Classificazione non supervisionata di provvedimenti: esperienze e criticità

Elena Palmieri

Dati

I dati presi in esame sono Atti amministrativi.

Si tratta di documenti lunghi in media **6-8 pagine** con la stessa pagina di apertura, ma la **struttura** del testo dei documenti è **variabile**.

REGIONE EMILIA-ROMAGNA
Atti amministrativi
GIUNTA REGIONALE
Atto del Dirigente DETERMINAZIONE
Num. 2028 del 02/02/2023 BOLOGNA

Proposta:	DPG/2023/1116 del 19/01/2023
Struttura proponente:	SETTORE AMMINISTRAZIONE E SISTEMA PARTECIPATE DIREZIONE GENERALE CURA DEL TERRITORIO E DELL'AMBIENTE
Oggetto:	RESTITUZIONE DEPOSITI CAUZIONALI A DIVERSI BENEFICIARI RELATIVO AL PROCEDIMENTO DEMANIALE MO05T0087
Autorità emanante:	IL RESPONSABILE - SETTORE AMMINISTRAZIONE E SISTEMA PARTECIPATE
Firmatario:	STEFANO ROTUNDO in qualità di Responsabile di settore

Visto di Regolarità Contabile - SPESE: ORSI MARINA espresso in data 31/01/2023

Responsabile del procedimento:	Stefano Rotundo

Classi

Le classi sono prese dal **Titolario delle Giunte Regionali**:

1 - ORGANI DI GOVERNO, AFFARI GIURIDICO-ISTITUZIONALI E COMUNICAZIONE
2 – ORGANIZZAZIONE, PATRIMONIO E RISORSE STRUMENTALI
3 – RISORSE UMANE
4 – RISORSE FINANZIARIE, GESTIONE CONTABILE E FISCALE
5 – SISTEMA INFORMATIVO
6 – PROGRAMMAZIONE, COORDINAMENTO E CONTROLLO
7 – AGRICOLTURA E ALLEVAMENTO
8 – ARTIGIANATO
9 – COMMERCIO
10 – INDUSTRIA E ATTIVITA' ESTRATTIVE
11 – TURISMO E STRUTTURE RICETTIVE
12 – PIANIFICAZIONE TERRITORIALE, URBANISTICA ED EDILIZIA
13 – INFRASTRUTTURE E TRASPORTI
14 – TUTELA DELL'AMBIENTE
15 – SANITA', IGIENE E VETERINARIA
16 – POLITICHE SOCIALI
17 – ISTRUZIONE, FORMAZIONE E LAVORO
18 – BENI E ATTIVITA' CULTURALI
19 – ATTIVITA' SPORTIVE E RICREATIVE

Problema

Vogliamo capire **automaticamente** a quale classe appartiene un documento.

↓

Non abbiamo un numero sufficiente di **documenti già classificati** per addestrare un classificatore in modo supervisionato.

↓

Utilizziamo **tecniche di classificazione non supervisionate**, i cui risultati vengono validati sui documenti classificati che possediamo.

Clustering

Le tecniche di clustering si basano su misure relative alla **somiglianza** tra gli elementi.

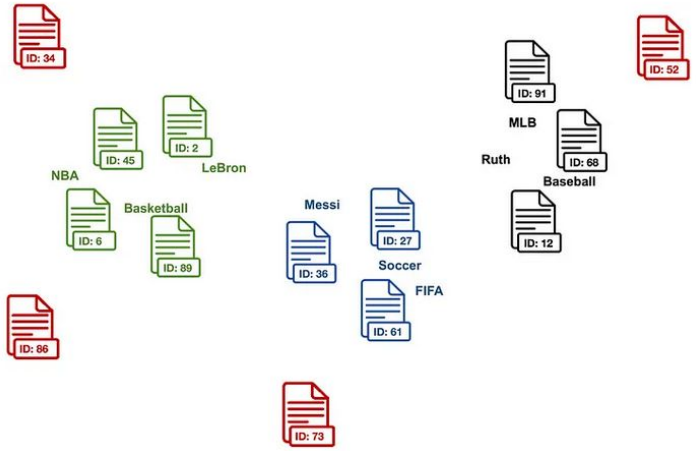
Intuitivamente, gli elementi più simili risulteranno essere più vicini tra loro in uno spazio multidimensionale.

L'appartenenza di un elemento ad un insieme dipende quindi da quanto l'elemento è distante dall'insieme stesso.

