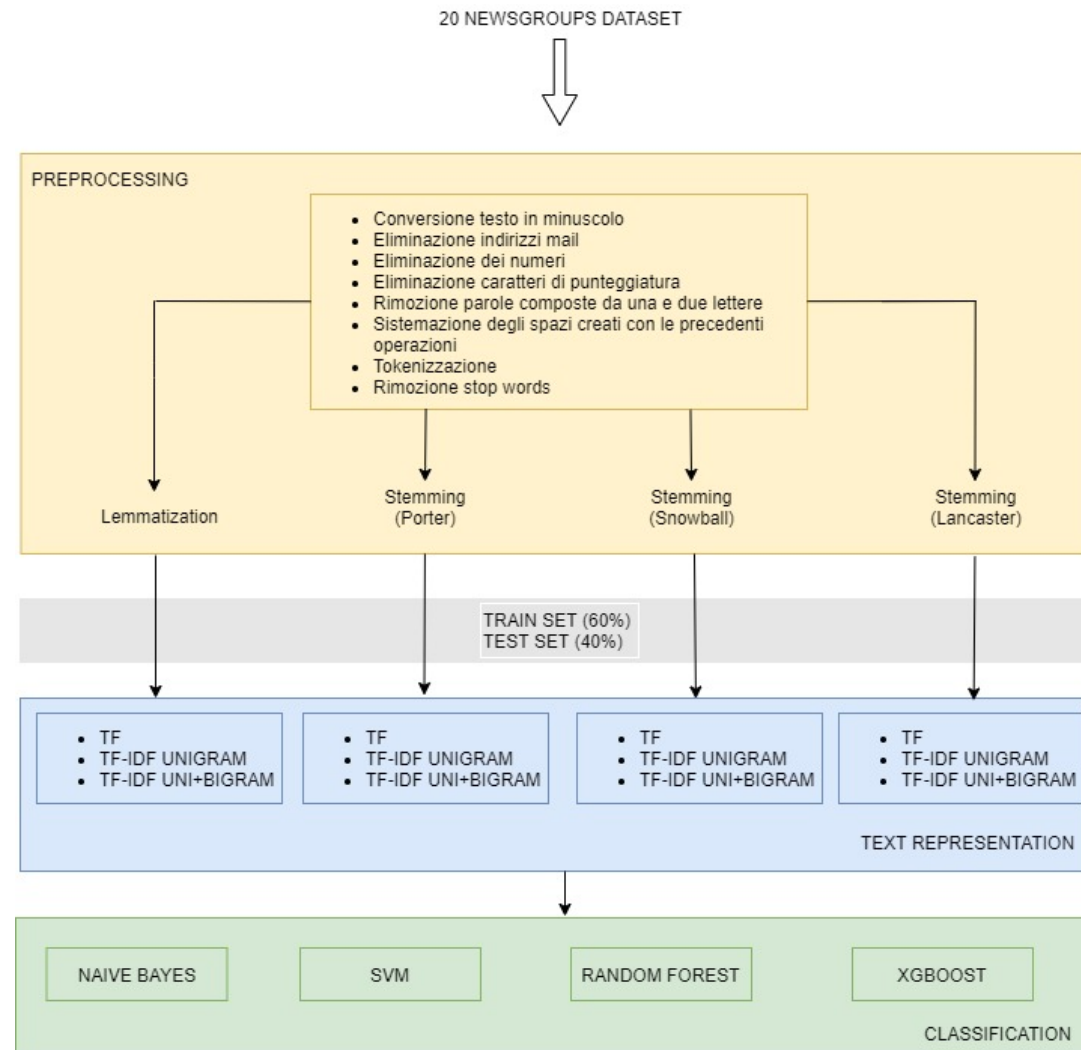


20 NEWS GROUP CLASSIFICATION

Anonella Zaccaria
Marco Savino

848647
793516

PIPELINE



DATASET *20 NEWSGROUPS*

19997 documenti divisi in 20 newsgroups:

- | | |
|-----------------------------|---------------------------|
| 1. alt.atheism | 11.rec.sport.hockey |
| 2. comp.graphics | 12.sci.crypt |
| 3. comp.os.ms-windows.misc | 13.sci.electronics |
| 4. comp.sys.ibm.pc.hardware | 14.sci.med |
| 5. comp.sys.mac.hardware | 15.sci.space |
| 6. comp.windows.x | 16.soc.religion.christian |
| 7. misc.forsale | 17.talk.politics.guns |
| 8. rec.autos | 18.talk.politics.mideast |
| 9. rec.motorcycles | 19.talk.politics.misc |
| 10.rec.sport.baseball | 20.talk.religion.misc |

PREPROCESSING

- Conversione testo in minuscolo
- Eliminazione indirizzi mail
- Eliminazione dei numeri
- Eliminazione caratteri di punteggiatura
- Rimozione parole composte da una e due lettere
- Sistemazione degli spazi creati con le precedenti operazioni
- Tokenizzazione
- Rimozione Stop Words
- Lemmatization
- Stemming: Porter, Snowball, Lancaster

