Università degli studi di Torino
Dipartimento di informatica

Progetto di Tecnologie del Linguaggio Naturale

Text categorization mediante modello VSM

Autori:
Christian Quaranta
Giulia Coucourde
Leonardo Magliolo

Anno Accademico 2022/2023

Introduzione:

Il progetto consiste nella realizzazione di un software in grado di classificare documenti mediante apprendimento automatico supervisionato che sfrutti il metodo di Rocchio con formalismo di rappresentazione Vector Space Model.
Il classificatore, in fase di apprendimento, dovrà tenere conto della vicinanza dei documenti rispetto al centroide positivo di riferimento ma anche della distanza relativa al centroide degli esempi negativi rispetto alla classe dei documenti considerati.

Dataset:

Il dataset su cui l'algoritmo è stato testato è composto da 200 documenti scritti in lingua italiana e partizionato in 10 classi differenti: 20 documenti per ogni classe.

Modello:

Definiti i seguenti insiemi:

- D: Set che racchiude tutti i documenti che è utile considerare
- T: Set che racchiude tutti i termini menzionati almeno una volta all'interno di qualche documento appartenente ad D
- C: Set che racchiude tutte le classi relative ai documenti che è utile considerare
- POS_i: Set che racchiude tutti i documenti appartenenti a D che sono classificati con l'i-esima classe.
- $NPOS_i$: Set che racchiude gli N documenti più positivi che non sono classificati con l'i-esima classe.
- W: insieme di tutti i possibili pesi w_{kj} associati all'k-esimo termine appartenente a T e alla j-esimo documento appartenente a D (determinati come il prodotto tra la frequenza del termine nel documento e la sua IDF)

Per ogni classe appartenente a C viene appreso un vettore $c_i: < f_{1i}, \dots, f_{|T|i} >$.

Per ogni componente f_{ki} viene determinato il valore della stessa mediante la seguente formula:

$$f_{ki} = \beta \cdot \sum_{d_j \in POS_i} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{d_j \in NPOS_i} \frac{w_{kj}}{|NPOS_i|}$$

Apprendimento:

Durante la fase di apprendimento è stato utile adoperare le seguenti tecniche per estrarre i valori dei vettori $< f_{1i}, \dots, f_{|T|i} >$:

- Bag of words: Utile per rappresentare un documento mediante le parole contenute, siccome il modello considerato non necessita di esaminare sequenze di parole.
- Rimozione delle stopwords: Rimozione delle parole che presentano una scarsa selettività relativa alle classi dei documenti in cui vengono citate.
- Lemmatizzazione: Parole diverse con lemmi comuni esprimono concetti simili, considerarle come componenti differenti all'interno dei vettori delle classi abbasserebbe l'accuracy del modello.
- Inverse Document Frequency: A parità di frequenza consente di valutare con maggior peso termini più selettivi a favore di un determinato sottoinsieme di classi.
- Parametri β , γ e N: Sono stati impostati i parametri β = 16 e γ = 4 (come da richiesta progettuale) ed N = 10.

Decoding:

La classe di un documento viene predetta sulla base della massimizzazione della similarità tra il vettore del documento stesso (d) e i dei vettori associati alle classi possibili:

Se $\hat{c} = argmax_{c_i} \left(\frac{d \cdot c_i}{|d| \cdot |c_i|} \right)$ allora la classe predetta del documento d sarà la stessa associata al vettore $\hat{c}$.

Testing:

E' stato scelto di testare le performance del modello sul dataset utilizzando la tecnica cross-validation (come suggerito nei requisiti del progetto).

In particolare, sono state effettuate 10 partizioni differenti del dataset composte rispettivamente dal 90% delle istanze (18 per classe) per i training set e il restate 10% (2 per classe) per il test set, per ognuna di esse è stata calcolata la matrice di confusione e l'accuracy del modello.

