Realizzazione di un Add-on per Google Docs per estrazione interattiva di pattern sintattici

Relatore
Alberto Bartoli

Correlatore
Eric Medvet

Candidato
Lorenzo Gasparini

Descrizione del problema

- Dato un documento di testo si vogliono *individuare* ed *estrarre* tutte le occorrenze di un determinato *pattern sintattico*
- **Esempi**: Indirizzi IP (`XXX.XXX.XXX.XXX`), Date (`DD/MM/YYYY`), E-mail (`alias@domain.ext`)

Motivazione di base

```
/(?:[0-9]{1,3}\.){3}[0-9]{1,3}/g
```

Figura: RegEx (semplificata) per indirizzo IP

- Gli strumenti per specificare ed estrarre pattern da documenti testuali esistono già (e.g. RegEx), perchè un nuovo approccio?
- La curva di apprendimento di tali strumenti è ripida, non sono alla portata degli utenti comuni
- **Idea**: l'utente fornisce degli **esempi** di entità da estrarre e da non estrarre, l'algoritmo **deduce** il pattern e lo estrae dal testo

Obiettivo della tesi

- Implementare un algoritmo di estrazione delle entità basato sulla sintassi, sotto forma di Add-on per Google Docs
- L'algoritmo è stato sviluppato nel laboratorio di Machine Learning, ed è risultato in media il migliore in un confronto con altri algoritmi su 10 dataset
- L'algoritmo si basa sull'**active learning**: genera un estrattore sulla base degli esempi forniti dall'utente; sceglie l'esempio che deve aggiungere l'utente per migliorare l'estrattore

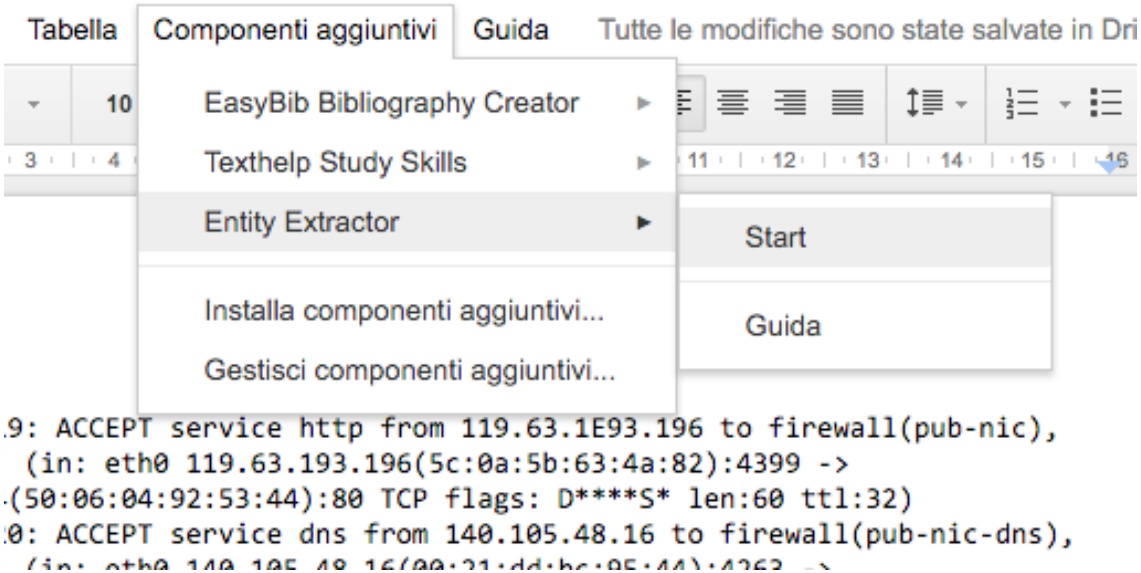
Google Docs

- Nasce nel 2006, è una piattaforma online di elaborazione testi
- Consente agli utenti di creare e modificare documenti direttamente nel browser e di collaborare con altri utenti in tempo reale

