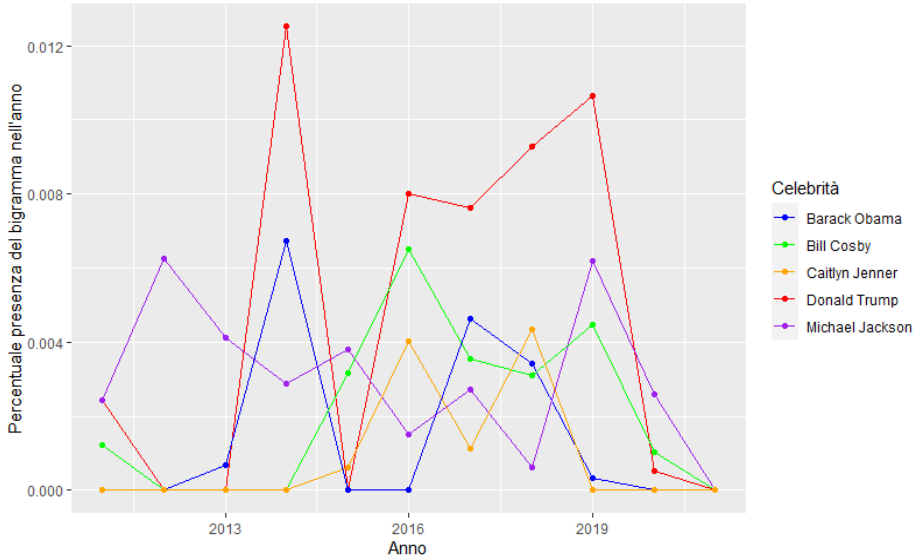
Dissecting Comedy: alle basi della comicità.

0. Introduzione

Il seguente progetto si pone come obiettivo l'analisi di 228 trascrizioni di spettacoli di stand-up comedy in lingua inglese dell'ultimo decennio, al fine di ricercare gli elementi caratterizzanti della comicità, attraverso strumenti di text mining. Per ottenere un dataset utilizzabile, è stato fatto scraping dal sito https://scrapsfromtheloft.com/stand-up-comedy-scripts/. Costruito il dataset, sono state rimosse le stop words da quest'ultimo, è stato praticato lo stemming, e infine costruito il wordcloud (*Figura 0.1*).

0.1 Wordcloud

Risulta subito evidente come parole volgari (*fuck, shit*) e colloquiali (*peopl, guy, time*) siano quelle prevalenti, seguite poi da un insieme eterogeneo di termini, rientranti nella categoria definibile come quotidianità (*love, talk, day, kid, friend, start, happen, life...*).
Successivamente l'attenzione viene spostata sullo studio dei bigrammi, in particolare sulla loro frequenza. I due bigrammi più frequenti riguardano l'etnia (*white people, black people*), e parte dei rimanenti sono invece esclamazioni (*god damn, jesus christ, holy shit…)*.
Tra i bigrammi si sono cercate le persone più citate: dalla politica si hanno Donald Trump, Bill Clinton e Barack Obama, dal mondo dello spettacolo Bill Cosby, Michael Jackson, Mick Jagger e Caitlyn Jenner (che compare sia come Caitlyn che come Bruce), mentre Anne Frank e Bin Laden riguardano fatti storici.
Riguardo alle celebrità, sono state scelte le 5 più ricorrenti e sono state visualizzate graficamente (*Figura 0.2*) le serie storiche delle frequenze con cui vengono citate durante gli anni presi in analisi:



0.2 Serie storica delle citazioni delle celebrità

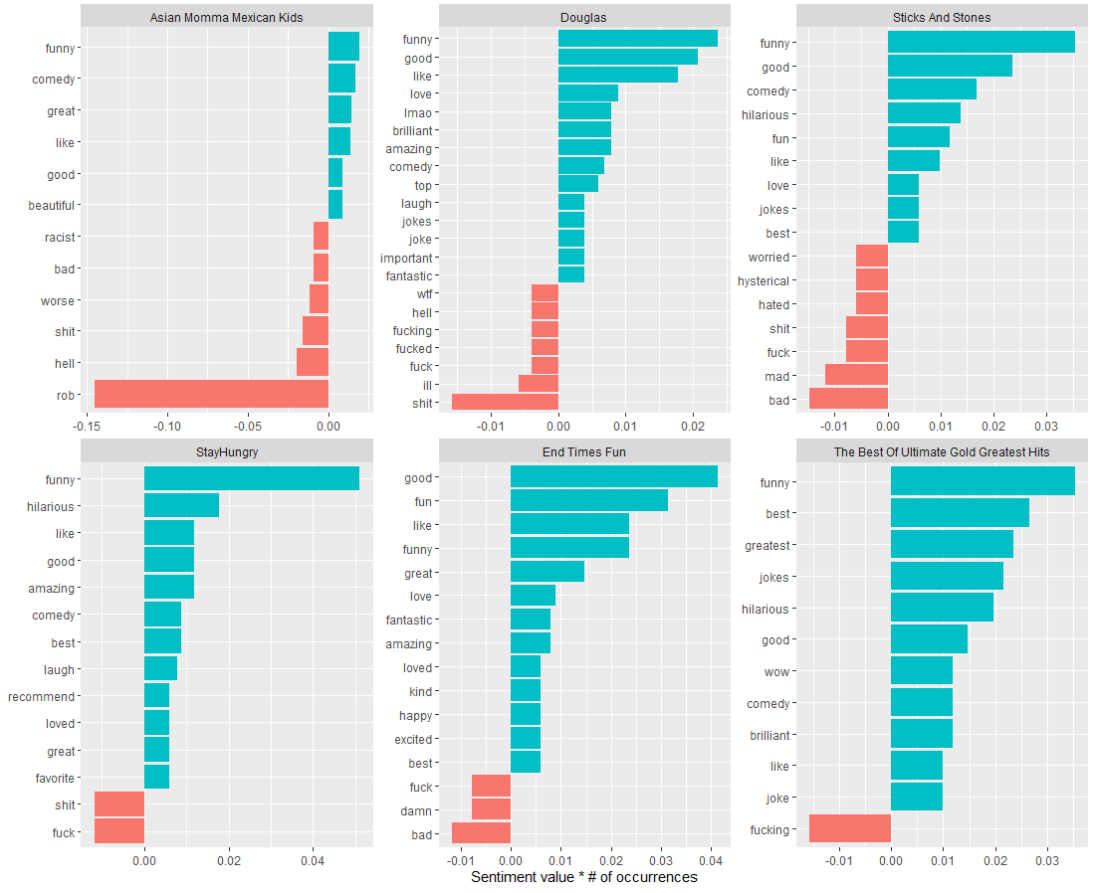
Risulta evidente come il numero di citazioni nell'anno vari molto a seconda di eventi di cronaca relativi al personaggio: Donald Trump, come facilmente pronosticabile, durante gli anni della sua presidenza, è spesso stato citato, così come Barack Obama, seppur in maniera minore. I bigrammi *Bill Cosby* e *Michael Jackson* hanno avuto una crescita causata dalle rispettive controversie: Cosby, nel 2014, è stato accusato di molestie, mentre nel 2019 è stato pubblicato *Leaving Neverland,* documentario riguardante le accuse di pedofilia rivolte a Jackson. Caitlyn Jenner ha iniziato ad essere citata nel 2015, dopo aver dichiarato di essere transessuale.
L'analisi è proseguita con l'osservazione dei trigrammi, ma questi non sono risultati informativi.

1. Metrica per la volgarità

Dato che la volgarità è un elemento ricorrente negli spettacoli, è stata costruita una metrica per dare un "punteggio di volgarità" a ciascun testo.
Per fare ciò si è considerata la blacklist di YouTube (ossia le parole inglesi bannate dai commenti di YouTube): prendendo il numero delle volte in cui le suddette parole sono presenti nello spettacolo e dividendolo per le parole totali dello show, si è ottenuta questa metrica, che è stata utilizzata sia per la classificazione sia per selezionare gli show da analizzare nella sentiment analysis.

