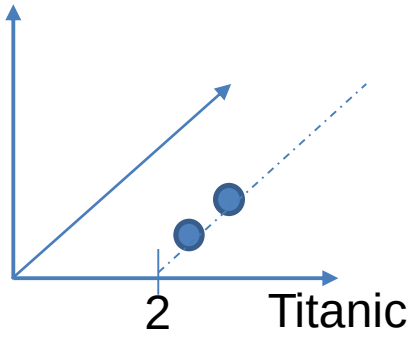
...Altri Problemi....

(1) ***Titanic*** si sa, è un prodotto cinematografico dalle grandi cifre. Grande merito del regista è attualizzare la rievocazione e il racconto di questo evento storico col pretesto di un gruppo di cacciatori di tesori perduti che, ai giorni nostri, decide di tornare sulle tracce del relitto per rinvenire un gioiello dal valore inestimabile, affondato, secondo la loro ricostruzione, insieme al ***Titanic***.

(2) Ma smettila di guardare il ***Titanic*** dai! Andiamo a vedere una bella mostra oggi! Non ti ho mai parlato di Van Gogh? Beh... Devi sapere che.... Autore di quasi 900 dipinti[1] e più di mille disegni, senza contare i numerosi schizzi non portati a terminee tanti appunti destinati probabilmente all'imitazione di disegni artistici di provenienza giapponese. Tanto geniale quanto incompreso in vita, van Gogh influenzò profondamente l'arte del XX secolo. Dopo aver trascorso molti anni soffrendo di frequenti disturbi mentali[2][3] morì all'età di 37 anni per una ferita da arma da fuoco, molto probabilmente auto-inflitta.[4] In quell'epoca i suoi lavori non erano molto conosciuti né tantomeno apprezzati. Van Gogh iniziò a disegnare da bambino, nonostante le continue pressioni del padre, pastore protestante che continuò ad impartirgli delle norme severe. Continuò comunque a disegnare finché non decise di diventare un pittore vero e proprio. Iniziò a dipingere tardi, all'età di ventisette anni, realizzando molte delle sue opere più note nel corso degli ultimi due anni di vita. I suoi soggetti consistevano in autoritratti, paesaggi, nature morte di fiori, dipinti con cipressi,rappresentazione di campi di grano e girasoli. La sua formazione si deve all'esempio del realismo paesaggistico dei pittori di Barbizon e del messaggio etico e sociale di Jean-François Millet. Van Gogh in età adulta lavorò per una ditta di mercanti d'arte, viaggiò tra L'Aia, Londra e Parigi. Per breve tempo si dedicò anche all'insegnamento; una delle sue aspirazioni iniziali fu quella di diventare un pastore e dal 1879 lavorò come missionario in una regione mineraria del Belgio, dove ritrasse persone della comunità locale. Nel 1885, dipinse la sua prima grande opera: I mangiatori di patate. La sua tavolozza, al momento costituita principalmente da cupi toni della terra, non mostra ancora nessun segno della colorazione viva che contraddistinguerà le sue successive opere. Nel marzo del 1886, si trasferì a Parigi dove scoprì gli impressionisti francesi. Più tardi, spostatosi nella Francia del sud, i suoi lavori furono influenzatidalla forte luce del sole che vi trovò. Meglio del tuo ***Titanic***.... O no?

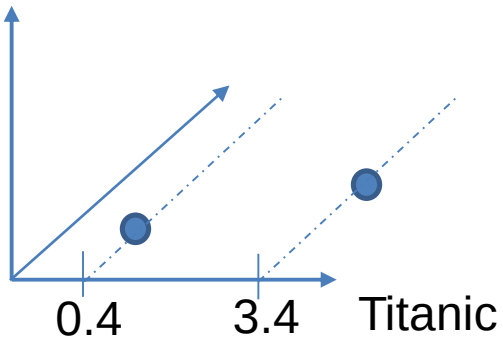
...Altri Problemi....

