Prova congiunta Master DataScience

Dopo aver scelto 3 film d’animazione a lungometraggio Disney per ciascuna decade dal 1950 al 2019 (compresi, se necessario, produzioni Pixar) Esempio 50: Cenerentola, Alice, Lilli 60: 101, Spada nella Roccia, Il libro della Giungla 70: Aristogatti, Robin Hood, Le avventure di Bianca e Bernie 80: Basil, Red & Toby, La sirenetta 90: La bella e la Bestia, Il Re Leone, Mulan 00: Bolt, Koda, Lilo & Stich 10: Oceania, Frozen, Frozen II Si dovrà:

• Eseguire lo scraping web da un sito qualunque pubblico di recensioni di almeno 100 recensioni in inglese per film scelto.

• Costruire un database molto semplice delle recensioni e del conteggio delle parole usate nelle recensioni (eliminare in questo caso le stopwords)

• Produrre una serie di risultati grafici e sintetici dell’analisi dei testi o Word Cloud o Grafici a barre o …

• Il programma in python dovrebbe essere in grado di poter, successivamente essere usato anche per altri film (usando lo stesso servizio di recensioni)

• Condurre una analisi di pre-processing sulle recensioni, operando una normalizzazione semplice e una morfologica (stemming e/o lemmatizzazione);

• Illustrare, mediante un’analisi descrittiva dei contenuti e la costruzione di mappe semantiche, i contenuti più rilevanti. Si suggerisce di applicare un’analisi delle corrispondenze lessicali sulla matrice bag of words e un’analisi latente semantica sulla matrice TFIDF;

• Clusterizzare le parole delle recensioni, applicando algoritmi gerarchici e non gerarchici sia sulle coordinate fattoriali sia sulla matrice Tfidf. Descrivere i principali risultati.

• I risultati dovranno essere presentati sia in forma globale che per decade La consegna dovrà contenere:

• Tutti i listati (R, python, …)

• Tutte le immagini prodotto

• Una relazione che raccolga tutto il lavoro svolto

• Un poster in formato PDF che riassumi graficamente i risultati Andrà aggiunto alla lista di consegna un diario delle modalità di interazione tra gli elementi del gruppo in termini di:

• piattaforme di comunicazione scelta

• tempo dedicato alle riunioni a distanza

• considerazione sui pro e contro delle modalità di tele-lavoro Consiglio: siate creativi.