# Metodi entropici per l'estrazione di informazione da testi scritti

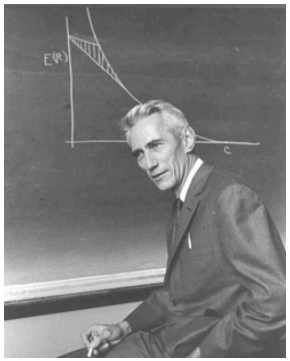
## Seminario per il corso di Fisica Matematica Applicata - Prof. Lenci

Chiara Basile - basile@dm.unibo.it

Dipartimento di Matematica
Università di Bologna

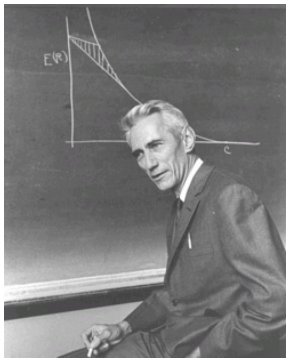
Bologna, 18/11/2009

# Andando alle origini...



1948: Claude E. Shannon scrive *A Mathematical Theory of Communication* fondando la teoria dell'informazione e dando in particolare la definizione di *entropia* di una variabile aleatoria e di una sorgente di informazione.

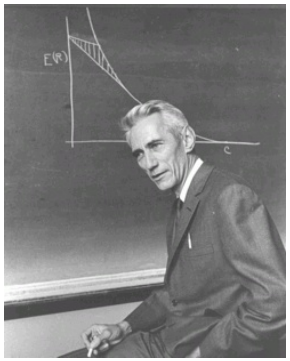
# Andando alle origini...



1948: Claude E. Shannon scrive *A Mathematical Theory of Communication* fondando la teoria dell'informazione e dando in particolare la definizione di *entropia* di una variabile aleatoria e di una sorgente di informazione.

*"La mia più grande preoccupazione era come chiamarla. Pensavo di chiamarla informazione, ma la parola era fin troppo usata, così decisi di chiamarla incertezza. Quando discussi della cosa con John Von Neumann, lui ebbe un'idea migliore. Mi disse che avrei dovuto chiamarla entropia, per due motivi: 'Innanzitutto, la tua funzione d'incertezza è già nota nella meccanica statistica con quel nome. In secondo luogo, e più significativamente, nessuno sa cosa sia con certezza l'entropia, così in una discussione sarai sempre in vantaggio.' "*

# Andando alle origini...



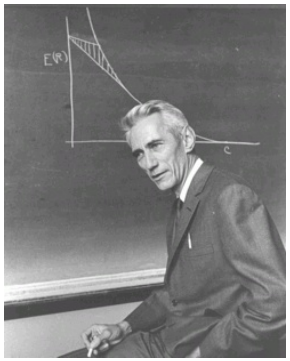
1948: Claude E. Shannon scrive *A Mathematical Theory of Communication* fondando la teoria dell'informazione e dando in particolare la definizione di *entropia* di una variabile aleatoria e di una sorgente di informazione.

*"La mia più grande preoccupazione era come chiamarla. Pensavo di chiamarla informazione, ma la parola era fin troppo usata, così decisi di chiamarla incertezza. Quando discussi della cosa con John Von Neumann, lui ebbe un'idea migliore. Mi disse che avrei dovuto chiamarla entropia, per due motivi: 'Innanzitutto, la tua funzione d'incertezza è già nota nella meccanica statistica con quel nome. In secondo luogo, e più significativamente, nessuno sa cosa sia con certezza l'entropia, così in una discussione sarai sempre in vantaggio.' "*

**Definizione (entropia di una v.a.).** Sia $X$ una variabile aleatoria che prende valori in $\mathcal{A} = \{a_1, \ldots, a_m\}$ e $\{p_i = P(X = a_i)\}_{i=1,\ldots,m}$ la sua distribuzione di probabilità. Si dice entropia di $X$ la quantità

$$H(X) = H(p_1, \ldots, p_m) = -K \sum_{i=1}^{m} p_i \log p_i.$$

# Andando alle origini...



1948: Claude E. Shannon scrive *A Mathematical Theory of Communication* fondando la teoria dell'informazione e dando in particolare la definizione di *entropia* di una variabile aleatoria e di una sorgente di informazione.

*"La mia più grande preoccupazione era come chiamarla. Pensavo di chiamarla informazione, ma la parola era fin troppo usata, così decisi di chiamarla incertezza. Quando discussi della cosa con John Von Neumann, lui ebbe un'idea migliore. Mi disse che avrei dovuto chiamarla entropia, per due motivi: 'Innanzitutto, la tua funzione d'incertezza è già nota nella meccanica statistica con quel nome. In secondo luogo, e più significativamente, nessuno sa cosa sia con certezza l'entropia, così in una discussione sarai sempre in vantaggio.' "*

**Definizione (entropia di una v.a.).** Sia $X$ una variabile aleatoria che prende valori in $\mathcal{A} = \{a_1, \ldots, a_m\}$ e $\{p_i = P(X = a_i)\}_{i=1,\ldots,m}$ la sua distribuzione di probabilità. Si dice entropia di $X$ la quantità

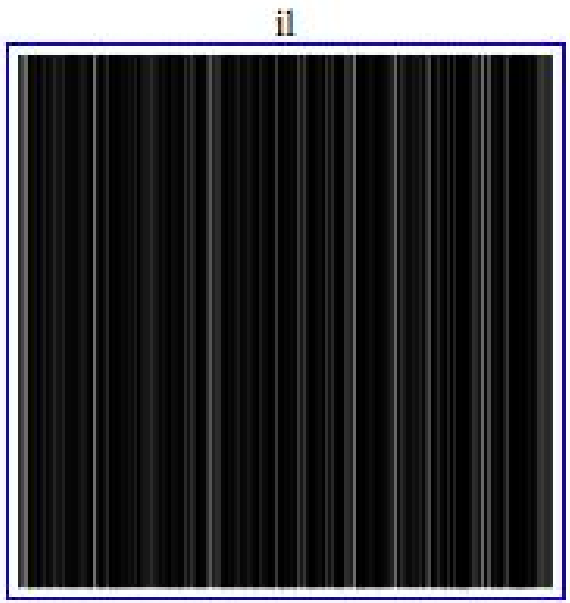
$$H(X) = H(p_1, \ldots, p_m) = -K \sum_{i=1}^{m} p_i \log p_i.$$

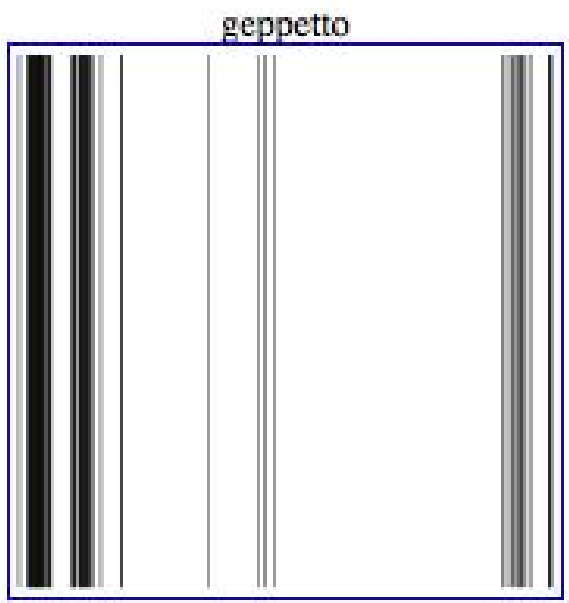
# Un'applicazione diretta: estrazione di parole chiave

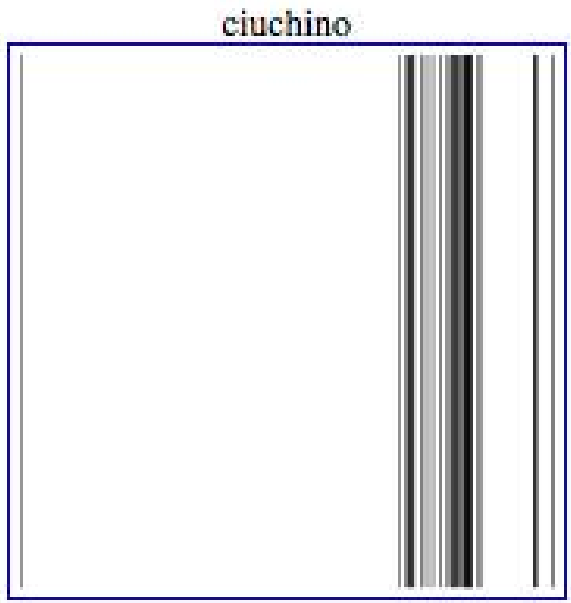
Quale sarà la distribuzione di `il` in *Pinocchio* di C. Collodi? E se invece prendessi `geppetto`, o `ciuchino`?

# Un'applicazione diretta: estrazione di parole chiave

Quale sarà la distribuzione di `il` in *Pinocchio* di C. Collodi? E se invece prendessi `geppetto`, o `ciuchino`?

