Procedure informatiche di tutela della trasparenza e riservatezza dei dati

Simone Marinai

Abstract: In questo contributo vengono inizialmente descritti i possibili tipi di anonimizzazione e vengono analizzati i formati di documenti su cui è necessario operare. Dopo aver analizzato lo stato dell'arte delle tecniche di anonimizzazione automatica di documenti viene descritto in dettaglio un prototipo di un applicativo di anonimizzazione semi-automatica di sentenze. Vengono infine analizzati i risultati sperimentali relativi all'utilizzo del prototipo nell'ambito del progetto Giustizia Agile.

1. Introduzione

La creazione di banche dati di merito richiede un bilanciamento fra vari interessi costituzionalmente rilevanti quali la tutela della riservatezza, la pubblicità del processo, la conoscibilità delle decisioni e la libertà di informazione. Al fine di tutelare la riservatezza delle parti in causa è necessario oscurare determinati dati personali dai documenti prima che questi vengano resi pubblici.

I dati che devono essere oscurati possono permettere di identificare le parti in causa sia direttamente (ad esempio quando sono presenti le generalità o l'indirizzo di residenza della persona coinvolta) che indirettamente (nel caso in cui la combinazione di informazioni generiche possa consentire l'identificazione della parte in causa a partire da informazione aggiuntiva). L'oscurazione di dati personali all'interno di documenti viene normalmente indicata come anonimizzazione (da *anonymization*) e in questo saggio utilizzeremo tale termine.

Dal punto di vista pratico, l'anonimizzazione dei dati personali presenti nei provvedimenti giurisdizionali deve essere effettuato su documenti di formato diverso a seconda dell'origine del provvedimento. In particolare, attualmente il processo civile è gestito prevalentemente per via digitale e sono impiegati applicativi (ad esempio il Sistema Informativo della Cognizione Civile Distrettuale, SICID) che supportano la gestione delle pratiche. Le sentenze possono essere depositate nell'applicativo SICID e una versione in formato MS-Word parzialmente anonimizzata può essere scaricata (utilizzando la funzione «epurazione dati sensibili» del programma) per una successiva anonimizzazione manuale.

Simone Marinai, University of Florence, Italy, simone.marinai@unifi.it, 0000-0002-6702-2277

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Simone Marinai, *Procedure informatiche di tutela della trasparenza e riservatezza dei dati*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0316-6.14, in Paola Lucarelli (edited by), *Giustizia sostenibile. Sfide organizzative e tecnologiche per una nuova professionalità*, pp. 213-228, 2023, published by Firenze University Press, ISBN 979-12-215-0316-6, DOI 10.36253/979-12-215-0316-6

Nel caso del processo penale, gestito prevalentemente sotto forma cartacea, al momento sono disponibili soltanto sentenze ottenute da scannerizzazioni e memorizzate in formato PDF.

In questo saggio faremo prima un cenno ai possibili tipi di anonimizzazione e ai possibili tipi di documenti su cui è necessario operare per poi passare a descrivere un prototipo di un applicativo di anonimizzazione semi-automatica che è stato utilizzato per elaborare sentenze nell'ambito del progetto Giustizia Agile.

2. Obiettivi dello studio

L'anonimizzazione di documenti è un'operazione complessa che normalmente richiede un notevole sforzo e tempo non trascurabile da parte di operatori umani che abbiano familiarità con il dominio di interesse. A meno di lavorare direttamente su documenti cartacei è poi necessario utilizzare appositi programmi per la modifica dei documenti. Identificare gli elementi di interesse all'interno dei documenti richiede una comprensione profonda del documento in oggetto e la messa in opera di opportune scelte pratiche che permettano di bilanciare la leggibilità con il rischio della possibilità di identificare (direttamente o indirettamente) le parti in causa. È evidente che un documento non oscurato massimizza la sua leggibilità e massimizza la trasparenza, ma rende possibile l'identificazione delle parti in causa. Per contro un documento in cui siano state oscurate tutte le parole rende impossibile l'identificazione delle parti, ma rende ovviamente al contempo impossibile comprendere l'oggetto del documento stesso. È quindi necessario pervenire ad un livello di oscurazione che permetta un bilanciamento tra leggibilità e protezione della riservatezza.

2.1 Definizioni utili per l'oscuramento dei dati

Per poter comprendere meglio i concetti di interesse, è utile dare una descrizione di alcuni termini che sono di interesse per questa discussione, seguendo (Lison et al. 2021). Un *identificatore diretto* è un insieme di variabili che sono uniche per un individuo o una entità (ad esempio una società). Appartengono a questa categoria i nomi (propri o societari), gli indirizzi, i numeri di telefono, i codici IBAN, i codici fiscali e le partite IVA. Tutte queste informazioni permettono di identificare direttamente l'entità. Un *quasi-identificatore* (*quasi-identifier*) è un insieme di informazioni che individualmente non consentono di risalire all'entità corrispondente, ma che in combinazione con altri quasi-identificatori e con informazione aggiuntiva possono permettere di risalire all'entità. Ad esempio, fanno parte di questa categoria informazioni quali il genere, la nazionalità, la città di residenza, la professione. La caratterizzazione di un'informazione come quasi-identificatore non è semplice, può essere soggettiva e dipende dal contesto. Nel caso specifico delle sentenze, in via ipotetica potrebbe essere aggiunto a queste informazioni il reato contestato ad un imputato: l'oscurazione di questa informazione renderebbe di fatto inutile la pubblicazione della sentenza, ma al tempo stesso la presenza di un reato specifico in combinazione con informazio-

ni relative al genere, nazionalità, città di residenza e professione potrebbe consentire in casi particolari di risalire alla persona coinvolta incrociando queste informazioni con notizie di cronaca apparse sui quotidiani.