	...	....	...	....	Titanic	....	....	....	....	....
(1)	1	1	3	1	**2**	0	0	0	3	4
(2)	3	1	1	1	**2**	1	1	1	1	4



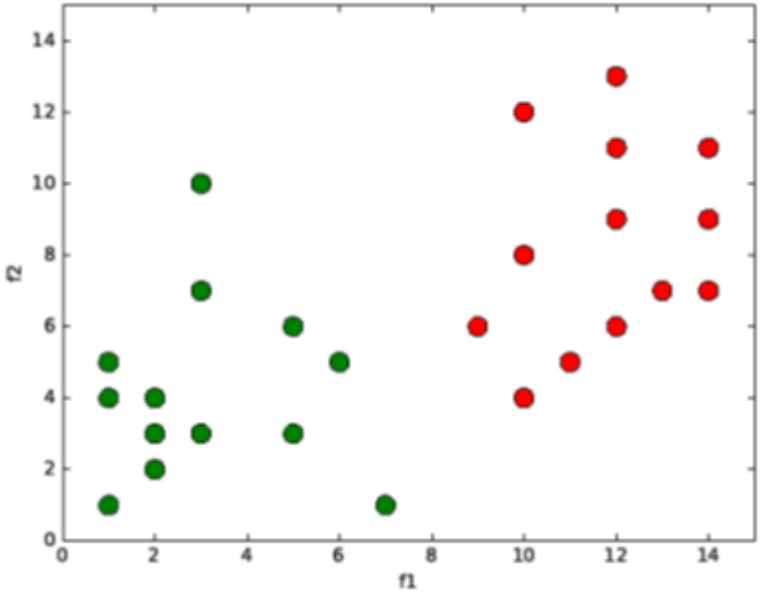
...Tf-Idf....

	...	....	...	....	Titanic	....	....	....	....	....
(1)	0.13	0.45	2.5	1	**3.4**	0	0	0	1.23	1.1
(2)	0.32	2.3	2.1	1.1	**0.4**	5.6	0.1	0.34	2.2	0.03

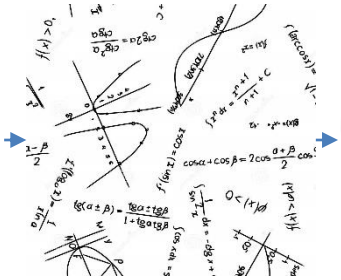


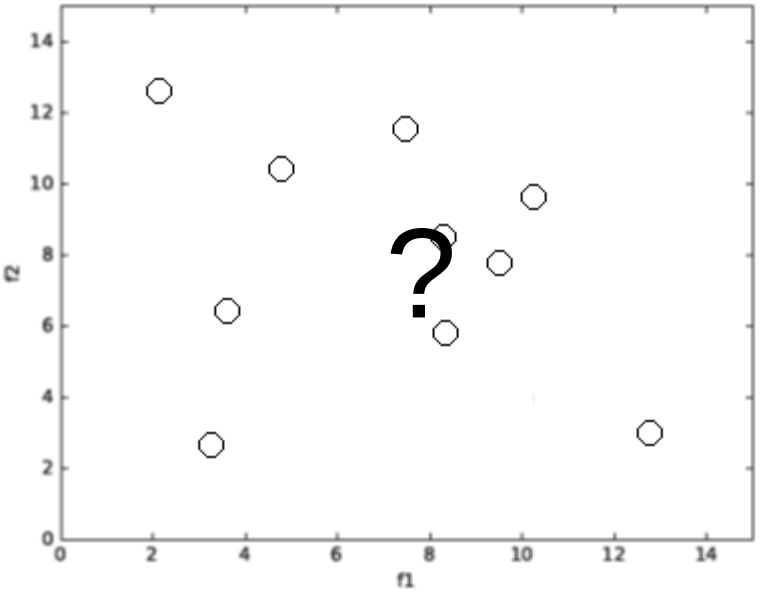
Automatic Text Classification

- **Supervised Learning** (Classificazione)



Training Set





Test Set

Spam Classification

	$word_1$	$word_2$	$word_3$	$word_4$	...	$word_n$	Classe
Training Set	0	0	1	1	...		SPAM
	0	1	1	0	...		SPAM
	1	1	1	0	...		NO-SPAM
	1	0	0	0	...		NO-SPAM
Test Set	**1**	**1**	**0**	**0**	**1**		**?**

Classification with Naive Bayes

X_1	X_2	X_3	X_4	...	X_n	Y
0	0	1	1	...		SPAM
0	1	1	0	...		SPAM
1	1	1	0	...		NO-SPAM
1	0	0	0	...		NO-SPAM
1	**1**	**0**	**0**	**1**		**?**

$P(Y=SPAM \mid X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \ldots) = ?$

$P(Y=NO\text{-}SPAM \mid X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \ldots) = ?$

1) P(Y=SPAM | X_1=1 ∩ X_2=1 ...)
2) P(Y=NO-SPAM | X_1=1 ∩ X_2=1 ...)

