UNIONE EUROPEA
Fondo Sociale Europeo
Fondo Europeo di Sviluppo Regionale
Agenzia per la Coesione Territoriale
Ministero della Giustizia
Direzione Generale per il Coordinamento delle Politiche di Coesione
PON GOVERNANCE E CAPACITÀ ISTITUZIONALE 2014-2020

UNI 4 JUSTICE

UNIVERSITAS PER LA GIUSTIZIA. PROGRAMMA PER LA QUALITÀ DEL SISTEMA GIUSTIZIA E PER L'EFFETTIVITÀ DEL GIUSTO PROCESSO

D4.2.4 LABORATORIO DI AI E *DATA SCIENCE* PER GIURISTI

TOPIC MODELING PER TESTI LEGALI

Lista degli autori

Nome	Ateneo
Coordinatore	
Prof. Luca Manzoni	UNITS
Prof. Alberto Bartoli	UNITS

Autori

Dott.ssa Martina Saletta	UNITS
Dott.ssa Chiara Gallese	UNITS

Sommario

Il presente report presenta l'attività di ricerca svolta nell'ambito dello studio e analisi di testi legali con tecniche di intelligenza artificiale. Il lavoro descritto è raccolto in un articolo scientifico al momento in fase di ultimazione, che sarà a breve inviato alla rivista "Artificial Intelligence and Law" edita da Springer per essere sottoposto al processo di peer review.

1. Introduzione

Negli ultimi anni, la capacità di analizzare automaticamente testi scritti in linguaggio naturale è diventata sempre più importante poiché consente di elaborare efficientemente e rapidamente grandi volumi di documenti e file. Queste tecniche, basate sull'elaborazione del linguaggio naturale (in gergo, NLP) e sull'apprendimento automatico, rivestono un ruolo chiave nel rilevare tendenze nascoste, sentimenti, temi e altre informazioni preziose dai dati testuali, fornendo utili intuizioni e predizioni. L'applicazione di queste metodologie può risultare particolarmente vantaggiosa nel contesto legale, assistendo i professionisti del diritto in diverse attività come l'analisi dei casi, il reperimento di documenti, il clustering o la classificazione. Tuttavia, fino ad ora, pochi lavori si sono concentrati specificamente sul dominio legale italiano.

In questo contesto, questo articolo propone uno studio mirato ad analizzare le sentenze della Corte Costituzionale italiana. In particolare, si è interessati a studiare quali temi sono stati trattati dalla Corte Costituzionale, come sono cambiati nel tempo e come sono correlati a periodi e eventi storici o culturali. A tal proposito, il *topic modeling* è una tecnica utile poiché permette di identificare in modo statistico i soggetti o i temi trattati in un vasto corpus di documenti in modo non supervisionato. Poiché l'attenzione è rivolta all'esame delle tendenze nel tempo, è necessario introdurre una dimensione temporale nell'analisi. Invece di analizzare i temi emergenti da sottoinsiemi di documenti prodotti in diversi intervalli di tempo, come fatto in altri studi, in questo lavoro sono stati considerati i temi nell'intero corpus e poi è stato esaminato come il numero di documenti correlati ai diversi temi sia cambiato nel tempo. Questo approccio semplifica

notevolmente il collegamento tra i temi emergenti dall'analisi non supervisionata e le categorie concettuali degli operatori umani, similmente a quanto fatto in lavori precedenti per lo studio della letteratura scientifica in un dato campo di ricerca.

In sintesi, i contributi proposti sono i seguenti:

- un metodo basato su intelligenza artificiale per lo studio dei topic che emergono in un corpus di documenti in cui la componente temporale è particolarmente significativa;
- un'applicazione efficace del metodo proposto per esaminare l'intero corpus delle sentenze della Corte Costituzionale italiana dal 1956 al 2022;
- una discussione approfondita su come i cambiamenti nell'apparizione di alcuni temi nel corso degli anni siano correlati a eventi significativi della storia italiana.

2. Setting sperimentale

Lo studio è consistito in (1) pulizia delle trascrizioni delle sentenze della Corte Costituzionale, (2) indagine algoritmica dei temi trattati nel corpus di testo e (3) analisi di come l'importanza relativa dei temi nel tempo sia correlata ad eventi storici noti.