Add-on per Google Docs

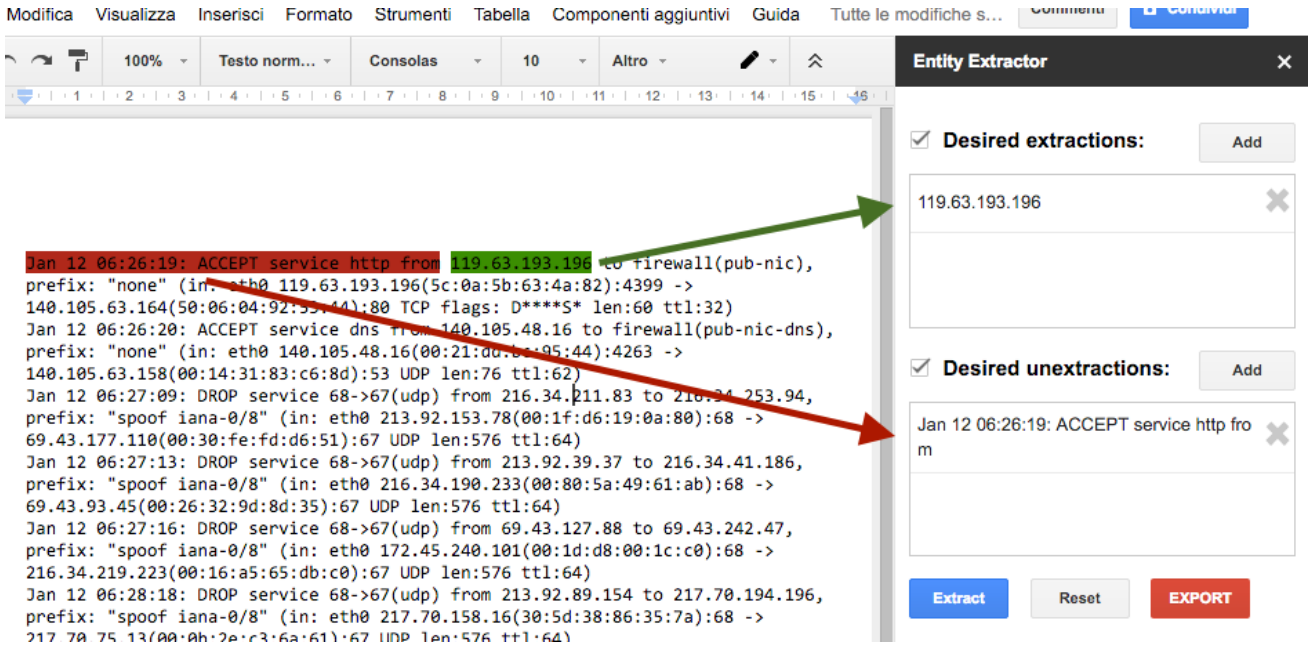
- Dal 2014 è possibile sviluppare componenti aggiuntivi che permettono di ampliare le funzionalità della piattaforma
- Un Add-on è composto da un insieme di file HTML/Javascript/CSS e di script *Google Apps Script*, il quale:
  - È basato su Javascript
  - Viene eseguito dai server di Google

Interfaccia Add-on



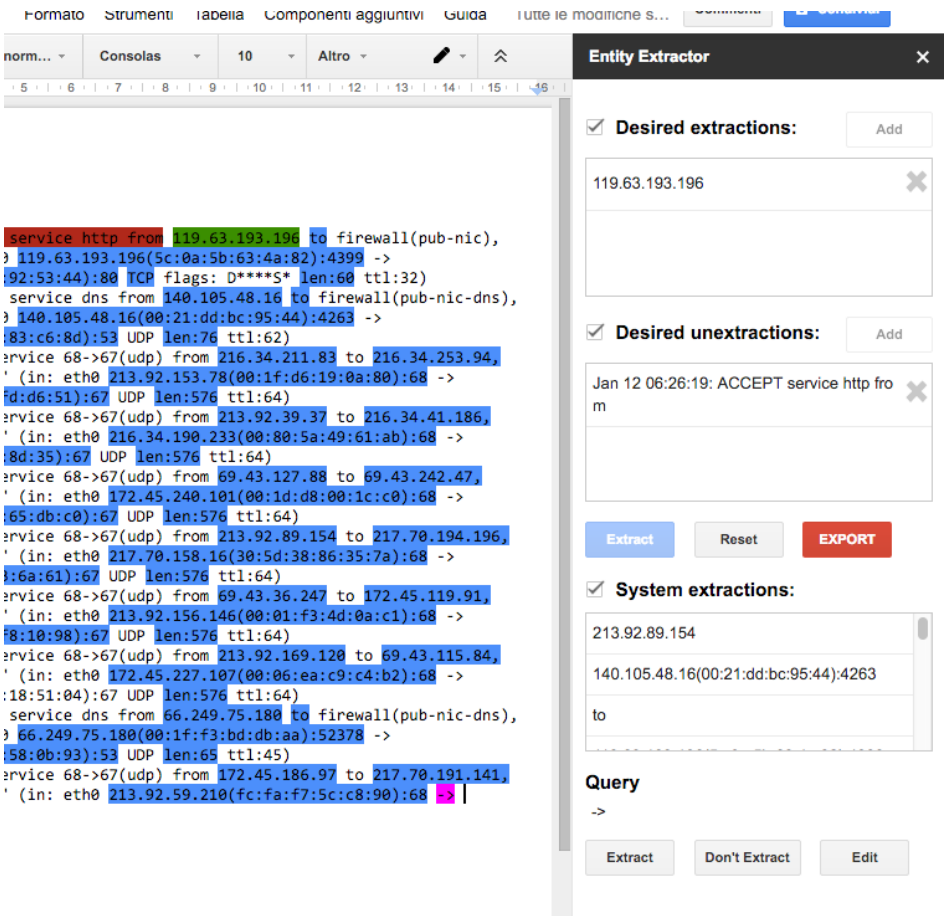
- Per aprire l'Add-on, si apre un documento di testo Google Docs e si preme *Start* nella relativa voce del menù