Dal punto di vista delle operazioni di oscuramento che si possono prendere in considerazione si parla prevalentemente di anonimizzazione, de-identificazione e pseudo-anonimizzazione. L'*anonimizzazione* è la rimozione completa e irreversibile di ogni informazione che possa portare all'identificazione di un soggetto sia direttamente che indirettamente. La *de-identificazione* è il processo di rimozione di specifici identificatori diretti da un documento o una collezione di dati. La *pseudo-anonimizzazione* è il processo di sostituzione di identificatori diretti e indiretti con opportuni valori codificati o pseudonimi (ad esempio sostituendo *Mario Rossi* con *Parte1*). La mappatura tra codici e identificatori reali può essere conservata separatamente rispetto al documento o cancellata, rendendo in tal caso irreversibile la trasformazione del documento.

2.2 Principali formati di file per documenti

I documenti da anonimizzare possono essere disponibili in vari formati a seconda della loro origine e in particolare dei programmi utilizzati per la loro generazione e/o acquisizione. La differenza principale è tra documenti facilmente modificabili dagli utenti (ad esempio documenti MS-Word) e documenti in formati non modificabili, ma il cui contenuto è facilmente ricercabile (ad esempio file PDF *digital-born*). Infine, possiamo considerare documenti ottenuti da un processo di digitalizzazione tramite scanner di documenti cartacei, che sono prevalentemente memorizzati e distribuiti come file PDF. Sebbene a prima vista gli ultimi due tipi di documenti appaiano analoghi, in realtà la differenza è sostanziale e ad esempio il testo all'interno di documenti digitalizzati talvolta non è ricercabile. L'estrazione di informazione da documenti PDF è tuttora oggetto di studio in particolar modo quando si ha a che fare con documenti complessi come gli articoli scientifici (Gemelli, Vivoli e Marinai 2022). Nel caso di sentenze l'informazione è prevalentemente di tipo testuale e quindi è possibile analizzare il testo all'interno dei documenti tramite tool ad accesso aperto liberamente disponibili.

Per consentire una migliore fruibilità del documento anonimizzato è necessario generare un documento che rispecchi la formattazione originale e in cui il testo non oscurato sia ricercabile per mezzo dei comuni strumenti di ricerca di testo.

Partendo da tali aspetti si è affrontato lo studio e l'implementazione di un prototipo finalizzato all'anonimizzazione di sentenze che potesse essere utilizzato per elaborare sentenze nell'ambito del progetto Giustizia Agile.

3. Stato dell'arte su tecniche automatiche di anonimizzazione

Sebbene esuli dagli scopi di questo lavoro una descrizione accurata delle principali tecniche di anonimizzazione è utile un breve richiamo ai diversi metodi che possono essere presi in considerazione per tale finalità.

Due sono le principali aree in cui è significativa la presenza di informazione personale anonimizzata e di conseguenza è cruciale lo sviluppo di tecniche automatiche di anonimizzazione: l'ambito clinico-medico nel caso della condivisione di dati clinici e l'ambito legale nel caso della condivisione di documenti con particolare riguardo alle sentenze (Lison et al. 2021; Csányi et al. 2021).

In generale, ci sono alcune differenze tra il tipo di informazione che deve essere gestita nei due casi.

Le informazioni cliniche si trovano prevalentemente, ma non esclusivamente, sotto forma di informazione strutturata (record in tabelle relazionali) che potrebbe comunque contenere elementi non-strutturati, ad esempio brevi testi in referti di visite specialistiche. In questo caso gli identificatori diretti sono spesso facilmente identificabili (ad esempio il nome del paziente è generalmente memorizzato in un apposito campo della base di dati) ed è rara la presenza di tali identificatori in testo libero (questo è il caso dell'indicazione del nome di un medico nel testo dell'anamnesi). La maggior difficoltà per l'anonimizzazione è in questo caso dovuta alla presenza di molte informazioni aggiuntive sul paziente (quasi-identificatori) che nel loro insieme possono talvolta permettere di identificare le entità coinvolte. In questo contesto è di interesse il concetto di *k-anonymity* (Chakaravarthy et al. 2008) per il quale una collezione di dati soddisfa la *k-anonymity* per $k>1$ quando sono presenti del dataset almeno k record per ogni combinazione di quasi-identificatori. Normalmente, in campo clinico l'informazione testuale è limitata a testi relativamente brevi in cui sono presenti molti acronimi o abbreviazioni e al contempo non mancano errori di digitazione.

