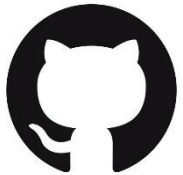
Homework 1

Alessio Marinucci

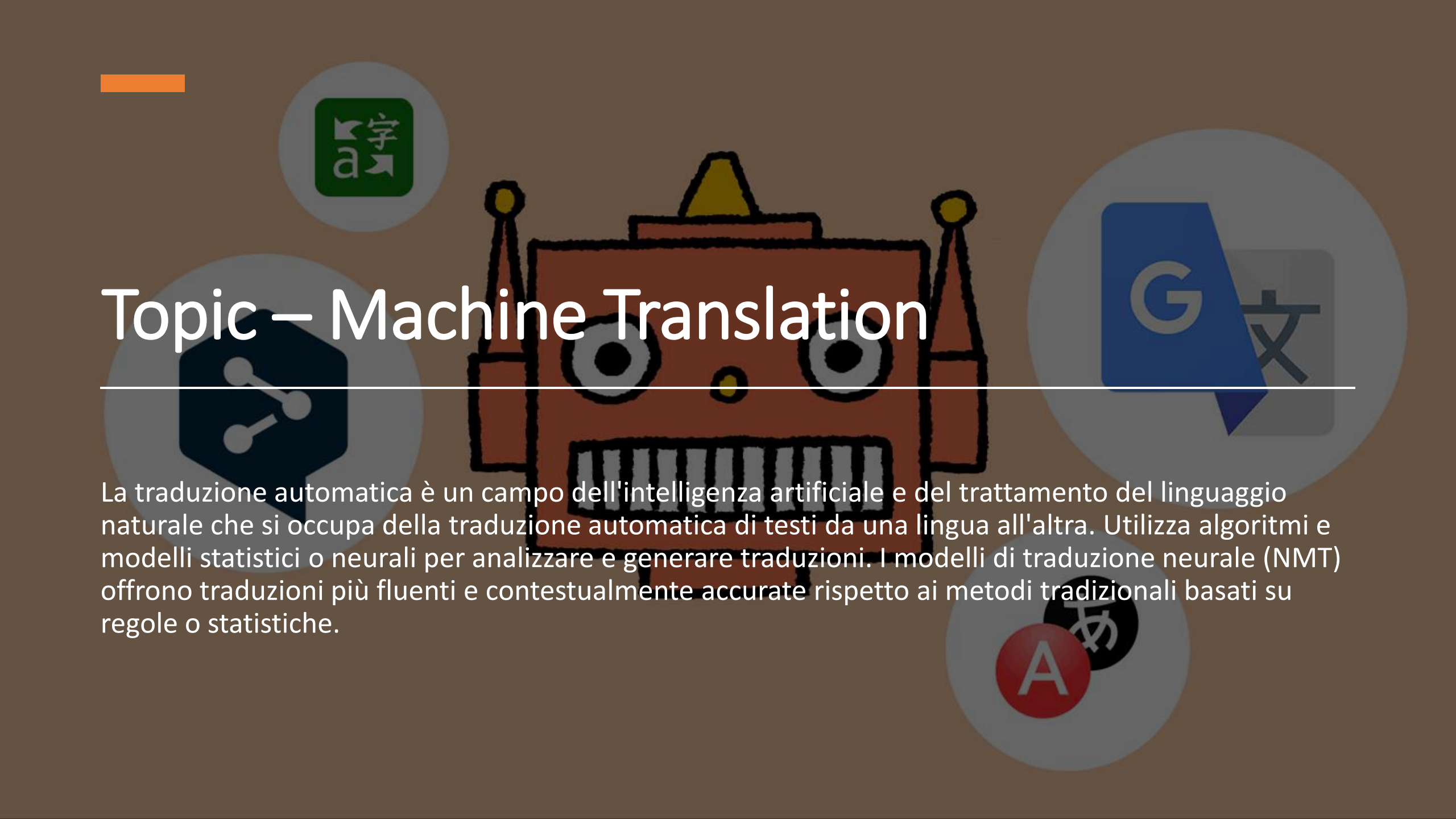
Riccardo Felici



https://github.com/DrRicky31/arXiv-Scraper

Obiettivo

- Il progetto richiede di scaricare dal portale arXiv.org un insieme di documenti in formato HTML da cui estrarre le seguenti informazioni:
  - Tabelle
  - Captions
  - Footnotes
  - Paragrafi con riferimento alle tabelle
- Dopo l'elaborazione i risultati di ogni estrazione sono raccolti all'interno di diversi file JSON.