Dataset diviso in Training set (60%) e Test set (40%)

TEXT REPRESENTATION

- Term Frequency (TF): la Term Frequency $t_{ft,d}$ del termine t nel documento d è definito come il numero di volte che t si verifica in d
- Term Frequency-Inverse Document Frequency (TF-IDF): il peso $tf-idf$ di un termine è il prodotto del suo peso tf e del suo peso idf
 - Unigram
 - Unigram + Bigram

Matrici costruite per:

- Dati lemmatizzati
- Dati stemmatizzati (Porter)
- Dati stemmatizzati (Snowball)
- Dati stemmatizzati (Lancaster)

→ 24 Matrici (12 Train + 12 Test)

TEXT CLASSIFICATION

- Multinomial NB
- SVM
- Random Forest
- XGBOOST

MULTINOMIAL NAIVE BAYES

	TRAIN (cross validation)		TEST	
	Accuracy	Time	Accuracy	Time
TF + LEM	0.86	383 ms	0.87	127 ms
TF + STEM (Porter)	0.85	351 ms	0.86	106 ms
TF + STEM (Snowball)	0.85	345 ms	0.86	113 ms
TF + STEM (Lancaster)	0.85	306 ms	0.85	111 ms
TF-IDF + LEM	0.86	252 ms	0.87	77.6 ms
TF-IDF + STEM (Porter)	0.85	215 ms	0.86	77.7 ms
TF-IDF + STEM (Snowball)	0.85	223 ms	0.86	77.3 ms
TF-IDF + STEM (Lancaster)	0.85	222 ms	0.86	76.6 ms
TF-IDF (bigram) + LEM	0.86	262 ms	0.87	101 ms
TF-IDF (bigram) + STEM (Porter)	0.86	267 ms	0.87	88.7 ms
TF-IDF (bigram) + STEM (Snowball)	0.86	261 ms	0.87	89.8 ms
TF-IDF (bigram) + STEM (Lancaster)	0.86	247 ms	0.86	88.2 ms

SUPPORT VECTOR MACHINE (SVM)

	TRAIN (cross validation)		TEST	
	Accuracy	Time	Accuracy	Time
TF + LEM	0.75	9 min 10 s	0.78	3 min 29 s
TF + STEM (Porter)	0.75	8 min 52 s	0.78	3 min 28 s
TF + STEM (Snowball)	0.75	8 min 55 s	0.78	3 min 27 s
TF + STEM (Lancaster)	0.74	8 min 37 s	0.77	3 min 20 s
TF-IDF + LEM	0.89	12 min 40 s	0.90	4 min 35 s
TF-IDF + STEM (Porter)	0.89	12 min 40 s	0.90	4 min 34 s
TF-IDF + STEM (Snowball)	0.89	12 min 27 s	0.90	4 min 29 s
TF-IDF + STEM (Lancaster)	0.88	12 min 26 s	0.89	4 min 18 s
TF-IDF (bigram) + LEM	0.89	14 min 12 s	0.90	5 min 4 s
TF-IDF (bigram) + STEM (Porter)	0.89	14 min 34 s	0.90	5 min 10 s
TF-IDF (bigram) + STEM (Snowball)	0.89	14 min 5 s	0.90	4 min 59 s
TF-IDF (bigram) + STEM (Lancaster)	0.89	13 min 44 s	0.89	5 min

RANDOM FOREST

	TRAIN (cross validation)		TEST	
	Accuracy	Time	Accuracy	Time
TF + LEM	0.83	2 min 46 s	0.84	32.6 s
TF + STEM (Porter)	0.82	2 min 25 s	0.84	30.6 s
TF + STEM (Snowball)	0.82	2 min 25 s	0.84	29.9 s
TF + STEM (Lancaster)	0.81	2 min 16 s	0.83	29.4 s
TF-IDF + LEM	0.82	1 min 31 s	0.83	23.8 s
TF-IDF + STEM (Porter)	0.81	1 min 33 s	0.83	23.7 s
TF-IDF + STEM (Snowball)	0.81	1 min 34 s	0.83	23.7 s
TF-IDF + STEM (Lancaster)	0.80	1 min 40 s	0.82	24.3 s
TF-IDF (bigram) + LEM	0.82	1 min 31 s	0.83	24.5 s
TF-IDF (bigram) + STEM (Porter)	0.81	1 min 31 s	0.83	24.8 s
TF-IDF (bigram) + STEM (Snowball)	0.82	1 min 31 s	0.83	24.4 s
TF-IDF (bigram) + STEM (Lancaster)	0.81	1 min 33 s	0.82	24.9 s