Nel campo legale la situazione è per qualche verso complementare. Le informazioni in forma strutturata sono probabilmente meno numerose, ma al contrario la parte testuale è preponderante. Ad esempio, considerando le sentenze, si possono avere documenti lunghi una sola pagina e altri la cui estensione è dell'ordine delle centinaia di pagine. In questo caso la prosa è accurata e pur seguendo regole generali può variare tra autori diversi. Oltre alla lunghezza dei documenti è utile tenere in considerazione il fatto che i documenti legali contengono una gran quantità di quasi-identificatori relativi alle parti coinvolte che potrebbero essere utilizzati per risalire all'identità reale (Csányi et al. 2021). Identificare queste informazioni in modo automatico non è semplice e allo stesso tempo, come detto precedentemente, la rimozione di eccessiva informazione potrebbe rendere ardua la comprensione del documento.

Nei prossimi paragrafi descriviamo brevemente i principali approcci che possono essere utilizzati per l'anonimizzazione automatica di testi.

3.1 Metodi basati su regole

I primi metodi per l'anonimizzazione di record medici hanno seguito strategie simili a quelle impiegate in vari strumenti per l'estrazione di informazioni (*information extraction*) da documenti semi strutturati (Witten 2004). In questi casi gli approcci sono basati sull'uso di dizionari, espressioni regolari e regole definite manualmente da esperti del dominio di interesse (ad es. Gupta, Saul and Gilbertson 2004).

Questi approcci consentono di raggiungere un'alta precisione, anche se i risultati sono difficilmente generalizzabili ad altri contesti. Tali soluzioni possono pertanto fornire risultati insoddisfacenti quando applicate a documenti dissimili da quelli per i quali sono state messe a punto.

In particolare, i *dizionari* contengono parole specifiche che devono essere preservate dall'oscurazione o che tipicamente precedono parole da oscurare (o meno). Ad esempio, nel caso di sentenze un nome di città che segua la parola 'Tribunale' non dovrà probabilmente essere oscurata.

Una *espressione regolare* può essere vista come una funzione che analizza sequenzialmente i simboli in una sequenza (normalmente caratteri in un testo) ed evidenzia la presenza di un pattern predefinito all'interno della sequenza. Inizialmente concepite ed utilizzate nell'ambito dello sviluppo di compilatori nei linguaggi di programmazione, le espressioni regolari trovano ampio utilizzo nel campo dell'elaborazione del linguaggio naturale (*Natural Language Processing*, NLP (Jurafsky and Martin 2000). Nell'analisi di testo le espressioni regolari permettono di definire in modo conciso e formalmente corretto forme alternative di un concetto, ad esempio cercando parole alternative (via o piazza) o varianti di un termine (via o viale, parte o parti).

Le *regole* permettono di combinare espressioni regolari e termini presenti in dizionari per stabilire se una determinata parola debba essere offuscata o meno. Ad esempio, una parola che inizi per lettera maiuscola e che sia presente in un dizionario di città dovrà essere preservata se segue 'Tribunale di', ma dovrà essere oscurata nel caso opposto.

È evidente che la definizione di opportune regole per l'anonimizzazione di testo è un'operazione che richiede l'intervento di esperti nel dominio di interesse.

La disponibilità di collezioni di documenti annotate da esperti consente lo sviluppo e l'utilizzo di tecniche di anonimizzazione basate su modelli di apprendimento automatico che sono descritti nel seguente paragrafo.

3.2 Metodi basati su tecniche di apprendimento automatico

Analogamente a quello che avviene in altri contesti applicativi, i sistemi basati su regole hanno limitazioni relative alla generalizzazione del loro comportamento nel caso di documenti diversi da quelli presi come riferimento dagli esperti che hanno contribuito alla definizione delle regole.

Se queste limitazioni sono critiche, l'utilizzo di tecniche di apprendimento automatico, più recentemente basate su approcci deep-learning, consente di aumentare la generalizzazione dei metodi sviluppati.

Approcci per l'anonimizzazione basati sull'apprendimento automatico sono stati sviluppati nell'ambito di tecniche di elaborazione del linguaggio naturale e in particolare utilizzando approcci per il *Named-Entity Recognition* (NER) ovvero il riconoscimento di entità con nome. Con il termine NER si intendono entità con nome, in cui l'entità può essere una persona, un luogo o un'azienda. In senso esteso si considerano anche numeri e date come entità da identificare.

Vista la natura degli identificatori che devono essere oscurati è abbastanza naturale che vari approcci per l'anonimizzazione si basino in modo significativo su tecniche di NER. In particolare, approcci per NER basati su apprendimento automatico sono stati utilizzati impiegando, ad esempio, *Support Vector Machine* (SVM), *Conditional Random Field* (CRF) e reti neurali ricorrenti (Garat and Wonsever 2022).

Più recentemente, lo sviluppo di *neural language model* e in particolare la ricerca su modelli di intelligenza artificiale basati su transformer hanno portato ad approcci in cui modelli linguistici pre-appresi (ad esempio BERT e sue varianti) sono stati impiegati per effettuare NER su documenti legali (Garat and Wonsever 2022; Di Martino et al. 2021).

Nella ricerca che stiamo portando avanti relativa all'anonimizzazione di sentenze abbiamo inizialmente preso in considerazione approcci per NER basati su transformer. Tuttavia, i risultati iniziali non sono stati soddisfacenti per i nostri obiettivi, sebbene siano stati utilizzati modelli pre-appresi su documenti legali italiani. Per ottenere risultati soddisfacenti nel tempo a disposizione per il progetto Giustizia Agile, abbiamo realizzato un primo prototipo basato su regole per l'anonimizzazione di sentenze di interesse. Vista la relativa omogeneità delle sentenze prese in considerazione i risultati ottenuti appaiono più che soddisfacenti come vedremo nell'analisi sperimentale. Nei seguenti paragrafi viene descritto il primo prototipo realizzato.

