

# Classificazione articoli ANSA con topic modelling



- Perché L.D.A. in text mining
- Scopo progetto

# Dataset

- [site:ansa.it/notizie/politica](https://www.ansa.it/notizie/politica)
- [site:ansa.it/notizie/economia](https://www.ansa.it/notizie/economia)
- [site:ansa.it/notizie/tecnologia](https://www.ansa.it/notizie/tecnologia)
- [site:ansa.it/notizie/cultura](https://www.ansa.it/notizie/cultura)
- ...



# Dataset - Struttura dati

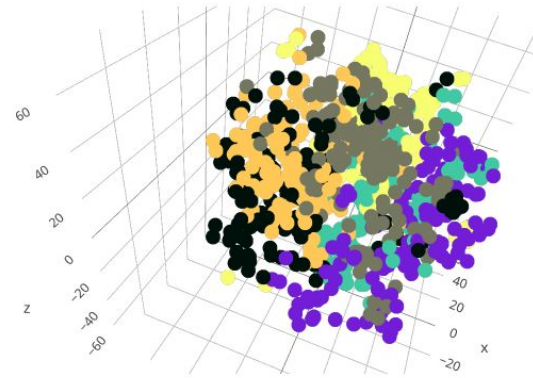
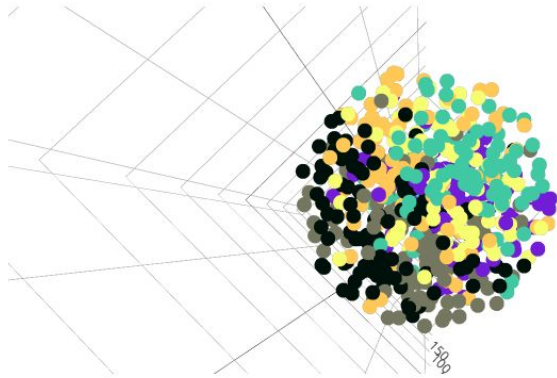
(Libri, Film, Cinema) ----> Cultura  
(Calcio, Basket, F1) ----> Sport  
....

```
{  
  "name": "Mario",  
  "surname": "Rossi",  
  "active": true,  
  "favoriteNumber": 42,  
  "birthday": {  
    "day": 1,  
    "month": 1,  
    "year": 2000  
  },  
  "languages": [ "it", "en" ]  
}
```

# Analisi preliminari

- tokenization
- rimozione stopwords
- rimozione tag html
- stemming

# Riduzione dimensionalita'



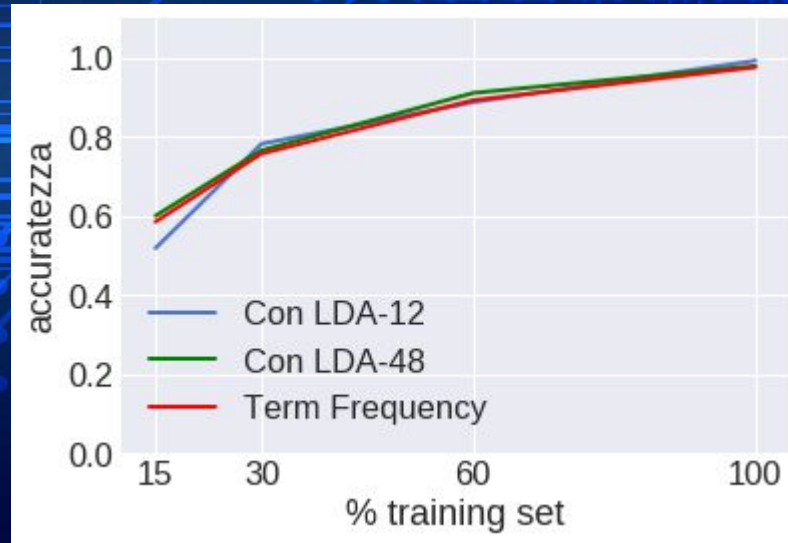
- trace 0
- trace 1
- trace 2
- trace 3
- trace 4
- trace 5



# Suddivisione dataset e configurazioni

- LDA-12
- LDA-48
- Term Frequency

# Variazione insieme training





# Modelli messi a confronto

Rappresentazione	Accuratezza
Dummy	17.8 %
LDA-12	<b>99.2 %</b>
LDA-48	97.8 %
Term Frequency	98.5 %

# Altre metriche di valutazione

Categoria	precision	recall	f1	support
Cronaca	1.00	1.00	1.00	107
Cultura	0.98	0.98	0.98	88
Economia	1.00	0.98	0.99	99
Politica	1.00	0.98	0.99	107
Sport	0.98	1.00	0.99	96
Tech	0.99	1.00	1.00	103

# Tempi

Rappresentazione	Tempi trasf.	Tempi class.
LDA-12	99.55 s	<b>0.02 s</b>
LDA-48	141.54 s	0.07 s
Term Frequency	<b>5.29 s</b>	1.93 s



Esperimenti

# Interpretazione risultati





Demo



Q & A





The End