Davide Capone
Matricola: 148324
148324@spes.uniud.it

Music Lyrics over Decades:

a NLP study

Introduzione

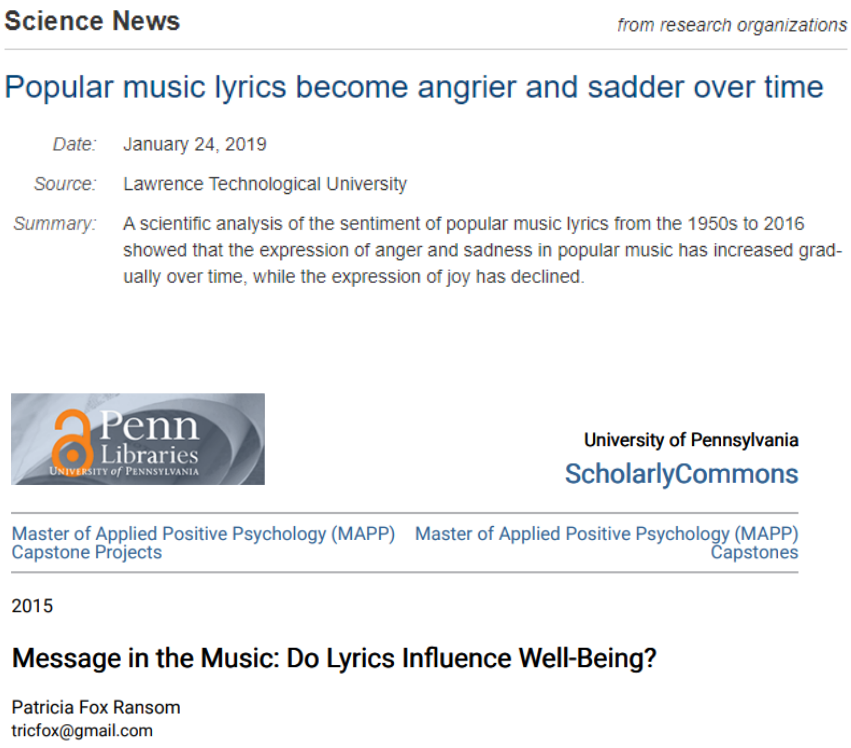
Descrizione del progetto

Questo progetto ha l'obiettivo di analizzare, mediante tecniche di Text Mining, come sono variati i testi musicali dal 1950 fino ai giorni nostri.
In particolare verranno considerate sette decadi principali:
anni *1950-1960, 1960-1970, 1970-1980, 1980-1990, 1990-2000, 2000-2010* e infine *2010-2020*.

Studiare com'è cambiata la musica, nel suo linguaggio, ci permette di descrivere quali sono stati i passaggi chiave nella storia e come la società si sta evolvendo.
Sappiamo inoltre, grazie a studi scientifici, che non solo è il contesto storico che caratterizza la musica, ma anche il viceversa: siamo di fatto influenzati, positivamente o negativamente da ciò che ascoltiamo.

Articoli e pubblicazioni introduttive al progetto

Riporto due articoli principali che introdurranno l'analisi:



Science News *from research organizations*

Popular music lyrics become angrier and sadder over time

Date: January 24, 2019

Source: Lawrence Technological University

Summary: A scientific analysis of the sentiment of popular music lyrics from the 1950s to 2016 showed that the expression of anger and sadness in popular music has increased gradually over time, while the expression of joy has declined.

Penn Libraries
University of Pennsylvania

University of Pennsylvania
ScholarlyCommons

Master of Applied Positive Psychology (MAPP) Capstone Projects

Master of Applied Positive Psychology (MAPP) Capstones

2015

Message in the Music: Do Lyrics Influence Well-Being?

Patricia Fox Ransom
tricfox@gmail.com

Il primo studio, condotto dalla *Lawrence Technological University*, afferma che le canzoni, con l'aumentare del tempo diventano sempre più cattive e tristi.
Il secondo invece, condotto dall'*University of Pennsylvania*, afferma che i testi delle canzoni possono influenzare il nostro stato psico-fisico (e in particolare il nostro stato di benessere).

Risulta chiaro, quindi, quanto sia importante effettuare un analisi su questo tema.

