Guida all'esecuzione del codice

Streaming

- Per questa fase considerare i file Producer e Consumer.
- La streaming avviene mediante Tweepy, un'istanza di *tweepy.Stream* stabilisce una sessione di streaming e instrada i messaggi all'istanza *StreamListener*.
- I tweet vengono raccolti in 24 ore.
- Dopo esser stati spediti al consumer vengono salvati in locale.
- Assicurarsi che sia il codice relativo al consumer sia il codice del producer stia girando

Tweet Cleaning e Text Analysis

- Da questo punto in poi i riferimenti sono relativi al file Codice_pt2
- Dalla colonna dei tweet riferita alla location viene estratto il Paese di appartenenza grazie alla funzione *geocode* della libreria geopy.
- I nomi degli stati vengono tradotti in inglese con la funzione *translate* della libreria *google_trans_new* (basata sull'API di Google Translate, ne esistono diverse versioni) e poi normalizzati con un dizionario ad hoc.
- Anche il testo dei tweet viene tradotto mediante la stessa funzione *translate.*
- Può capitare che la funzione ritorni un errore, sembrerebbe essere un problema di aggiornamento, se in un futuro non sarà ancora risolto dagli sviluppatori recuperare file *google_trans_new.py* contenente il codice e sostituire la voce response = (decoded_line + ']') con response = (decoded_line) (abitualmente riga 151 e 233)
- I valori di polarity e subjectivity vengono calcolati mediante la funzione *TextBlob*
- Le funzioni di pulizia, tokenizzazione e stemming sono create appositamente e hanno come unico input il testo del tweet.
- Anche la funzione textanalysis è stata creata ad hoc, essa restituisce come output un dataframe contenente le 20 parole più frequenti, i 20 bigrammi e trigrammi più frequenti e salva su file una word cloud, ricevendo come input il paese di cui si vuole calcolare il tutto e il percorso che porta alla cartella in cui si vuole che venga salvata la word cloud.
- Questa parte si conclude con un ciclo che permette la creazione dei due dataset che verranno poi successivamente utilizzati per la creazione delle visualizzazioni, uno contenente le 300 parole più utilizzate per ogni paese e l'altro contenente i 50 bigrammi e trigrammi più utilizzati per ogni paese
- Infine, si crea un ultimo dataset contenente, per ogni paese, la media del valore di *polarity*, e degli score (*pos, neg e neu*) e le rispettive deviazioni standard.

Nuove Fonti: Cleaning e Integration I

- Il dato proveniente da WorldBank (Socio_economic_inds.csv) è formato da una colonna country, una colonna contenente il nome degli indicatori e le colonne relative agli anni da 2010 a 2020.
- Da questo viene estratto per ogni paese il valore dell'indice più recente ottenendo un dataset che associa ad ogni stato un set di indicatori con il valore relativo più recente. In seguito, verranno selezionati solo gli indici più funzionali.
- La pulizia dei dati ambientali (epi2020results20200604.csv) consiste semplicemente nel selezionare gli indicatori opportuni.
- Segue il join dei due dataset.
- Il dato ottenuto conterrà per ogni stato un valore relativo ad ogni indicatore (possono essere presenti valori mancanti). Ad esso vengono aggiunte due colonne prop_tweet e prop_user create mediante l'utilizzo del dataset Social il quale contiene informazioni circa il numero di utenti social, la percentuale di utenti twitter e il numero di tweet per paese ottenuti dalla streaming. Il primo indicatore è stato ottenuto rapportando il numero di tweet pubblicati per stato al numero di utenti twitter nello stesso, il secondo invece è ottenuto nel medesimo modo ma considerando il numero di user che ha effettuato una pubblicazione nel periodo di tempo scelto. Entrambi i valori sono per milione di abitanti.

Integration II e creazione DB

- Si uniscono il dato relativo alla sentiment con quello relativo agli indicatori.
- Si unisce il dato relativo ai tweet con il dato relativo agli n-grammi e alle parole più frequenti
- Mediante creazione di colonne si pongono le basi per il passaggio al formato .json il quale viene poi caricato sul cloud MongoDB Atlas.
- Relativamente al caricamento su Atlas, fare riferimento alla guida.