Topic – Machine Translation

La traduzione automatica è un campo dell'intelligenza artificiale e del trattamento del linguaggio naturale che si occupa della traduzione automatica di testi da una lingua all'altra. Utilizza algoritmi e modelli statistici o neurali per analizzare e generare traduzioni. I modelli di traduzione neurale (NMT) offrono traduzioni più fluenti e contestualmente accurate rispetto ai metodi tradizionali basati su regole o statistiche.

Libreria lxml

Perchè l'abbiamo scelta?

Performance: Più veloce nella maggior parte dei casi rispetto a xml.etree.ElementTree.

Approccio più efficiente: punta a massimizzare l'efficienza durante il parsing e la manipolazione di documenti HTML, con una gestione efficiente della memoria.

XPath: Fornisce un supporto avanzato per XPath (fondamentale in questo contesto).

Validazione schema: Supporta la validazione degli schemi HTML, consentendo di verificare che un documento segua uno schema specifico.

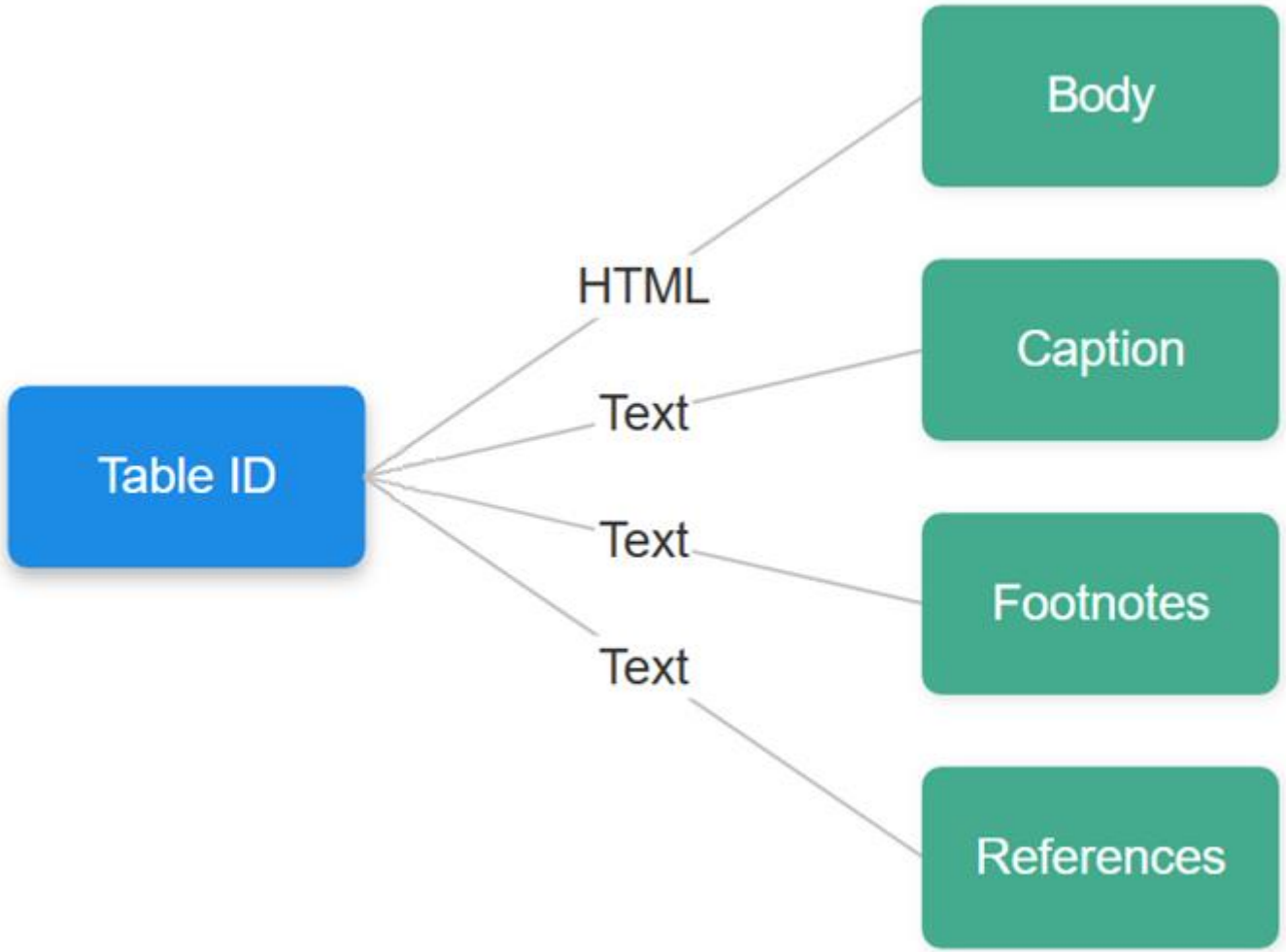
Query XPath

- Tabelle
  - ID: *‘//table/@id’*
  - ‘//table[@id="{table_id}"]’
- Caption
  - ‘//table[@id="{table_id}"]/ancestor::figure//figcaption/text()’
- Footnotes
  - ‘//table[@id="{table_id}"]/ancestor::figure//sup/span/text()’
- Paragrafi
  - Figure ID: ‘//table[@id="{table_id}"]/ancestor::figure/@id’
  - //*[substring(@href, string-length(@href) - string-length('#{figure_id[0]}') + 1) = '#{figure_id[0]}']/ancestor::p

Struttura JSON

Per ciascuna tabella sono estratte le seguenti informazioni, associate ad un Table ID:

- Corpo della tabella in formato HTML
- Caption in formato testuale
- Lista di footnote in formato testuale
- Lista dei paragrafi contenenti dei riferimenti alla tabella, in formato testuale.



Data cleaning

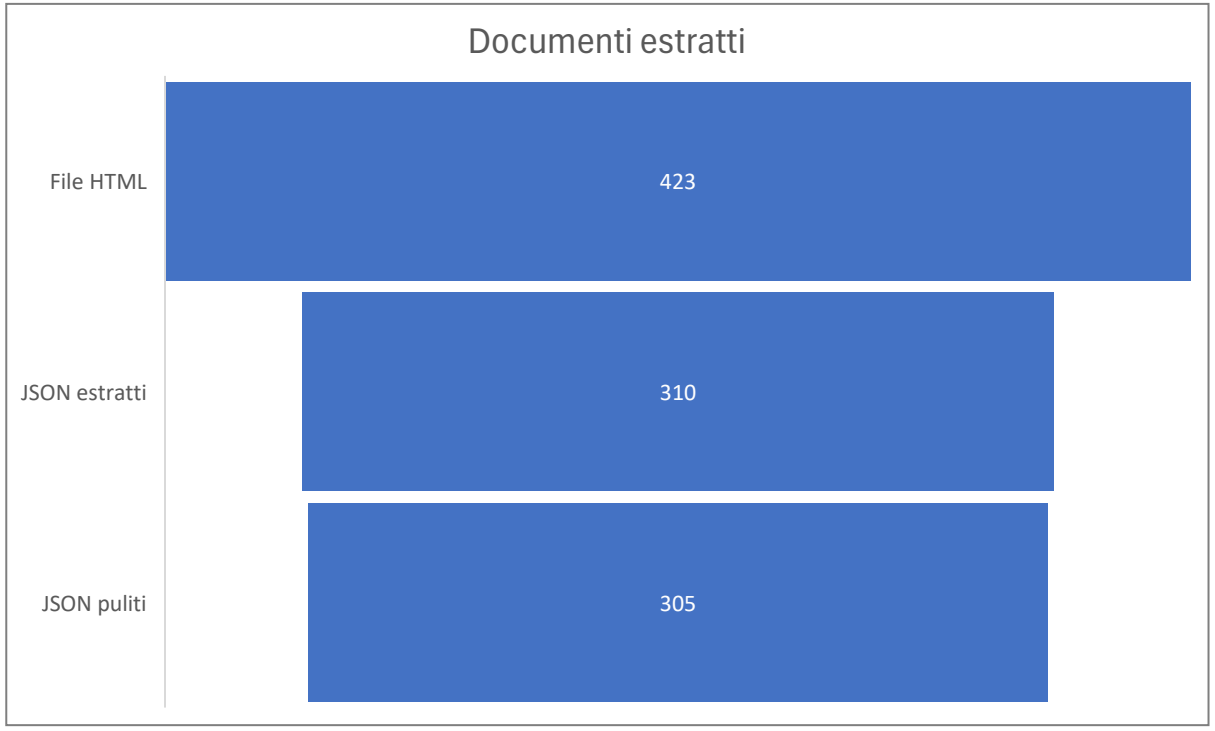
Controllo del contenuto HTML per identificare i documenti contenenti errori 403 o associati a paper che non hanno la versione HTML.