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

1) $P(Y=SPAM \mid X_1=1 \cap X_2=1 \ldots) = \frac{P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM)\,P(Y=SPAM)}{P(X_1=1 \cap X_2=1 \ldots)}$

2) $P(Y=NO\text{-}SPAM \mid X_1=1 \cap X_2=1 \ldots) = \frac{P(X_1=1 \cap X_2=1 \ldots \mid Y=NO\text{-}SPAM)\,P(Y=NO\text{-}SPAM)}{P(X_1=1 \cap X_2=1 \ldots)}$

1) $P(Y=SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM)\,P(Y=SPAM)$

2) $P(Y=NO\text{-}SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=NO\text{-}SPAM)\,P(Y=NO\text{-}SPAM)$

A priori probability

$$P(Y=SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM)\, P(Y=SPAM)$$

$$P(Y=NO-SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=NO-SPAM)\, P(Y=NO-SPAM)$$

X_1	X_2	X_3	X_4	...	X_n	Y
0	0	1	1	...		SPAM
0	1	1	0	...		SPAM
1	1	1	0	...		NO-SPAM
1	0	0	0	...		NO-SPAM
1	**1**	**0**	**0**	**1**		**?**

$P(Y=SPAM) = 2/4 = 0.5$ $P(Y=NO-SPAM) = 2/4 = 0.5$

A priori probability

$$P(Y=SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM)\, P(Y=SPAM)$$

$$P(Y=NO-SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=NO-SPAM)\, P(Y=NO-SPAM)$$

X_1	X_2	X_3	X_4	...	X_n	Y
0	0	1	1	...		SPAM
0	1	1	0	...		SPAM
1	1	1	0	...		NO-SPAM
1	0	0	0	...		NO-SPAM
1	**1**	**0**	**0**	**1**		**?**

$P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM) = ??$ $P(X_1=1 \cap X_2=1 \ldots \mid Y=NO-SPAM) = ??$

$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \mid Y=SPAM)$

??

$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0)$

??

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0)$$

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0) = P(X_1=1 \mid X_2=1 \cap X_3=0 \cap X_4=0) * P(X_2=1 \mid X_3=0 \cap X_4=0) * P(X_3=0 \mid X_4=0) * P(X_4=0)$$

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \mid Y=SPAM)$$

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \mid Y=SPAM) = P(X_1=1 \mid X_2=1 \cap X_3=0 \cap X_4=0 \cap Y=SPAM) *$$
$$P(X_2=1 \mid X_3=0 \cap X_4=0 \cap Y=SPAM) *$$
$$P(X_3=0 \mid X_4=0 \cap Y=SPAM) *$$
$$P(X_4=0 \mid Y=SPAM) * P(Y=SPAM)$$

Likelihood

$$P(Y=SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM)\, P(Y=SPAM)$$

$$P(Y=NO-SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=NO-SPAM)\, P(Y=NO-SPAM)$$

X_1	X_2	X_3	X_4	...	X_n	Y
0	0	1	1	...		SPAM
0	1	1	0	...		SPAM
1	1	1	0	...		NO-SPAM
1	0	0	0	...		NO-SPAM
1	**1**	**0**	**0**	**1**		**?**

$P(X_1=1 \mid X_2=1 \cap X_3=0 \cap X_4=0 \cap Y=SPAM)*$
$P(X_2=1 \mid X_3=0 \cap X_4=0 \cap Y=SPAM)*$
$P(X_3=0 \mid X_4=0 \cap Y=SPAM) *$
$P(X_4=0 \mid Y=SPAM) * P(Y=SPAM)$

$P(X_1=1 \mid X_2=1 \cap X_3=0 \cap X_4=0 \cap Y=NO-SPAM)*$
$P(X_2=1 \mid X_3=0 \cap X_4=0 \cap Y= NO-SPAM)*$
$P(X_3=0 \mid X_4=0 \cap Y=NO-SPAM) *$
$P(X_4=0 \mid Y=NO-SPAM) * P(Y=NO-SPAM)$

$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \mid Y=SPAM)$

The Naive Bayes Assumtpion

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0) = P(X_1=1) * P(X_2=1) * P(X_3=0) * P(X_4=0)$$