4. Un prototipo per anonimizzare sentenze

Lo sviluppo del prototipo di applicativo per l'anonimizzazione di sentenze ha seguito varie fasi, elencate nel seguito e descritte nei successivi paragrafi.

- Analisi dei requisiti ovvero delle caratteristiche dei documenti da anonimizzare e delle necessità di oscuramento dei dati;
- sviluppo di un primo prototipo per l'anonimizzazione di sentenze, basato su regole comprensibili anche da utenti senza competenze informatiche;
- impiego del prototipo per l'anonimizzazione di sentenze nell'ambito del progetto Giustizia Agile;
- sviluppo di un programma per l'anonimizzazione «manuale» di parole che non siano state identificate automaticamente.

4.1 Analisi dei requisiti

Una prima fase dello studio ha comportato una interazione con i giuristi dell'Università degli Studi di Firenze che hanno partecipato al progetto Giustizia Agile per determinare le principali necessità e peculiarità del processo di anonimizzazione desiderato. In tale studio sono state analizzate in particolare le caratteristiche dei documenti gestiti tramite l'applicativo SICID e le limitazioni del processo di anonimizzazione presente in tale sistema. Sono state inoltre analizzate in dettaglio le caratteristiche dei principali documenti presenti in ambito civile e penale.

A seguito di tale studio sono stati identificati i dati che è necessario epurare secondo il seguente schema:

- nomi di tutte le parti (genitori, minori, coniugi, nonni, fratelli, parenti in generale, famiglie affidatarie ecc.) con date di nascita ed eventuali codici fiscali;
- relativamente agli atti citati, devono essere epurati i dati (es. registrazione Agenzia Entrate, repertorio notarile, atti anagrafe, atti di matrimonio con luogo e data);
- I Sezione: atti di matrimonio con luogo e data; numeri sentenze di separazione e divorzio appellate e quant'altro eventualmente richiamato: solo per le cause di famiglia e riconoscimento paternità, danno endo-familiare, nullità matrimonio, interdizione e inabilitazione, divorzio, separazione, famiglia di fatto, tutte le modifiche, ads, delibazione sentenze straniere e sacra rota in materia di famiglia e diritti della persona (stranieri, cittadinanza, permesso per ricongiungimento per ragioni familiari) e tutte le cause trattate dalla sezione minori, il tutto purché definito con sentenza;
- identificazione catastale;
- nomi dei testi;
- targhe autoveicoli;
- dati giudiziari ma solo se pertinenti alle parti (decreti ingiuntivi);
- indirizzo luoghi lavoro, residenza, domicilio (e anche nome degli istituti e delle case-famiglia ecc. e dei Servizi Sociali, SERD SERT, Ufsmia e UFSMA).

Idealmente le parti epurate dovrebbero essere sostituite secondo il seguente schema.

- Nomi delle parti in causa sostituiti con 'XX', 'YY', 'ZZ' …
- Nomi dei testi sostituiti con 'T1', 'T2', …
- Date e luoghi, dati anagrafici, dati giudiziari pertinenti alle parti e ulteriori dati da oscurare sostituiti con '-------'

Al momento, il prototipo del programma di anonimizzazione gestisce i dati indicati nello schema precedente in modo abbastanza grossolano gestendo il testo da un punto di vista sintattico piuttosto che semantico. Inoltre, per il momento non permette di sostituire i dati oscurati differenziando le parti e i testi. È importante osservare che vengono anonimizzate parole (o insiemi di parole successive) e non entità e quindi ad esempio per la persona 'Mario Rossi' vengono anonimizzate separatamente le parole 'Mario' e 'Rossi'.

4.2 Sviluppo del prototipo

Il prototipo per l'anonimizzazione di sentenze è stato implementato in linguaggio Python utilizzando librerie open-source per estrarre informazioni da documenti PDF e per generare i documenti anonimizzati. In Figura 1 è riassunta la pipeline di elaborazione dei documenti descritta nel seguito.

Nell'ottica di semplificare l'implementazione del programma e il suo utilizzo, vengono elaborate con questa procedura sia sentenze derivanti da file MS-

Word e successivamente salvate in formato PDF (tipicamente provenienti dal programma SICID) che sentenze scannerizzate e memorizzate ancora in formato PDF (presenti nel caso di sentenze penali).