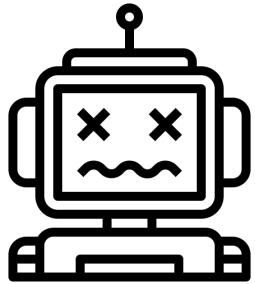
Lbl2Vec

Lbl2Vec è una tecnica di classificazione simile al clustering, che però introduce la possibilità di **descrivere le classi** tramite l'utilizzo di parole chiave.

Basketball	Soccer	Baseball
NBA	FIFA	MLB
Basketball	Soccer	Baseball
LeBron	Messi	Ruth
...	...	...



Risultati

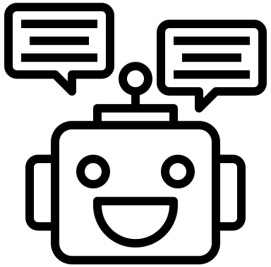


Questi classificatori non sono risultati efficaci per risolvere il problema.

Perché?

- I documenti sono molto lunghi.
- Tutti i documenti hanno una parte consistente di premesse legali che risultano “tutte uguali” per questo tipo di classificatori.

Classificazione Zero-Shot



Idea: Chiediamo ad un Large Language Model (LLM) di classificare i documenti.

Cosa vuol dire Zero-Shot?
Il modello non ha mai visto queste classi documenti prima d'ora.

I LLM hanno però già visto degli esempi di classificazione (e.g., filtro anti-spam).

Quindi, dando al modello una descrizione delle classi, questo potrebbe essere in grado di classificare correttamente i nostri documenti.

Classificazione Zero-Shot

La classificazione effettuata tramite un LLM spesso **non è sostenibile**:

- Far funzionare un LLM in un computer locale è molto oneroso in termini di tempo, per classificare un documento possono essere necessari diversi minuti.

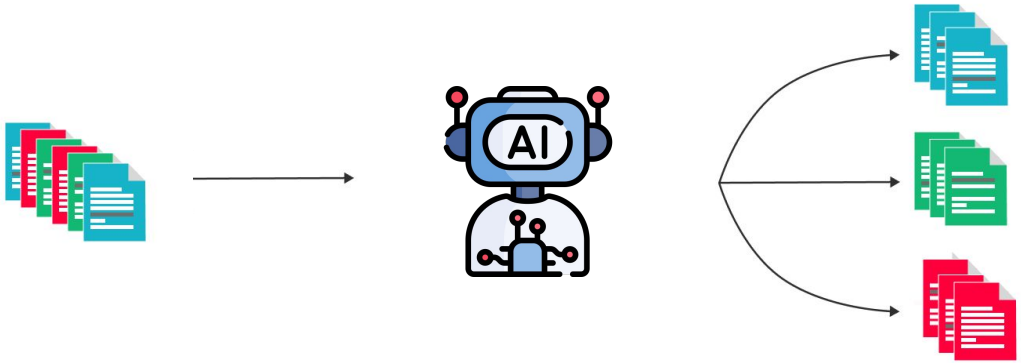
- Utilizzare LLM come GPT richiede l'iscrizione ad un servizio, che può risultare molto costoso. In genere il prezzo sale con il numero di chiamate, quindi con il numero di documenti che si vogliono classificare.

- Il nostro classificatore sarebbe dipendente da un servizio esterno.

Classificazione Zero-Shot

Idea: Utilizziamo GPT per classificare un numero di documenti che ci consentano di allenare un nostro modello.

In questo modo si limitano i costi e si crea un **modello leggero e veloce**, specializzato nella classificazione dei nostri documenti secondo le nostre classi.



Classificazione di documenti con GPT

I documenti classificati utilizzando GPT formano un dataset che chiameremo *silver*, che verrà utilizzato per creare il nostro modello.

Caratteristiche del dataset *silver*:

- Numero di documenti elevato
- Ogni classe deve essere rappresentata dallo stesso numero di documenti
- Assenza di documenti ripetuti

Dataset *silver* - Criticità

- La **classificazione di GPT** potrebbe **non** essere **corretta**.

  → Validazione della qualità della classificazione tramite l'aiuto di esperti.

- Ottenere lo **stesso numero di documenti per ogni classe**.

  → Classificare un numero sufficientemente alto di documenti in modo da averne un numero minimo per ogni classe, poi eliminare alcuni dei documenti appartenenti alle classi più rappresentate.

Creazione di un Modello - Risultati Preliminari

Utilizzando il dataset *silver* per la creazione di un modello molto leggero, i risultati sono incoraggianti:

- La classificazione di nuovi documenti risulta **molto veloce**.
- Le classi assegnate dal modello risultano corrette nella maggior parte dei casi.

TODO: I risultati del modello devono essere validati dagli esperti.