Guida operativa.

Il primo script da lanciare per poter riprodurre questo lavoro è “Prodkafkatwitter” contenente il producer Kafka, bisogna all’interno di quest’ultimo inserire le proprie chiavi per l’API di Twitter, ottenibili registrandosi come developer.

Il producer effettuerà lo streaming dei tweets relativi agli argomenti indicati come chiavi di ricerca.

Appena possibile va avviato anche lo script “Conskafkatwitter+MONGO” che andrà a mettersi in ascolto con gli argomenti di nostro interesse.

Entrambi gli script vanno collegati a Nifi che renderà il processo totalmente automatizzato, i tweet verranno inseriti in pile e sarà effettuato lo storage direttamente su MONGODB.

Sul template di Nifi, si fa partire il processore ConsumeKafkaRecord, che collegandosi allo script del Kafka Consumer raccoglie i tweet delle serie tv prese in analisi e li mette su un processore di tipo Connection, che funge da coda. Parallelamente si fa partire il processore PutMongoRecord, che invece si occupa di inserire i tweet in coda su Mongo. In caso di fallimento entrambi i processori sono connessi ad un Funnel di Nifi.

Una volta raccolta la collection di tweets di nostro interesse, va effettuata la pulizia di questi ultimi, lanciando lo script “puliziaDBMongo”.

Verranno quindi rimossi tutti i tweet che presentano valori identici all’interno della colonna Text e sempre sulla stessa colonna viene poi effettuata la ricerca delle nostre parole chiave, tramite il comando str.contains, in modo da escludere tutti i tweets che pur non riguardando i nostri argomenti sono stati acquisiti da Tweepy, in quanto questa libreria effettua la ricerca su ogni campo dei tweet e non limitandosi soltanto alla voce Text, che è quella di maggior contenuto informativo.

Successivamente va lanciato “Sentiment_NLTK” che, come si intuisce dal nome, serve a calcolare lo score della sentiment analysis dei tweets raccolti.

Qui tramite un'unica funzione, vengono rimosse dal testo dei tweets: le emoticon, la punteggiatura, simboli come # e RT, i collegamenti ipertestuali ed un corpus di parole che NLTK chiama “Stopwords” cioè quelle parole che fungono da congiunzione ma non aggiungono significato alla frase.

Quindi vengono spezzettate le frasi in singole parole, tramite la funzione tokenizer di NLTK, e viene calcolato lo score di ogni tweet, come media dello score delle parole che lo compongono.

Alla fine di questo script avremo due campi denominati “score1” e “score2”, il primo contenente un dizionario di valori (negative, positive, neutral e compound) ed il secondo contenente un valore di sintesi che utilizzeremo nel prosieguo del nostro lavoro.

Procediamo ora all’integrazione dei database NETFLIX e IMDB ed alla successiva integrazione dei vari tweets con le serie tv a cui si riferiscono. Per fare ciò bisogna lanciare “Integrazione”.

Alla fine di questa fase, otterremo un database JSON con struttura annidata al cui interno troviamo le caratteristiche delle serie tv e per ognuna un ulteriore campo chiamato “tweet” al cui interno troveremo tutti i tweet che lo riguardano.

Aggiungiamo dunque il campo avg_score ad ogni serie tv, calcolandolo come media del valore di score2 relativo a tutti i tweet annidati, bisogna quindi lanciare lo script “add_AVG_Score”.

Da questo momento il database è completo, e possono essere eseguite le query, basta semplicemente lanciarne gli script.