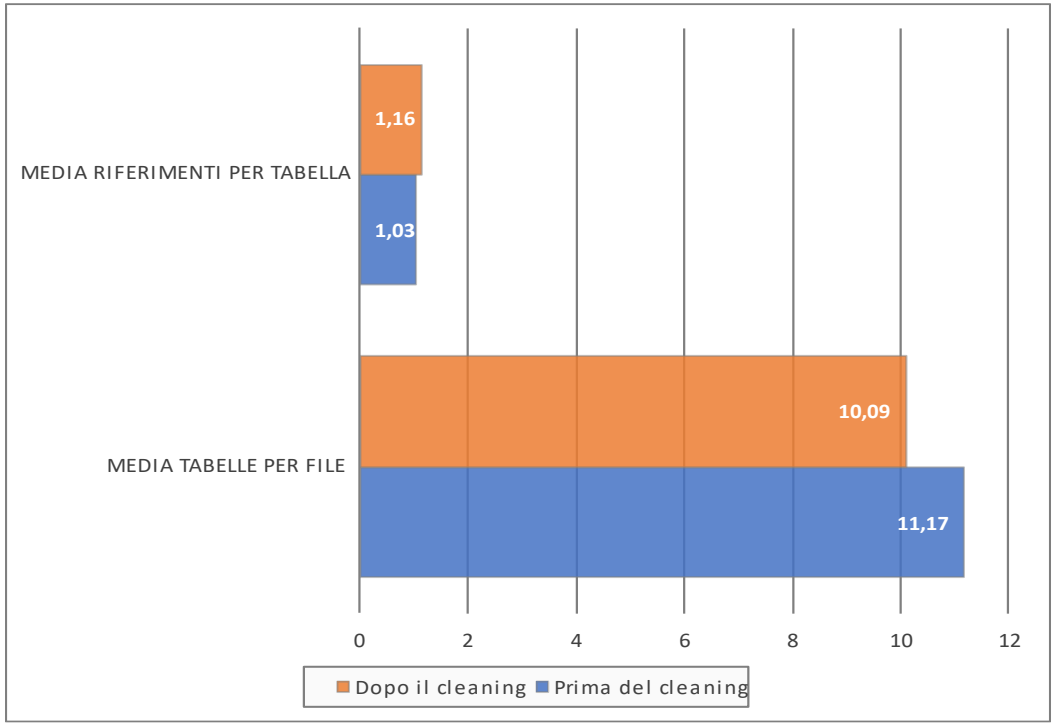
Filtraggio dei JSON selezionando solo gli elementi con chiavi che contengono il carattere "T".

Eliminazione dei JSON vuoti.