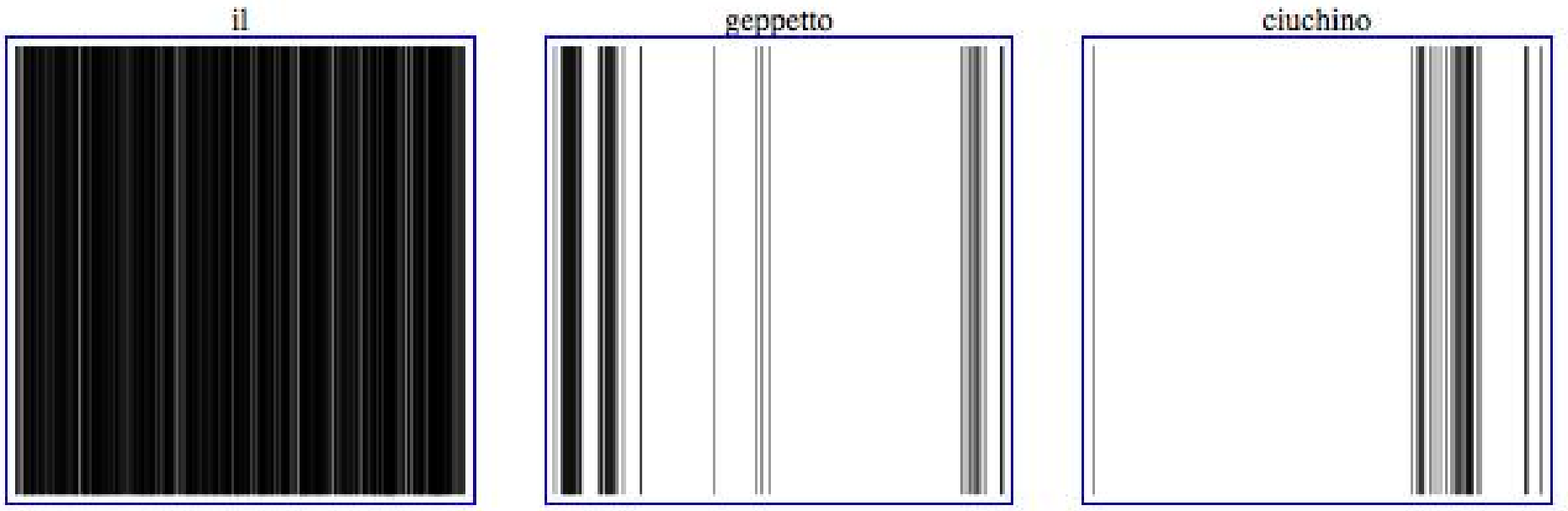


# Un'applicazione diretta: estrazione di parole chiave

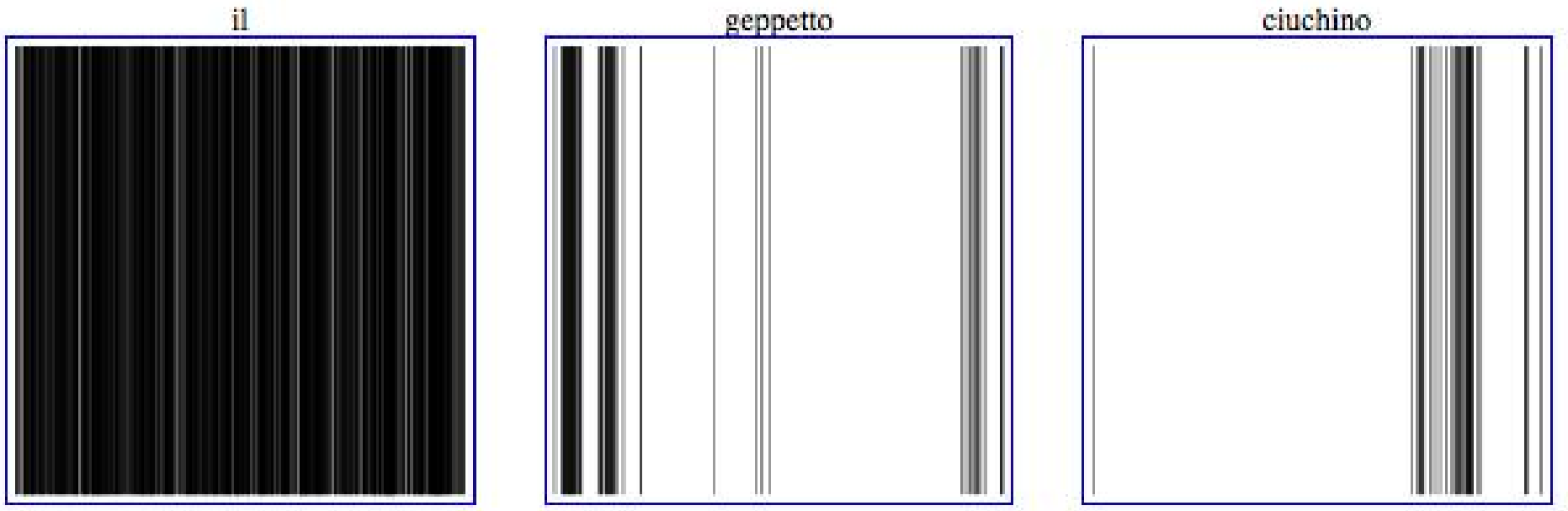
Quale sarà la distribuzione di `il` in *Pinocchio* di C. Collodi? E se invece prendessi `geppetto`, o `ciuchino`?




La sola frequenza non basta...

# Un'applicazione diretta: estrazione di parole chiave

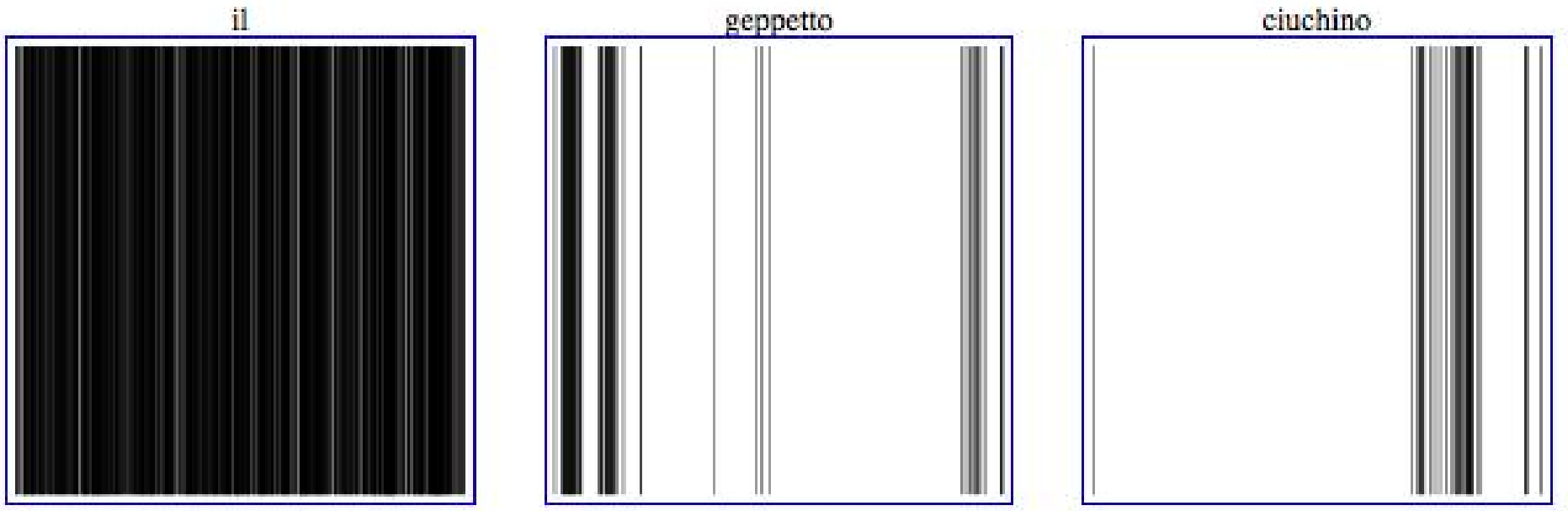
Quale sarà la distribuzione di `il` in *Pinocchio* di C. Collodi? E se invece prendessi `geppetto`, o `ciuchino`?




La sola frequenza non basta... invece l'idea è: le parole "importanti" sono quelle con distribuzione meno uniforme.

# Un’applicazione diretta: estrazione di parole chiave

Quale sarà la distribuzione di `il` in *Pinocchio* di C. Collodi? E se invece prendessi `geppetto`, o `ciuchino`?




La sola frequenza non basta... invece l’idea è: le parole “importanti” sono quelle con distribuzione meno uniforme.

Come cogliere questa non uniformità?

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

Idea dell'algoritmo:

- divido il testo in $P$ parti (i capitoli, oppure taglio a lunghezza fissata)

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

Idea dell'algoritmo:

- divido il testo in $P$ parti (i capitoli, oppure taglio a lunghezza fissata)
- calcolo la distribuzione $\{p_i(w) = \frac{f_i(w)}{\sum_{i=1}^{P} f_i(w)}\}_i$ per ogni parola $w$ ($f_i(w)$ = frequenza empirica di $w$ nella $i-$esima parte)

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

Idea dell'algoritmo:

- divido il testo in $P$ parti (i capitoli, oppure taglio a lunghezza fissata)
- calcolo la distribuzione $\{p_i(w) = \frac{f_i(w)}{\sum_{i=1}^{P} f_i(w)}\}_i$ per ogni parola $w$ ($f_i(w)$ = frequenza empirica di $w$ nella $i-$esima parte)
- calcolo l'entropia $S(w) = -\frac{1}{\log P} \sum_{i=1}^{P} p_i(w) \log p_i(w)$

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

Idea dell'algoritmo:

- divido il testo in $P$ parti (i capitoli, oppure taglio a lunghezza fissata)
- calcolo la distribuzione $\{p_i(w) = \frac{f_i(w)}{\sum_{i=1}^{P} f_i(w)}\}_i$ per ogni parola $w$ ($f_i(w)$ = frequenza empirica di $w$ nella $i-$esima parte)
- calcolo l'entropia $S(w) = -\frac{1}{\log P} \sum_{i=1}^{P} p_i(w) \log p_i(w)$
- divido $1 - S(w)$ per $1 -$ l'entropia di un testo con le stesse frequenze, ma parole mescolate in modo casuale $\Rightarrow$ considero la deviazione da un testo random

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

Idea dell'algoritmo:

- divido il testo in $P$ parti (i capitoli, oppure taglio a lunghezza fissata)
- calcolo la distribuzione $\{p_i(w) = \frac{f_i(w)}{\sum_{i=1}^{P} f_i(w)}\}_i$ per ogni parola $w$ ($f_i(w)$ = frequenza empirica di $w$ nella $i-$esima parte)
- calcolo l'entropia $S(w) = -\frac{1}{\log P} \sum_{i=1}^{P} p_i(w) \log p_i(w)$
- divido $1 - S(w)$ per $1 -$ l'entropia di un testo con le stesse frequenze, ma parole mescolate in modo casuale $\Rightarrow$ considero la deviazione da un testo random

`il` $\rightarrow$ 1.11 `geppetto` $\rightarrow$ 17.31 `ciuchino` $\rightarrow$ 12.34

# Estrazione di parole chiave

Ad esempio calcolando l'entropia delle distribuzioni delle parole!

