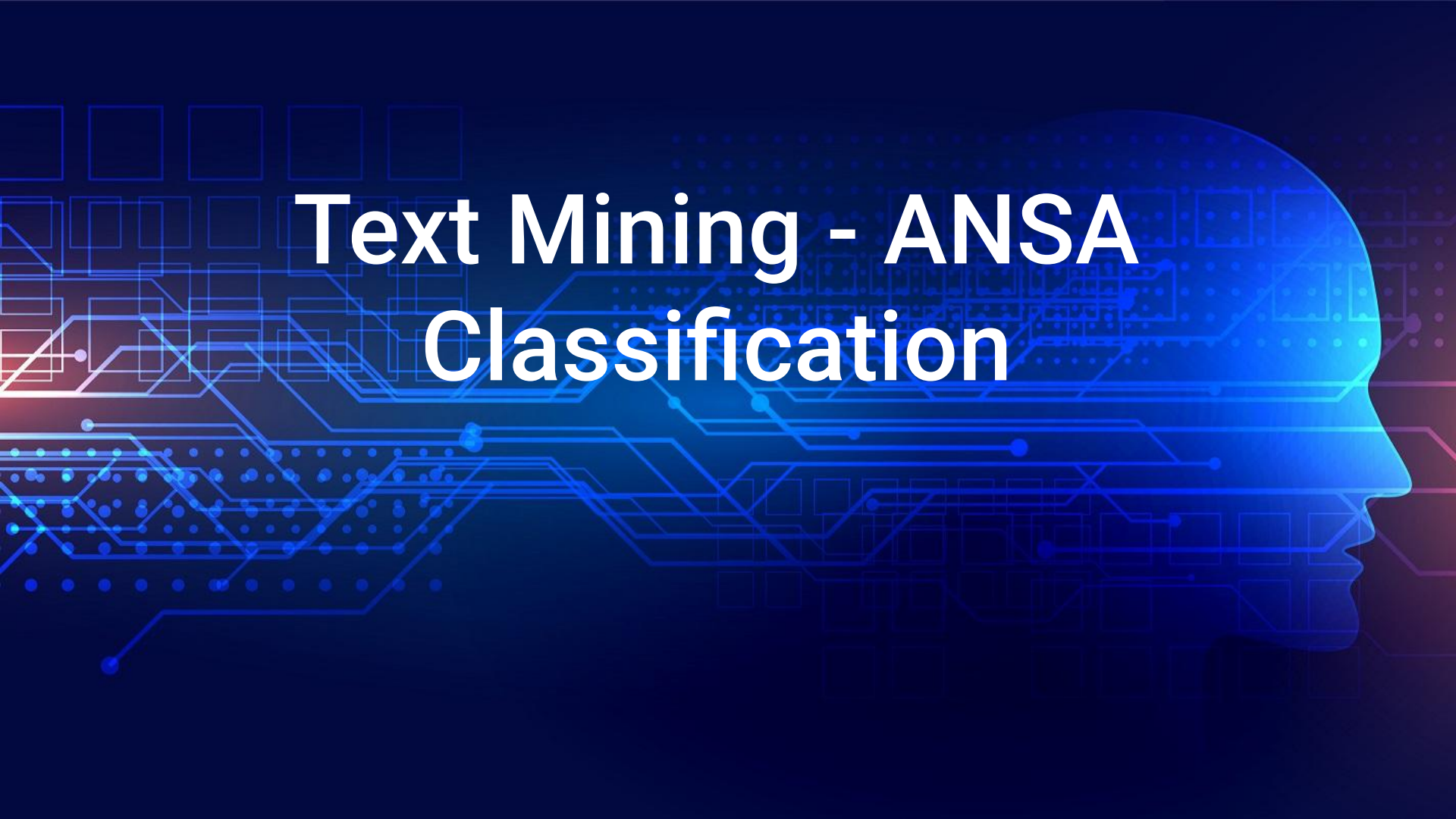

Text Mining - ANSA
Classification

- Perche’ L.D.A. in text mining
- Scopo progetto

Dataset

- site:ansa.it/notizie/politica
- site:ansa.it/notizie/economia
- site:ansa.it/notizie/tecnologia
- site:ansa.it/notizie/cultura

...

