A pipeline for data management, knowledge extraction and semantic analysis of unstructured legal judgments

Chiara **Bonfanti**, Michele **Colombino**, Giorgia **Iacobellis**, Rachele **Mignone**, Ivan **Spada**, Laurentiu Jr Marius **Zaharia**, Marinella **Quaranta**, Marianna **Molinari**, Susanna **Marta**, Ilaria Angela **Amantea**, Davide **Audrito**

Supervisione: Luigi **Di Caro,** Emilio **Sulis** e Guido **Boella**

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

UNIVERSITÀ DI TORINO

NEXT GENERATION UPP: NUOVI SCHEMI COLLABORATIVI TRA UNIVERSITÀ E UFFICI GIUDIZIARI PER IL MIGLIORAMENTO DELL'EFFICIENZA E DELLE PRESTAZIONI DELLA GIUSTIZIA NELL'ITALIA NORD-OVEST
Progetto finanziato con il contributo dell'Unione europea, Programma Operativo Nazionale Governance e Capacità Istituzionale 2014-2020, Fondo Sociale europeo e Fondo europeo di sviluppo regionale
Codice progetto: CUP D19J22000240006.

NEXT GENERATION UPP

nuovi schemi collaborativi tra Università e uffici giudiziari per il miglioramento dell'efficienza e delle prestazioni della giustizia nell'Italia Nord Ovest

Linea 1.3

Definizione del catalogo delle attività e delle procedure per l'attivazione ed il potenziamento degli Uffici per il processo.

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

UNIVERSITÀ DI TORINO

NEXT GENERATION UPP: NUOVI SCHEMI COLLABORATIVI TRA UNIVERSITÀ E UFFICI GIUDIZIARI PER IL MIGLIORAMENTO DELL'EFFICIENZA E DELLE PRESTAZIONI DELLA GIUSTIZIA NELL'ITALIA NORD-OVEST
Progetto finanziato con il contributo dell'Unione europea, Programma Operativo Nazionale Governance e Capacità Istituzionale 2014-2020, Fondo Sociale europeo e Fondo europeo di sviluppo regionale
Codice progetto: CUP D19J22000240006.

Obiettivi

Pipeline per la definizione di un servizio di archiviazione, consultazione, classificazione automatica e ricerca semantica di sentenze che faciliti alcune delle attività dei magistrati e degli UPP.

Principali Tasks:

- Estrazione automatica e segmentazione di sentenze
- Classificazione automatica delle sentenze

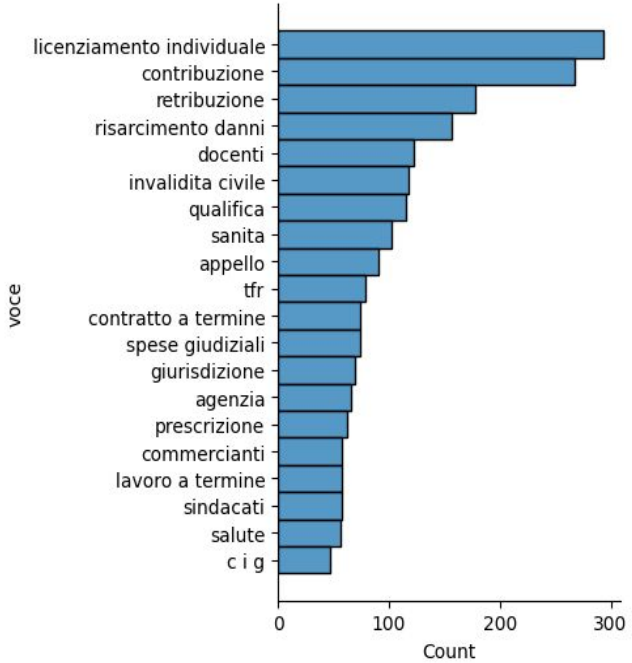
Future work:

- Criteri di similarità
- Estrazione dei principi di diritto

Estrazione e segmentazione

Rappresentazione in formato JSON

- Metadati:
  - Codice/anno sentenza
  - Codice/anno NRG
  - Voce di classificazione (dove presente)
- Contenuto
  - Oggetto
  - Conclusioni
  - Fatto e decisioni
  - P.Q.M



- **27477** sentenze, **I° e II°** grado, sezione **lavoro**, Torino.
- **4804** sentenze etichettate
- Distribuzione delle voci **fortemente sbilanciata**

Classificazione automatica

Datasets:

- Equilibrato: **800** sentenze su **10 voci** (dataset1)
- Non Equilibrato: **1872** sentenze su **15 voci** (dataset2)
- TF, TF-IDF
- Italian-Legal-BERT
- Doc2Vec

Modelli di machine learning:

- Support Vector Machine
- Logistic regression
- Random forest classifier

Risultati

- **98% accuracy,** dataset1 Random forest e Doc2Vec
- **95,5% accuracy**, dataset2 Logistic Regression e Doc2Vec

Test 2 - corpus 15 classes				
Dataset	**Random forest**	**SVM**	**Logistic Regression**	**Enseble Voting**
Average accuracy				
TF	0.784	0.776	0.802	0.816
TF-IDF	0.784	0.805	0.794	0.808
Ita-legal BERT	0.722	0.714	0.786	0.741
Doc2Vec	0.914	0.954	0.962	0.957
Average precision				
TF	0.859	0.829	0.791	0.765
TF-IDF	0.865	0.859	0.837	0.853
Ita-legal BERT	0.773	0.835	0.766	0.853
Doc2Vec	0.943	0.966	0.972	0.965
Average recall				
TF	0.730	0.723	0.785	0.788
TF-IDF	0.726	0.744	0.737	0.750
Ita-legal BERT	0.640	0.595	0.748	0.750
Doc2Vec	0.878	0.945	0.955	0.955

Future work

Criteri di similarità:

L'individuazione di criteri di similarità comporta diversi benefici:

- per migliorare la classificazione:
  - per **esito simile**: conforme o difforme
  - per **citazioni** a leggi e ad altre sentenze
- per definire uno strumento di **ricerca semantica** di sentenze

Estrazione dei principi di diritto:

Come arricchimento e supporto dei task di classificazione e similarità, i principi di diritto, espressi dalla corte di Cassazione, aiutano alla creazione di una nuova **metrica** semantica **di confronto** fra sentenze.

Grazie per l'attenzione

UNIVERSITÀ
DI TORINO

UniTO
Dipartimento di Informatica