Il linguaggio del rap

Possibilità di un'analisi multidisciplinare: web-scraping, text-mining, data-visualization

Stefano PERNA, Raffaele GUARASCI, Alessandro MAISTO, Pierluigi VITALE
Università di Salerno, Fisciano (SA), Italia

intro

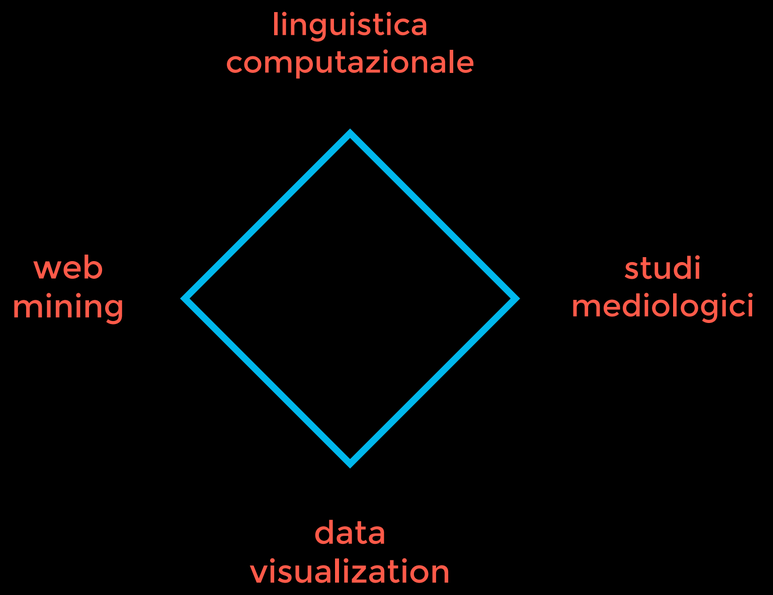
Obiettivi

- Costruzione di una risorsa terminologica e di strumenti interattivi di consultazione per l'analisi dell'universo linguistico e semantico di una parte importante della musica popolare in lingua italiana, il rap

- Elaborazione di nuovi metodi e strumenti per l'estrazione e l'analisi degli User Generated Contents diffusi nel Social Web

intro

Approccio



background

User Generated Contents

UGC denotes any form of content such as blogs, wikis, discussion forums, posts, chats, tweets, podcasting, pins, digital images, video, audio files, and other forms of media that was created by users of an online system or service, often made available via social media Web sites. The content refers to news, encyclopedias and other reference works, movie and product reviews, problem processing, posting of consumer items and comments on them, accounts of (personal) happenings and events, fan fiction, trip planners, crowdfunding and others, and the content is often created through interaction with other users.

Moens, Li, Chua 2014

background

User Generated Contents

- Gli UGC rappresentano una straordinaria occasione per la costruzione di collezioni di dati e di risorse di conoscenza altrimenti difficilmente ottenibili

MIR & User Generated Contents

- Nel settore del Music Information Retrieval (MIR), ad esempio, la presenza sul web di grandi quantità di informazioni pubblicate dagli utenti e dagli appassionati di musica su siti personali, webmagazines, blog, social media ha aperto la strada a nuovi campi e prospettive di ricerca [Schedl, Sordo, Koenigstein, Weinsberg 2014]

- In particolare la grande opera di trascrizione su web dei testi delle canzoni da parte dei fan ha reso disponibili enormi quantità di testi digitalizzati pronti per l'analisi automatica e il text-mining [Mahedero et al., 2005; Kleedorfer et al., 2008; Hu et al., 2009; Hirjee and Brown, 2009; Hirjee and Brown 2010; Malmi et. al, 2015]

musixmatch TOP LYRICS COMMUNITY APPS

SIGN IN SIGN UP

Words Matter.

Discover the world's largest lyrics catalog

Type song title, artist or lyrics

Top lyrics

Faded
Alan Walker

"The monsters running wild inside of me"

♡ 16160

Stressed Out
twenty one pilots

"When our momma sang us to sleep but now ...

♡ 23814

LATEST ON POP GENIUS

16 days ago
sereinik

Zayn's Going His Own Direction

The hearts of a million and one teenagers broke in early 2015 when the news that One Direction's own Zayn Malik was leaving the band for his own solo project. The English artist has reemerged as ZAYN, and he's revamped to be far slicker than ever before. Whether he's channeling JT on "BeFoUr" or pushing ethnic Urdu into the mainstream with "Flower", ZAYN proves that he is truly a mind of his own with his debut album *Mind of Mine*.

a month ago
Slickk

ABOUT POP GENIUS

Welcome to Pop Genius! We break down the catchiest tunes of all time, from Phil Spector drama to Madonna to Lana. Join the discussion in our forum, and follow us on Twitter and Facebook.

