



Università degli Studi di Bari Aldo Moro  
Dipartimento di Informatica

Anno Accademico 2013/2014

# **Piattaforma cloud per l'analisi di big data provenienti da social network**

*Tecnologie per l'individuazione di contenuti a sfondo discriminatorio*

Relatori:

Prof. Pasquale Lops

Dott. Pierpaolo Basile

Laureando:

Gianvito Taneburgo



## ■ Grande mole di dati **non strutturati**:

- video
- immagini
- log
- email
- sensori
- like +1
- ecc.

Unità di misura	Simbolo	Valore (byte)
gigabyte	GB	$10^9$
terabyte	TB	$10^{12}$
petabyte	PB	$10^{15}$
exabyte	EB	$10^{18}$

- Diffusione Internet
- Aumento velocità media download/upload
- Aumento potere computazionale
- Aumento capacità di memorizzazione
- Sviluppo dispositivi e reti mobili

*Big Data means the data is large enough that you have to think about it in order to gain insights from it.*

- Sviluppare applicazioni per estrarre **valore** dai **dati grezzi**

- **Brand reputation analysis**
- Recommender system
- Targeted advertising
- Query log mining



- Brand reputation analysis
- Recommender system
- Targeted advertising
- Query log mining

## More Top Picks for You



Windows 7 Home Premium SP1  
64bit...  
Windows  
★★★★☆ (1,268)  
~~\$128.70~~ **\$102.95**



Asus P8Z77-V LK Intel Z77 DDR3  
LGA...  
★★★★☆ (100)  
**\$186.12**



Intel Core i7-3770K Quad-Core...  
★★★★★ (345)  
~~\$400.00~~ **\$325.99**



Corsair Vengeance 8 GB DDR3 1600  
MHz...  
★★★★☆ (547)



Gigabyte Intel Z77 LGA 1155 AMD...  
★★★★☆ (126)



Samsung 840 Series 2.5 inch  
120GB...  
★★★★☆ (1,483)  
**\$249.98**

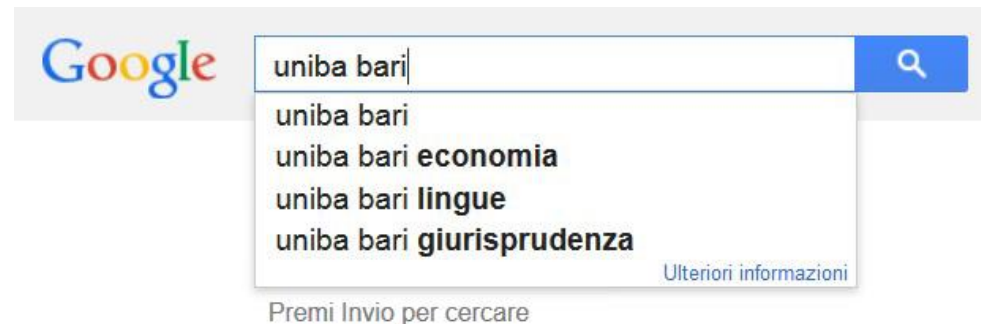
> [View your shopping cart](#)

- Brand reputation analysis
- Recommender system
- **Targeted advertising**
- Query log mining





- Brand reputation analysis
- Recommender system
- Targeted advertising
- Query log mining



ed informatica?