Idea dell'algoritmo:

- divido il testo in $P$ parti (i capitoli, oppure taglio a lunghezza fissata)
- calcolo la distribuzione $\{p_i(w) = \frac{f_i(w)}{\sum_{i=1}^{P} f_i(w)}\}_i$ per ogni parola $w$ ($f_i(w)$ = frequenza empirica di $w$ nella $i-$esima parte)
- calcolo l'entropia $S(w) = -\frac{1}{\log P} \sum_{i=1}^{P} p_i(w) \log p_i(w)$
- divido $1 - S(w)$ per $1 -$ l'entropia di un testo con le stesse frequenze, ma parole mescolate in modo casuale $\Rightarrow$ considero la deviazione da un testo random

`il` $\longrightarrow$ 1.11 `geppetto` $\longrightarrow$ 17.31 `ciuchino` $\longrightarrow$ 12.34

[M.A. Montemurro, D. Zanette. Entropic analysis of the role of words in literary texts. *Advances in Complex Systems* **5** (2002)]
[J.P. Herrera, P.A. Pury. Statistical keyword detection in literary corpora. *The European Physical Journal B* **63** (2008)]

# Qualche esempio...

Collodi, *Pinocchio*

| Parola | Frequenza | Rank | " Entropia P=10" |
|---|---|---|---|
| "geppetto" | 72 | 74 | 17.3098 |
| "lucignolo" | 34 | 169 | 12.9177 |
| "ciuchino" | 40 | 146 | 12.3454 |
| "carro" | 20 | 270 | 9.35142 |
| "monete" | 34 | 171 | 8.84606 |
| "volpe" | 44 | 132 | 8.55535 |
| "assassini" | 26 | 219 | 8.42977 |
| "scuola" | 51 | 114 | 8.05374 |
| "babbo" | 74 | 71 | 7.67954 |
| "cane" | 60 | 95 | 7.23255 |
| "fata" | 80 | 68 | 7.16696 |
| "pescatore" | 14 | 373 | 7.1636 |
| "burattinaio" | 17 | 316 | 7.02224 |
| "pesce" | 52 | 111 | 6.90386 |
| "campo" | 29 | 197 | 6.8975 |
| "gatto" | 43 | 137 | 6.88186 |
| "paese" | 45 | 128 | 6.84757 |
| "omino" | 16 | 339 | 6.84734 |
| "legno" | 64 | 87 | 6.80002 |
| "orecchi" | 25 | 232 | 6.54047 |
| "zecchini" | 16 | 341 | 6.53928 |
| "mare" | 41 | 143 | 6.23485 |
| "oro" | 30 | 191 | 6.18603 |

Darwin, *On the origin of species*

| Parola | Frequenza | Rank | " Entropia P=10" |
|---|---|---|---|
| "hybrids" | 120 | 165 | 48.8041 |
| "sterility" | 78 | 261 | 31.7779 |
| "formations" | 90 | 225 | 28.2131 |
| "fertility" | 81 | 249 | 27.5185 |
| "islands" | 131 | 151 | 25.5163 |
| "breeds" | 124 | 161 | 24.3746 |
| "cells" | 47 | 425 | 22.9738 |
| "instincts" | 73 | 281 | 22.24 |
| "instinct" | 58 | 355 | 21.5692 |
| "bees" | 65 | 312 | 21.1503 |
| "selection" | 383 | 55 | 21.0818 |
| "varieties" | 396 | 54 | 21.0391 |
| "wax" | 39 | 510 | 19.9557 |
| "rudimentary" | 71 | 287 | 19.3286 |
| "seeds" | 101 | 199 | 19.214 |
| "plants" | 297 | 64 | 19.1113 |
| "groups" | 173 | 115 | 18.6104 |
| "organs" | 124 | 162 | 17.007 |
| "organ" | 81 | 250 | 16.9788 |
| "forms" | 397 | 53 | 16.6224 |
| "crosses" | 48 | 420 | 16.5062 |
| "formation" | 80 | 254 | 16.4638 |
| "slaves" | 34 | 576 | 16.3948 |

# Un approccio indipendente dalla lingua?

| Parola | Frequenza | Rank | " Entropia P=10" |
|---|---|---|---|
| "geppetto" | 72 | 74 | 17.3098 |
| "lucignolo" | 34 | 169 | 12.9177 |
| "ciuchino" | 40 | 146 | 12.3454 |
| "carro" | 20 | 270 | 9.35142 |
| "monete" | 34 | 171 | 8.84606 |
| "volpe" | 44 | 132 | 8.55535 |
| "assassini" | 26 | 219 | 8.42977 |
| "scuola" | 51 | 114 | 8.05374 |
| "babbo" | 74 | 71 | 7.67954 |
| "cane" | 60 | 95 | 7.23255 |
| "fata" | 80 | 68 | 7.16696 |
| "pescatore" | 14 | 373 | 7.1636 |
| "burattinaio" | 17 | 316 | 7.02224 |
| "pesce" | 52 | 111 | 6.90386 |
| "campo" | 29 | 197 | 6.8975 |
| "gatto" | 43 | 137 | 6.88186 |
| "paese" | 45 | 128 | 6.84757 |
| "omino" | 16 | 339 | 6.84734 |
| "legno" | 64 | 87 | 6.80002 |
| "orecchi" | 25 | 232 | 6.54047 |
| "zecchini" | 16 | 341 | 6.53928 |
| "mare" | 41 | 143 | 6.23485 |
| "oro" | 30 | 191 | 6.18603 |

| Parola | Frequenza | Rank | " Entropia P=10" |
|---|---|---|---|
| "geppetto" | 75 | 80 | 18.4227 |
| "donkey" | 58 | 107 | 15.285 |
| "lamp" | 34 | 187 | 12.2753 |
| "wick" | 34 | 186 | 12.2753 |
| "father" | 103 | 62 | 11.2329 |
| "gold" | 53 | 117 | 11.0371 |
| "fox" | 44 | 144 | 9.04447 |
| "mastro" | 21 | 290 | 8.9697 |
| "fish" | 40 | 159 | 8.31116 |
| "fire" | 28 | 219 | 8.25638 |
| "fisherman" | 16 | 350 | 8.18697 |
| "pieces" | 39 | 166 | 7.93893 |
| "boys" | 65 | 95 | 7.85438 |
| "pi" | 15 | 371 | 7.67528 |
| "cricket" | 36 | 177 | 7.64817 |
| "fairy" | 94 | 67 | 7.4408 |
| "field" | 28 | 213 | 7.18605 |
| "eater" | 17 | 336 | 7.02224 |
| "cat" | 43 | 148 | 6.95288 |
| "pigeon" | 15 | 367 | 6.85885 |
| "sea" | 43 | 149 | 6.78834 |
| "it" | 447 | 12 | 6.76192 |
| "harlequin" | 13 | 429 | 6.65191 |
| "shark" | 29 | 207 | 6.63758 |

# Tornando alla teoria...

Cos'è una sorgente di informazione?

# Tornando alla teoria...

Cos'è una sorgente di informazione? Shannon:

sorgente di informazione = processo stocastico stazionario ergodico

# Tornando alla teoria...

Cos'è una sorgente di informazione? Shannon:

sorgente di informazione = processo stocastico stazionario ergodico

**Definizione.** Dato uno spazio di probabilità $(\Omega, \mathcal{S}, P)$ e un alfabeto finito $\mathcal{A}$, un processo stocastico è una successione infinita $\{X_n\} = \{X_1, X_2, \ldots, X_n, \ldots\}$ di variabili aleatorie definite da $\Omega$ in $\mathcal{A}$.

# Tornando alla teoria...

Cos'è una sorgente di informazione? Shannon:

sorgente di informazione = processo stocastico stazionario ergodico

**Definizione.** Dato uno spazio di probabilità $(\Omega, \mathcal{S}, P)$ e un alfabeto finito $\mathcal{A}$, un processo stocastico è una successione infinita $\{X_n\} = \{X_1, X_2, \ldots, X_n, \ldots\}$ di variabili aleatorie definite da $\Omega$ in $\mathcal{A}$.
Il processo si dice stazionario se

$$P(X_1 = a_1, \ldots, X_n = a_n) = P(X_{1+k} = a_1, \ldots, X_{n+k} = a_n)$$

$\forall\ a_1, \ldots, a_n \in \mathcal{A},\ \forall\ k, n \in \mathbb{N}$.