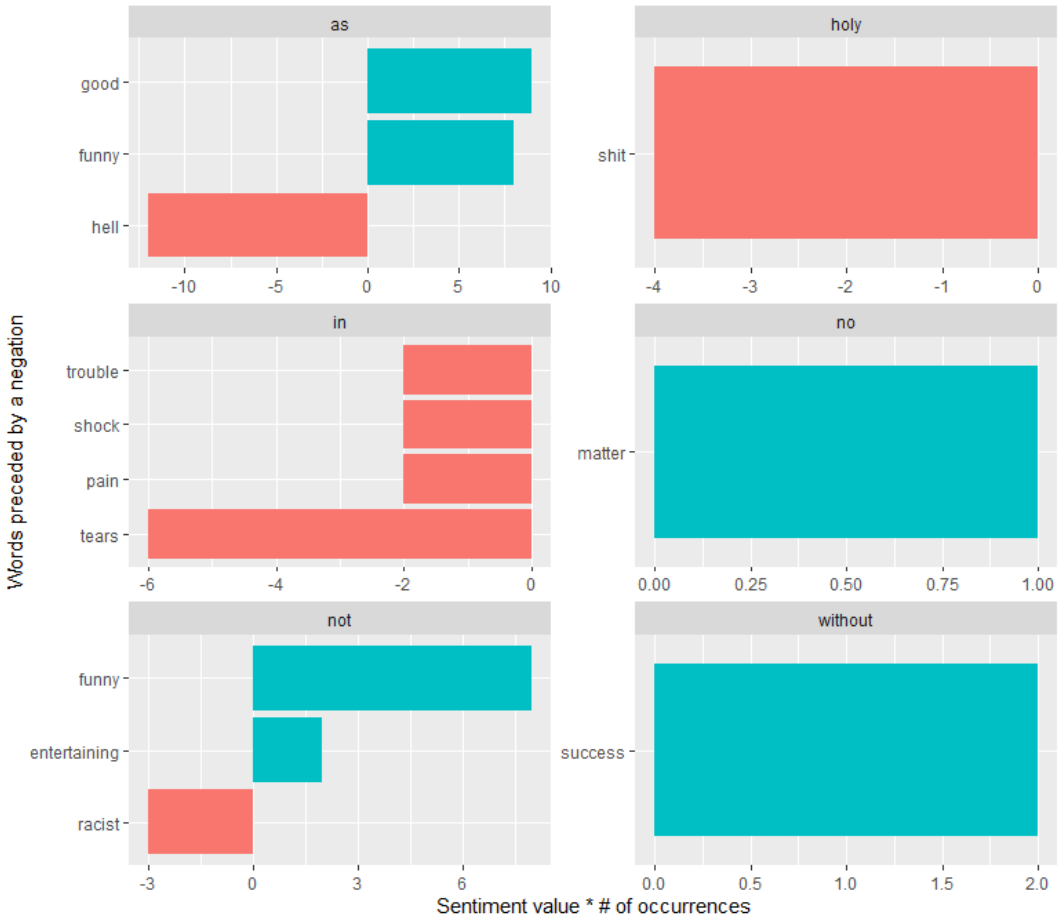
2. Sentiment analysis

È stato ritenuto interessante fare una sentiment analysis dei tweet che parlano dei vari spettacoli.
Considerando la difficoltà nel fare lo scraping da Twitter, è stato deciso di selezionare un numero ristretto di spettacoli, optando per i tre più volgari e i tre meno volgari, ricavati sfruttando la metrica relativa alla volgarità descritta nel precedente paragrafo.
Per ciascuno spettacolo sono stati presi tweet dal giorno della pubblicazione dello show, fino al mese successivo, senza considerare risposte ad altri tweet o tweet che contenessero link.
È stata condotta un'analisi sulle parole che danno maggior contributo al *sentiment value* (per calcolarlo è stato utilizzato l'*afinn sentiment lexicon*, con un intero tra -5, valenza negativa, e +5, valenza positiva) per vedere se ci fosse qualche parola non informativa che contribuisse erroneamente. Nella *Figura 2.1* si nota come una parola che influenza molto negativamente è *rob.*