Obiettivi

- variazione del sentiment: è possibile riscontrare un aumento del sentiment negativo negli anni?
- quali tipologie di sentiment sono ora maggiormente riscontrabili? e quali in declino?
- quali sono le parole che caratterizzano maggiormente il nuovo millennio? quali invece sono fortemente in contrasto con la prima decade?
- la distribuzione delle parole utilizzate è rimasto invariato?

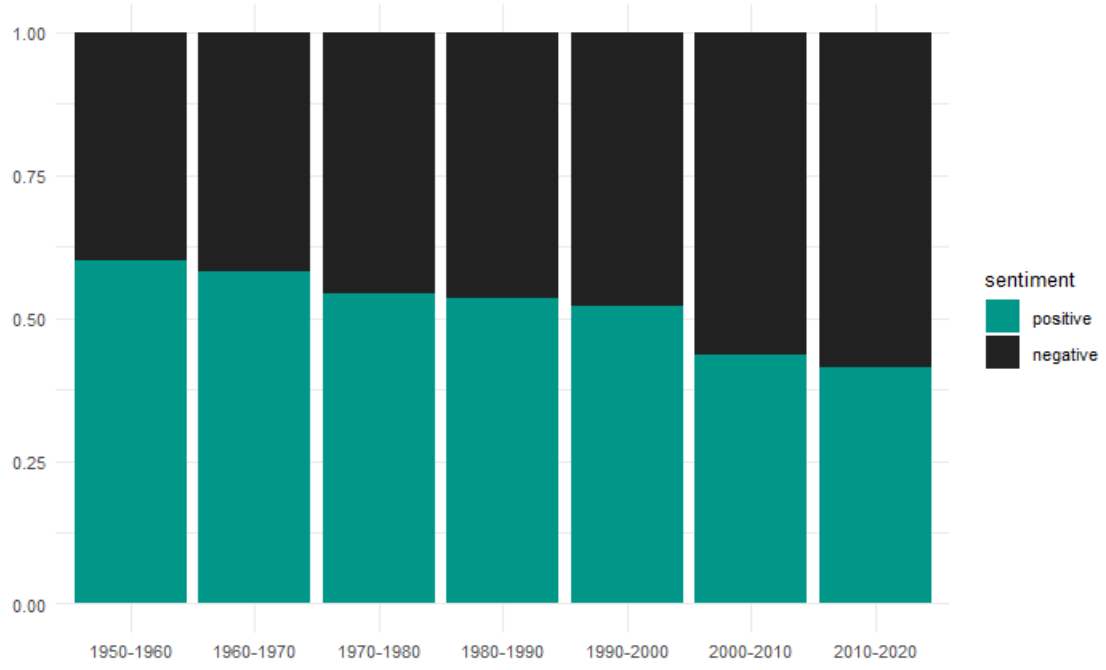
Music lyrics among decades

Descrizione dei dataset utilizzati

Il primo dataset utilizzato è reperibile dalla piattaforma *Kaggle*. Raccoglie oltre 25000 testi musicali in lingua inglese con descrizione di artista e anno di uscita.
Tutte le canzoni fanno riferimento a musica popolare e/o tipiche di quella decade.
Per facilitare l'analisi si è scelto mantenere nel dataset contenente l'intero listato di parole (*token*, approccio *tidy*) solo quelle che trovano una corrispondenza in un vocabolario di lingua inglese (o parole comunemente usate anche in tale lingua).
Sono state aggiunte al vocabolario varie parolacce conosciute (vedremo infatti che l'uso di termini non appropriati risulta essere in certe decadi molto frequente).

Variazione del sentiment negli anni

Nella prima analisi viene studiata la variazione del sentiment negli anni.
In particolare verificando se esiste un significativo aumento del sentiment negativo o di quello positivo. Avendo come obiettivo principale la conferma del primo articolo riportato nell'introduzione.