# Tornando alla teoria...

Cos'è una sorgente di informazione? Shannon:

sorgente di informazione = processo stocastico stazionario ergodico

**Definizione.** Dato uno spazio di probabilità $(\Omega, \mathcal{S}, P)$ e un alfabeto finito $\mathcal{A}$, un processo stocastico è una successione infinita $\{X_n\} = \{X_1, X_2, \ldots, X_n, \ldots\}$ di variabili aleatorie definite da $\Omega$ in $\mathcal{A}$.
Il processo si dice stazionario se

$$P(X_1 = a_1, \ldots, X_n = a_n) = P(X_{1+k} = a_1, \ldots, X_{n+k} = a_n)$$

$\forall\ a_1, \ldots, a_n \in \mathcal{A},\ \forall\ k, n \in \mathbb{N}$.

*Intuitivamente*, il processo si dice ergodico se quasi ogni sequenza (infinita) da esso prodotta ha le stesse proprietà statistiche, cioè è un "buon rappresentante" della distribuzione della sorgente.

# Tornando alla teoria...

Cos'è una sorgente di informazione? Shannon:

sorgente di informazione = processo stocastico stazionario ergodico

**Definizione.** Dato uno spazio di probabilità $(\Omega, \mathcal{S}, P)$ e un alfabeto finito $\mathcal{A}$, un processo stocastico è una successione infinita $\{X_n\} = \{X_1, X_2, \ldots, X_n, \ldots\}$ di variabili aleatorie definite da $\Omega$ in $\mathcal{A}$.
Il processo si dice stazionario se

$$P(X_1 = a_1, \ldots, X_n = a_n) = P(X_{1+k} = a_1, \ldots, X_{n+k} = a_n)$$

$\forall\ a_1, \ldots, a_n \in \mathcal{A},\ \forall\ k, n \in \mathbb{N}$.

*Intuitivamente*, il processo si dice ergodico se quasi ogni sequenza (infinita) da esso prodotta ha le stesse proprietà statistiche, cioè è un "buon rappresentante" della distribuzione della sorgente.

▸ sorgenti Markoviane

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

E se volessi calcolarla, non avendo a disposizione sequenze infinite?

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

E se volessi calcolarla, non avendo a disposizione sequenze infinite?

Ci vengono in aiuto due cose:

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

E se volessi calcolarla, non avendo a disposizione sequenze infinite?

Ci vengono in aiuto due cose:

- il Teorema di Shannon-McMillan-Breiman (AEP):

$$\lim_{n\to\infty} -\frac{1}{n} \log P(a_1^n) = h(\{X_n\})$$

  per q.o. sequenza $a \in \mathcal{A}^{\mathbb{N}}$, quando $\{X_n\}$ è una sorgente stazionaria ergodica

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

E se volessi calcolarla, non avendo a disposizione sequenze infinite?

Ci vengono in aiuto due cose:

- il Teorema di Shannon-McMillan-Breiman (AEP):

$$\lim_{n\to\infty} -\frac{1}{n} \log P(a_1^n) = h(\{X_n\})$$

  per q.o. sequenza $a \in \mathcal{A}^{\mathbb{N}}$, quando $\{X_n\}$ è una sorgente stazionaria ergodica

  $\Rightarrow$ posso usare una sola sequenza come rappresentante dell'intera sorgente! (ergodicità...)

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

E se volessi calcolarla, non avendo a disposizione sequenze infinite?

Ci vengono in aiuto due cose:

- il Teorema di Shannon-McMillan-Breiman (AEP):

$$\lim_{n\to\infty} -\frac{1}{n} \log P(a_1^n) = h(\{X_n\})$$

  per q.o. sequenza $a \in \mathcal{A}^{\mathbb{N}}$, quando $\{X_n\}$ è una sorgente stazionaria ergodica

  $\Rightarrow$ posso usare una sola sequenza come rappresentante dell'intera sorgente! (ergodicità...)

- i compressori di dati

# Entropia di sorgenti d'informazione

Entropia (o *entropy rate*) di una sorgente di informazione:

$$h(\{X_n\}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}.$$

E se volessi calcolarla, non avendo a disposizione sequenze infinite?

Ci vengono in aiuto due cose:

- il Teorema di Shannon-McMillan-Breiman (AEP):

$$\lim_{n\to\infty} -\frac{1}{n} \log P(a_1^n) = h(\{X_n\})$$

  per q.o. sequenza $a \in \mathcal{A}^{\mathbb{N}}$, quando $\{X_n\}$ è una sorgente stazionaria ergodica

  $\Rightarrow$ posso usare una sola sequenza come rappresentante dell'intera sorgente! (ergodicità...)

- i compressori di dati  gzip  bzip2 ...

# Entropia e compressione

Intuitivamente: un compressore di dati è una funzione

$$\mathcal{C} : \mathcal{A}^* \to \mathcal{B}^*$$

che sfrutta le ripetizioni nella sequenza finita $a \in \mathcal{A}^*$ per ridurne la dimensione: si spera cioè che sia quasi sempre $|\mathcal{C}(a)| < |a|$.

# Entropia e compressione

Intuitivamente: un compressore di dati è una funzione

$$\mathcal{C} : \mathcal{A}^* \rightarrow \mathcal{B}^*$$

che sfrutta le ripetizioni nella sequenza finita $a \in \mathcal{A}^*$ per ridurne la dimensione: si spera cioè che sia quasi sempre $|\mathcal{C}(a)| < |a|$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| `hey_di` | `d` | `d` | `l` | `e` | `_diddle` | `_` | `t` | `he` | `_` |
| `hey_di` | (1,2) | (1,1) | `l` | (1,8) | (7,7) | (1,7) | `t` | (2,19) | (1,4) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| `ca` | `t` | `_` | `a` | `n` | `d` | `_the_` | `f` | `iddle` |
| `ca` | (1,6) | (1,4) | (1,3) | `n` | (1,14) | (5,12) | `f` | (5,23) |

# Entropia e compressione

Intuitivamente: un compressore di dati è una funzione

$$\mathcal{C} : \mathcal{A}^* \rightarrow \mathcal{B}^*$$

che sfrutta le ripetizioni nella sequenza finita $a \in \mathcal{A}^*$ per ridurne la dimensione: si spera cioè che sia quasi sempre $|\mathcal{C}(a)| < |a|$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| `hey_di` | `d` | `d` | `l` | `e` | `_diddle` | `_` | `t` | `he` | `_` |
| `hey_di` | (1,2) | (1,1) | `l` | (1,8) | (7,7) | (1,7) | `t` | (2,19) | (1,4) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| `ca` | `t` | `_` | `a` | `n` | `d` | `_the_` | `f` | `iddle` |
| `ca` | (1,6) | (1,4) | (1,3) | `n` | (1,14) | (5,12) | `f` | (5,23) |

Questo “usare le ripetizioni” ha un rapporto con l’entropia...

# Entropia e compressione

Intuitivamente: un compressore di dati è una funzione

$$\mathcal{C} : \mathcal{A}^* \to \mathcal{B}^*$$

che sfrutta le ripetizioni nella sequenza finita $a \in \mathcal{A}^*$ per ridurne la dimensione: si spera cioè che sia quasi sempre $|\mathcal{C}(a)| < |a|$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| hey_di | d | d | l | e | _diddle | _ | t | he | _ |
| hey_di | (1,2) | (1,1) | l | (1,8) | (7,7) | (1,7) | t | (2,19) | (1,4) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ca | t | _ | a | n | d | _the_ | f | iddle |
| ca | (1,6) | (1,4) | (1,3) | n | (1,14) | (5,12) | f | (5,23) |

Questo “usare le ripetizioni” ha un rapporto con l’entropia...

$\mathcal{C}$ si dice universale se (non rigorosamente...) per ogni processo ergodico

$$\lim_{n\to\infty} \frac{|\mathcal{C}(a_1^n)|}{n} = h(\{X_1^n\}).$$

# Entropia e compressione

Intuitivamente: un compressore di dati è una funzione

$$\mathcal{C} : \mathcal{A}^* \to \mathcal{B}^*$$

che sfrutta le ripetizioni nella sequenza finita $a \in \mathcal{A}^*$ per ridurne la dimensione: si spera cioè che sia quasi sempre $|\mathcal{C}(a)| < |a|$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| hey_di | d | d | l | e | _diddle | _ | t | he | _ |
| hey_di | (1,2) | (1,1) | l | (1,8) | (7,7) | (1,7) | t | (2,19) | (1,4) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ca | t | _ | a | n | d | _the_ | f | iddle |
| ca | (1,6) | (1,4) | (1,3) | n | (1,14) | (5,12) | f | (5,23) |

Questo “usare le ripetizioni” ha un rapporto con l’entropia...

$\mathcal{C}$ si dice universale se (non rigorosamente...) per ogni processo ergodico

$$\lim_{n\to\infty} \frac{|\mathcal{C}(a_1^n)|}{n} = h(\{X_1^n\}).$$

# Autori come sorgenti di informazione

sorgente d'informazione

...L'indifferenza è il peso morto della storia...