Strumenti adatti ai dati



**Memorizzare** big data

**Elaborare** big data



## Tecnologie **scalabili**



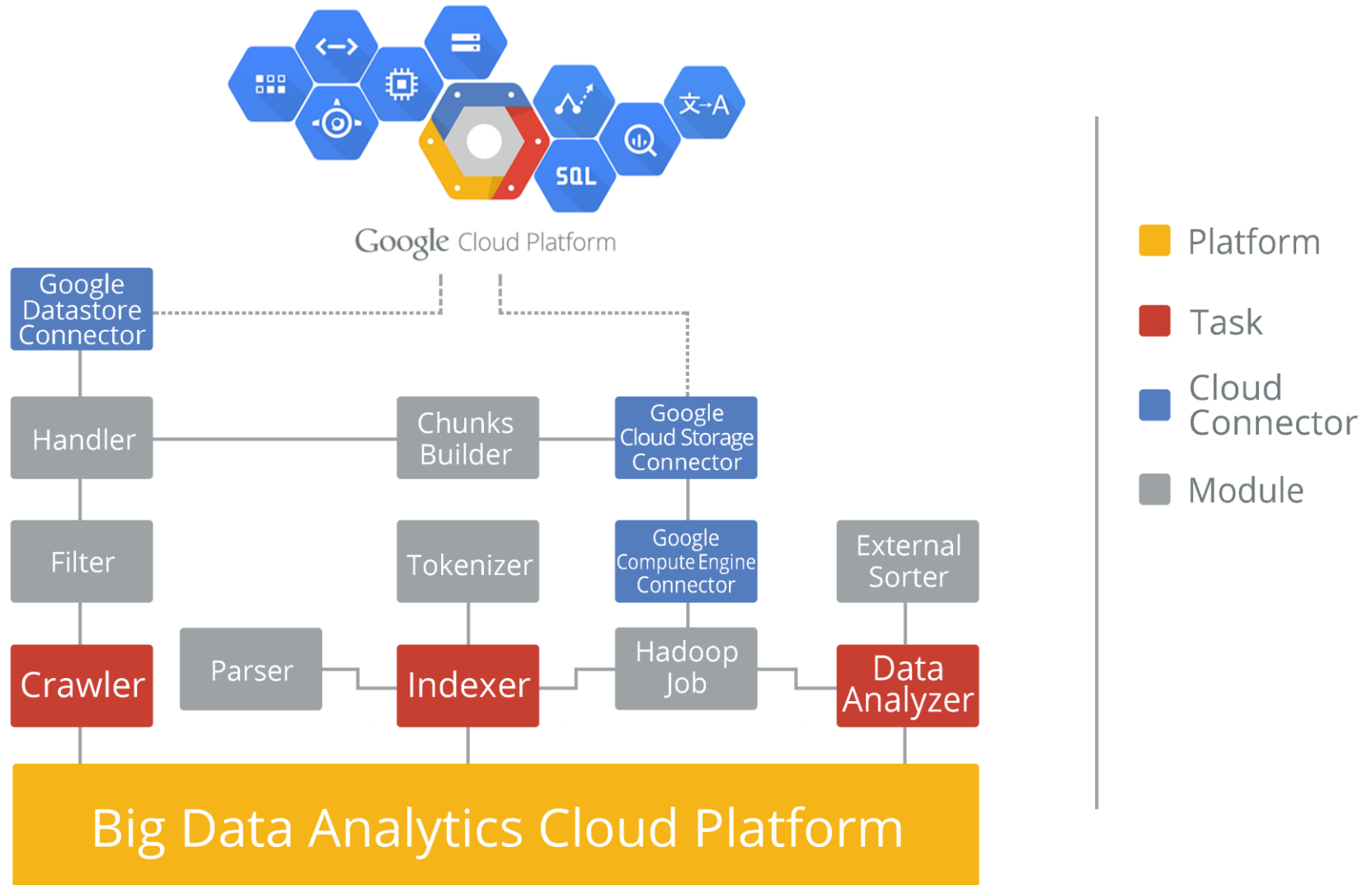
## Piattaforme di **cloud** computing



Google Cloud Platform

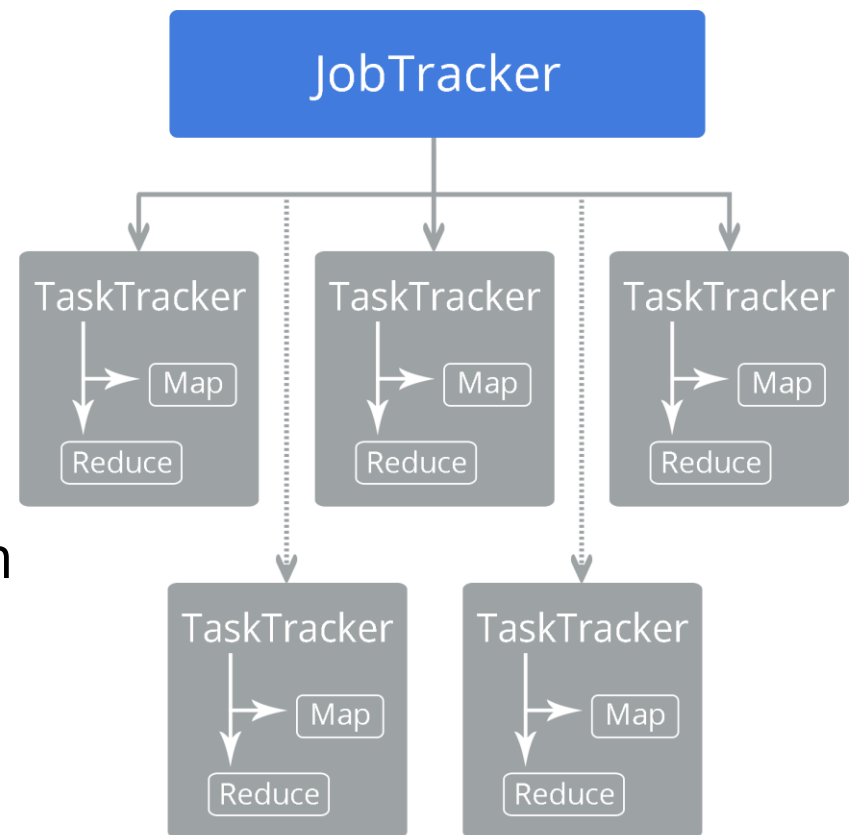
SOFTLAYER®  
an IBM Company

# Architettura della piattaforma



Framework per il **calcolo distribuito** tra nodi di un cluster

- Architettura master/slave
- Modello di programmazione MapReduce
  - ▣ **map**: elaborazione locale
  - ▣ **reduce**: aggrega i risultati
- Hadoop Distributed File System
- Commodity hardware
- Scalabilità orizzontale



- Map:

$(K, V) \rightarrow \text{list}(I, P)$

- Reduce:

$(I, \text{list}(P)) \rightarrow \text{list}(P)$

Esempio: indicizzazione di un documento.

- Map:

$(\text{int}, \text{String}) \rightarrow \text{list}(\text{String}, \text{int})$

```
map (int lineNumber, String line) :
```

```
    for each word w in line:
```

```
        Emit Intermediate (w, 1);
```

- Reduce:

$(\text{String}, \text{list}(\text{int})) \rightarrow (\text{String}, \text{int})$

```
reduce (String w, List<Integer> values) :
```

```
    long count = 0;
```

```
    for each value v in values:
```

```
        count += v;
```

```
    Emit (w, count);
```

«

*Il link per la tesi è questo <http://goo.gl/Gojmx2> ed il link per la presentazione è questo <http://goo.gl/7ozuaA>.*

*Il contatto è questo [taneburgo@gmail.com](mailto:taneburgo@gmail.com)*

»

Termine	Frequenza	Termine	Frequenza
Il	2	<a href="http://goo.gl/Gojmx2">http://goo.gl/Gojmx2</a>	1
link	2	ed	1
per	2	il	1
la	2	presentazione	1
tesi	1	<a href="http://goo.gl/7ozuaA">http://goo.gl/7ozuaA</a>	1
è	3	contatto	1
questo	3	<a href="mailto:taneburgo@gmail.com">taneburgo@gmail.com</a>	1

## ■ Map input:



*Il link per la tesi è questo <http://goo.gl/Gojmx2>  
ed il link per la presentazione è questo  
<http://goo.gl/7ozuaA>.*