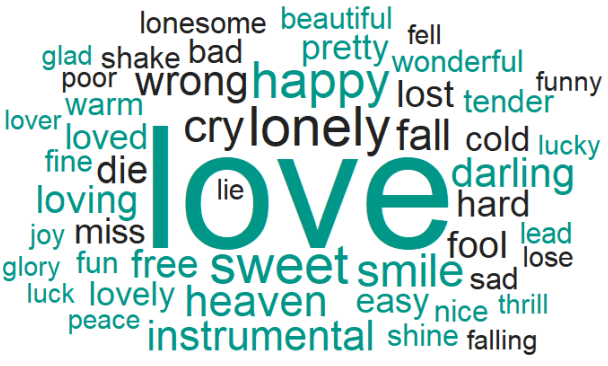


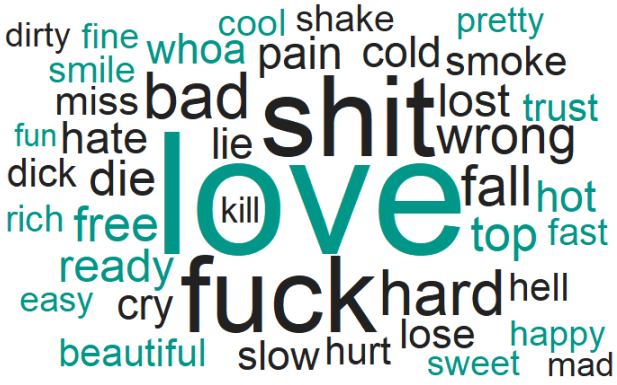
Il grafico mostra chiaramente come il sentiment negativo sia in significativa crescita.
In particolare all'inizio del ventunesimo secolo questo trend viene accentuato: il sentiment negativo risulta essere prevalente su quello positivo in termini relativi percentuali.

Interessante notare che, a partire dal *1970* e verso fine anni *'90,* il sentiment positivo e quello negativo risultano essere bilanciati.

Wordclouds

La variazione del sentiment studiata nel paragrafo precedente, implica un uso sicuramente diverso tra le parole nella prima e nell'ultimo decade (*1950-1960*, *2010-2020*).
Possiamo visualizzare queste differenze attraverso due *wordclouds*:

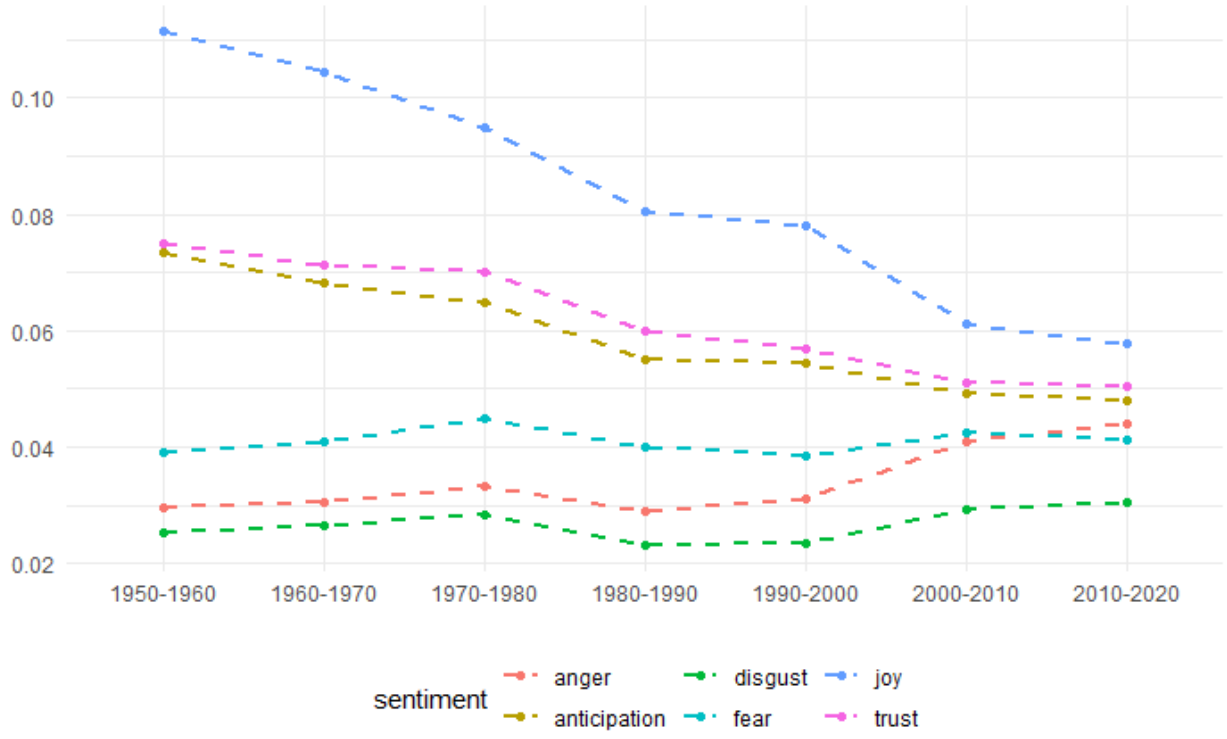




La prima wordcloud fa riferimento alla decade ‘50-’60 mentre la seconda alla decade ‘10-’20 del ventunesimo secolo.
Sebbene la parola più ricorrente risulti essere ‘*love*’ in entrambi i casi, si può notare come emergano prepotentemente parole come: *dick, shit, fuck, ..* nell’ultima decade.
Quest’aspetto è sicuramente dato dalla forte presenza delle canzoni rap e/o trap, che dominano le attuali classifiche e che sono in genere contraddistinte da un uso di parole spesso inappropriate.
Inoltre possiamo notare come emergano, nell’ultimo wordcloud, aspetti più introspettivi, legati alla persona: *pain, hurt, cry:* fanno intuire che i testi trattino maggiormente temi quali delusioni amorose, ciò contribuisce negativamente in termini di sentiment.

Tipologie di sentiment dominanti e trend

Interessante è stato analizzare la variazione delle tipologie di sentiment negli anni:



Si può notare come il sentiment inerente la gioia (*joy*) e la fiducia (*trust*) siano in netta decrescita rispetto alle decadi passate.

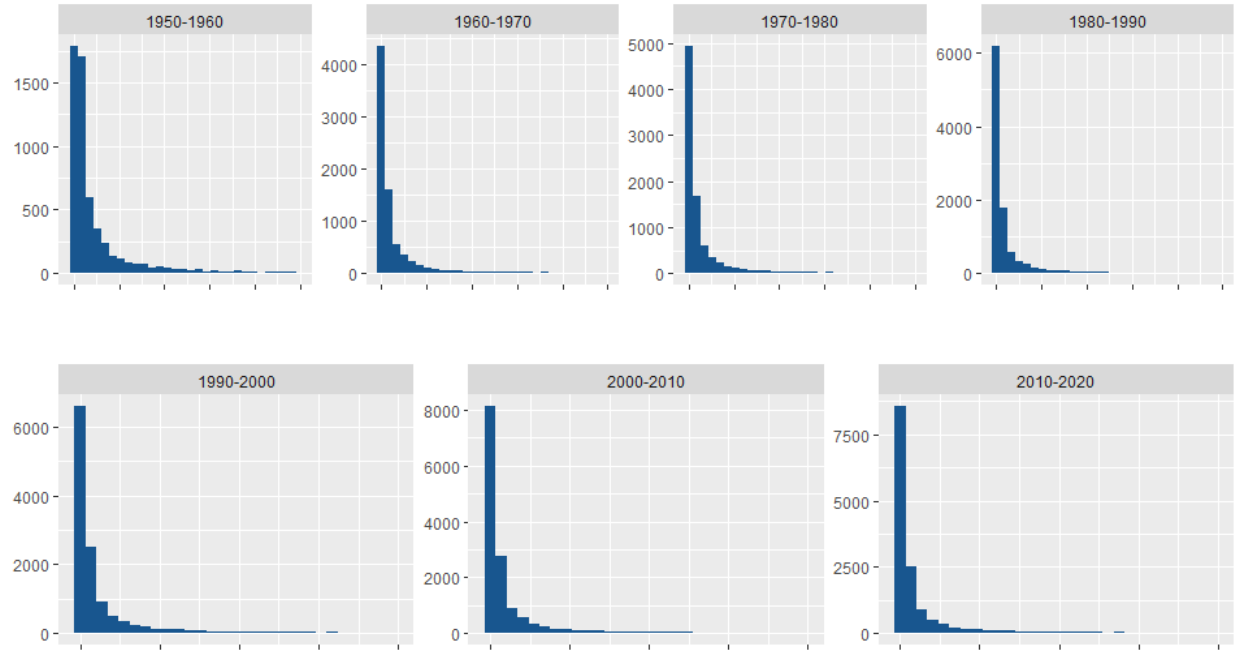
Contrariamente a quest'ultimo aspetto, il sentiment inerente la rabbia (anger) sta subendo una sostanziale crescita a partire dalla decade *1990-2000*.