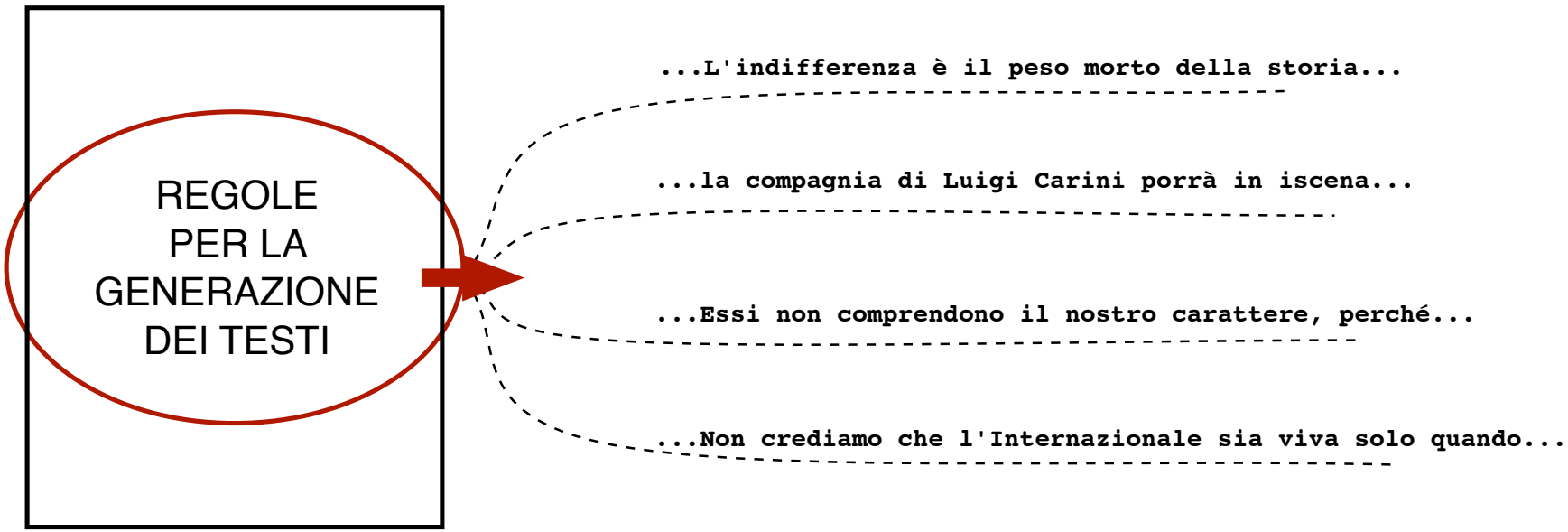
...la compagnia di Luigi Carini porrà in iscena...

...Essi non comprendono il nostro carattere, perché...

...Non crediamo che l'Internazionale sia viva solo quando...

produzioni della sorgente

# Autori come sorgenti di informazione

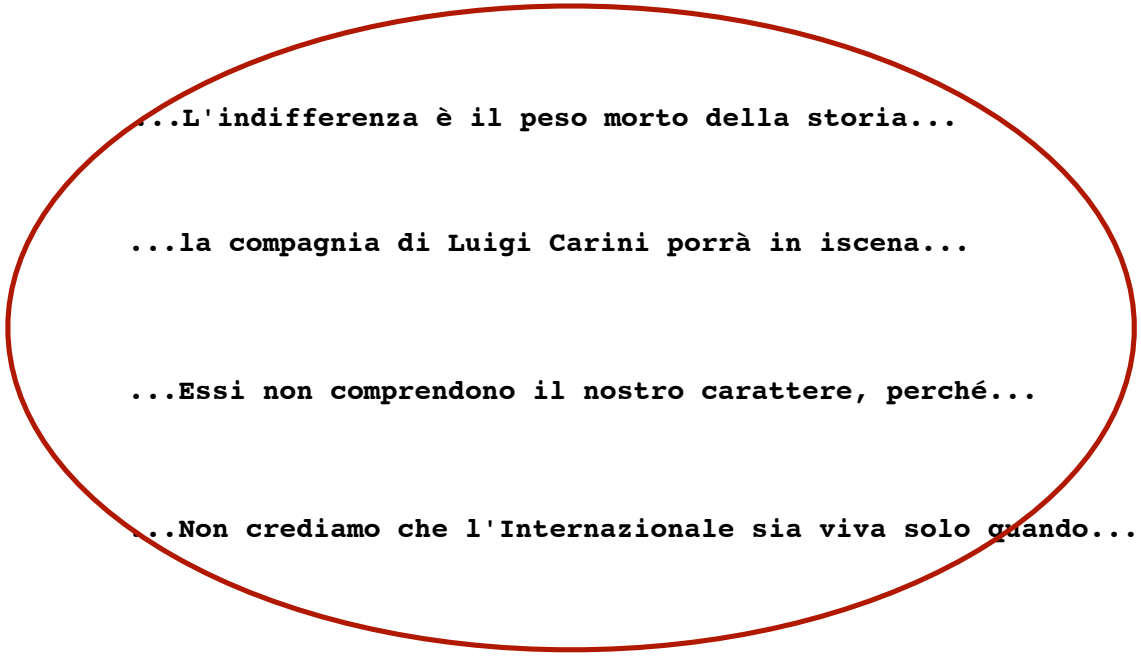



sorgente d'informazione

produzioni della sorgente

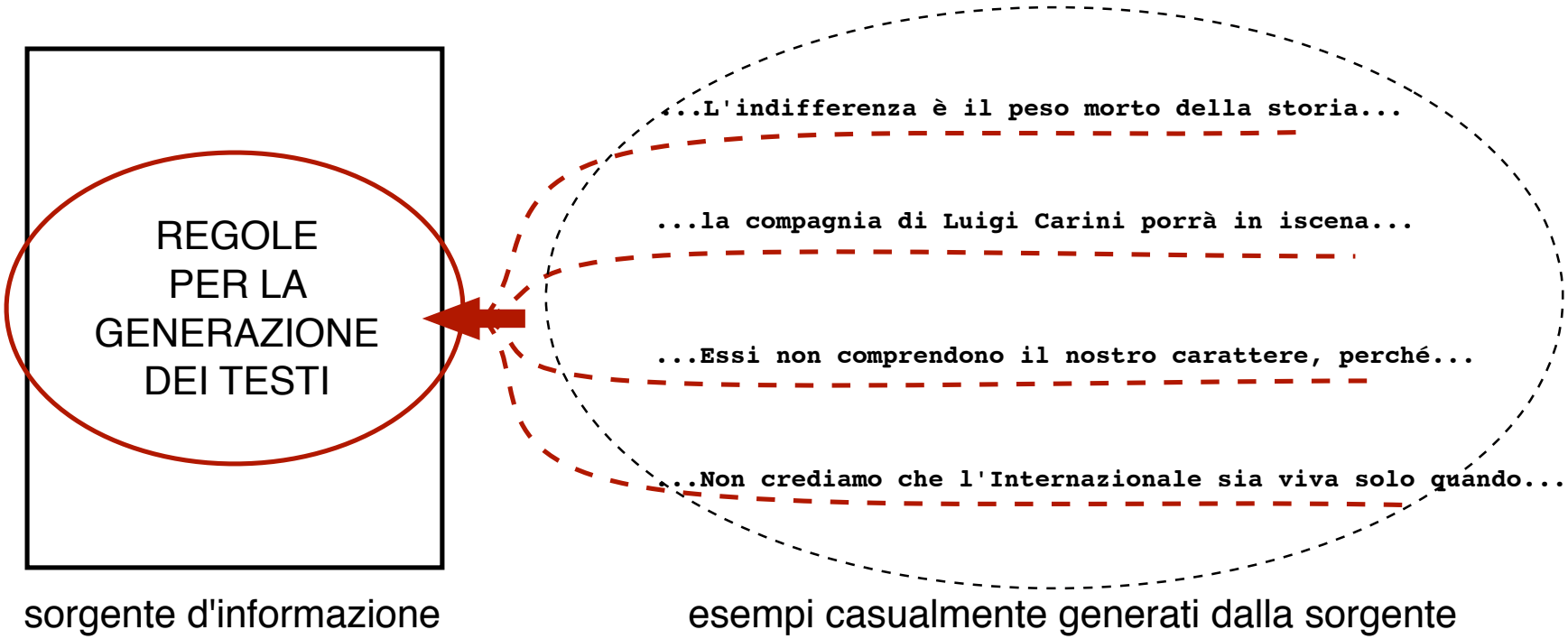
# Autori come sorgenti di informazione


sorgente d'informazione




esempi casualmente generati dalla sorgente

# Autori come sorgenti di informazione




sorgente d'informazione

esempi casualmente generati dalla sorgente

# Confronto tra sorgenti

Per riconoscere l'autore l'entropia della singola sorgente non basta!

| author | work | compression rate |
|---|---|---|
| Dante | Commedia | 3.2 |
| | De Vulgari Eloquentia | 3.0 |
| | Convivio | 2.7 |
| Boccaccio | Decamerone | 2.8 |
| Petrarca | Canzoniere | 3.1 |

# Confronto tra sorgenti

Per riconoscere l'autore l'entropia della singola sorgente non basta!

| author | work | compression rate |
|---|---|---|
| Dante | Commedia | 3.2 |
| | De Vulgari Eloquentia | 3.0 |
| | Convivio | 2.7 |
| Boccaccio | Decamerone | 2.8 |
| Petrarca | Canzoniere | 3.1 |

Quel che ci serve è l' entropia relativa tra due sorgenti $P$ e $Q$:

$$d(Q \parallel P) := \limsup_{n\to\infty} \frac{1}{n} \sum_{a\in\mathcal{A}^n} Q(a) \log \frac{Q(a)}{P(a)},$$

# Confronto tra sorgenti

Per riconoscere l'autore l'entropia della singola sorgente non basta!

| author | work | compression rate |
|---|---|---|
| Dante | Commedia | 3.2 |
| | De Vulgari Eloquentia | 3.0 |
| | Convivio | 2.7 |
| Boccaccio | Decamerone | 2.8 |
| Petrarca | Canzoniere | 3.1 |

Quel che ci serve è l' entropia relativa tra due sorgenti $P$ e $Q$:

$$d(Q \parallel P) := \limsup_{n\to\infty} \frac{1}{n} \sum_{a\in\mathcal{A}^n} Q(a) \log \frac{Q(a)}{P(a)},$$

che esprime "di quanto sbaglio" nel considerare $a$, generata dalla sorgente $Q$, come se fosse stata generata da $P$.