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \mid Y=SPAM) = P(X_1=1 \mid Y=SPAM) * P(X_2=1 \mid Y=SPAM) * P(X_3=0 \mid Y=SPAM) * P(X_4=0 \mid Y=SPAM)$$

Likelihood

$$P(Y=SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=SPAM)\, P(Y=SPAM)$$

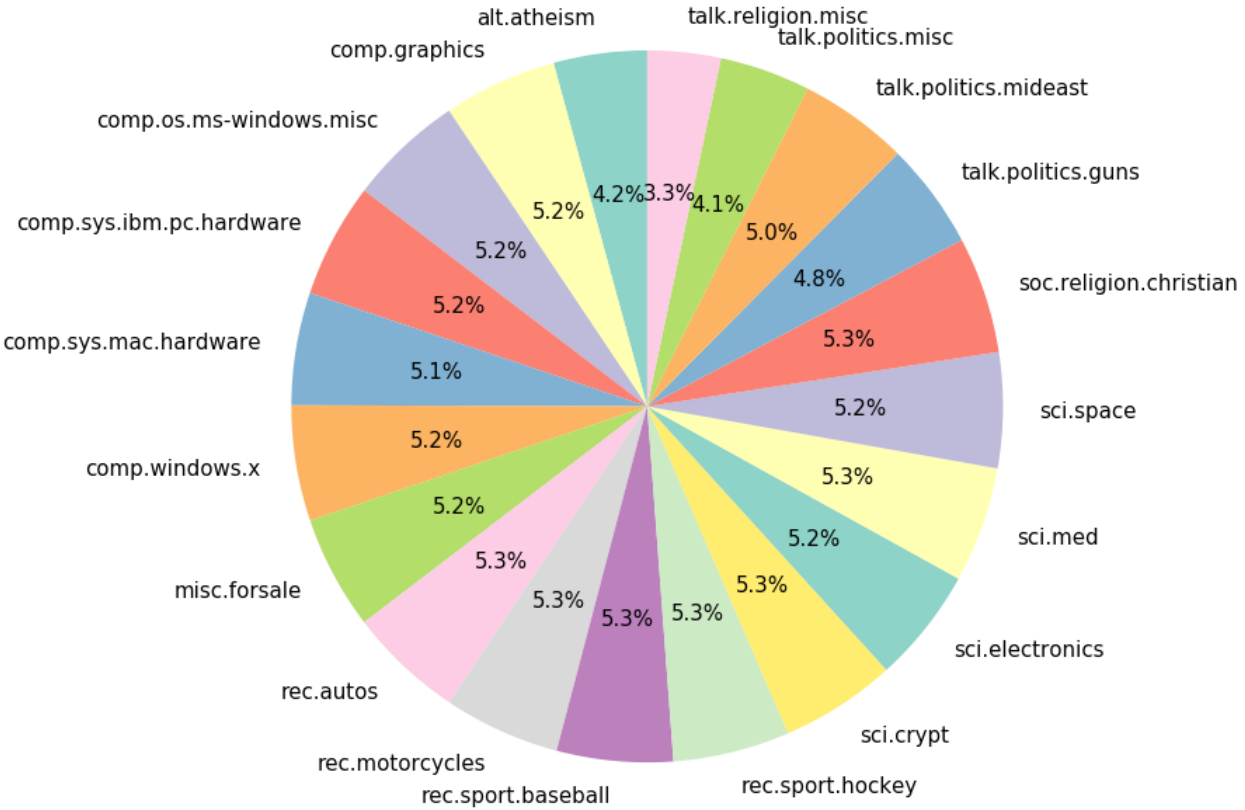
$$P(Y=NO-SPAM \mid X_1=1 \cap X_2=1 \ldots) = P(X_1=1 \cap X_2=1 \ldots \mid Y=NO-SPAM)\, P(Y=NO-SPAM)$$

X_1	X_2	X_3	X_4	...	X_n	Y
0	0	1	1	...		SPAM
0	1	1	0	...		SPAM
1	1	1	0	...		NO-SPAM
1	0	0	0	...		NO-SPAM
1	**1**	**0**	**0**	**1**		**?**

$$P(X_1=1 \cap X_2=1 \cap X_3=0 \cap X_4=0 \mid Y=SPAM) =$$
$$P(X_1=1 \mid Y=SPAM) * P(X_2=1 \mid Y=SPAM) *$$
$$P(X_3=0 \mid Y=SPAM) * P(X_4=0 \mid Y=SPAM)$$

$$P(X_1=1 \mid Y=SPAM) = 0$$
$$P(X_2=1 \mid Y=SPAM) = 0.5$$
$$P(X_3=0 \mid Y=SPAM) = \ldots$$
$$P(X_4=0 \mid Y=SPAM) = 0.5$$

