UNIVERSITÀ DEGLI STUDI DI CAGLIARI

FACOLTÀ DI SCIENZE

Corso di Laurea Triennale in Informatica

Analisi ed estrazioni di informazioni da referti medici: Un caso di studio e implementazione nel progetto SIMIOR

Relatore
Dott. Gianmarco Cherchi

Studente
Lorenzo Ludovico Concas
Matr. N. 65315

ANNO ACCADEMICO 2022/2023

In the working world, most digital documents are structured following the *de facto* standard called *Portable Document Format*, better known as PDF. Different environments ranging from schools to public administration, up to the national health service use this type of document to exchange information both with end users and between internal parties.

It is sometimes useful to be able to analyze these documents to extrapolate information, in this thesis two methods of data extraction will be analyzed and an implementation of a solution to this problem by analyzing the work done on the SIMIOR project.

Nel mondo lavorativo la maggior parte dei documenti digitali sono strutturati seguendo lo standard *de facto* chiamato *Portable Document Format*, meglio conosciuto come PDF. Ambienti diversi che variano dalla scuola all'amministrazione pubblica, fino al servizio sanitario nazionale utilizzano questa tipologia di documenti per scambiare informazioni sia con gli utenti finali sia fra le parti interne.

Risulta utile talvolta poter analizzare questi documenti per estrapolare informazioni, in questa tesi verranno analizzate due modalità di estrazione dei dati e un implementazione di una soluzione a questo problema analizzando il lavoro svolto sul progetto SIMIOR.

Indice

Capitolo 1

Introduzione

1.1 Cenni storici

Il *Portable Document Format* è nato nel 1993 da Adobe Inc. con lo scopo di creare un formato file che permettesse di rappresentare documenti testuali o con immagini indipendentemente dall'hardware o dal software che usato per generare o visualizzare il documento stesso. Al momento la sua versione attuale è la 2.0 del 2017 ma lo standard più diffuso è la 1.4 di cui esiste la norma ISO 32000 stabilita nel 2008. Essendo un formato libero, possono essere creati programmi che leggono e scrivono documenti PDF senza la necessità di pagare diritti di utilizzo del brevetto, tuttavia l'accesso ai documenti contenenti le specifiche ISO sono a pagamento. Nonostante il successo di questa tecnologia, il PDF non è considerato come adeguato alla conservazione sul lungo termine perché la riproduzione può dipendere da sorgenti esterne quali Font o oggetti esterni (link, immagini allegate esternamente ecc).

1.1.1 La struttura del PDF

I PDF sono espressi tramite una sequenza di caratteri ASCII a 7 bit, pertanto le strutture di questo formato seguono tale codifica; fanno eccezione possibili oggetti binari che mantengono il loro flusso di bit originale senza alterazioni. Ogni file inizia con un intestazione leggibile indicante la versione del formato:

$$\%\text{PDF}-1.4$$

La parte costituente e fondamentale di un PDF sono dei blocchi costruttivi chiamati "COS" *"Carousel" Object Structure* (nome del progetto originale diventato in seguito Adobe Acrobat), questi blocchi nonostante il nome non sono dei veri e propri "oggetti" come nei linguaggi *object-oriented*, ma rappresentano la struttura dei documenti. Esistono 9 tipi di oggetti *COS*, ognuno con un compito specifico:

- Oggetto *null*: è indicato come la semplice sequenza di caratteri `"null"`, è comunemente utilizzato per indicare la mancanza di un valore
- Oggetti booleani: sono la rappresentazione diretta dei due valori dell'algebra booleana, sono indicati come `"true"` e `"false"`, alcuni PDF scritti scorrettamente presentano variazioni nelle maiuscole (es `TRUE`, `True`, `FALSE` o `False`)
- Oggetti numerici : possono essere di due tipi, interi o reali, equivalenti ai rispettivi tipi matematici.
  I tipi interi consistono in uno o più cifre decimali, precedute opzionalmente dal segno e rappresentano un valore in base 10. Di seguito alcuni esempi
  - 1
  - -2134
  - 96

  I tipi reali presentano in più rispetto agli interi il punto come separazione fra parte intera e parte decimale, non è supportata la notazione esponenziale ne le radici non decimali. Alcuni esempi
  - 0.0099
  - .200
  - -3.1415
- Oggetti di tipo "Nome": Sono una sequenza di caratteri univoca in formato UTF-8 (escluso il carattere ASCII `null`) preceduta sempre da una barra obliqua (/). E' usato per definire un set di valori fissi.
- Oggetti stringa: delle semplici serie di byte scritti come caratteri, racchiusi fra parentesi tonde o come dati in formato esadecimale racchiusi fra parentesi angolari
  -