# Confronto tra sorgenti

Per riconoscere l'autore l'entropia della singola sorgente non basta!

| author | work | compression rate |
|---|---|---|
| Dante | Commedia | 3.2 |
| | De Vulgari Eloquentia | 3.0 |
| | Convivio | 2.7 |
| Boccaccio | Decamerone | 2.8 |
| Petrarca | Canzoniere | 3.1 |

Quel che ci serve è l' entropia relativa tra due sorgenti $P$ e $Q$:

$$d(Q \parallel P) := \limsup_{n\to\infty} \frac{1}{n} \sum_{a\in\mathcal{A}^n} Q(a) \log \frac{Q(a)}{P(a)},$$

che esprime "di quanto sbaglio" nel considerare $a$, generata dalla sorgente $Q$, come se fosse stata generata da $P$.

E' una sorta di distanza tra sorgenti!

# Confronto tra sorgenti

Per riconoscere l'autore l'entropia della singola sorgente non basta!

| author | work | compression rate |
|---|---|---|
| Dante | Commedia | 3.2 |
| | De Vulgari Eloquentia | 3.0 |
| | Convivio | 2.7 |
| Boccaccio | Decamerone | 2.8 |
| Petrarca | Canzoniere | 3.1 |

Quel che ci serve è l' entropia relativa tra due sorgenti $P$ e $Q$:

$$d(Q \parallel P) := \limsup_{n\to\infty} \frac{1}{n} \sum_{a\in\mathcal{A}^n} Q(a) \log \frac{Q(a)}{P(a)},$$

che esprime "di quanto sbaglio" nel considerare $a$, generata dalla sorgente $Q$, come se fosse stata generata da $P$.

E' una sorta di distanza tra sorgenti! E se volessi calcolarla?

# Entropia relativa e compressione

Teorema di Ziv & Merhav, 1993:

$$D(Q \parallel P) = \lim_{n \to \infty} \frac{1}{n} \left[ c(y|x) \log n - c(y) \log c(y) \right]$$

# Entropia relativa e compressione

Teorema di Ziv & Merhav, 1993:

$$D(Q \parallel P) = \lim_{n \to \infty} \frac{1}{n} \left[ c(y|x) \log n - c(y) \log c(y) \right],$$

dove:

$$\begin{aligned} x &= \text{sequenza lunga } n \text{ generata da } P; \\ y &= \text{sequenza lunga } n \text{ generata da } Q; \end{aligned}$$

# Entropia relativa e compressione

Teorema di Ziv & Merhav, 1993:

$$D(Q \parallel P) = \lim_{n\to\infty} \frac{1}{n} \left[ c(y|x) \log n - c(y) \log c(y) \right],$$

dove:

$$\begin{aligned} x &= \text{sequenza lunga } n \text{ generata da } P; \\ y &= \text{sequenza lunga } n \text{ generata da } Q; \\ c(y) &\approx \text{lunghezza della versione compressa di } y \text{ con un} \\ &\quad \text{compressore LZ [Lempel \& Ziv, 1976, 1977, 1978];} \\ c(y|x) &\approx \text{lunghezza della versione compressa di } y \text{ ottenuta} \\ &\quad \text{solo con sottosequenze trovate in } x. \end{aligned}$$

# Entropia relativa e compressione

Teorema di Ziv & Merhav, 1993:

$$D(Q \parallel P) = \lim_{n\to\infty} \frac{1}{n} \left[ c(y|x) \log n - c(y) \log c(y) \right],$$

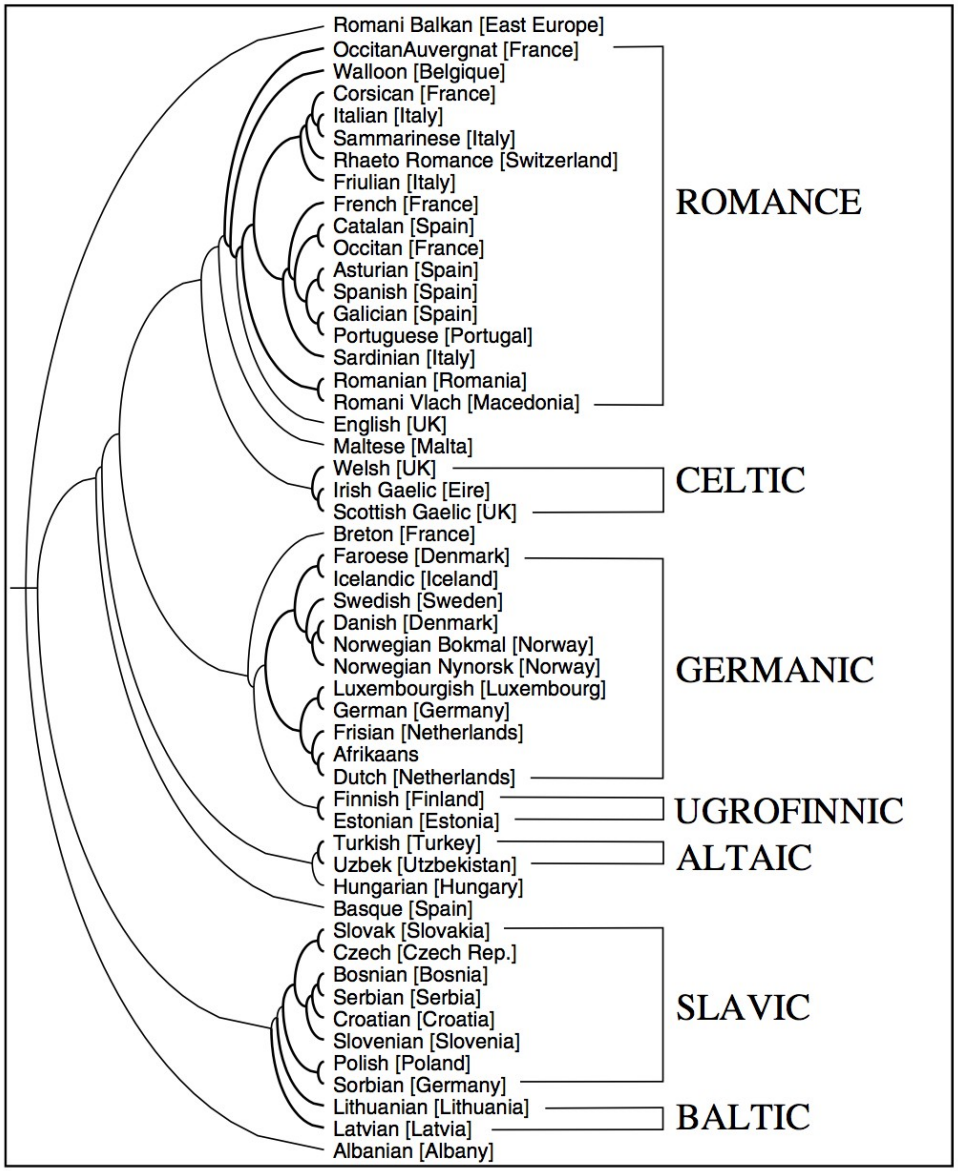
dove:

$$\begin{aligned} x &= \text{sequenza lunga } n \text{ generata da } P; \\ y &= \text{sequenza lunga } n \text{ generata da } Q; \\ c(y) &\approx \text{lunghezza della versione compressa di } y \text{ con un} \\ &\quad \text{compressore LZ [Lempel \& Ziv, 1976, 1977, 1978];} \\ c(y|x) &\approx \text{lunghezza della versione compressa di } y \text{ ottenuta} \\ &\quad \text{solo con sottosequenze trovate in } x. \end{aligned}$$

Posso quindi usare la compressione per approssimare $d$: comprimiamo $y$ utilizzando però $x$ come "dizionario".

[D. Benedetto, E. Caglioti, V. Loreto. Language trees and zipping *Physical Review Letters* **88** (2002)]

# Un esempio: alberi filolinguistici (BCL)




[D. Benedetto, E. Caglioti, V. Loreto. Language trees and zipping *Physical Review Letters* **88** (2002)]

# Antonio Gramsci

(Ales, 1891 - Roma, 1937)
Politico, intellettuale e giornalista, tra i fondatori del Partito Comunista d'Italia.

# Antonio Gramsci

(Ales, 1891 - Roma, 1937)
Politico, intellettuale e giornalista, tra i fondatori del Partito Comunista d'Italia.

Scrisse migliaia di articoli pubblicati su diversi giornali (*L'Ordine Nuovo, Avanti!, ...*)

# Antonio Gramsci

(Ales, 1891 - Roma, 1937)
Politico, intellettuale e giornalista, tra i fondatori del Partito Comunista d'Italia.

Scrisse migliaia di articoli pubblicati su diversi giornali (*L'Ordine Nuovo, Avanti!, ...*), gran parte dei quali anonimi.

# Antonio Gramsci

(Ales, 1891 - Roma, 1937)
Politico, intellettuale e giornalista, tra i fondatori del Partito Comunista d'Italia.

Scrisse migliaia di articoli pubblicati su diversi giornali (*L'Ordine Nuovo, Avanti!, ...*), gran parte dei quali anonimi.
Lo stesso fecero i suoi colleghi e compagni: Amedeo Bordiga, Palmiro Togliatti, Angelo Tasca...