Dataset - Struttura dati

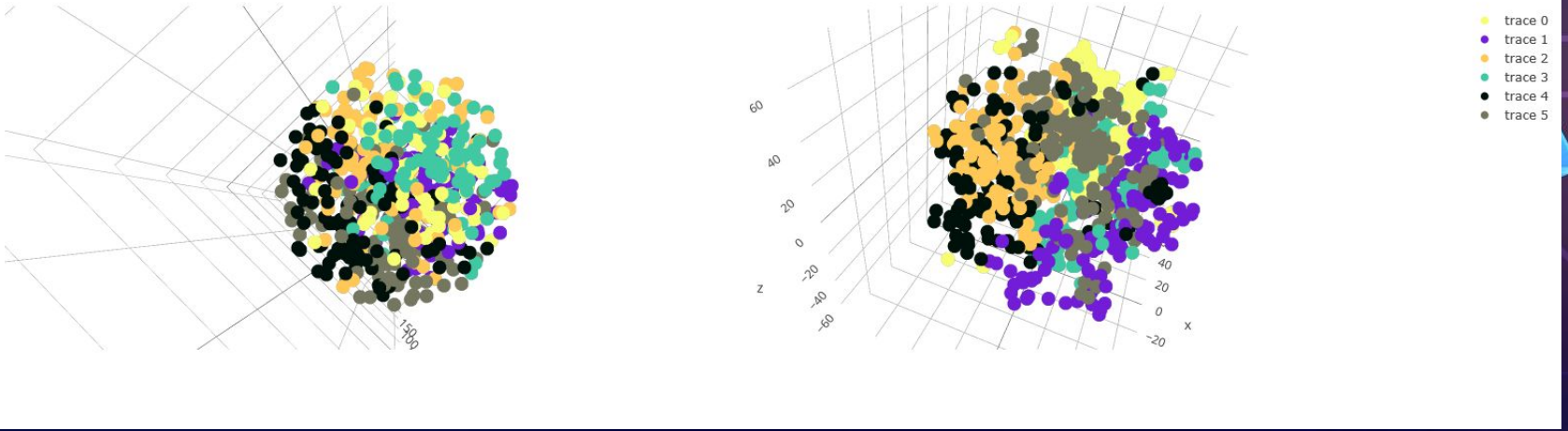
(Libri, Film, Cinema) ----> Cultura
(Calcio, Basket, F1) ----> Sport
....


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

Preprocessamento

- tokenization
- rimozione stopword
- rimozione tag html
- stemming

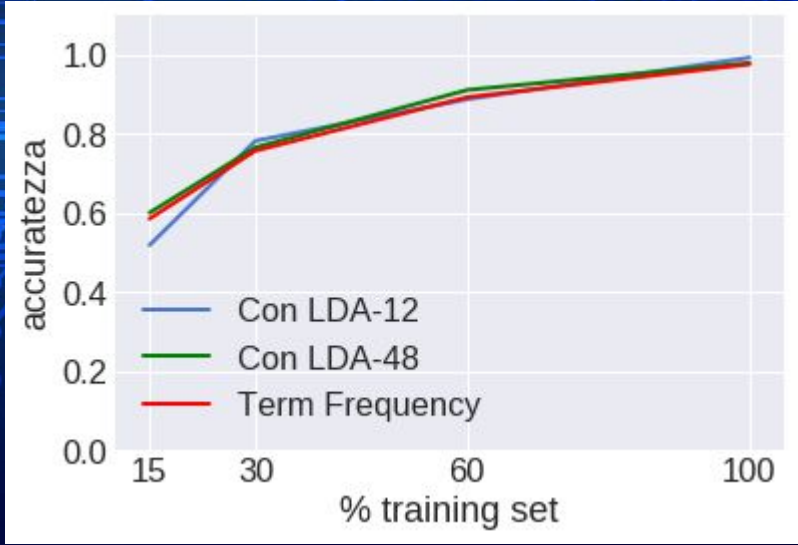
Riduzione dimensionalita’



Suddivisione dataset e configurazioni

- LDA-12
- LDA-48
- Term Frequency

Variazione insieme training



Modelli messi a confronto

Rappresentazione	Accuratezza
Dummy	17.8 %
LDA-12	**99.2 %**
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	**0.02 s**
LDA-48	141.54 s	0.07 s
Term Frequency	**5.29 s**	1.93 s

Interpretazione risultati


Q & A


The End