Dataset

Il dataset studiato è la collezione completa delle sentenze della Corte Costituzionale italiana degli anni tra il 1956 e il 2022, che consiste in circa 21.500 mila sentenze.
Ogni sentenza è contenuta in un file XML, la cui parte rilevante è rappresentata dalla concatenazione dei campi “collegio”, “epigrafe”, “testo” e “dispositivo”.
Per preparare i dati, abbiamo prima rimosso tutte le etichette XML e quindi eseguito le seguenti operazioni di pre-processing:

1. Decodifica di tutti i riferimenti di markup numerici nei corrispondenti caratteri ASCII.
2. Sostituzione delle abbreviazioni specifiche del dominio con le corrispondenti espressioni complete. Per questo passaggio, abbiamo utilizzato un elenco di circa 2.000 abbreviazioni comunemente usate nei testi giuridici italiani e ottenute da progetti correlati.
3. Rimozione dei caratteri numerici, della punteggiatura e delle stop words.
4. Conversione in minuscolo.
5. Riduzione delle parole flesse alle loro forme radice. Questo passaggio è stato eseguito utilizzando la versione italiana dell'algoritmo di stemming Porter (Porter, 1980).
6. Rimozione dei nomi e dei cognomi dei giudici che, nella storia, sono stati membri della Corte Costituzionale italiana.
7. Rimozione dei termini che sono apparsi in più del 50% dei documenti, poiché i termini molto frequenti sono generalmente considerati di scarso valore informativo, simili alle cosiddette "stop words" (Baeza-Yates e Ribeiro-Neto, 1999).

Il corpus risultante presentava un vocabolario composto da circa 64.000 parole.

Topic modeling

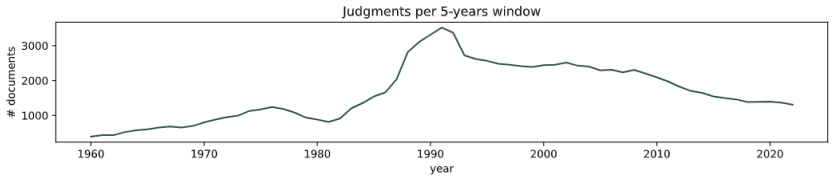
Abbiamo quindi eseguito il topic modeling sul corpus utilizzando il modello Latent Dirichlet Allocation (LDA) (Blei et al. 2003).

Nel topic modeling, uno degli elementi cruciali è lo stabilire quale sia il numero "ottimo" di topic da considerare. Per individuare tale numero, abbiamo considerato i valori di "coherence" (R̊oder et al, 2015) e di "silhouette" (Rousseeuw, 1987) calcolati sui vettori tf-idf (Sammut e Webb, 2010), ottenendo 15 come numero ottimo di topic.

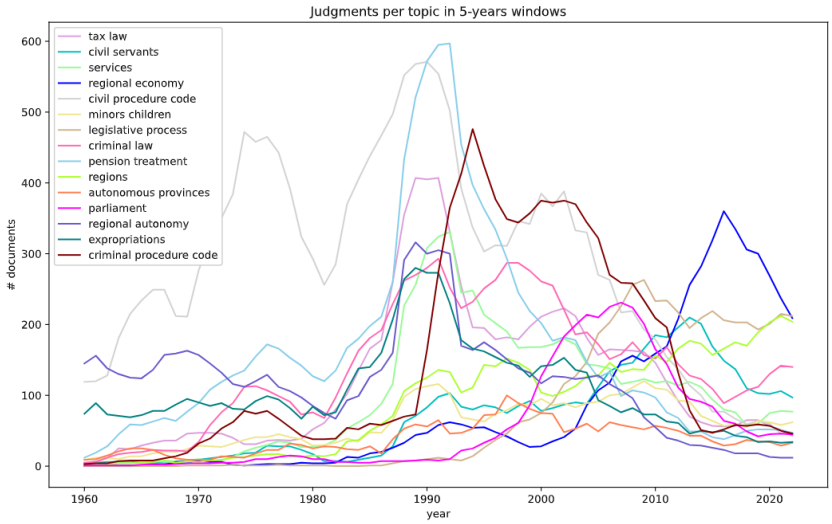
Analisi temporale

Abbiamo esaminato l'evoluzione dei topic nel tempo, in intervalli temporali di 5 anni con spostamenti di 1 anno, come segue. Ad ogni passo, abbiamo conteggiato il numero totale di documenti (Figura a) e il numero di documenti appartenenti a ciascuno dei 15 temi modellati (Figura