I file delle sentenze scannerizzate sono di fatto sotto forma di immagini anche se contenuti in file PDF. A seconda del programma utilizzato per la scannerizzazione al momento della digitalizzazione è possibile che il testo sia accessibile (ricercabile dentro ad un programma di visualizzazione di PDF, ad esempio Adobe Acrobat) o meno. In ogni caso non è direttamente modificabile se non utilizzando programmi specifici per gestire documenti PDF. Per poter accedere alle informazioni testuali del provvedimento è in questo caso necessario utilizzare un programma per il riconoscimento di testo da immagini (OCR: *Optical Character Recognition*). A seguito di una analisi delle alternative è stato deciso di impiegare il software open-source *Tesseract* (Smith 2007). I programmi per OCR permettono di convertire automaticamente testo stampato (talvolta anche testo scritto a mano) in documenti in formato modificabile. *Tesseract* OCR, progetto attualmente sponsorizzato da Google, è un motore di riconoscimento ottico dei caratteri open-source rilasciato dal 2005 a partire da software precedentemente sviluppato da HP. *Tesseract* è uno dei motori più popolari e ampiamente utilizzati al mondo, grazie alla sua accuratezza e alla disponibilità gratuita del codice sorgente e di modelli per il riconoscimento di testo in varie lingue (incluso ovviamente l'italiano). È scritto in C++ con un'interfaccia di programmazione in diversi linguaggi, tra cui Python, linguaggio di programmazione utilizzato in questo progetto.

Dopo aver riconosciuto il testo all'interno delle sentenze il documento viene nuovamente salvato tramite la libreria di *Tesseract* in formato PDF senza modificare l'immagine, ma sovrapponendo a questa il testo riconosciuto con un font trasparente. Questo consente ad appositi programmi di estrarre il testo con la corrispondente posizione nella pagina per procedere all'effettiva anonimizzazione.

Come detto precedentemente in questo modo è possibile elaborare con la stessa procedura sia documenti scannerizzati che documenti PDF *digital-born* in cui il testo è ricercabile. Anche in presenza di documenti in formato Word è sempre possibile salvare tali documenti in PDF in modo da utilizzare un solo programma di anonimizzazione.

Il testo nei file PDF (indipendentemente dall'origine) viene estratto utilizzando la libreria Python open-source *PyMuPDF* che consente di identificare dal documento le parole con le informazioni sulla loro posizione nella pagina per poterle successivamente eventualmente oscurare.

Una volta identificate le parole da anonimizzare, secondo le regole descritte successivamente, vengono oscurate tramite rettangoli bianchi ed infine viene impiegato nuovamente il software *Tesseract* OCR per rendere cercabile la parte visibile del documento analizzato. Una prossima versione del programma consentirà di inserire diversi identificativi in sostituzione delle parole anonimizzate come richiesto dalle specifiche sopra riassunte.

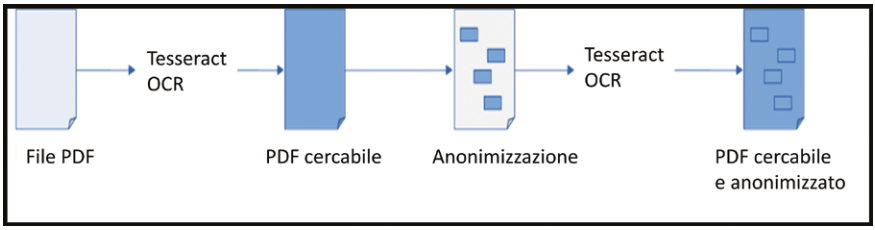

Figura 1 – Pipeline di elaborazione dei documenti.

4.2.1 Estrazione del testo dalle sentenze

Tramite l'uso della libreria Python *PyMuPDF*, è possibile estrarre il testo dal documento in formato PDF, mantenendo l'informazione sulla posizione della parola all'interno della pagina (*bounding box*) e andarla a oscurare inserendo sopra la parola un rettangolo bianco bordato di nero.

Tutti i termini estratti dal file PDF vengono preliminarmente ripuliti da caratteri speciali così da non incorrere in errori dovuti ad erroneo riconoscimento da parte del programma di OCR. Ad esempio, il termine 'ROSSI,' diventa 'ROSSI', rimuovendo la virgola attaccata alla parola.

4.2.2 Ricerca dati da epurare

La ricerca di parole da anonimizzare nel testo estratto da *PyMuPDF* avviene utilizzando regole descritte tramite Espressioni Regolari, per identificare all'interno del testo pattern ben definiti.

Le principali regole implementate riguardano la ricerca di:

- parole con la prima lettera maiuscola;
- parole tutte maiuscole;
- date in vari formati, anche con il mese scritto a parole;
- sentenze;
- codici fiscali;
- numeri di telefono;
- CAP;
- targhe di veicoli;
- partite IVA;
- indirizzi e-mail.

Queste regole sono state scelte per rispondere all'esigenza di andare ad oscurare le informazioni nella sentenza, secondo quanto indicato nelle specifiche descritte dai giuristi membri del progetto. Le prime due voci (parole con la prima lettera maiuscola e parole tutte maiuscole) permettono in genere di identificare nomi di persone, enti o indirizzi. Le date seguono svariati formati che possono comprendere solo cifre o includere i nomi dei mesi per esteso o abbreviati. Le

regole prese in considerazione permettono di gestire tutti i casi presenti nelle sentenze analizzate nello studio.

Oltre all'uso delle espressioni regolari, una parte del codice effettua un controllo semantico del testo, definendo delle regole legate alla successione delle parole significative nella frase. Più nello specifico, esistono regole semantiche che prevedono l'anonimizzazione di parole che normalmente non dovrebbero essere oscurate, ma seguono certi termini specifici. Altre regole lasciano in chiaro parole che normalmente dovrebbero essere oscurate, ma seguono altri termini specifici, ed altre ancora che non oscurano parole anche se nel testo appaiono in maiuscolo.