## ■ Map output:

(Il, 1)	(il, 1)
(link, 1)	(link, 1)
(per, 1)	(per, 1)
(la, 1)	(la, 1)
(tesi, 1)	(presentazione, 1)
(è, 1)	(è, 1)
(questo, 1)	(questo, 1)
( <a href="http://goo.gl/Gojmx2">http://goo.gl/Gojmx2</a> , 1)	( <a href="http://goo.gl/7ozuaA">http://goo.gl/7ozuaA</a> , 1)
(ed, 1)	

## ■ Map input:

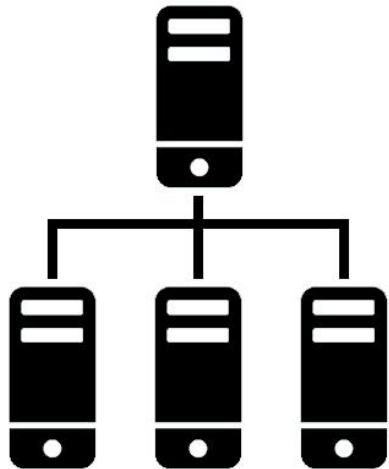


*Il contatto è questo  
[taneburgo@gmail.com](mailto:taneburgo@gmail.com)*

## ■ Map output:

(Il, 1)
(contatto, 1)
(è, 1)
(questo, 1)
( <a href="mailto:taneburgo@gmail.com">taneburgo@gmail.com</a> , 1)





## ■ Reduce input:

(ll, [1, 1])

(link, [1, 1])

(per, [1, 1])

(la, [1, 1])

(tesi, [1])

(è, [1, 1, 1])

(questo, [1, 1, 1])

(<http://goo.gl/Gojmx2>, [1])

(ed, [1])

(il, [1])

(presentazione, [1])

(<http://goo.gl/7ozuaA>, [1])

(contatto, [1])

([taneburgo@gmail.com](mailto:taneburgo@gmail.com), [1])

## ■ Reduce output:

(ll, 2)

(link, 2)

(per, 2)

(la, 2)

(tesi, 1)

(è, 3)

(questo, 3)

(<http://goo.gl/Gojmx2>, 1)

(ed, 1)

(il, 1)

(presentazione, 1)

(<http://goo.gl/7ozuaA>, 1)

(contatto, 1)

([taneburgo@gmail.com](mailto:taneburgo@gmail.com), 1)



## ■ Map input:

*Il link per la tesi è questo <http://goo.gl/Gojmx2>  
ed il link per la presentazione è questo  
<http://goo.gl/7ozuaA>.*

## ■ Map output:

(Il, 1)	(il, 1)
(link, 1)	(link, 1)
(per, 1)	(per, 1)
(la, 1)	(la, 1)
(tesi, 1)	(presentazione, 1)
(è, 1)	(è, 1)
(questo, 1)	(questo, 1)
( <a href="http://goo.gl/Gojmx2">http://goo.gl/Gojmx2</a> , 1)	( <a href="http://goo.gl/7ozuaA">http://goo.gl/7ozuaA</a> , 1)
(ed, 1)	



## ■ Map input:

*Il contatto è questo  
[taneburgo@gmail.com](mailto:taneburgo@gmail.com)*

## ■ Map output:

(Il, 1)
(contatto, 1)
(è, 1)
(questo, 1)
( <a href="mailto:taneburgo@gmail.com">taneburgo@gmail.com</a> , 1)



- Combine input =  
Map output



- Combine input =  
Map output

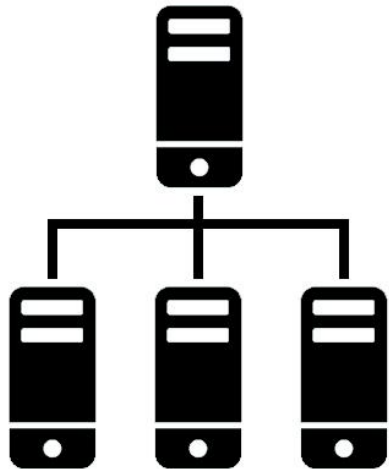
- Combine output:

(Il, 1)  
(link, 2)  
(per, 2)  
(la, 2)  
(tesi, 1)  
(è, 2)  
(questo, 2)  
(<http://goo.gl/Gojmx2>, 1)  
(ed, 1)

(il, 1)  
(presentazione, 1)  
(<http://goo.gl/7ozuaA>, 1)

- Combine output:

(Il, 1)  
(contatto, 1)  
(è, 1)  
(questo, 1)  
(taneburgo@gmail.com, 1)



## ■ Reduce input:

(ll, [1, 1])

(link, [2])

(per, [2])

(la, [2])

(tesi, [1])

(è, [2, 1])

(questo, [2, 1])

(<http://goo.gl/Gojmx2>, [1])

(ed, [1])

(il, [1])

(presentazione, [1])

(<http://goo.gl/7ozuaA>, [1])

(contatto, [1])

([taneburgo@gmail.com](mailto:taneburgo@gmail.com), [1])

## ■ Reduce output:

(ll, 2)

(link, 2)

(per, 2)

(la, 2)

(tesi, 1)

(è, 3)

(questo, 3)

(<http://goo.gl/Gojmx2>, 1)

(ed, 1)

(il, 1)

(presentazione, 1)

(<http://goo.gl/7ozuaA>, 1)

(contatto, 1)