Interfaccia Add-on



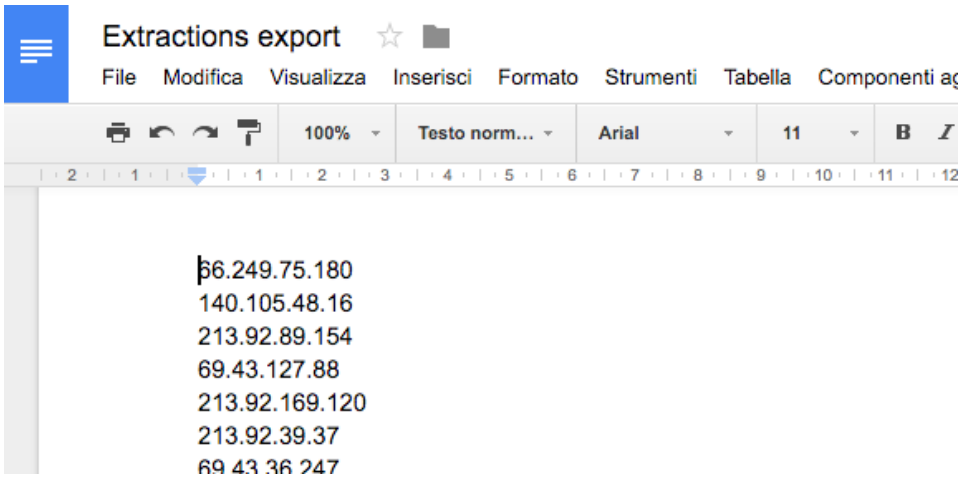
- L'aggiunta di nuovi esempi avviene mediante la selezione del testo ed il click sul relativo pulsante *Add*
- Gli esempi vengono evidenziati con colori diversi per un'indicazione visuale istantanea

Interfaccia Add-on



- *Extract* avvia la costruzione dell'estrattore in base agli esempi forniti
- Il comportamento dell'estrattore è mostrato dal sistema
- Viene formulata una query allo scopo di ottenere un nuovo esempio (*active learning*)

Esportazione estrazioni



- Cliccando su Export è possibile esportare le attuali estrazioni suggerite in un nuovo documento
- Si aprirà una finestra modale con un link al nuovo documento contenente le estrazioni

Dettagli implementativi

Nome file	Righe	Contenuto
EntityExtractor.gs	561	Motore dell'algoritmo
Main.gs	507	Gestione interazione utente
Set.gs	54	Struttura dati insieme
Sidebar.css.html	67	CSS Sidebar
Sidebar.html	61	HTML Sidebar
Sidebar.js.html	322	Javascript Sidebar
Store.gs	53	Gestore memorizzazione dati server
TextRange.gs	34	Struttura dati annotazione

Limiti della piattaforma e sviluppi futuri

- Google Docs è una piattaforma proprietaria che presenta dei limiti intriseci:
  - Le evidenziature sono permanenti, modificano la struttura del documento
  - Non è possibile gestire l'evento di chiusura dell'Add-on, impedendo l'esecuzione di azioni di pulizia del documento
  - Lo spazio di archiviazione lato server è ristretto, ciò rende difficile implementare meccanismi di cache atti a diminuire il carico computazionale dell'algoritmo
- Soluzione:
  - Migrazione a piattaforma web standalone o GUI desktop

Demo

Dimostrazione

Fine

Grazie per l'attenzione.