La decrescita del sentiment dell'anticipazione (*anticipation*) fa pensare che probabilmente si stanno vivendo tempi maggiormente incerti (causa crisi economiche, aumento povertà, ecc...) e si hanno meno aspettative sul futuro.
Di fatto anche *Umberto Galimberti*, un noto filosofo e psicoanalista, in uno dei suoi discorsi ha affermato che:
"Il futuro non dà più certezze come nel passato [...] Viviamo nell'epoca nichilista. Ci parlano di speranza, di augurio di un futuro diverso. Tutte parole negative, che non danno certezze nel domani".

Si può infine notare che nella decade *1970-1980* il sentiment inerente la rabbia (*anger*) si sia leggermente alzato: questo potrebbe essere dato dal fatto che in quegli anni riscossero molto successo generi aventi tematiche molto forti: l'hard-rock ne è un esempio.

Term and document term frequency

Distribuzioni di frequenza delle parole

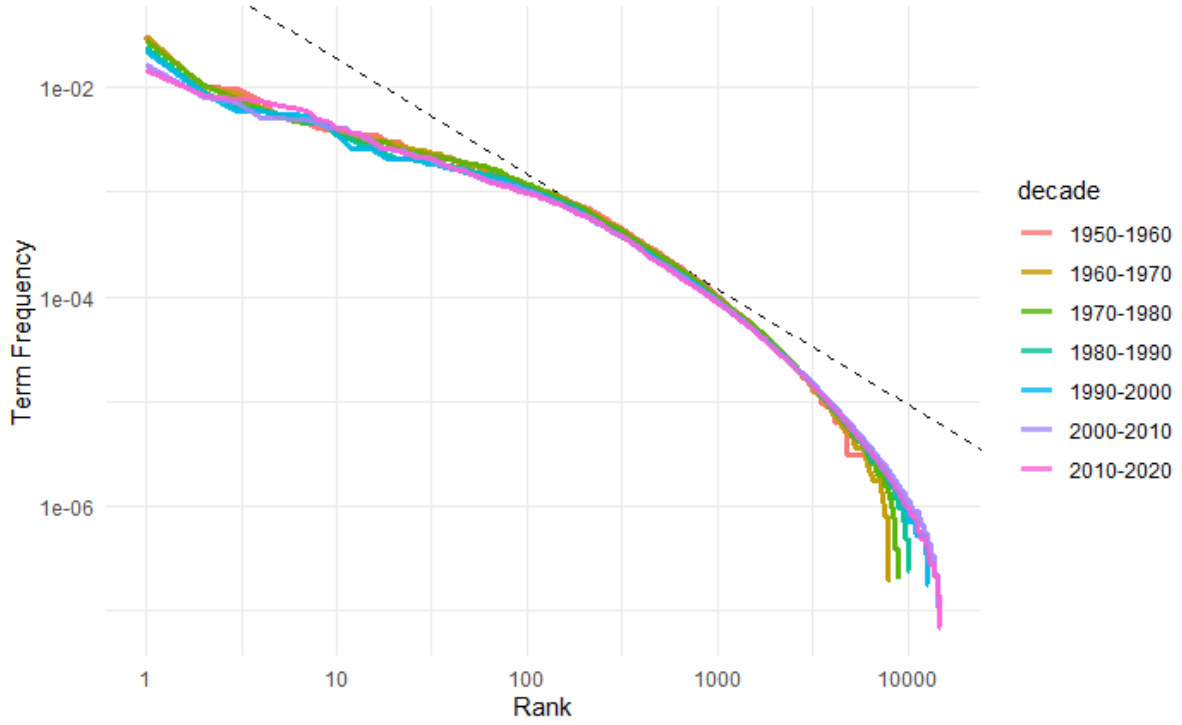


In tutte e sette le decadi possiamo notare come le distribuzioni di frequenza siano sostanzialmente delle *power-law:* pochi termini vengono ripetuti molte volte e tanti termini che invece appaino raramente.

