IGSG-Marker, il software per la strutturazione automatica degli atti normativi in lingua italiana

Lorenzo Bacci

Descrizione

Il software open-source IGSG Marker consente l'analisi strutturale automatica di testi legislativi in lingua italiana non precedentemente annotati, ha licenza Apache 2.0, è consegnato come libreria Java ed è presente sul seguente repository GitLab:
https://gitlab.com/IGSG/MARKER/marker

L'*output* del Marker consiste in un documento *XML* composto dal testo presente nel documento passato come *input* arricchito con le annotazioni riguardanti la strutturazione ed alcuni metadati.

Lo *schema XML* scelto non segue un vero e proprio standard (nuovo o esistente), ma è orientato all'uso applicativo, di immediata comprensione e allo stesso tempo in grado di catturare tutta l'informazione possibile legata alla struttura presente nel testo in *input*, compresi alcuni elementi di formattazione, e facilmente trasformabile in altri formati prendendo solo ciò che necessario in funzione dell'applicazione.

Alcuni esempi di possibili applicazioni dell'*output XML* del Marker:

- trasformazione in formato *HTML* per presentare il testo normativo evidenziandone la struttura e le partizioni, in maniera graficamente più informativa e chiara rispetto al testo originale piatto;
- trasformazione in un altro formato o standard *XML* per documenti giuridici per integrazione con altre filiere applicative, *storage* dei documenti strutturati, etc.;
- identificazione di specifiche porzioni di testo in contesti di *Information Retrieval*;
- integrazione in generici ambienti redazionali o in applicazioni per l'acquisizione di testi normativi;
- fornisce la base per applicazioni che intendono estrarre ulteriore semantica dai testi normativi o che intendono fare analisi e ragionamenti su di essi (ad es.: applicare automaticamente gli emendamenti contenuti nelle *virgolette*, eseguire il *diff* strutturale di due versioni dello stesso testo normativo, fare statistiche sulle partizioni, trovare errori redazionali in un *corpus* normativo come errori di sequenza nelle partizioni e citazioni a partizioni inesistenti, facilitare il calcolo di *similarità* testuale e concettuale tra singole partizioni di atti normativi diversi, etc.).

Marker implementa al suo interno numerosi scanner lessicali compilati con JFlex (https://jflex.de) che contribuiscono a mantenere alta l'efficienza dell'analisi, permettendo l'utilizzo del software anche in servizi *on-the-fly*.