**1.Wordle twitter data**

Contesto

Con il gioco Wordle che emerge come nuova tendenza in questi giorni, la maggior parte degli utenti si preoccupa sempre di dove si trovano gli altri e quale tentativo di indovinare è stato il più comune. Quindi ecco un enorme set di dati di tweet che le persone hanno condiviso. Questi dati contengono un facile accesso a quale ID Wordle ha quale distribuzione dei tentativi di ipotesi, quale posizione è la più attiva nella riproduzione di Wordle e anche da quale fonte stanno riproducendo.

Come i dati sono stati collezionati

I dati vengono aggiornati ogni giorno e si riflettono sul precedente Wordle ID. I dati vengono raccolti dai tweet dell'intera giornata. Poiché il set di dati costituirebbe una parte importante se eseguito durante l'intera giornata e pertanto i tweet vengono eliminati per 20.000 tweet. Una singola richiesta potrebbe ricevere 100 tweet. e quasi 180 richieste in 15 minuti, ottenendo i dati in circa 30 minuti per il mondo dell'ID.

Inoltre, questi dati si riferiscono al giorno in cui viene riprodotto il gioco. Per es. Wordle 254 è stato riprodotto il 28 febbraio 2022, quindi i tweet vengono recuperati da quel giorno.

Avvio dei tweet da WORDLE ID 254.

Ispirazione

Sebbene MKBHD sia stata la prima ispirazione per giocare a Wordle, questo account è stato l'ispirazione per creare un set di dati e aiutare gli altri

Domande:

- qual è la media dei tentativi degli utenti per vincere?

- da quale piattaforma venga eseguite più partire (Android, Iphone, Web App)?

- gli utenti che condividono tanti tweet tendono a vincere o a perdere?

- in quale paese si gioca di più?

- in base ai tweet se una certa parola del giorno era più facile o difficile per gli utenti da trovare (ha richiesto più o meno tentativi)?

- predizione: il prossimo tweet sarà di vincita o perdita?

Nota. La colonna Location va molto pulita

Contro: Una sola persona ci ha fatto una analisi sopra.

**2.Superstore Dataset**

Contesto

Con le crescenti richieste e la concorrenza spietata sul mercato, un gigante dei supermercati sta cercando le tue conoscenze per capire cosa funziona meglio per loro. Vorrebbero capire quali prodotti, regioni, categorie e segmenti di clientela dovrebbero prendere di mira o evitare.

Puoi anche fare un ulteriore passo avanti e provare a costruire un modello di regressione per prevedere le vendite o il profitto.

Impazzisci con il set di dati, ma assicurati anche di fornire alcuni approfondimenti aziendali per migliorare.

Metadati

ID riga => ID univoco per ogni riga.

ID ordine => ID ordine univoco per ciascun cliente.

Data ordine => Data ordine del prodotto.

Data di spedizione => Data di spedizione del Prodotto.

Modalità di spedizione=> Modalità di spedizione specificata dal Cliente.

ID cliente => ID univoco per identificare ciascun cliente.

Nome cliente => Nome del cliente.

Segmento => Il segmento a cui appartiene il Cliente.

Paese => Paese di residenza del Cliente.

Città => Città di residenza del Cliente.

Stato => Stato di residenza del Cliente.

Codice Postale => Codice Postale di ogni Cliente.

Regione => Regione di appartenenza del Cliente.

ID Prodotto => ID univoco del Prodotto.

Categoria => Categoria del prodotto ordinato.

Sottocategoria => Sottocategoria del prodotto ordinato.

Nome prodotto => Nome del prodotto

Vendite => Vendite del Prodotto.

Quantità => Quantità del Prodotto.

Sconto => Sconto previsto.

Profitto => Profitto/Perdita subita.

Pro: ci sono 1112 analisi già fatte

Gli ordini vanno dal 2014 al 2017.

Domande:

-Per ogni paese qual è la categoria di prodotti più frequentemente acquistata?

-se una persona ha acquistato una volta, tende a fare altri ordini con quale frequenza e quali no?

-negli anni che andamento ha avuto il numero di ordini?

- negli anni che andamento ha avuto il numero di articoli per ordine?

- negli anni che andamento ha avuto il valore complessivo di acquisto per ordine?

- negli anni che popolarità hanno avuto certe categorie? È aumentata o diminuita?

- cambia il numero di articoli acquistati se presente uno sconto?

- qual è la relazione tra sconto sull’ordine e perdita nel profitto?

- per ciascun segment quali sono generalmente gli sconti previsti?

**3.TED Talks**

Contesto

Questo set di dati è stato creato per gli studenti principianti di analisi dei dati che possono esplorare il campo con dati reali. L'uso dei dati dei discorsi TED li aiuterà ad analizzare i discorsi e potranno anche guardare i discorsi del loro autore preferito con l'aiuto del set di dati.