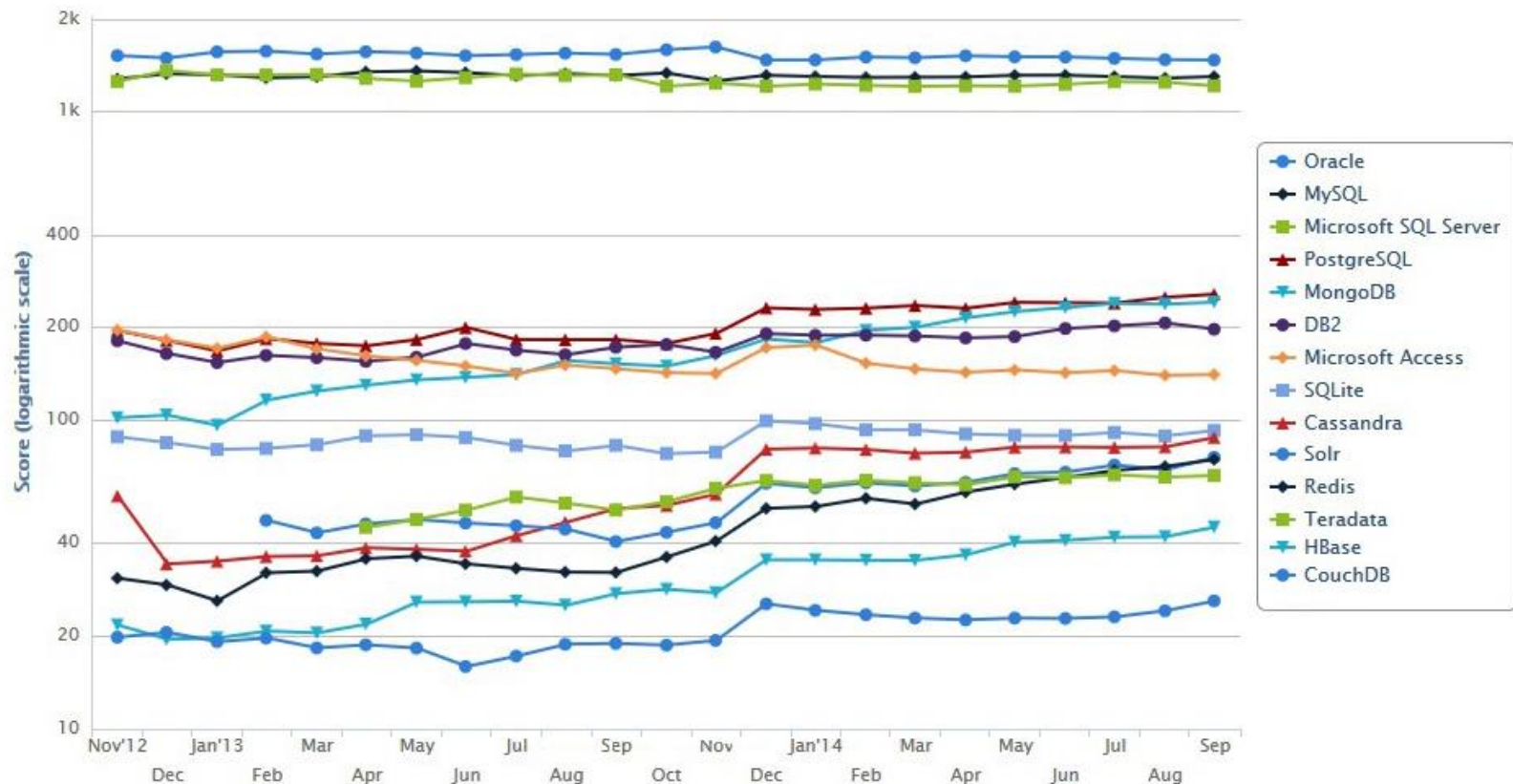
([taneburgo@gmail.com](mailto:taneburgo@gmail.com), 1)

## Vantaggi

- scalabilità orizzontale
- modello flessibile dei dati

## Svantaggi

- transazioni ACID → BASE
- teorema CAP



*Classifica dei DBMS per numero di installazioni*



## Google Cloud Platform



- Infrastructure-as-a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)
- Database-as-a-Service (DaaS)



## Google Cloud Platform



- Compute Engine: IaaS
- Cloud Storage: IaaS
- Datastore: DaaS

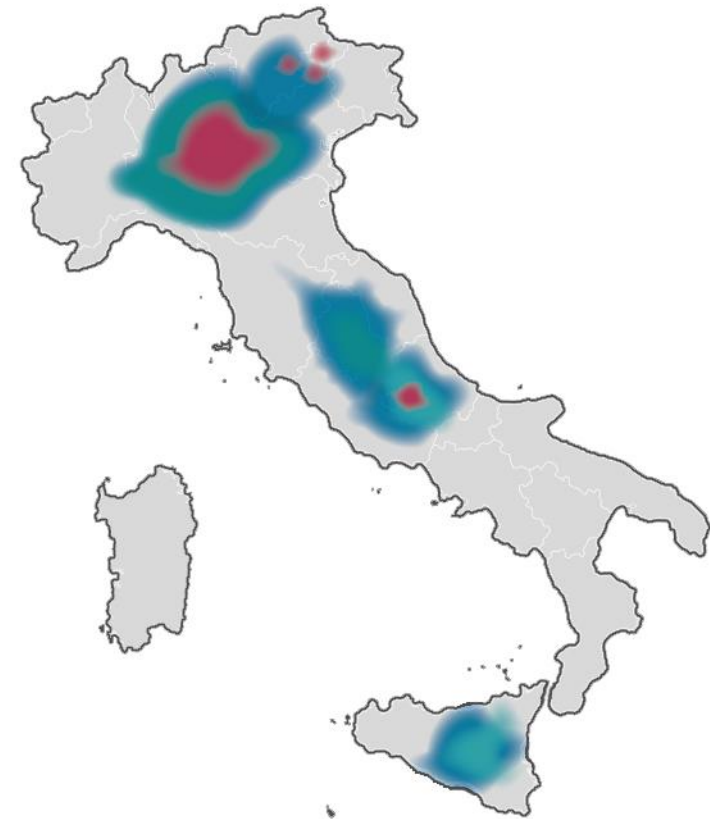


# La «Mappa dell'Intolleranza»

- Quanto siamo razzisti?
- Quanto siamo omofobi?
- Quanto discriminiamo il prossimo?