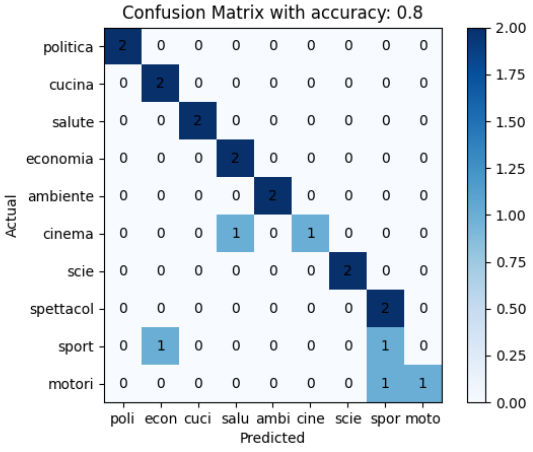
Librerie utilizzate:

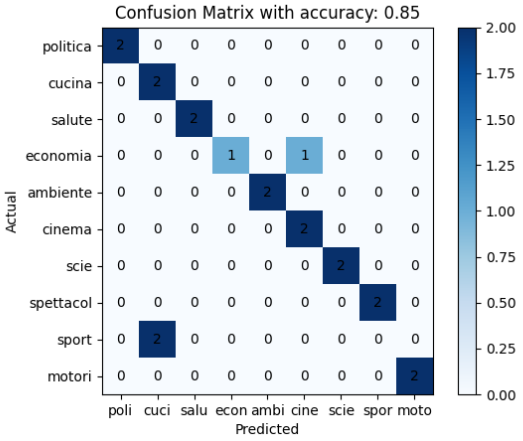
In aggiunta alle funzionalità native messe a disposizione da Python, è stato scelto di utilizzare le seguenti librerie a supporto della realizzazione del progetto:

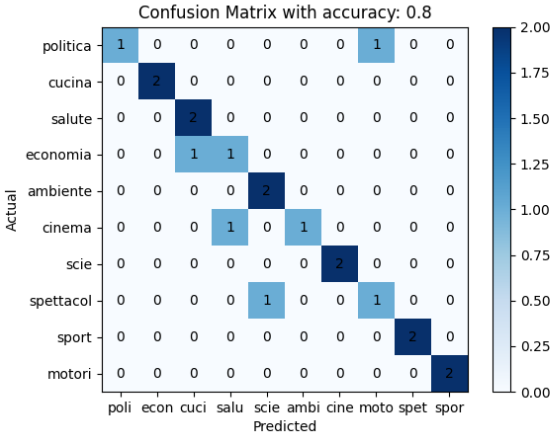
- Spacy: Utilizzata per la rimozione delle stopwords ed effettuare le operazioni di lemmatizzazione. Il modello che è utilizzato per Spacy è : "it_core_news_lg".
- Scikit-learn: Utilizzata per partizionare il dataset, calcolare la metrica cosine-similarity e rappresentare le matrici di confusione.
- Pandas e Numpy: Utilizzate per la rappresentazione dei vettori e delle matrici utili all'implementazione del modello.

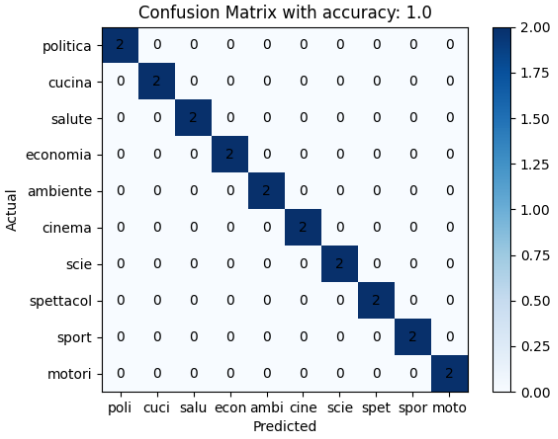
Risultati ottenuti:

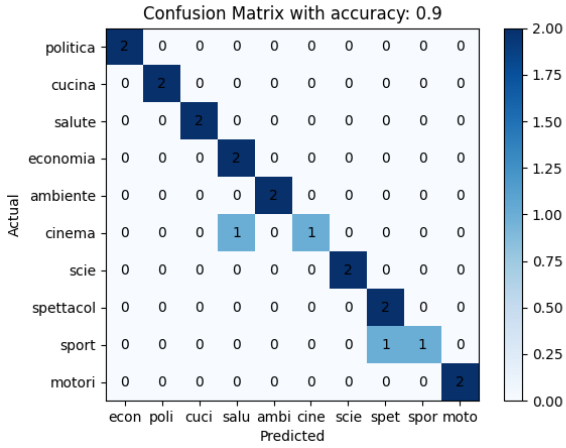
Vengono riportate di seguito le matrici di confusione associate ad ognuno dei 10 test effettuati per la cross-validation con relativi valori di accuracy:

