ABSTRACT

Master Universitario di I Livello in: "ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE" A.A. 2022/2023

Titolo della tesi: I tweet dei politici italiani: un' applicazione per la loro lettura automatica e per la loro interpretazione analitica

Autore: Matilde Fanizza

Abstract

Il presente progetto è dedicato allo stage presso Metis Ricerche S.r.l., un'azienda specializzata nell'acquisizione dati e gestione dell'informazione da fonti pubbliche e private.☐Lo scopo principale è comprendere il cambio di gestione da Twitter Inc. a X Corp. e analizzare le differenze tra le due aziende (com'è noto X Corp. ha sostituito Twitter Inc.). Questo studio mira alla comprensione dei pensieri di alcuni rappresentanti delle principali forze politiche italiane, provenienti dall'estrazione dei loro tweet più recenti.

L'obiettivo finale consiste nell'identificare similitudini e differenze delle posizioni dei politici oggetto di studio. Inizialmente, è stato approfondito l'uso di alcune librerie Python, 'Tweepy', 'JSON'(JavaScript Object Notation), 'OS'(Operating System), 'Re'(Regular Expression), 'Pandas' ed 'NLTK'(Natural Language ToolKit), usate per permettere il collegamento all'applicazione X Corp., per estrarre i tweet e analizzarli.

In seguito, viene creato un dizionario mediante l'importazione di dati da un file Excel esterno che contiene i riferimenti degli account X di ciascun politico e di ciascun partito a loro riferito. Successivamente, viene creata su Python una funzione progettata per estrarre i tweet associati ad un politico specifico utilizzando il suo nome utente X come input (presente nel file esterno sopracitato). I problemi riscontrati, derivanti dalle strategie di commercio implementate con la comparsa di X Corp., riguardano sia i limiti di estrazione dei tweet che i richiami della suddetta funzione. Per tale ragione, sono state implementate delle operazioni, testate attraverso sperimentazioni, che risolvono questi problemi. Tra queste, ad esempio, la schedulazione del codice Python ogni ora. Durante la prima esecuzione di questa funzione, i codici ID di riferimento dei tweet estratti, formati da 10 cifre, vengono registrati in un file di testo esterno. In

questo modo, nelle esecuzioni successive dell'operazione, saranno estratti solamente i tweet con il codice ID più recente (che corrisponde al numero maggiore) rispetto a quelli precedentemente memorizzati. Questo approccio consente di superare il problema associato ai limiti di estrazione.□Durante la fase di salvataggio dei tweet estratti, viene implementato un processo di data quality per semplificare le fasi successive. Tra le operazioni citate, si include la separazione di ciascuna parola nei tweet, il conteggio delle parole e l'eliminazione di alcune mediante una lista di esclusione (contenente congiunzioni, articoli e preposizioni). Si passa poi alla creazione di una Wordcloud ottenuta con l'omonima libreria, per individuare a livello visivo le parole più frequenti.□Successivamente vengono eseguite ulteriori analisi, al fine di individuare le correlazioni dei pensieri dei politici tramite le parole usate. La parte finale del progetto si focalizza sull'essenziale ruolo dell'etica nell'analisi dei dati sociali. Questa componente contribuisce non solo all'integrità della ricerca, ma anche a preservare un ambiente informativo e non distorto.□Le risorse tecniche comprendono strumenti, tecnologie e competenze specifiche che saranno impiegate durante tutte le fasi del lavoro.

Il progetto ha richieso un approfondimento del linguaggio Python in riferimento sopratutto alle librerie citate. Vengono utilizzati file Excel esterni per l'importazione e l'esportazione di dati riguardanti i codici degli account tweet dei politici e, tramite procedure SAS, sono state condotte analisi di correlazione e clustering e sono stati generati report e grafici basati sulle informazioni estratte dai tweet. Attraverso l'elaborazione dei dati, l'impiego di algoritmi di analisi testuale e la creazione di visualizzazioni, il progetto mira a fornire una visione dettagliata delle conversazioni online dei politici italiani su una piattaforma chiamata X. Durante l'implementazione del progetto, è emerso che Python si è dimostrato uno strumento potente e flessibile per l'estrazione e l'analisi dei dati provenienti da Twitter. Le visualizzazioni create hanno contribuito a identificare i temi principali e le tendenze nelle conversazioni dei politici italiani, offrendo una panoramica dettagliata del panorama politico attraverso il social media. I punti forti del progetto riguardano principalmente la schedulazione automatica impiegata per la fase di estrazione dei tweet, che garantisce analisi precise e continuative, l'integrazione con molteplici strumenti di analisi, che offre una visione completa delle similitudini e delle differenze tra i politici e l'aspetto etico e della privacy nell'estrazione e nell'analisi dei dati provenienti da X, che garantisce che le pratiche di estrazione rispettino le normative sulla privacy e i diritti degli utenti. Grazie a Metis Ricerche S.r.l è stato possibile capire come funziona lavorare su progetti come questo e sopratutto come imparare ad affrontare i problemi e le sfide che possono emergere durante il percorso.