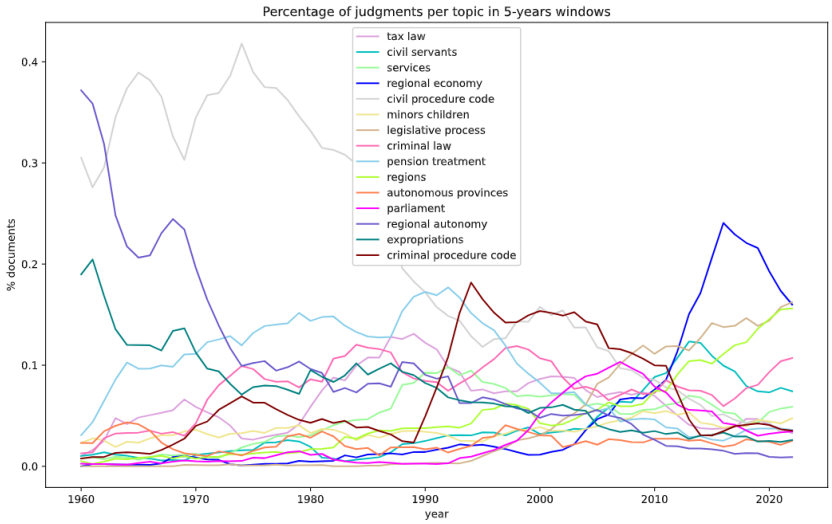
b). Poiché il numero di documenti varia negli anni, abbiamo anche misurato il rapporto dei documenti appartenenti a un determinato tema (Figura c).
Le figure mostrano che i topic presentano diverse tendenze, suggerendo interessanti spunti di riflessione. In generale, è interessante esaminare queste tendenze, poiché sono evidenza di topic che aumentano o diminuiscono la loro importanza nel corso degli anni, o che, in determinati



(a)



(b)



(c)

momenti storici, emergono per motivi che meritano di essere compresi.

3. Conclusione

In conclusione, lo studio descritto in questo report evidenzia l'importanza crescente dell'analisi automatica di testi mediante tecniche di elaborazione del linguaggio naturale (NLP), con particolare attenzione alle moderne tecniche di intelligenza artificiale che consentono di identificare, in modo automatico, i topic all'interno di grandi collezioni di documenti in modo non supervisionato.
Lo studio si è focalizzato sulle tematiche giuridiche emergenti dalle sentenze della Corte Costituzionale Italiana dal 1956 al 2022, analizzando inoltre l'evoluzione temporale di tali argomenti in termini di comparsa, prevalenza e cambiamenti nel corso del tempo. Per migliorare l'accuratezza dell'analisi, sono stati combinati metodi quantitativi e qualitativi. Inizialmente, è stata quantificata la presenza delle tematiche giuridiche, esaminandone la frequenza, la rilevanza e le variazioni nel corso del tempo. Successivamente, è stata effettuata un'analisi manuale per comprendere come le variazioni nella rilevanza di tali tematiche in determinati periodi fossero correlate a eventi storici.
Questo lavoro contribuisce all'evoluzione del campo di azione dell'intelligenza artificiale applicata al settore legale, dimostrando il potenziale delle tecniche NLP nell'analisi di ampi dataset giuridici. Le informazioni ottenute da questa ricerca costituiscono una base solida per futuri studi sulla dinamica del discorso legale, consentendo anche analisi comparative tra diverse giurisdizioni, lingue e periodi storici.

BIBLIOGRAFIA

Baeza-Yates RA, Ribeiro-Neto BA (1999) Modern Information Retrieval. ACM Press / Addison-Wesley.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J. Mach. Learn. Res. 3:993–1022

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137.

R ̈oder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM. ACM, pp 399–40.

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53 – 65.

Sammut C, Webb GI (eds) (2010) TF–IDF, Springer US, Boston, MA, pp 986–987