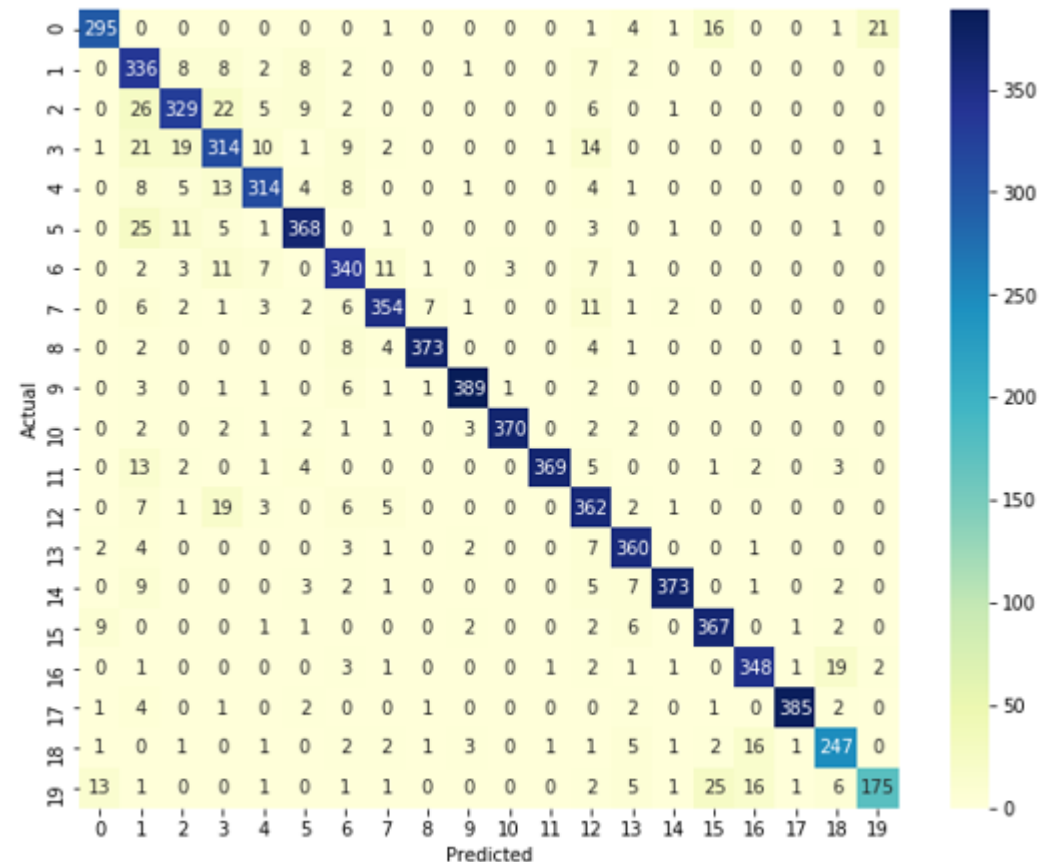
XGBOOST

	TRAIN (cross validation)		TEST	
	Accuracy	Time	Accuracy	Time
TF + LEM	0.80	15 min 44 s	0.80	3 min 20 s
TF + STEM (Porter)	0.80	13 min 33 s	0.79	2 min 58 s
TF + STEM (Snowball)	0.79	13 min 24 s	0.79	2 min 56 s
TF + STEM (Lancaster)	0.78	12 min 10 s	0.79	2 min 40 s
TF-IDF + LEM	0.79	13 min 33 s	0.79	3 min 20 s
TF-IDF + STEM (Porter)	0.79	13 min 58 s	0.79	3 min 27 s
TF-IDF + STEM (Snowball)	0.79	14 min 14 s	0.79	3 min 30 s
TF-IDF + STEM (Lancaster)	0.78	14 min 25 s	0.78	3 min 32 s
TF-IDF (bigram) + LEM	0.80	15 min 32 s	0.80	3 min 51 s
TF-IDF (bigram) + STEM (Porter)	0.80	15 min 45 s	0.80	3 min 49 s
TF-IDF (bigram) + STEM (Snowball)	0.80	15 min 30 s	0.80	3 min 49 s
TF-IDF (bigram) + STEM (Lancaster)	0.79	15 min 42 s	0.79	3 min 53 s

CONCLUSIONI

Il modello migliore in termini di *Recall*, *Precision* ed *F1-score* risulta essere **Support Vector Machine** nelle rappresentazioni tf-idf

	precision	recall	f1-score	support
0	0.92	0.87	0.89	340
1	0.71	0.90	0.80	374
2	0.86	0.82	0.84	400
3	0.79	0.80	0.79	393
4	0.89	0.88	0.89	358
5	0.91	0.88	0.90	416
6	0.85	0.88	0.87	386
7	0.92	0.89	0.91	396
8	0.97	0.95	0.96	393
9	0.97	0.96	0.96	405
10	0.99	0.96	0.97	386
11	0.99	0.92	0.96	400
12	0.81	0.89	0.85	406
13	0.90	0.95	0.92	380
14	0.98	0.93	0.95	403
15	0.89	0.94	0.91	391
16	0.91	0.92	0.91	380
17	0.99	0.96	0.98	399
18	0.87	0.87	0.87	285
19	0.88	0.71	0.78	248
accuracy			0.90	7539
macro avg	0.90	0.89	0.90	7539
weighted avg	0.90	0.90	0.90	7539



GRAZIE PER L'ATTENZIONE