Statistiche

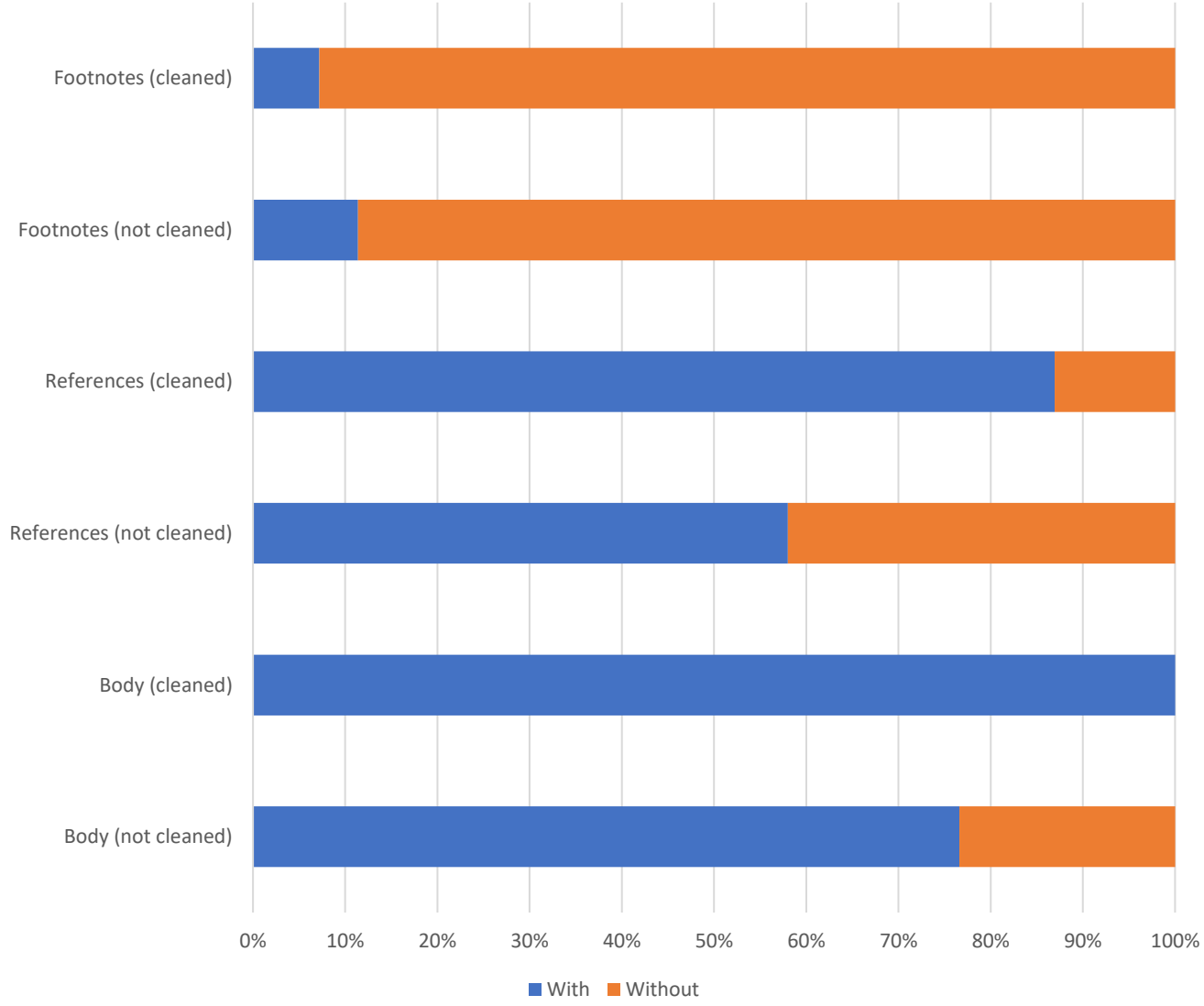
Analisi sui documenti estratti e relativo numero di tabelle per file