In particolare, sono oggetto di anonimizzazione tutte quelle parole che seguono determinati termini come:

- 'c/c', 'fattura', 'fatt', 'repertorio', 'rep', 'racc', 'data', 'cf', 'iva', 'piva', per andare ad oscurare dati relativi alle parti;
- 'mappa', 'particella', 'part', 'sub', 'cat', per andare ad oscurare dati relativi ad identificazioni catastali.

Come controllo finale, per assicurare una corretta identificazione di tutte le parole da oscurare, vengono inizialmente raccolte in un dizionario tutte le parole che rispondono alle regole sopra definite e successivamente viene processato nuovamente il documento. Per ogni termine presente nel documento viene controllato se questo è contenuto nel dizionario di parole da oscurare e in caso positivo viene inserito il blocco di copertura al di sopra della parola.

4.2.3 Visualizzazione di determinate parole

Per preservare quanto possibile la leggibilità del documento anonimizzato non vengono oscurate determinate parole in deroga a quanto sopra descritto. In particolare, non sono oggetto di anonimizzazione tutte quelle parole che seguono termini come:

- 'corte', 'appello', 'tribunale', 'cassazione', 'cass', 'suprema', 'legittimità', 'cassa', 'dm', 'previdenza', 'costituzionale', 'consiglio', 'stato', 'cgue', 'cedu', 'conti', 'sentenza', 'sentenze', 'udienza', 'ludienza', 'delludienza', 'alludienza', 'pqm', 'repubblica', 'popolo', 'nrg', 'rg', 'repert', 'registrato', 'registrata', 'pubbl', 'legge', 'dl', 'decreto', 'legislativo', 'dlgs', 'lgs', 'testo', 'tu', 'regolamento', 'reg', 'direttiva', 'dir', 'proc', per lasciare in chiaro leggi o numeri di sentenze;
- 'dallavv', 'dallavvto', 'dellavv', 'dellavvto', 'avvto', 'avv', 'avvocato', 'giudice', per lasciare in chiaro il nome proprio di avvocati e giudici.

Si osservi che alcune delle parole sopra indicate derivano dal processo di eliminazione di caratteri speciali (ad esempio 'dall'avv' diventa 'dallavv').

Inoltre, non sono oggetto di anonimizzazione i seguenti termini legati al contesto giuridico:

- 'attore', 'convenuto', 'ricorrente', 'resistente', 'reclamante', 'reclamato', 'appellante', 'appellato', 'terzo', 'chiamato', 'cassazione', 'cass', 'corte', 'appello', 'decreto', 'ingiuntivo', 'di', 'sentenza', 'sent', 'ordinanza', 'ord', 'sezione', 'sez', 'decreto', 'legge', 'dl', 'legislativo', 'dlgs', 'giurisdizione', 'competenza', 'legittimazione', 'attiva', 'passiva', 'nullità', 'risoluzione', 'recesso'.

4.2.4 Esempi di applicazione delle regole

Le regole precedentemente menzionate sono state scelte per rispondere all'esigenza di oscurare le informazioni nella sentenza legate a vari fattori. Nel seguito illustriamo alcune delle regole per mezzo di esempi fittizi corrispondenti alla forma di sentenze reali, ma contenenti dati immaginari per non esporre ritagli di sentenze.

1. Tutti i nomi propri, solitamente presenti nella sentenza con la prima lettera maiuscola o tutti in maiuscolo, devono essere oscurati, tranne quelli degli avvocati e giudici citati nella sentenza in oggetto.
2. Tutte le informazioni relative alle parti, oltre al nome proprio, come indirizzi di residenza o domicilio, codice fiscale, nazionalità e simili.

Un esempio di applicazione di queste regole è riportato in Figura 2.

AZIENDA S.R.L. (C.F. 01234567890) con sede in Firenze in via dei Preti, 135

[ ] (C.F. [ ]) con sede in [ ] in via [ ]

Figura 2 – Nell'esempio, relativo ad un'azienda, vengono oscurati il nome, il codice fiscale e l'indirizzo della sede della parte.

3. Tutti gli atti e sentenze citate prettamente legate alle parti in causa come per esempio sentenze di separazione, divorzio, matrimonio, minorili e simili (Figura 3).

MARIO ROSSI, nato a FIRENZE (FI) il 01/01/1980; trascritto nei registri degli atti di matrimonio del Comune di Firenze al nr. 12 Parte II Serie A anno 2000;

[ ], nato a [ ] ([ ]) il [ ]; trascritto nei registri degli atti di matrimonio del Comune di [ ] al nr. [ ] Parte [ ] Serie [ ] anno [ ];

Figura 3 – Nell'esempio vengono oscurati il nome e la data di nascita della parte. Inoltre, vengono oscurati i dati relativi agli atti del matrimonio, compreso il comune di registrazione. Viene tuttavia lasciata in chiaro la struttura della frase.

4. Tutte le date rappresentate in forma breve (GG.MM.AA o GG.MM.AAAA), in forma estesa (GG mese AAAA) e con caratteri speciali diversi dal '.' (GG-MM-AA o GG/MM/AA). Esempio in Figura 4.