Nel 2013:

- 45% dei giovani si è dichiarato xenofobo o diffidente degli stranieri
- 92% delle persone LGBT discriminate per l'orientamento sessuale
- 25% degli omosessuali e 6.743.000 donne vittime di violenze

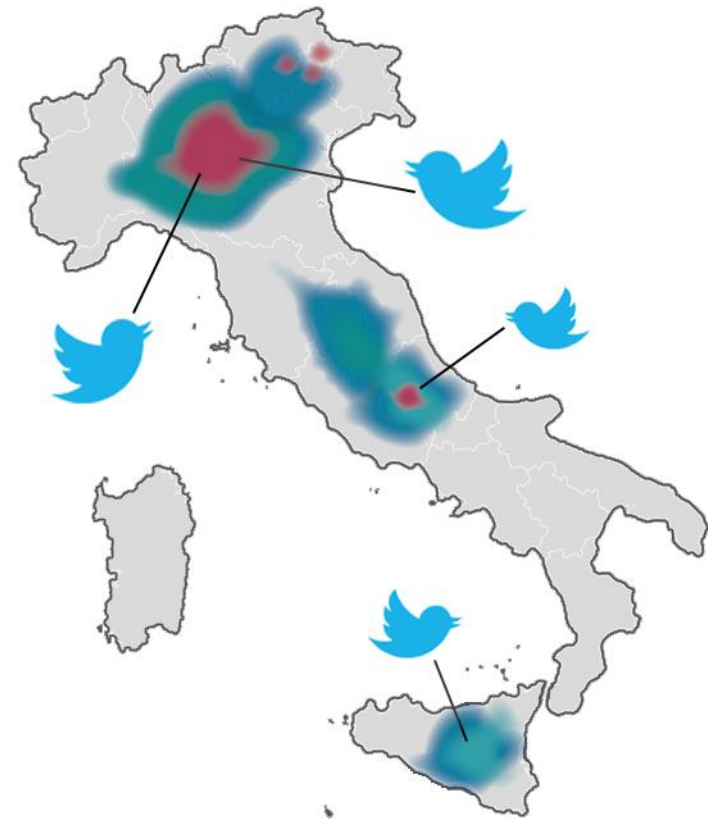


# La «Mappa dell'Intolleranza»

- Quanto siamo razzisti?
- Quanto siamo omofobi?
- Quanto discriminiamo il prossimo?

Nel 2013:

- 45% dei giovani si è dichiarato xenofobo o diffidente degli stranieri
- 92% delle persone LGBT discriminate per l'orientamento sessuale
- 25% degli omosessuali e 6.743.000 donne vittime di violenze



# Raccolta preliminare dei tweet

OMOFOBIA	RAZZISMO	DISABILITÀ	MISOGINIA	ANTISEMITISMO
finocchio	neg*o	nano	baldr**ca	ebreo ai forni
ricch***e	terrone	storpio	zoc**la	rabbino
fr***o	zingaro	spastico	boc***nara	giudeo
rotti***lo	muso giallo	zoppo	tro**na	ebreo di me**a
cul***one	crucco	cerebroleso	mign**ta	

*Esempi di seed per classi di discriminazione*



# 1<sup>st</sup> Research question

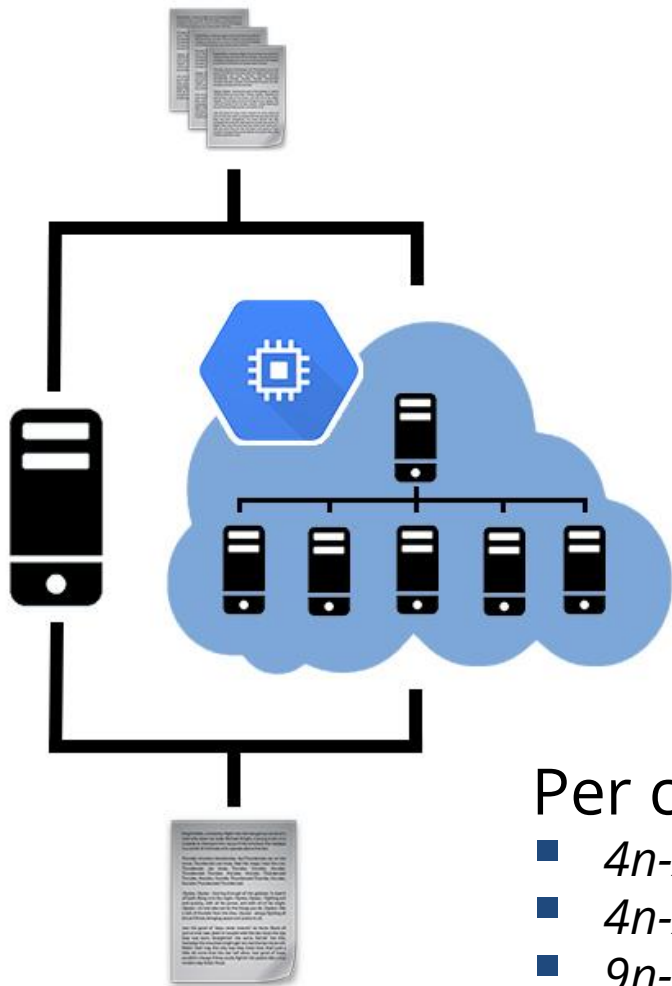
Quantificare vantaggi e svantaggi di una soluzione cloud per l'elaborazione distribuita di **grandi** e **piccoli** dataset