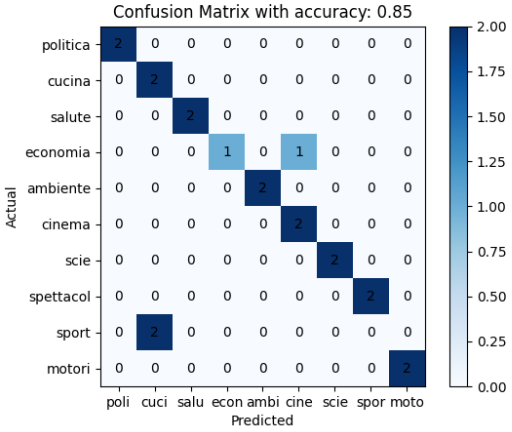


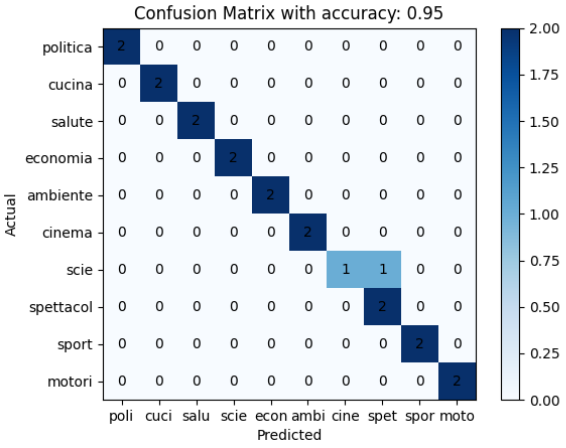


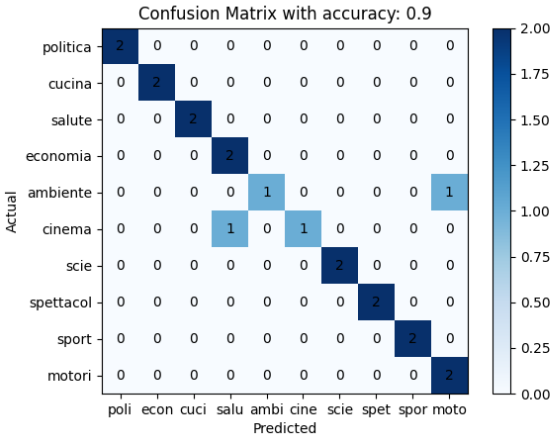

Confusion Matrix with accuracy: 0.8
politica 1 0 0 0 0 0 0 1 0 0
cucina 0 2 0 0 0 0 0 0 0 0
salute 0 0 2 0 0 0 0 0 0 0
economia 0 0 1 1 0 0 0 0 0 0
ambiente 0 0 0 0 2 0 0 0 0 0
cinema 0 0 0 1 0 1 0 0 0 0
scie 0 0 0 0 0 0 2 0 0 0
spettacol 0 0 0 0 1 0 0 1 0 0
sport 0 0 0 0 0 0 0 0 2 0
motori 0 0 0 0 0 0 0 0 0 2
Actual
poli econ cuci salu scie ambi cine moto spet spor
Predicted
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00

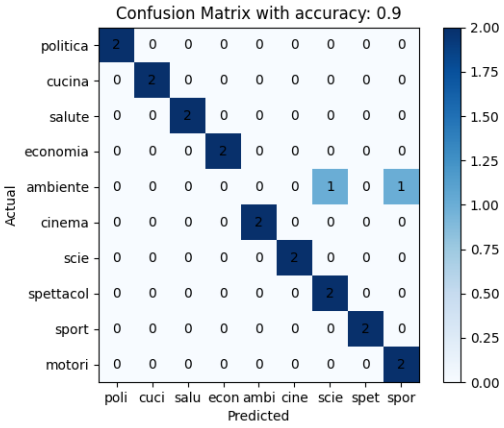

Confusion Matrix with accuracy: 1.0
politica 2 0 0 0 0 0 0 0 0 0
cucina 0 2 0 0 0 0 0 0 0 0
salute 0 0 2 0 0 0 0 0 0 0
economia 0 0 0 2 0 0 0 0 0 0
ambiente 0 0 0 0 2 0 0 0 0 0
cinema 0 0 0 0 0 2 0 0 0 0
scie 0 0 0 0 0 0 2 0 0 0
spettacol 0 0 0 0 0 0 0 2 0 0
sport 0 0 0 0 0 0 0 0 2 0
motori 0 0 0 0 0 0 0 0 0 2
Actual
poli cuci salu econ ambi cine scie spet spor moto
Predicted
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00