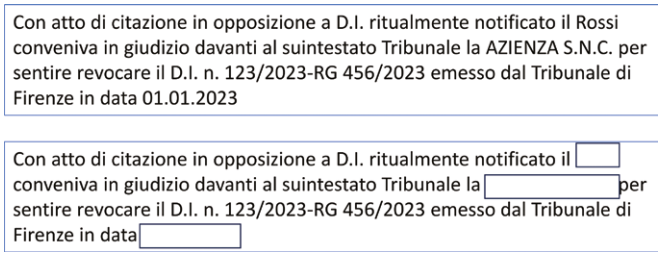

Figura 4 – Nell'esempio viene oscurata la data al termine della frase.

5. Tutte quelle informazioni identificative di beni legati alle parti, come per esempio targhe di veicoli o simili (Figura 5).

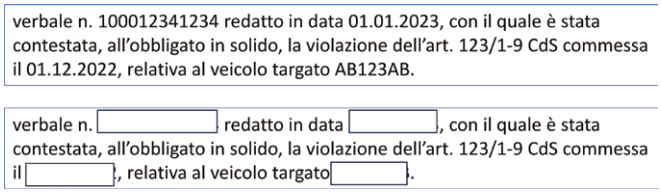

Figura 5 – Nell'esempio vengono oscurati il numero di verbale, la targa del veicolo e le date in cui è stata commessa la violazione e quella in cui è stata contestata.

6. Tutte le località (esempio in Figura 6).

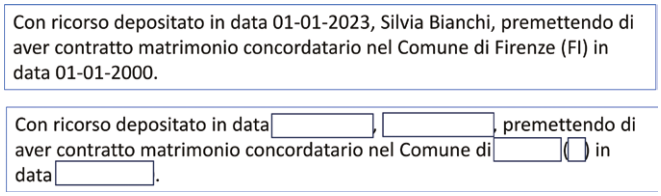

Figura 6 – Nell'esempio viene oscurato il nome della città (Firenze), ma viene conservata la frase 'nel Comune di' per migliorare la leggibilità della sentenza anonimizzata.

7. Tutte le identificazioni catastali (esempio in Figura 7).

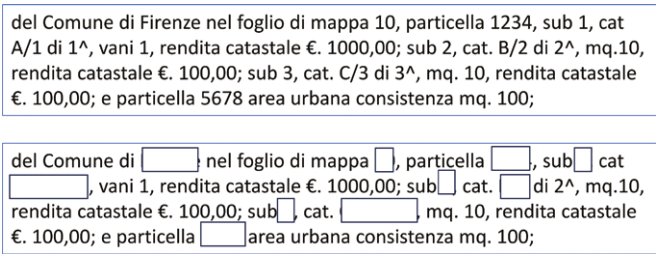

Figura 7 – Nell'esempio vengono oscurati i dati catastali (comunque fittizi e non reali).

4.2.5 Generazione documento anonimizzato

Al termine dell'anonimizzazione il file PDF prodotto contiene immagini delle sentenze sulle quali sono stati applicati rettangoli bianchi in corrispondenza dei dati da epurare. Trattandosi di immagini, come descritto in precedenza, non è possibile cercare del testo all'interno di questi file rendendo più difficile la consultazione dei documenti anonimizzati. Ritenendo la facilità di consultazione un obiettivo fondamentale dell'applicativo realizzato, si è deciso di eseguire nuovamente l'OCR sul risultato dell'anonimizzazione in modo che l'output prodotto dall'applicativo fosse un documento anonimizzato e comunque consultabile con agilità.

Un esempio è riportato in Figura 8.

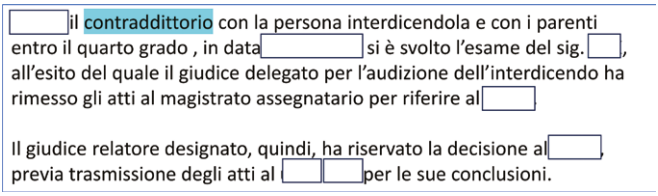

Figura 8 – Esempio di risultato finale in cui si evidenzia la possibilità di cercare del testo all'interno del documento (ad es. 'contraddittorio').

4.3 Valutazione sperimentale del prototipo

Il prototipo sviluppato è stato impiegato per anonimizzare sentenze in campo civile provenienti prevalentemente dal Tribunale di Prato. In totale sono stati elaborati 705 file PDF corrispondenti ad altrettante sentenze. La lunghezza media delle sentenze è di 9,63 pagine; la sentenza più corta ha 3

pagine, mentre la più lunga ne ha 55. Il numero medio di parole per ogni sentenza è pari a 3311 delle quali in media ne sono state anonimizzate 305. Il tempo medio di elaborazione (comprensiva di tutte le fasi) è stato di 4,1 secondi per ogni pagina.

Le sentenze elaborate dal sistema sono state successivamente analizzate dai giuristi partecipanti al progetto. A seguito di questa analisi è risultato che 642 sentenze sono state anonimizzate correttamente dal programma, mentre per le restanti 63 sentenze è stata necessaria una correzione manuale nella quali sono state corrette mediamente 2,4 parole per ogni sentenza. Riteniamo che questi risultati siano più che soddisfacenti e in linea con le aspettative degli utenti.