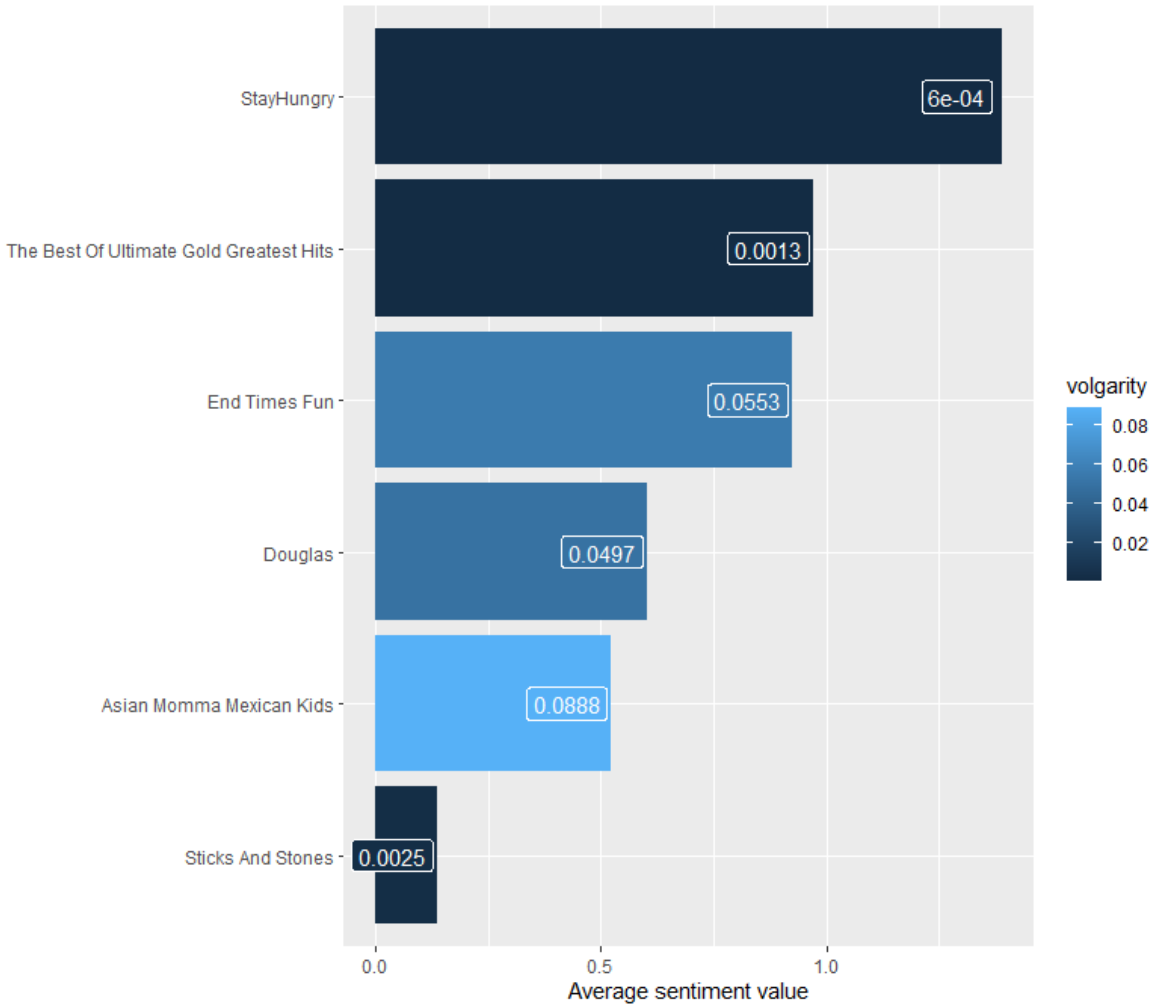
2.1 Parole che influiscono di più sulla sentiment analysis per spettacolo

rob non è altro che il nome dell'autore dello spettacolo *Asian Momma, Mexican Kids*, e non è stato perciò considerato. Altri termini quali *shit*, *fucking* e le parole usate per fare la ricerca su Twitter non sono state considerate perché non informative o ambigue per il calcolo del sentiment value.
Successivamente è stata fatta un'analisi dei bigrammi contenenti una negazione o che potrebbero avere una valenza contraria rispetto a quanto considerato dal dizionario ontologico usato per la sentiment (il dizionario *afinn*) (*Figura 2.2*).



2.2 Analisi dei bigrammi

Questi sono poi stati sostituiti con unigrammi con sentiment value uguale in modulo ma di segno opposto, ad esempio *not funny* viene sostituito da *fraud* (*funny* ha valore +4, *fraud* -4).
Dopo aver effettuato tutte queste correzioni, dalla *Figura 2.3* si vede come non sembra esserci un legame tra volgarità dello spettacolo e quanto quest'ultimo venga apprezzato dal pubblico. Ovviamente sono tutte analisi meramente qualitative, considerando che sono stati presi in esame solamente 6 spettacoli, campione troppo piccolo per fare una qualsiasi affermazione non qualitativa.
Interessante vedere come il valore sia maggiore di zero per tutti gli spettacoli, indicando come il pubblico di Twitter americano sembri apprezzare la stand-up comedy, sia quando questa è volgare che quando non lo è.



