ARTICLE TITLE

FABRIZIO COMINETTI, DAVIDE ABETE, RUBEN AGAZZI

CONTENTS

ABSTRACT

Il progetto realizzato consiste nella creazione di una base di dati a grafo contenente le relazioni fra i vari prodotti Marvel. All'interno del database sono presenti relazioni come ad esempio fra personaggi e fumetti, personaggi e personaggi, etc...

1 INTRODUZIONE

Lo scopo del progetto consiste nella realizzazione di un dataset riguardante vari prodotti Marvel, fra cui personaggi, film, serie tv e fumetti. L'obiettivo del progetto è di ottenere una sorta di "rete sociale" di super eroi o personaggi marvel, i quali sono collegati ai rispettivi fumetti, film , serie tv, o rispettivi collegamenti interpersonali fra personaggi e personaggi.

2 FONTI DATI

Per l'ottenimento dei dati sono stati utilizzati due metodi diversi: l'utilizzo di una Web API e Web Scraping.

2.1 Web Scraping

Per la parte riguardante il Web Scraping come sorgente dati è stata utilizzata la Marvel Cinematic Universe Wiki.

Lo scraping è stato effettuato eseguendo un notebook python. I passi eseguiti dal notebook per completare lo scraping sono:

1. Ottenimento dalla pagina relativa a tutti i personaggi della wiki i nomi dei personaggi con i relativi link alla pagina personale.
2. per ogni personaggio aprire la pagina personale e ottenere sempre tramite scraping le informazioni rilevanti, come ad esempio: biografia, lista di film in cui è presente il personaggio, lista di serie tv in cui è presente il personaggio e relazioni con altri personaggi.
3. Per ogni film trovato nelle pagine dei personaggi aprire la pagina relativa al film e ottenere tramite scraping le informazioni del film come ad esempio la trama, i registi, scrittori, compositori, incasso, data di uscita e durata del film.
4. Per ogni serie trovata nelle pagine dei personaggi aprire la pagina relativa alla serie e ottenere tramite scraping le informazioni della serie in questione, come ad esempio la trama, i registi, i produttori e i compositori
5. Salvataggio temporaneo all'interno di file csv di film, serie tv e personaggi per essere processati in seguito.

2.2 Web Api

Per l'ottenimento dei dati tramite web API è stata utilizzato il servizio formito dalla Marvel che mette a disposizione una sua Web Api per ottenere i dati relativi a personaggi e fumetti. L'ottenimento dei dati è avvenuto tramite esecuzione di un notebook python. L'esecuzione consiste nei seguenti passi:

1. Ottenimento dei dati dei personaggi in formato JSON tramite chiamata al relativo endpoint della web api a gruppi di 100 personaggi, in quanto è il limite imposto dagli sviluppatori del servizio.
2. Salvataggio su file CSV dei dati relativi ai personaggi ottenuti.
3. Ottenimento dei dati dei fumetti in formato JSON tramite chiamata al relativo endpoint della web api a gruppi di 100 fumetti, in quanto è il limite imposto dagli sviluppatori del servizio.
4. Salvataggio su file CSV dei dati relativi ai personaggi ottenuti.

3 DATA EXPLORATION

La fase di data exploration è stata realizzata sempre utilizzando python con alcune librerie grafiche per la creazione di visualizzazioni.

3.1 Data Quality

Le principali problematiche riscontrate nella qualità dei dati sono state riscontrate nelle biografie dei personaggi e nelle relazioni tra personaggi ottenute tramite web scraping;

3.1.1 Data Quality e Cleaning Biografia

Per quanto riguarda la qualità dei dati delle biografie le principali problematiche sono:

1. Il notebook salva i vari paragrafi delle biografie come una lista di stringhe, quindi prima di tutto la biografia viene trasformata in una stringa, concatenando le varie stringhe presenti nella lista e inserendo un carattere newLine fra un elemento e l'altro.
2. Il notebook inoltre salvava alcuni paragrafi della biografia più volte, per questo prima di concatenare la stringa viene effettuato un controllo per vedere se la porzione di stringa è già stata aggiunta alla stringa finale
3. Infine viene fatto un escaping dei caratteri speciali, come ad esempio il carattere " o ', in modo da non avere problemi nell'inserimento nel database dei dati.

3.1.2 Completezza

4 DATA CLEANING

5 DATA INTEGRATION

6 DATABASE

6.1 Schema matching