Contenuto

Questo set di dati contiene 6 diverse caratteristiche di ciascun discorso disponibile sul sito web di TED che puoi trovare di seguito

* Titolo: Titolo del discorso
* autore: Autore di Talk
* data: Data in cui ha avuto luogo la conversazione
* visualizzazioni: numero di visualizzazioni del discorso
* Mi piace: numero di Mi piace del discorso
* link: Link della conferenza da ted.com

Ringraziamenti

I dati sono stati estratti dal sito web ufficiale di TED e sono disponibili sotto la licenza Creative Commons.

Questo set di dati può aiutarti a rispondere alle seguenti domande:

* Trovare i discorsi TED più popolari
* Trovare i relatori dei discorsi TED più popolari (in termini di numero di discorsi)
* Analisi mensile della frequenza dei discorsi TED
* Analisi annuale della frequenza dei TED talk
* Trovare i discorsi TED del tuo autore preferito
* Trovare i discorsi TED con la visione migliore per apprezzare la razione
* Trovare discorsi TED in base ai tag (come il clima)
* Trovare i relatori dei TED talk più popolari (in termini di numero di visualizzazioni)

Nota: si può fare clustering sulle parole chiavi dei titoli per trovare la categoria di appartenenza.

**4.9000+ Movies Dataset**

Contesto

Questo set di dati è stato creato per creare un sistema di raccomandazione di film utilizzando modelli di elaborazione del linguaggio naturale e apprendimento automatico esclusivamente per gli studenti che desiderano entrare nel mondo della scienza dei dati e dell'apprendimento automatico.

Contenuto

Caratteristiche del set di dati:

* Release\_Date: data in cui è stato rilasciato il film.
* Titolo: nome del film.
* Panoramica: Breve riassunto del film.
* Popolarità: è una metrica molto importante calcolata dagli sviluppatori TMDB in base al numero di visualizzazioni giornaliere, voti giornalieri, numero di utenti contrassegnati come "preferiti" e "watchlist" per i dati, data di rilascio e altri parametri.
* Vote\_Count: voti totali ricevuti dagli spettatori.
* Vote\_Average: valutazione media basata sul conteggio dei voti e sul numero di spettatori su 10.
* Lingua\_originale: lingua originale dei film. La versione doppiata non è considerata la lingua originale.
* Genere: categorie in cui può essere classificato il film.
* Poster\_Url: URL della locandina del film.

Domande:

* **Il genere più in voga (popolare) nei vari anni? Il genere più apprezzato (voti) nei vari anni? Il genere più apprezzato è anche quello più popolare?**
* Numeri di film rilasciati per anno?
* Il numero di voti negli anni è aumentato o diminuito?
* I film in lingue inglese sono quelli con più voti (perché accessibili a un pubblico più ampio)?
* I film in lingua non inglese (più di nicchia e quindi più di autore) sono più apprezzato di quelli in lingua inglese?

**5.Google Play Store Apps (\*)**

Informazioni sul set di dati

Contesto

Sebbene molti set di dati pubblici (su Kaggle e simili) forniscano dati sull'App Store di Apple, non ci sono molti set di dati equivalenti disponibili per le app di Google Play Store ovunque sul Web. Scavando più a fondo, ho scoperto che la pagina dell'App Store di iTunes implementa una struttura simile ad un'appendice ben indicizzata per consentire un web scraping semplice e facile. D'altra parte, Google Play Store utilizza sofisticate tecniche moderne (come il caricamento dinamico della pagina) utilizzando JQuery, rendendo lo scraping più impegnativo.

Contenuto

Ogni app (riga) ha valori per categoria, valutazione, dimensioni e altro.

Ispirazione

I dati delle app del Play Store hanno un enorme potenziale per portare al successo le aziende che creano app. È possibile ricavare informazioni utili affinché gli sviluppatori possano lavorare e conquistare il mercato Android!

Domanda:

* La versione di Android minima supportata dall’app influenza il numero di installazioni della stessa?
* Un’app a pagamento (dato che una persona pondera meglio il fatto di scaricarla o meno) ha generalmente un rating più alto di una app gratuita? È vero che questo rating più alto è determinato da un numero minore di recensioni (o forse meglio installazioni) rispetto a una app gratuita? Ha una app gratuita più installazioni di una a pagamento? Se si ha anche un numero di recensioni più alto rispetto a una a pagamento? Le persone sono più propense a recensire un’app a pagamento piuttosto che una gratuita o viceversa?
* A quale genere appartengono le app più scaricate? Qual è la correlazione tra genere e il rating?
* Che relazione c’è tra la data di ultimo update e il numero di installazioni o il rating?
* È aumentato il numero di app con target ‘children’ dato che l’età di accesso al telefono in questi tempi diminuisce?
* Le recensioni più lunghe sono tendenzialmente quelle negative o quelle positive?