Naive Bayes Classifier in Weka



Information Gain: An example

Outlook	Temperature	Humidity	Windy	Play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Entropia

- Consideriamo una **variabile aleatoria X** che possa assumere i valori $x_1, x_2, x_3, x_4 \ldots x_n$ con probabilità $p_1, p_2, p_3, p_4 \ldots p_n$. La quantità media di informazione (o di **sorpresa!!**) detta anche ENTROPIA è:

$$H(X) = -\sum_{i=1}^{n} p_i * log_2 p_i$$

Information Gain: Entropia del Dataset

Play
yes
yes
yes
yes
yes
yes
no
yes
no
no
no
no
yes
yes

9 yes, 5 no

$$E[play] = -(\frac{9}{14} log_2 \frac{9}{14} + \frac{5}{14} log_2 \frac{5}{14})$$

$$E[play] = 0.94$$

Information Gain

Outlook	Play
overcast	yes
overcast	yes
overcast	yes
overcast	yes
rainy	yes
rainy	yes
rainy	no
rainy	yes
rainy	no
sunny	no
sunny	no
sunny	no
sunny	yes
sunny	yes

- **Overcast**: 4 records, 4 are "yes", 0 are "no"

$$-\left(\frac{4}{4}log_2\frac{4}{4}\right) = \mathbf{0}$$

- **Rainy**: 5 records, 3 are "yes", 2 are "no"

$$-\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = \mathbf{0.97}$$

- **Sunny**: 5 records, 2 are "yes", 3 are "no"

$$-\left(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\right) = \mathbf{0.97}$$

Information Gain: Entropia di Outlook

Outlook	Play
overcast	yes
overcast	yes
overcast	yes
overcast	yes
rainy	yes
rainy	yes
rainy	no
rainy	yes
rainy	no
sunny	no
sunny	no
sunny	no
sunny	yes
sunny	yes

Expected New Entropy (Outlook)

$$\frac{4}{14} * 0 + \frac{5}{14} * 0.97 + \frac{5}{14} * 0.97 = \mathbf{0.69}$$

Information Gain (Outlook)

0.94 - 0.69 = 0.25

Information Gain

Outlook	Temperature	Humidity	Windy	Play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

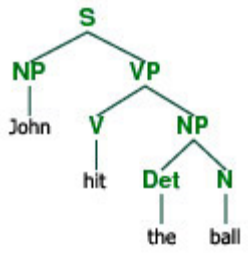
Attribute	Information Gain
outlook	0.94 - 0.69 = 0.25
temperature	0.94 - 0.91 = 0.03
humidity	0.94 - 0.79 = 0.15
windy	0.94 - 0.89 = 0.05

Information Gain Weka

Part of speech Tagging

- Is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech (identification of words as nouns, verbs, adjective, adverbs etc.)

- Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times (…dog…)



[‘Hello’ , ‘every’ , ‘one’ , ‘!’ , ‘How’ , ‘are’ , ‘you’ , ‘?’ ,
‘I’ , ‘love’ , ‘this’ , ‘sunny’ , ‘day’ , ‘.’]

POS

[(‘Hello’ , ‘UH’) ,(‘every’ , ‘JJ’) , (‘one’ , ‘NN’) , (‘!’ , ‘.’) , (‘How’ ,
‘WRB’) , (‘are’ , ‘VBP’) , (‘you’ , ‘PRP’) , (‘?’ , ‘.’) , (‘I’ , ‘PRP’) , (‘love’ ,
‘VB’) , (‘this’ , ‘DT’) , (‘sunny’ , ‘JJ’) , (‘day’ , ‘NN’) , (‘.’ , ‘.’)]

Part of speech Tagging Python

Pattern

- Individuare alcune pattern lessicali all'interno della frase, e duplicarli nella stessa frase.
- Che effetto produce sul training set espresso con il modello bag of word?

Dan Jurafsky

Laplace (add-1) smoothing for Naïve Bayes

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$