# Il Progetto Gramsci

## Antonio Gramsci

(Ales, 1891 - Roma, 1937)
Politico, intellettuale e giornalista, tra i fondatori del Partito Comunista d'Italia.

Scrisse migliaia di articoli pubblicati su diversi giornali (*L'Ordine Nuovo, Avanti!, ...*), gran parte dei quali anonimi.
Lo stesso fecero i suoi colleghi e compagni: Amedeo Bordiga, Palmiro Togliatti, Angelo Tasca...

Progetto Gramsci (con *Istituto Fondazione Gramsci*, Roma): riconoscimento degli articoli gramsciani in vista di un'Edizione Nazionale dell'opera di Gramsci.

# Il Progetto Gramsci

## Antonio Gramsci

(Ales, 1891 - Roma, 1937)
Politico, intellettuale e giornalista, tra i fondatori del Partito Comunista d'Italia.

Scrisse migliaia di articoli pubblicati su diversi giornali (*L'Ordine Nuovo, Avanti!, ...*), gran parte dei quali anonimi.
Lo stesso fecero i suoi colleghi e compagni: Amedeo Bordiga, Palmiro Togliatti, Angelo Tasca...

Progetto Gramsci (con *Istituto Fondazione Gramsci*, Roma): riconoscimento degli articoli gramsciani in vista di un'Edizione Nazionale dell'opera di Gramsci.

[C. Basile, D. Benedetto, E. Caglioti, M. Degli Esposti. An example of mathematical authorship attribution. *Journal of Mathematical Physics*. **49** (2008)]
[C. Basile, M. Lana. L'attribuzione di testi con metodi quantitativi: riconoscimento di testi gramsciani. *AIDAinformazioni*, **1-2** (2008)]

# Il metodo: distanze di similarità

Il metodo usato per il Progetto Gramsci si compone di alcuni passi:

1. misura di quantità significative nei testi $\longrightarrow$ $n-$grammi, cioè sequenze qualsiasi di $n$ caratteri consecutivi (nessuna struttura grammaticale o sintattica viene considerata);

# Il metodo: distanze di similarità

Il metodo usato per il Progetto Gramsci si compone di alcuni passi:

1. misura di quantità significative nei testi $\longrightarrow$ $n-$grammi, cioè sequenze qualsiasi di $n$ caratteri consecutivi (nessuna struttura grammaticale o sintattica viene considerata);
2. calcolo di (pseudo-)distanze di similarità che esprimano la vicinanza tra la sorgente nota (i.e. alcuni testi di riferimento, gramsciani e non) e i testi da attribuire;

# Il metodo: distanze di similarità

Il metodo usato per il Progetto Gramsci si compone di alcuni passi:

1. misura di quantità significative nei testi $\longrightarrow$ $n-$grammi, cioè sequenze qualsiasi di $n$ caratteri consecutivi (nessuna struttura grammaticale o sintattica viene considerata);
2. calcolo di (pseudo-)distanze di similarità che esprimano la vicinanza tra la sorgente nota (i.e. alcuni testi di riferimento, gramsciani e non) e i testi da attribuire;
3. attribuzione ad un dato autore sulla base della distanza "media" del testo incognito dai testi di riferimento di quell'autore.

# Il metodo: distanze di similarità

Il metodo usato per il Progetto Gramsci si compone di alcuni passi:

1. misura di quantità significative nei testi $\longrightarrow$ $n-$grammi, cioè sequenze qualsiasi di $n$ caratteri consecutivi (nessuna struttura grammaticale o sintattica viene considerata);
2. calcolo di (pseudo-)distanze di similarità che esprimano la vicinanza tra la sorgente nota (i.e. alcuni testi di riferimento, gramsciani e non) e i testi da attribuire;
3. attribuzione ad un dato autore sulla base della distanza "media" del testo incognito dai testi di riferimento di quell'autore.

Abbiamo scelto di usare due distanze:

# Il metodo: distanze di similarità

Il metodo usato per il Progetto Gramsci si compone di alcuni passi:

1. misura di quantità significative nei testi $\longrightarrow$ $n-$grammi, cioè sequenze qualsiasi di $n$ caratteri consecutivi (nessuna struttura grammaticale o sintattica viene considerata);
2. calcolo di (pseudo-)distanze di similarità che esprimano la vicinanza tra la sorgente nota (i.e. alcuni testi di riferimento, gramsciani e non) e i testi da attribuire;
3. attribuzione ad un dato autore sulla base della distanza "media" del testo incognito dai testi di riferimento di quell'autore.

Abbiamo scelto di usare due distanze:

- una basata sulla statistica degli $n-$grammi di lunghezza fissata;

# Il metodo: distanze di similarità

Il metodo usato per il Progetto Gramsci si compone di alcuni passi:

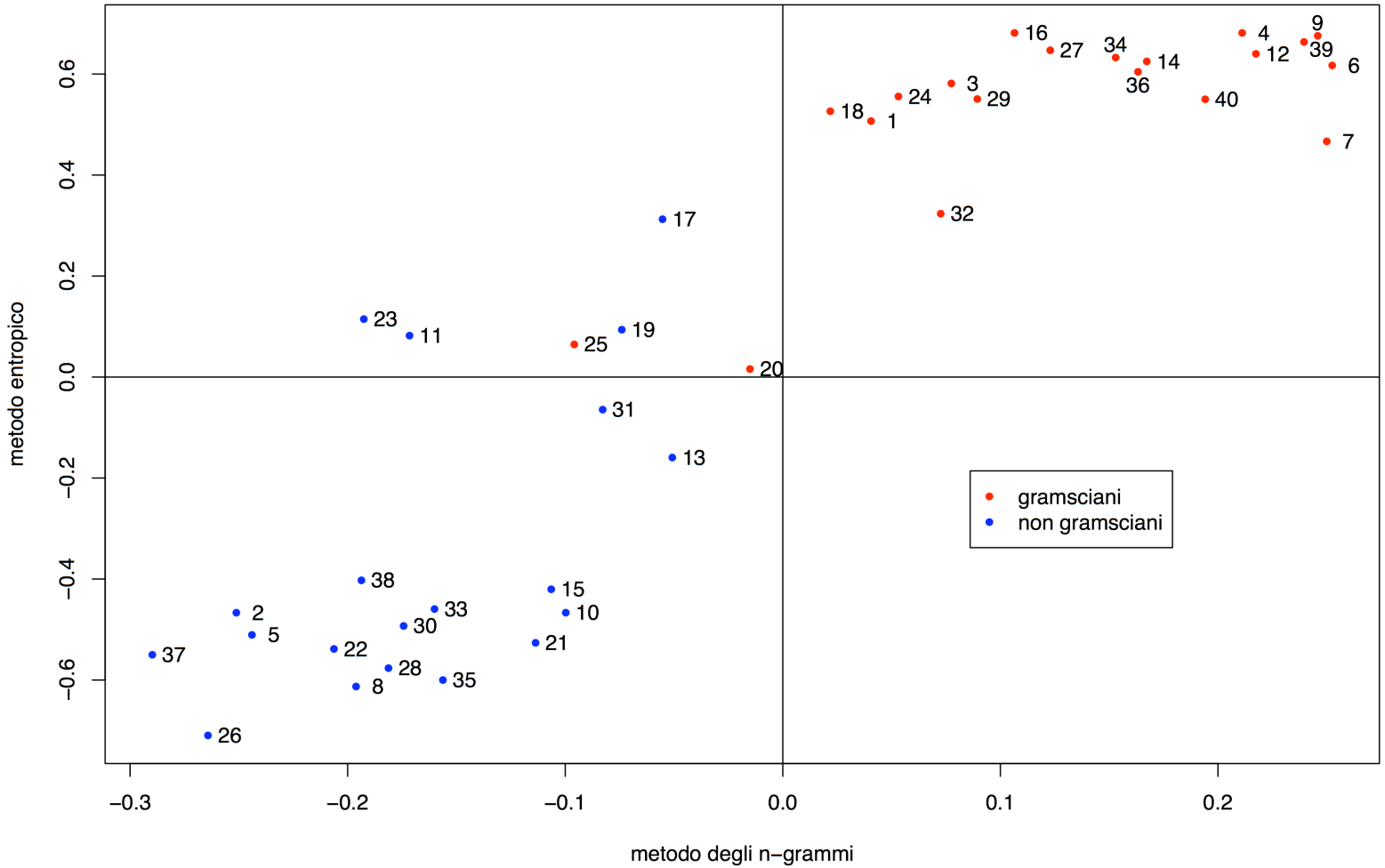
1. misura di quantità significative nei testi $\longrightarrow$ $n-$grammi, cioè sequenze qualsiasi di $n$ caratteri consecutivi (nessuna struttura grammaticale o sintattica viene considerata);
2. calcolo di (pseudo-)distanze di similarità che esprimano la vicinanza tra la sorgente nota (i.e. alcuni testi di riferimento, gramsciani e non) e i testi da attribuire;
3. attribuzione ad un dato autore sulla base della distanza "media" del testo incognito dai testi di riferimento di quell'autore.

Abbiamo scelto di usare due distanze:

- una basata sulla statistica degli $n-$grammi di lunghezza fissata;
- una basata sul Teorema di Ziv e Merhav, cioè un'approssimazione dell'entropia relativa tramite compressori LZ (BCL).