Possiamo quindi verificare, mediante grafico, la *legge di Zipf*.

Conferma della legge di Zipf

Ho classificato le parole tra le più popolari e le più rare per ogni decade (effettuando quindi un *ranking)*. La legge di zipfs afferma che la frequenza che una parola appaia è proporzionale al suo rank.



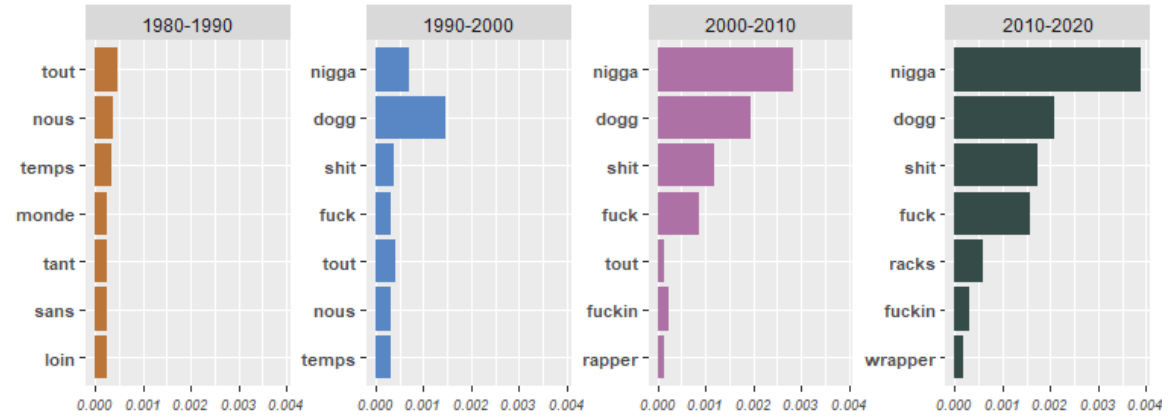
La linea tratteggiata rappresenta la curva teorica in funzione del rank.

La curva rispetta la legge di Zipf maggiormente nei ranghi centrali, ma meno nei ranghi bassi: probabilmente perché sono state rimosse le *stop-words,* le quali tendono ad essere ripetute molte volte, e in loro presenza, nei ranghi bassi la distribuzione si avvicinerebbe quindi al modello teorico.

Term frequency - Inverse Document Frequency

Viene applicato *tf-idf* per valutare quali sono state le parole più caratterizzanti per ogni decade. In questo caso i documenti sono l'insieme delle parole per decade.

Risultati significativi sono riscontrabili nelle ultime quattro decadi:



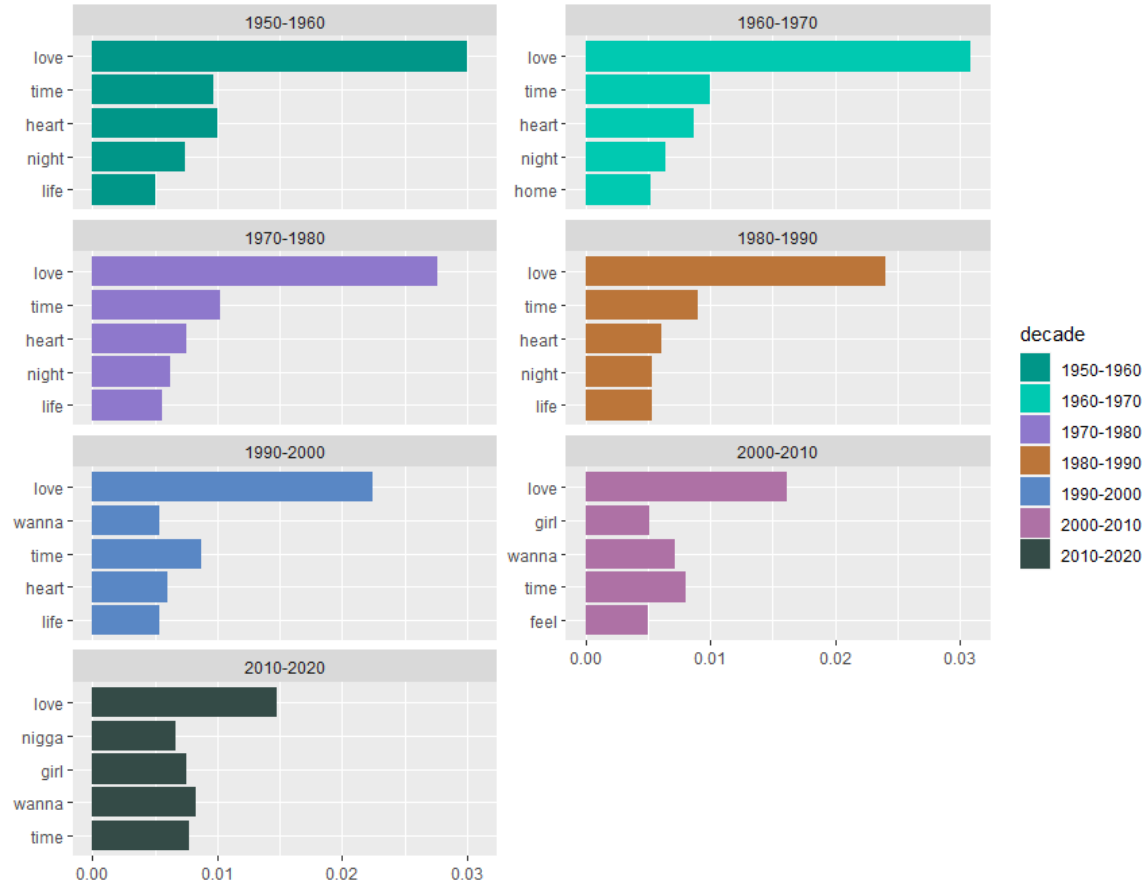
E' interessante notare come le parole in lingua francese siano molto caratterizzanti per la decade *1980-1990*: a seguito di un indagine sul web ho potuto verificare che, in quegli anni, le canzoni pop-dance francesi ebbero una forte crescita su larga scala.

Emerge inoltre che parole quali *nigga, dogg, shit, fuck* siano sempre più frequenti negli anni. Anche in questo caso è evidente che l'influenza principale è data da generi rap e trap.

La parola *racks* deriva principalmente dalla cultura trap: questa parola è stata inventata in quei anni e viene utilizzata per simboleggiare lo stato di ricchezza della cantante. Abbiamo un aspetto di vanità rilevante in quella decade.

Termini utilizzati maggiormente

Si è voluto rappresentare i cinque termini maggiormenti utilizzati per ogni decade:



Rappresentando la frequenza in termini relativi (rispetto al numero totale di parole per decade), ci permette di fare confronti tra decadi sull'uso dei termini maggiormente utilizzati.

Il termine *love*, risulta essere sempre il più gettonato: di fatto la maggior parte delle canzoni raccontano una storia d'amore in un modo o nell'altro.

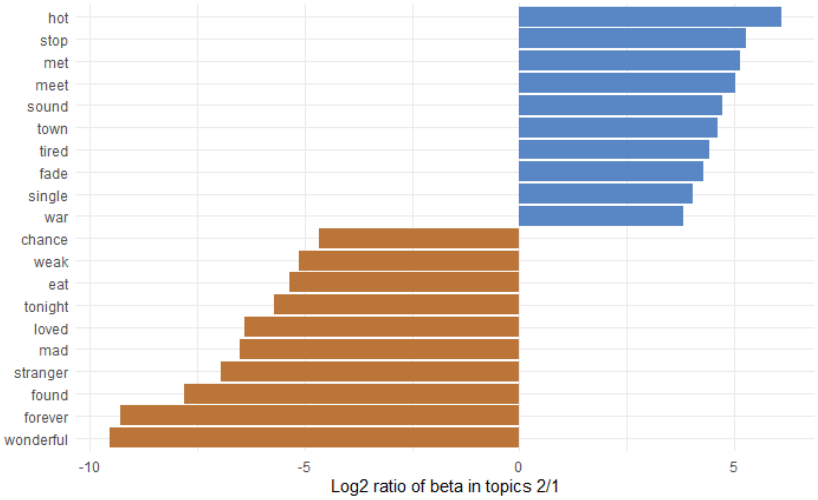
Ma la tendenza nell'usare la parola love sta decrescendo nel tempo. Questo fa pensare che in futuro potrebbe non essere la più frequente e che probabilmente le canzoni si stanno sempre di più svincolando dai racconti d'amore.