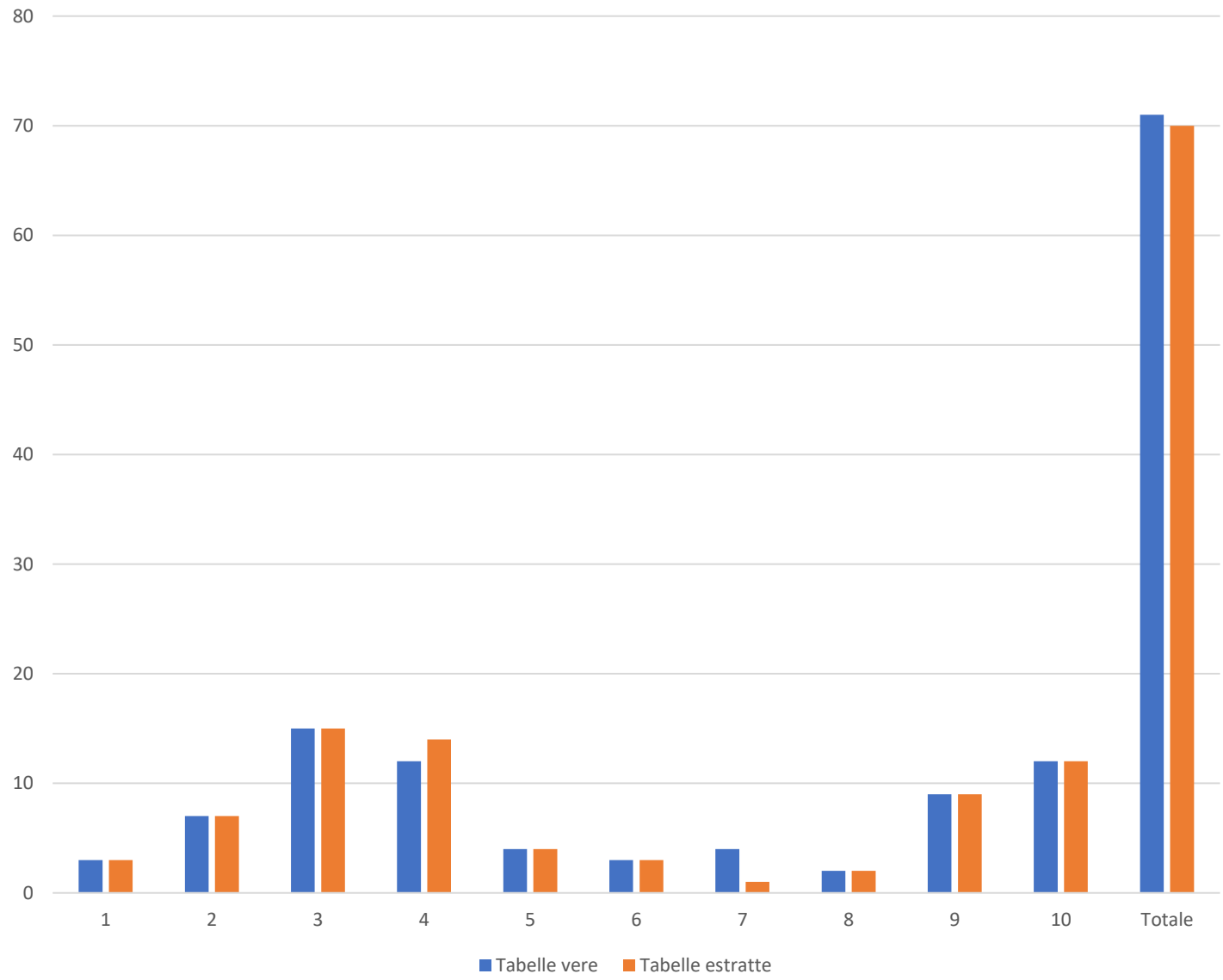
Statistiche

Percentuale di elementi estratti dalle tabelle prima e dopo il processo di cleaning



Statistiche

Verifica della corretta estrazione delle tabelle, confrontando con una ground-truth di 10 elementi



Conclusioni

Problemi incontrati

- Download file da arXiv
- Ricerca espressioni xpath per le reference

Nozioni imparate

- Automatizzazione per l’estrazione di dati da web
- Costruzione di regole xpath

GRAZIE PER L’ATTENZIONE!