Collezione	Contenuto	Dimensione
itWaC	1.870.000 documenti	11,16 GB
omofobia	22.564 tweet	1.738 kB
disabilità	29.793 tweet	2.242 kB
misoginia	249.425 tweet	17.935 kB

*Dati di input dell'esperimento*

Confronto dei tempi medi di indicizzazione con Hadoop tra:

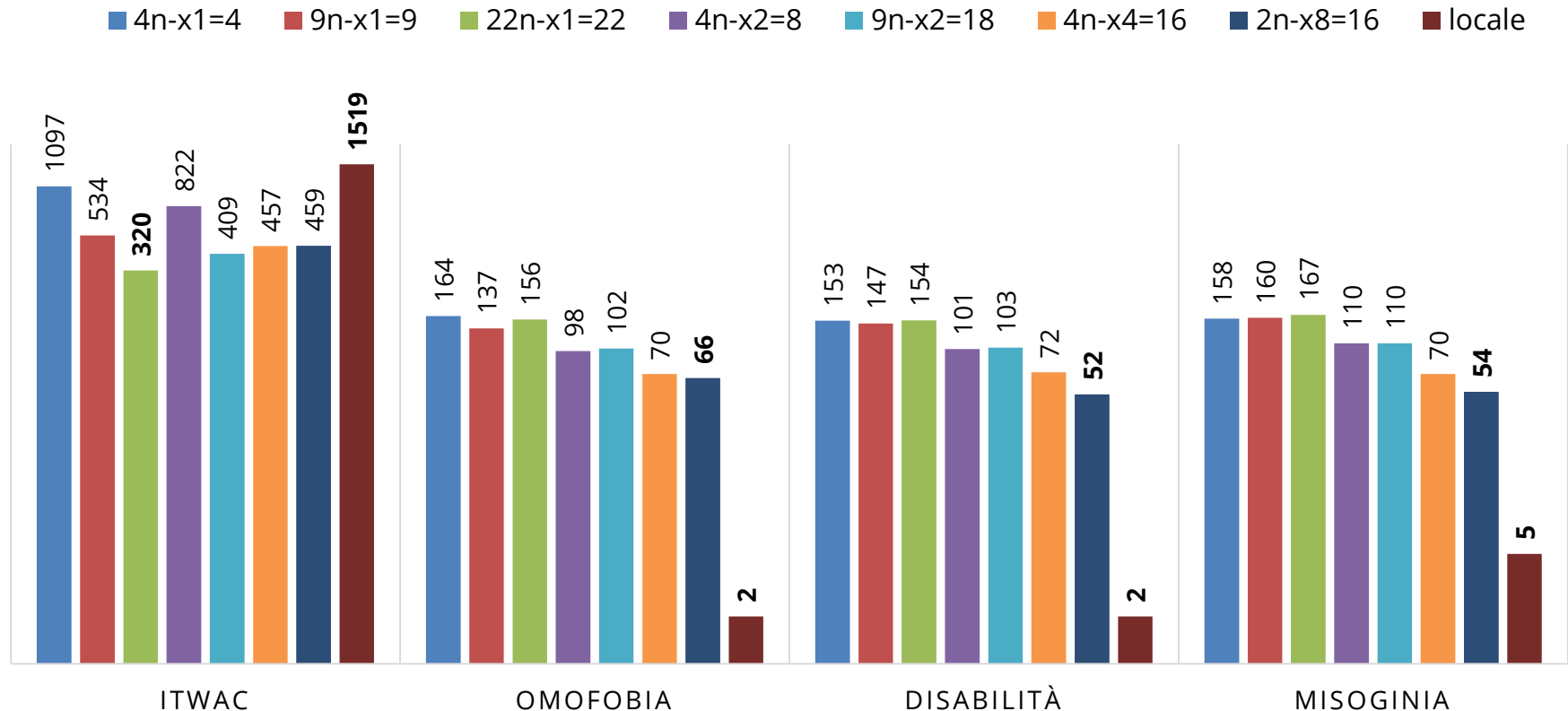
- macchina locale (cluster in modalità pseudo-distribuita)
- cluster su Compute Engine differenti tra loro per
  - numero di nodi interconnessi
  - tipo di macchine virtuali



Per convenzione:

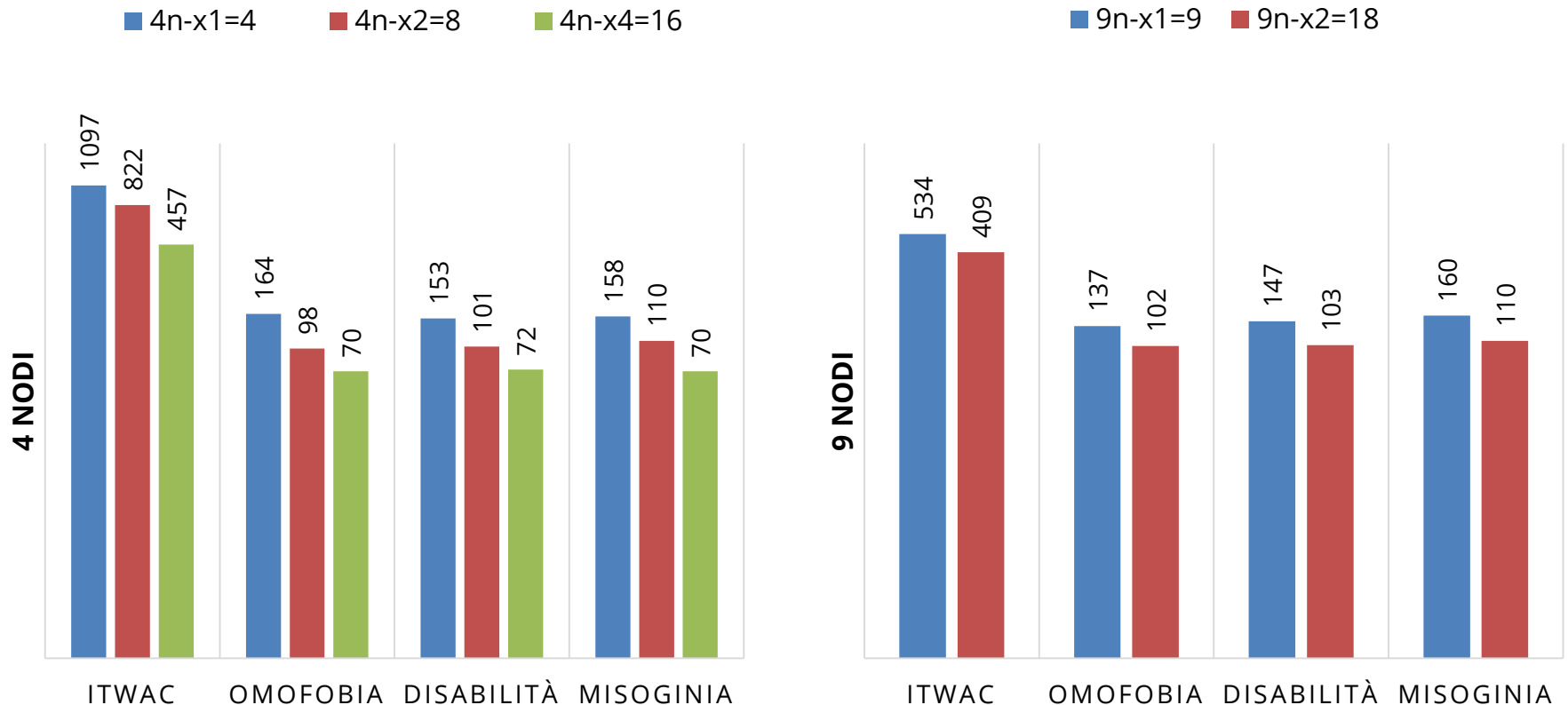
- $4n-x1=4$ : cluster con 4 nodi da 1 CPU ciascuno (4 tot)
- $4n-x2=8$ : cluster con 4 nodi da 2 CPU ciascuno (8 tot)
- $9n-x2=18$ : cluster con 9 nodi da 2 CPU ciascuno (18 tot)
- ...
- $\#nodi-tipo\_VM=CPU\_cluster$

## TEMPI MEDI DI INDICIZZAZIONE (s)



Grandi dataset: miglior cluster outperforms macchina locale  
 Piccoli dataset: macchina locale outperforms miglior cluster

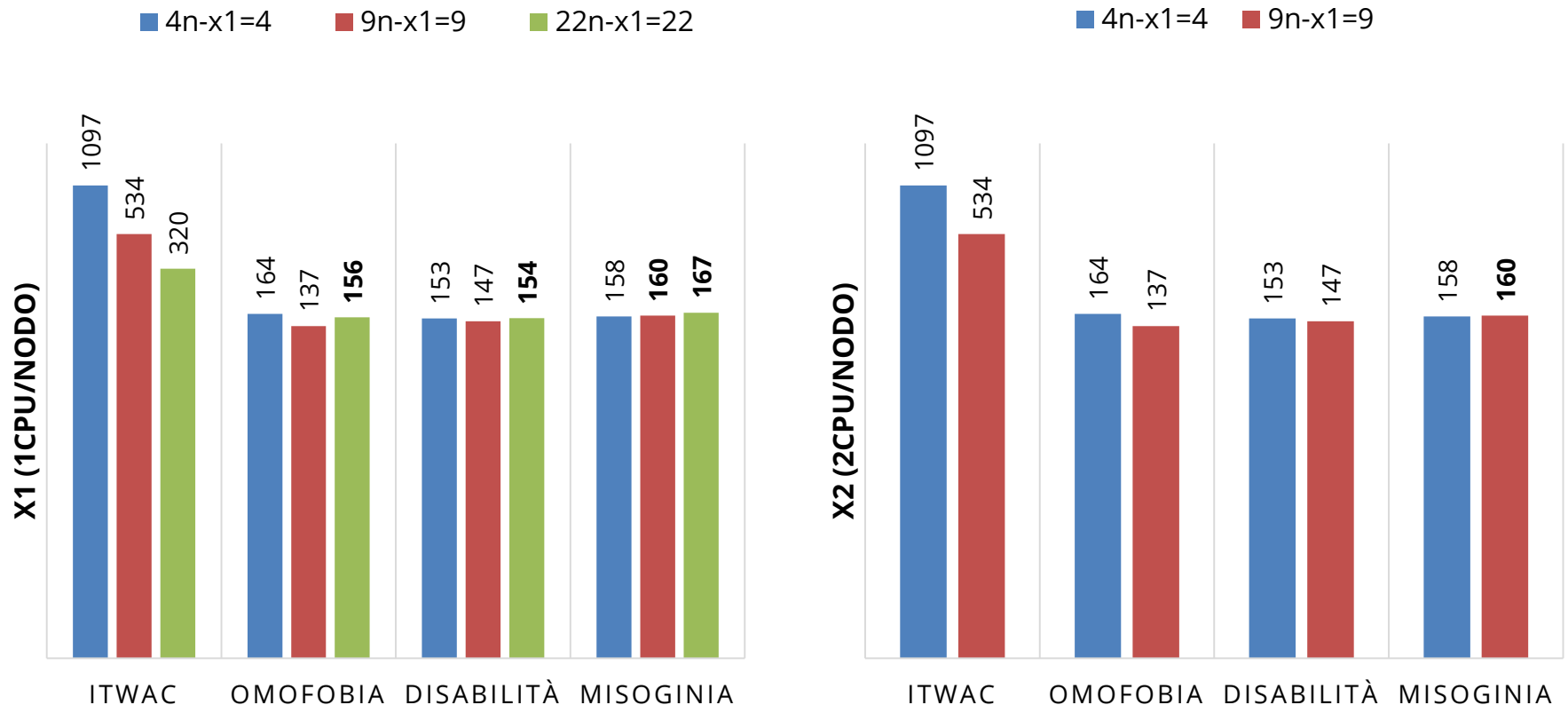
## TEMPI MEDI DI INDICIZZAZIONE (s)