Possiamo notare inoltre che le decadi *'50-'60-'70-'80-'90-'00* risultano essere molto simili e solo recentemente (ultime due decadi) possiamo notare dei cambiamenti significativi.

Topic modelling

Riporto unicamente le decadi per cui ho ottenuto risultati significativi.

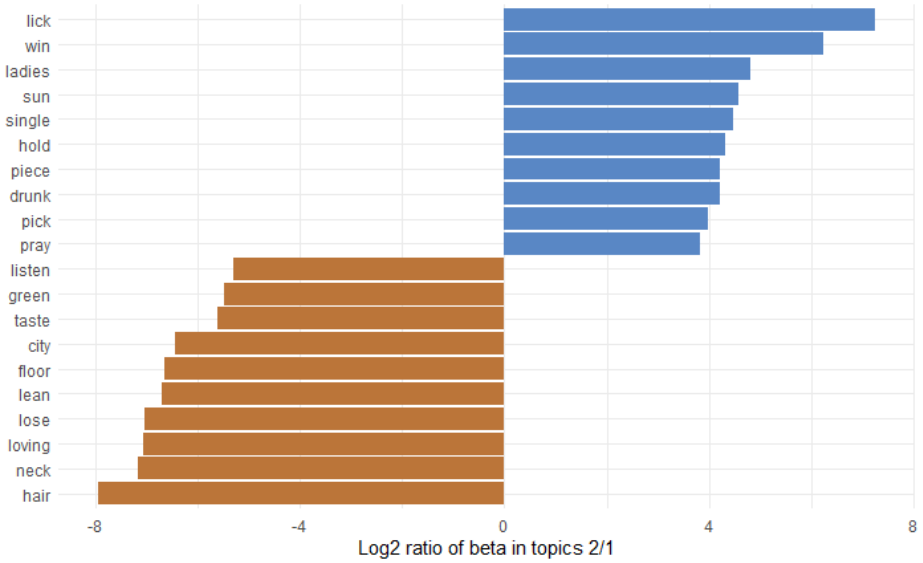
1950-1960 topics



Il grafico raffigura due topic principali per la decade *1950-1960*.
Risultano essere prevalenti argomenti che incitano alla fine della guerra e alla pace (parole quali: *war, stop, tired, sound*) infatti la decade si colloca alla fine della Seconda Guerra Mondiale.

Il secondo topic sembra trattare di storie classiche d'amore (parole quali: *wonderful, found, forever, tonight, loved*).

2010-2020 topics



Un topic principale per questa decade sembra trattare feste (*drunk*, *ladies, single*).

Conclusioni

Il sentiment nelle canzoni è diventato sempre più negativo nel tempo.
Le differenze tra la prima e l'ultima decade analizzate sono significative:

- 1950-1960 rappresenta la decade con maggior sentiment positivo nei testi musicali
- la decade *2000-2010* registra il più alto livello di negatività nella storia (considerando gli ultimi 70 anni).
- l'uso di termini inappropriati e/o che simboleggiano tratti di vanità sono in aumento
- anche gli argomenti differiscono molto: a causa di continui cambiamenti nella società

Fonti

- Dataset principali: https://www.kaggle.com/datasets/terminate9298/songs-lyrics?select=songs_details.csv
- Lista di 'Bad Words': https://www.kaggle.com/datasets/nicapotato/bad-bad-words
- Umberto Galimberti parla del disagio giovanile: "Il futuro non dà più certezze come nel passato": https://www.qdpnews.it/comuni/sernaglia-della-battaglia/umberto-galimberti-parla-a-sernaglia-del-disagio-giovanile-il-futuro-non-garantisce-piu-certezze-come-nel-passato/
- L'invasione della musica pop-dance francese degli anni '80: https://www.capital.it/articoli/anni-80-musica-francese-lio-plastic-bertrand-stephanie-di-monaco/
- Message in the Music: Do Lyrics Influence Well-Being: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1094&context=mapp_capstone
- Popular music lyrics become angrier and sadder over time: https://www.sciencedaily.com/releases/2019/01/190124124737.htm