# Progetto Gramsci: i risultati

- asse orizzontale: voto ottenuto con la distanza degli 8-grammi
- asse verticale: voto ottenuto con la distanza entropica



# Eccetera eccetera...

Plagio: *la pratica di prendere il lavoro e le idee di qualcun'altro e di passarle come proprie.*

# Eccetera eccetera...

Plagio: *la pratica di prendere il lavoro e le idee di qualcun'altro e di passarle come proprie.* (Oxford Dictionary)

# Eccetera eccetera...

Plagio: *la pratica di prendere il lavoro e le idee di qualcun'altro e di passarle come proprie.* (Oxford Dictionary)

Supponiamo di avere un testo sospettato di plagio, e un dataset enorme di possibili fonti (es. il web). Come selezionare pochi testi rilevanti su cui fare un'analisi dettagliata?

# Eccetera eccetera...

Plagio: *la pratica di prendere il lavoro e le idee di qualcun'altro e di passarle come proprie.* (Oxford Dictionary)

Supponiamo di avere un testo sospettato di plagio, e un dataset enorme di possibili fonti (es. il web). Come selezionare pochi testi rilevanti su cui fare un'analisi dettagliata?

Alcune proposte:

- usare l'entropia relativa tra le distribuzioni delle frequenze delle parole nel testo sospetto e in quelli di riferimento;

# Eccetera eccetera...

Plagio: *la pratica di prendere il lavoro e le idee di qualcun'altro e di passarle come proprie.* (Oxford Dictionary)

Supponiamo di avere un testo sospettato di plagio, e un dataset enorme di possibili fonti (es. il web). Come selezionare pochi testi rilevanti su cui fare un'analisi dettagliata?

Alcune proposte:

- usare l'entropia relativa tra le distribuzioni delle frequenze delle parole nel testo sospetto e in quelli di riferimento;
- confrontare con qualche misura di similarità i dizionari degli $n-$grammi dei due testi...

# Eccetera eccetera...

Plagio: *la pratica di prendere il lavoro e le idee di qualcun'altro e di passarle come proprie.* (Oxford Dictionary)

Supponiamo di avere un testo sospettato di plagio, e un dataset enorme di possibili fonti (es. il web). Come selezionare pochi testi rilevanti su cui fare un'analisi dettagliata?

Alcune proposte:

- usare l'entropia relativa tra le distribuzioni delle frequenze delle parole nel testo sospetto e in quelli di riferimento;
- confrontare con qualche misura di similarità i dizionari degli $n-$grammi dei due testi...


[A. Barrón-Cedeño, P. Rosso, J.M. Benedí. Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance. *Proceedings of CICling 2009*]
[C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, M. Degli Esposti. A plagiarism detection procedure in three steps: selection, matches and "squares". *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* (2009).]

# Eccetera eccetera...

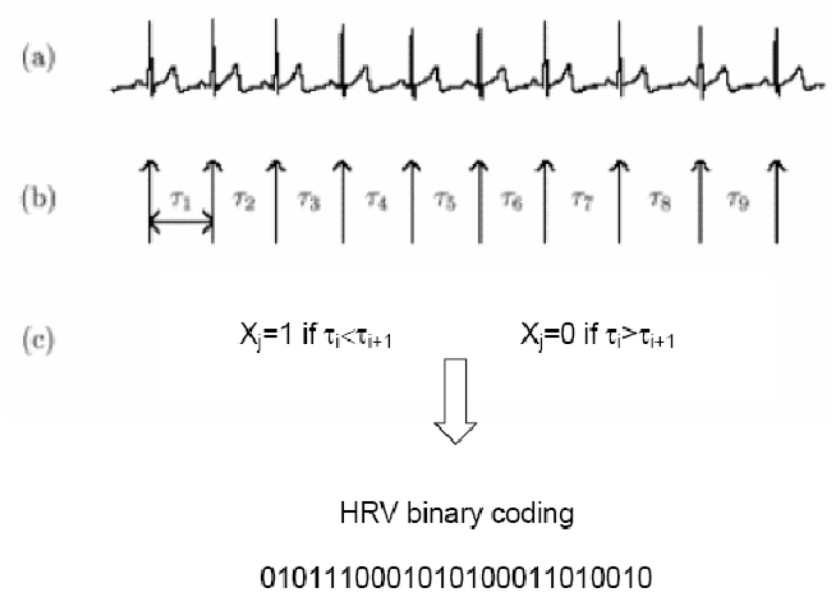
I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)




[M. Degli Esposti, C. Farinelli, M. Manca, A. Tolomelli. A similarity measure for biological signals: new applications to HRV analysis. *Japan Journal of Biostatistics* **1** (2007)]

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)
- suddivisione per genere di brani musicali

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)
- suddivisione per genere di brani musicali
- estrazione di informazione medica da referti ospedalieri, codifica, associazione a casi simili e letteratura medica rilevante

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)
- suddivisione per genere di brani musicali
- estrazione di informazione medica da referti ospedalieri, codifica, associazione a casi simili e letteratura medica rilevante
- creazione automatica di riassunti di romanzi o articoli scientifici

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)
- suddivisione per genere di brani musicali
- estrazione di informazione medica da referti ospedalieri, codifica, associazione a casi simili e letteratura medica rilevante
- creazione automatica di riassunti di romanzi o articoli scientifici
- classificazione di immagini

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)
- suddivisione per genere di brani musicali
- estrazione di informazione medica da referti ospedalieri, codifica, associazione a casi simili e letteratura medica rilevante
- creazione automatica di riassunti di romanzi o articoli scientifici
- classificazione di immagini
- confronto di testi sulla base dei loro grafi degli $n-$grammi...

# Eccetera eccetera...

I problemi di estrazione automatica di informazioni da sequenze simboliche sono praticamente infiniti: quasi tutto è codificabile come sequenza!

Qualche esempio in ordine sparso:

- classificazione di sequenze biologiche (HRV) o genetiche (DNA)
- suddivisione per genere di brani musicali
- estrazione di informazione medica da referti ospedalieri, codifica, associazione a casi simili e letteratura medica rilevante
- creazione automatica di riassunti di romanzi o articoli scientifici
- classificazione di immagini
- confronto di testi sulla base dei loro grafi degli $n-$grammi...

Ce n'è per tutti i gusti!!

# Sorgenti Markoviane

Shannon fa l'ulteriore ipotesi semplificativa che ogni sorgente "reale" sia Markoviana, cioè che per ogni $n \in \mathbb{N}, a_1, \ldots, a_{n+1} \in \mathcal{A}$:

$$P(X_{n+1} = a_{n+1} | X_1^n = a_1^n) = P(X_{n+1} = a_{n+1} | X_n = a_n),$$

ovvero la sorgente ha memoria finita nel tempo: "ricorda" solo l'ultimo valore che ha emesso.

# Sorgenti Markoviane

Shannon fa l'ulteriore ipotesi semplificativa che ogni sorgente "reale" sia Markoviana, cioè che per ogni $n \in \mathbb{N}, a_1, \ldots, a_{n+1} \in \mathcal{A}$:

$$P(X_{n+1} = a_{n+1} | X_1^n = a_1^n) = P(X_{n+1} = a_{n+1} | X_n = a_n),$$

ovvero la sorgente ha memoria finita nel tempo: "ricorda" solo l'ultimo valore che ha emesso.
Ciò si generalizza facilmente a memoria $k, k \in \mathbb{N}$.

# Sorgenti Markoviane

Shannon fa l'ulteriore ipotesi semplificativa che ogni sorgente "reale" sia Markoviana, cioè che per ogni $n \in \mathbb{N}, a_1, \ldots, a_{n+1} \in \mathcal{A}$:

$$P(X_{n+1} = a_{n+1} | X_1^n = a_1^n) = P(X_{n+1} = a_{n+1} | X_n = a_n),$$

ovvero la sorgente ha memoria finita nel tempo: "ricorda" solo l'ultimo valore che ha emesso.
Ciò si generalizza facilmente a memoria $k, k \in \mathbb{N}$.

▸ perché sorgenti Markoviane?

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

$k = 0$ `ttkdnnc,t ou u m hvioega t,tna keseilra`

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

$k = 0$ `ttkdnnc,t ou u m hvioega t,tna keseilra`

$k = 1$ `fin my then i win blo his owe 'se a pe p`

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

$k = 0$ `ttkdnnc,t ou u m hvioega t,tna keseilra`

$k = 1$ `fin my then i win blo his owe 'se a pe p`

$k = 4$ `where as added, in recollections, may now how him going artific chap`

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

$k = 0$ `ttkdnnc,t ou u m hvioega t,tna keseilra`

$k = 1$ `fin my then i win blo his owe 'se a pe p`

$k = 4$ `where as added, in recollections, may now how him going artific chap`

$k = 10$ `by this time, estella left me stand aside, to see if she could be easier for the wash; that's a blazing fire.`

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

$k = 0$ `ttkdnnc,t ou u m hvioega t,tna keseilra`

$k = 1$ `fin my then i win blo his owe 'se a pe p`

$k = 4$ `where as added, in recollections, may now how him going artific chap`

$k = 10$ `by this time, estella left me stand aside, to see if she could be easier for the wash; that's a blazing fire.`

# Testi artificiali

Costruiamo la distribuzione Markoviana calcolando empiricamente le frequenze in *Oliver Twist, David Copperfield, Great Expectations* e *A Tale of Two Cities* di Charles Dickens, per un totale di $\sim$ 4.5 milioni di caratteri.

$k = 0$ `ttkdnnc,t ou u m hvioega t,tna keseilra`

$k = 1$ `fin my then i win blo his owe 'se a pe p`

$k = 4$ `where as added, in recollections, may now how him going artific chap`

$k = 10$ `by this time, estella left me stand aside, to see if she could be easier for the wash; that's a blazing fire.`

back