Nota: OLAP(genere, tipo, n

**6. Goodreads-books**

Contesto

Il motivo principale per creare questo set di dati è il requisito di un set di dati di libri buono e pulito. Essendo io stesso un allibratore (vedi cosa ho fatto lì?) Avevo cercato set di dati sui libri nello stesso Kaggle - e ho scoperto che mentre la maggior parte dei set di dati aveva una buona quantità di libri elencati, c'erano a) colonne principali mancanti o b) dati gravemente impuri. Voglio dire, non puoi determinare quanto sia bello un libro solo da alcune recensioni di testo, andiamo! Ciò di cui avevo bisogno erano numeri, numeri interi solidi e numeri in virgola mobile che dicessero quante persone hanno apprezzato o odiato il libro, quanto gli è piaciuto e cose del genere. Anche il buon set di dati che ho trovato era ben pulito, conteneva un numero di file interconnessi, il che aumentava il fastidio. Ciò mi ha spinto a utilizzare l'API Goodreads per ottenere un set di dati ben pulito, con solo le funzionalità promettenti (meno quelle ridondanti) e il risultato è il set di dati in cui ti trovi adesso.

Ispirazione

Il motivo alla base della creazione di questo set di dati è piuttosto semplice: sto elencando i libri per tutti gli amanti dei libri, indipendentemente dalla lingua, dalla pubblicazione e tutto il resto. Quindi vai avanti e usalo a tuo piacimento, scopri quale libro dovresti leggere dopo (ci sono pochissimi sistemi gratuiti di raccomandazione dei contenuti che suggeriscono i libri l'ultima volta che ho controllato), quali sono i dettagli di ogni libro che hai letto, crea una parola cloud dai libri che desideri leggere: tutti i possibili approcci per esplorare questo set di dati sono benvenuti.

Ho iniziato a creare questo set di dati il ​​25 maggio 2019 e intendo aggiornarlo frequentemente.

Domande:

* I libri lunghi hanno più o meno recensioni dei
* Qual è stato l’andamento inteso come average rating per casa editrice dei libri pubblicati negli anni? Qual è stato l’andamento inteso come autore rating per casa editrice dei libri pubblicati negli anni? Le pagine dei libri tendono ad aumentare o a diminuire negli anni (la nostra ipotesi è che negli ultimi anni si pubblichino libri sempre più corti)?

Previsione: una casa editrice pubblicherà un buon libro come prossimo?

* Il numero di recensioni per libri vecchi e nuovi è diversa?

Note. Le informazioni mancanti le si possono ricavare dall’ISBN

**7. Video Game Sales**

Contenuto

Troverai i tuoi videogiochi preferiti e quanto erano popolari in base alle loro vendite.

* Quale console ha avuto i giochi più di successo (perché più vendute)?
* Quale console più in voga nei vari anni in base alle vendite? Quale console più in voga in un paese in base alle vendite nei vari paesi?
* Il Role-playing è il genere più venduto in giappone?
* Quanti videogiochi per anno sono stati rilasciati nei vari anni? Quanti videogiochi saranno rilasciati in futuro?
* Quanti videogiochi sono rilasciati per certe piattaforme negli anni? Es. mi aspetto che i giochi rilasciati per il NES siano nulli o pochi rispetto agli anni 90?
* Le vendite totali dei giochi per una console sono aumentate o diminuite rispetto alle console della generazione precedente (dello stesso produttore). Es. le vendite totali del 3ds hanno fatto meglio di quelle del ds, gambeboy rispetto al gameboy color, PSV rispetto PSP ecc. per capire quali console hanno costituito un flop rispetto alle precedenti o successive.
* Confronto tra i produttori nei vari anni (es. xbox vs playstation)
* Quale produttore vende di pù nelle varie regioni (NA, EU, JP).

**8. 160k Spotify songs from 1921 to 2020 (Sorted)**

Informazioni sul set di dati

Il set di dati di Spotify (intitolato data.csv) è costituito da oltre 160.000 tracce ordinate per nome, dal 1921 al 2020 trovate in Spotify a giugno 2020. Raccolti dall'utente Kaggle e dal data scientist turco Yamaç Eren Ay, i dati sono stati recuperati e tabulati dal API Web di Spotify. Ogni riga nel set di dati corrisponde a una traccia, con variabili come titolo, artista e anno posizionate nelle rispettive colonne. A parte le variabili fondamentali, sono stati estratti anche gli elementi musicali di ogni traccia, come il tempo, la ballabilità e la tonalità; l'algoritmo per questi valori è stato generato da Spotify sulla base di una serie di parametri tecnici.

Analisi esplorativa dei dati (EDA)

* Studiare le correlazioni tra le variabili nei dati Spotify.
* L'evoluzione dei diversi elementi musicali nel corso degli anni.
* Il divario tra canzoni esplicite e non esplicite nel corso degli anni.
* Determinare se esiste una differenza significativa nella popolarità tra le canzoni esplicite e quelle non esplicite.
* Trovare le emozioni più frequenti nelle tracce di Spotify e analizzarne gli elementi musicali in base alla modalità e alla tonalità della traccia.
* Determinazione delle classificazioni delle tracce Spotify tramite K-Means Clustering.

Domande:

* Le canzoni più recenti durano meno di quelle old e very old? Ci si aspetta un andamento decrescente negli anni.
* Le canzoni più ballabili sono quelle che sono state pubblicate tra gli anni 80 e 90 (categorizzate come old)?
* Qual è la correlazione tra il fatto che la instrumentalness tenda a 1 e il fatto che nel titolo compaia il termine ‘noise’ o ‘rain’.