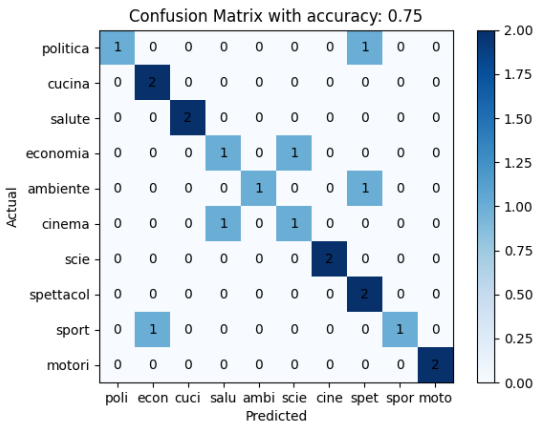

Confusion Matrix with accuracy: 0.9
politica 2 0 0 0 0 0 0 0 0 0
cucina 0 2 0 0 0 0 0 0 0 0
salute 0 0 2 0 0 0 0 0 0 0
economia 0 0 0 2 0 0 0 0 0 0
ambiente 0 0 0 0 2 0 0 0 0 0
cinema 0 0 0 1 0 1 0 0 0 0
scie 0 0 0 0 0 0 2 0 0 0
spettacol 0 0 0 0 0 0 0 2 0 0
sport 0 0 0 0 0 0 0 1 1 0
motori 0 0 0 0 0 0 0 0 0 2
Actual
econ poli cuci salu ambi cine scie spet spor moto
Predicted
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00


Confusion Matrix with accuracy: 0.85
politica 2 0 0 0 0 0 0 0 0 0
cucina 0 2 0 0 0 0 0 0 0 0
salute 0 0 2 0 0 0 0 0 0 0
economia 0 0 0 1 0 1 0 0 0 0
ambiente 0 0 0 0 2 0 0 0 0 0
cinema 0 0 0 0 0 2 0 0 0 0
scie 0 0 0 0 0 0 2 0 0 0
spettacol 0 0 0 0 0 0 0 2 0 0
sport 0 2 0 0 0 0 0 0 0 0
motori 0 0 0 0 0 0 0 0 0 2
Actual
poli cuci salu econ ambi cine scie spor moto
Predicted
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00


Confusion Matrix with accuracy: 0.95
politica 2 0 0 0 0 0 0 0 0 0
cucina 0 2 0 0 0 0 0 0 0 0
salute 0 0 2 0 0 0 0 0 0 0
economia 0 0 0 2 0 0 0 0 0 0
ambiente 0 0 0 0 2 0 0 0 0 0
cinema 0 0 0 0 0 2 0 0 0 0
scie 0 0 0 0 0 0 1 1 0 0
spettacol 0 0 0 0 0 0 0 2 0 0
sport 0 0 0 0 0 0 0 0 2 0
motori 0 0 0 0 0 0 0 0 0 2
Actual
poli cuci salu scie econ ambi cine spet spor moto
Predicted
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00


Confusion Matrix with accuracy: 0.9
politica 2 0 0 0 0 0 0 0 0 0
cucina 0 2 0 0 0 0 0 0 0 0
salute 0 0 2 0 0 0 0 0 0 0
economia 0 0 0 2 0 0 0 0 0 0
ambiente 0 0 0 0 1 0 0 0 0 1
cinema 0 0 0 1 0 1 0 0 0 0
scie 0 0 0 0 0 0 2 0 0 0
spettacol 0 0 0 0 0 0 0 2 0 0
sport 0 0 0 0 0 0 0 0 2 0
motori 0 0 0 0 0 0 0 0 0 2
Actual
poli econ cuci salu ambi cine scie spet spor moto
Predicted
2.00 1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00





L’accuracy media riportata ammonta a 0.87 e la varianza delle accuracy è 0.0057.

Il numero di errori medio per classe invece risulta piuttosto variabile:

- Spettacoli: 0.8
- Politica: 0.5
- Economia: 0.4
- Motori:0.4
- Ambiente: 0.2
- Salute: 0.1
- Scientifico: 0.1
- Sport: 0.1
- Cucina: 0

L’ipotesi è che alcune categorie si prestino meglio a descrivere scenari che ne riguardano altre:

- Gli spettacoli descrivono spesso situazioni che riguardano ambiti esterni alla categoria di appartenenza stessa.
- La politica necessita di regolamentare aspetti della vita dell’uomo esterni alla categoria stessa, per cui è possibile che vengano usati termini specifici comparsi di frequente in documenti appartenenti ad altre classi.
- Un ragionamento simile può esser applicato in ambito economico.