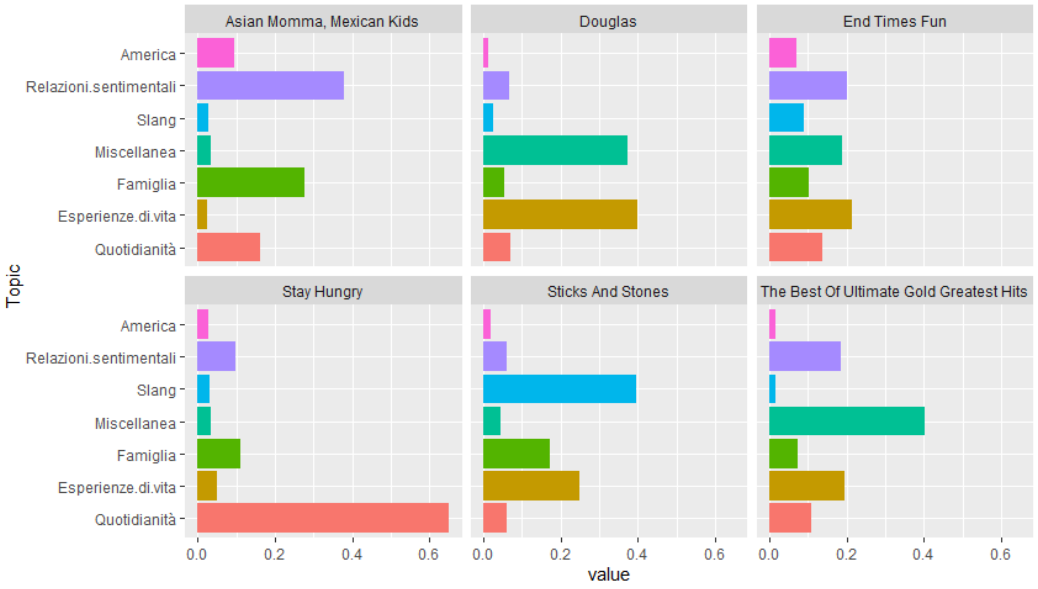
2.3 Spettacoli rispetto al sentiment value per tweet e livello volgarità (colore)

3. Classificazione

Costruita la metrica della volgarità, si è tentato di classificare i testi in base a quest'ultima misura, utilizzando gli stilemi di ogni singolo spettacolo come esplicative. Tuttavia, trattando la volgarità come dicotomica, l'albero ottenuto non è risultato informativo, avendo divisioni date da termini generici.

4. Topic modeling

Un altro scopo dell'analisi è stato quello di individuare gli argomenti trattati nelle stand-up comedy. In prima battuta, si è provveduto alla modifica dei testi dei 228 show in modo da tenere in considerazione alcuni bigrammi particolarmente informativi e maggiormente presenti nei testi come, ad esempio, *Donald Trump*, *white people*, *black people*. In seconda battuta, così come è stato fatto nell'analisi esplorativa, sono stati divisi i testi in modo che ogni parola risultasse un'osservazione, sono state eliminate le stop words, e applicato lo stemming alle parole.
Con l'insieme degli stilemi a disposizione si è andati a calcolare la frequenza di ogni stilema in ogni stand-up comedy e, a partire dalle frequenze, è stata creata la Document-Term Matrix (DTM). Quest'ultima ci ha permesso, mediante l'algoritmo di Gibbs Sampling e con un'interpretazione ex-post, di individuare sette topic: *America*, *Relazioni sentimentali*, *Slang*, *Miscellanea*, *Famiglia*, *Esperienze di vita* e *Quotidianità*.
Infine, sono state calcolate le proporzioni dei topic in ogni testo ed è stata focalizzata l'attenzione sui sei testi analizzati nella *sentiment*. La *Figura 4.1* mostra la differenza degli argomenti trattati nelle tre stand-up meno volgari: in *Stay Hungry* si parla notevolmente della quotidianità, in *Sticks And Stones* vi è un costante uso di slang e una presenza non banale delle esperienze di vita, mentre in *The Best Of Ultimate Gold Greatest Hits* prevale una miscellanea di argomenti.



4.1 Proporzione dei topic nei tre testi più volgari e nei tre testi meno volgari

D'altro canto, tra i tre show più volgari, *Asian Momma, Mexican Kids* affronta maggiormente i temi della famiglia e delle relazioni sentimentali, *Douglas* è incentrato sulle esperienze di vita e sulla miscellanea, mentre *End Times Fun* non ha un topic prevalente.

5. Conclusione

Lo studio mostra come, alla base degli spettacoli, vi sia una forte influenza della vita ordinaria, e un frequente uso della volgarità. Questo è riscontrabile nei termini più utilizzati e nei topic trovati tramite la *Latent Dirichlet Allocation* (LDA): il topos della vita quotidiana probabilmente mette in relazione il pubblico e il comico, mentre la volgarità provoca una reazione non sempre negativa nello spettatore (come visto nella *Sentiment Analysis*), talvolta smuovendo convinzioni e valori tradizionali.