Mantenendo costante il numero di nodi e migliorando il tipo di macchina virtuale, i tempi di indicizzazione diminuiscono sempre



## TEMPI MEDI DI INDICIZZAZIONE (s)



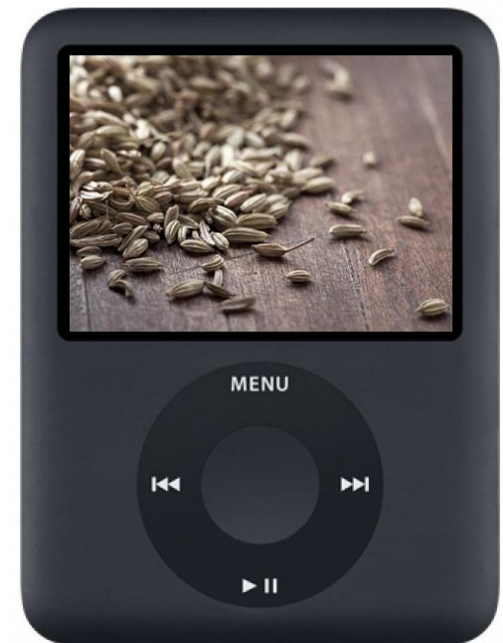
Mantenendo costante il tipo di macchina virtuale ed aumentando il numero di nodi, i tempi di indicizzazione diminuiscono solo su grandi dataset

## Language model

- *RT #Vogliadi #solo5ingredienti @CraftMarmalade: Frollini all'olio d'oliva e semi di finocchio ;D <http://t.co/GZd7loEcbx>*
  - *retweet, hashtag, menzioni, emoticon, URL, abbreviazioni, ecc.*

## Falsi positivi

- *Angolo del risparmio: iPod **nano** con il 30% di sconto! <http://t.com/VwnCRLhCLH>*
- *Cuoco omofobo espelle dalla cucina anche i semi di **finocchio***



## 2<sup>nd</sup> Research question

Migliorare i seed ricercando ed analizzando i termini più significativi nei tweet

Collezione	Cardinalità vocabolario
itWaC	4.431.080 termini
omofobia	54.507 termini
disabilità	72.569 termini
misoginia	280.412 termini

*Dati di input dell'esperimento*

- **Divergenza di Kullback-Leibler:** misura la prossimità tra due distribuzioni di probabilità discrete:
  - ▢  $C$ : corpus itWaC
  - ▢  $T$ : collezione di tweet di una classe
- **Pointwise Kullback-Leibler divergence:** è il contributo dato da una parola  $s$  alla divergenza complessiva di  $T$  da  $C$ :

$$\delta_s(T\|C) = P_T(s) \ln \frac{P_T(s)}{P_C(s)}$$

- ▢ Correzione di Laplace:  $P_D(i) = \frac{f_{D,i} + 1}{f_D + |D|}$
  - ▢ Calcolo distribuito con Hadoop: **14 milioni di probabilità!**
- Le parole con PKLD maggiore sono quelle più significative nella collezione  $T$  quando questa viene confrontata con  $C$

Nelle prime posizioni dei ranking emergono:

- bestemmie, imprecazioni, turpiloquio
- personaggi famosi (politici, giovani cantanti, ecc...)
- trasmissioni televisive
- termini complementari ai seed in espressioni d'uso comune (nano: Biancaneve, iPod, giardino; finocchio: semi, tisana)
- almeno **due nuovi seed** tra i primi 300 termini più significativi per ogni classe

# Ranking

OMOFOBIA		DISABILITA'		MISOGINIA	
Harry	cazzo	@Ty_il_nano	iPhone	@justinbieber	<b>madonna</b>
culo	<b>tisana</b>	<b>@matteorenzi</b>	@justinbieber	Vaffanculo	Profilo
merda	isterica	Renzi	malefico	inferno	Sesso
psiconano	foto	ottavo	Apple	Brutta	mamma
<b>#tvoi</b>	@GayOggi	<b>iPod</b>	cazzo	depressa	tette
<b>#GF13</b>	<b>prezzemolo</b>	<b>giardino</b>	Biancaneve	<b>vergine</b>	'rca
Justin	Tisana	merda	<b>Brunetta</b>	Bocca	<b>inchiavabile</b>

- Sistemi distribuiti per memorizzare ed elaborare dati:
  - ▣ controproducenti per piccoli dataset
  - ▣ indispensabili per big data
- Piattaforme di cloud computing:
  - ▣ migliorano le performance delle tecnologie scalabili
  - ▣ facili da configurare ed utilizzare
  - ▣ economicamente vantaggiose (costo esperimenti: 21€)
- Raccolta ed analisi dei tweet:
  - ▣ rimozione di seed poco significativi e filtraggio in base al contesto
  - ▣ aggiunta di nuovi seed



- Piattaforma d'analisi dei dati:
  - ▣ aggiungere task di data visualization
  - ▣ integrare altri prodotti cloud di Google (es: Big Query)
- Elaborazione dei tweet:
  - ▣ indagare sulla semantica dei messaggi (es: sentiment analysis)
  - ▣ apprendere automaticamente language model dei tweet
- Riconoscimento di contenuti discriminatori:
  - ▣ costruire classificatori per i tweet
  - ▣ studiare la personalità degli individui per ottenere dei profili dei discriminatori (omofobo, razzista, antisemita, ecc.)
  - ▣ analizzare la struttura del grafo del social network

Domande?

Grazie per l'attenzione