Per testare l'applicativo con sentenze acquisite tramite digitalizzazione sono state elaborate alcune sentenze provenienti dall'Ufficio GIP del Tribunale di Firenze. In totale sono stati elaborati 1136 file PDF corrispondenti ad altrettante sentenze. La lunghezza media delle sentenze è di 7,24 pagine; la sentenza più corta ha 2 pagine, mentre la più lunga ne ha 226. Il numero medio di parole per ogni sentenza è pari a 1575 delle quali in media ne sono state anonimizzate 271.

4.4 Applicativo per la correzione manuale di sentenze anonimizzate

Al momento si ritiene che l'anonimizzazione automatica non possa essere perfetta e che sia sempre possibile che siano presenti Falsi Positivi (FP) e Falsi Negativi (FN) anche se auspicabilmente in misura ridotta. I FP sono parole anonimizzate che non lo dovevano essere, mentre i FN sono parole da anonimizzare che non lo sono state. I FP rendono il documento meno leggibile, mentre i FN sono sicuramente problematici dal punto di vista della tutela della riservatezza delle parti coinvolte.

Come si è visto dal test dell'applicativo, il processo di anonimizzazione non è perfetto e talvolta si producono sentenze all'interno delle quali rimangono in chiaro dei dati che dovrebbero essere oscurati. Fortunatamente questo ha coinvolto soltanto il 10% delle sentenze elaborate. Si ritiene che questa percentuale si potrà ridurre ulteriormente con successivi miglioramenti del programma integrando il sistema basato su regole con tecniche di apprendimento automatico.

Per gestire le sentenze con errori di anonimizzazione è stato modificato l'applicativo per permettere all'utente di correggere questi errori manualmente, specificando quali siano i termini da oscurare che siano rimasti dopo il primo controllo.

In particolare, l'utente dopo aver letto il documento anonimizzato, deve compilare un file di testo specificando per ogni sentenza le parole che dovranno essere nascoste. Ogni riga del file di correzione dovrà contenere la coppia di valori come nel seguente formato (nome del file; parola da oscurare). L'applicativo di anonimizzazione analizza il documento seguendo la procedura sopra descritta. Tuttavia quando viene identificata una delle parole contenute nel file questa viene oscurata a prescindere dall'insieme di regole che sarebbero normalmente applicate.

5. Conclusioni e sviluppi futuri

Sebbene il prototipo sviluppato sia limitato sotto vari aspetti, si ritiene che i risultati raggiunti siano soddisfacenti per un primo utilizzo e per effettuare l'anonimizzazione di sentenze nell'ambito del progetto Giustizia Agile. Sono allo studio miglioramenti del programma che possano gestire alcune limitazioni. In particolare, si prevedono tre aree principali di miglioramento: permettere di gestire formati diversi in ingresso e uscita oltre al formato PDF attualmente considerato (ad esempio salvando il file anonimizzato in formato MS-Word); gestire l'anonimizzazione differenziata delle entità, ad esempio indicando con XX1 tutte le occorrenze di una determinata parte; studiare il possibile utilizzo di tecniche avanzate per l'anonimizzazione di sentenze basate su algoritmi allo stato dell'arte di NLP e di apprendimento automatico[1].

Riferimenti bibliografici

Chakaravarthy, Venkatesan T., Himanshu Gupta, Prasan Roy, and Mukesh K. Mohania. 2008. "Efficient techniques for document sanitization." In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM 2008, 843-52. Napa Valley, California, USA.

Csányi, Gergely Márk, Nagy Dániel, Vági Renátó, Vadász János Pál, and Tamás Orosz. 2021. "Challenges and Open Problems of Legal Document Anonymization." *Symmetry* 13, 8: 1490.

Di Martino, B., Marulli F., Lupi P., e A. Cataldi. 2021. "A machine learning based methodology for automatic annotation and anonymisation of privacy-related items in textual documents for justice domain." In *Complex, Intelligent and Software Intensive Systems: Proceedings of the 14th International Conference on Complex, Intelligent and Software Intensive Systems*, CISIS-2020, 530-39. Springer International Publishing.

Garat, Diego, and Dina Wonsever. 2022. "Automatic Curation of Court Documents: Anonymizing Personal Data." *Information* 13, 1: 2.

Gemelli, Andrea, Vivoli Emanuele, e Simone Marinai. 2022. "Graph neural networks and representation embedding for table extraction in PDF documents." In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1719-726. IEEE.

Gupta, D., Saul M., and J. Gilbertson. 2004. "Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research." *American journal of clinical pathology* 121, 2: 176-86.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st. ed. Upper Saddle River (NJ): Prentice Hall PTR.

Lison, P., Pilán I., Sánchez D., Batet M., and L. Øvrelid. 2021. "Anonymisation models for text data: State of the art, challenges and future directions." In *Proceedings of*

[1] Ringraziamo per la collaborazione alla realizzazione del progetto e del presente contributo Lisa Cresti e Simone Giovannini, borsisti presso il Laboratorio di Intelligenza Artificiale del Dipartimento di Ingegneria dell'Informazione (DINFO) dell'Università degli Studi di Firenze.

the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 4188-203.

Smith, R. 2007. "An overview of the Tesseract OCR engine." In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2, 629-33. IEEE.

Witten, Ian H. 2004. "Text Mining." In *The Practical Handbook of Internet Computing*, edited by Munindar P. Singh. New York: Chapman and Hall/CRC.