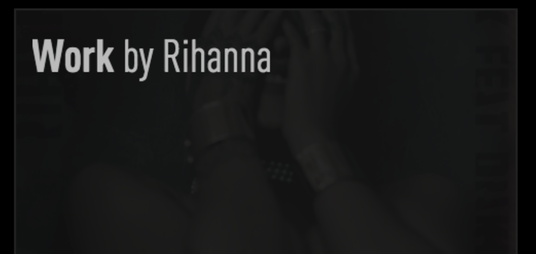
HOT ON POP GENIUS

One Dance by Drake

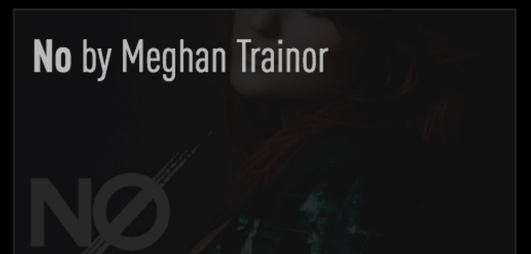
7 Years by Lukas Graham

Low Life by Future

Work by Rihanna



No by Meghan Trainor



SONGLYRICS
know the words®

Search artists, albums, and songs

Top Albums

View All »

Mind of Mine
ZAYN

More Issues Than Vogue
K. Michelle

Anti
Rihanna

Purpose
Justin Bieber

25
Adele

Featured Stories

All Music News »

Top 25 Lyrics

Top 100 of All Time »

1		**Love Yourself** Justin Bieber
2		**Work (feat. Drake)** Rihanna

In the Know

All Music News »

RIFF'D: Ben Harper & the Innocent Criminals' 'Call it What it Is'

rap

Il rap è tra i generi musicali più vitali e di maggiore impatto degli ultimi decenni (Lena, 1995; Toop, 1999; Forman and Neal, 2004; Pinkney, 2007), estesosi ormai ben oltre gli originari confini statunitensi per divenire fenomeno globale (Androutsopoulos and Arno 2003; Osumare, 2007; Alim et al., 2008) all'interno del quale è possibile riscontrare una ricchissima produzione testuale ed un alto tasso di innovazione e sperimentazione di forme linguistiche (Cutler, 2007; Bradley, 2009; Terkourafi, 2010).

rap

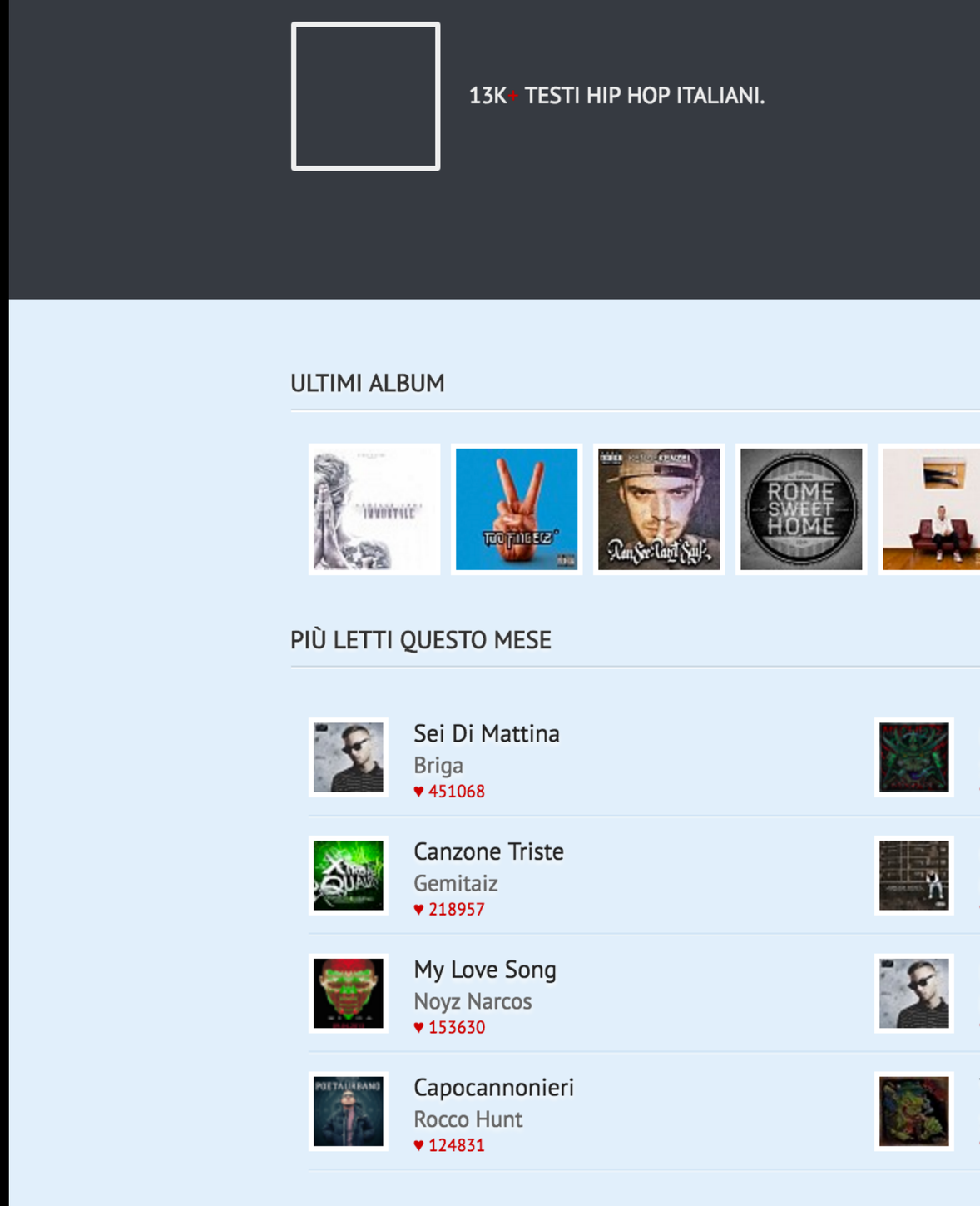
L’idea alla base del presente lavoro è quella di elaborare una risorsa che renda possibile ottenere una “cartografia” della lingua del rap, che permetta di osservare e analizzare nel suo complesso un settore della produzione culturale contemporanea estremamente diffuso e popolare anche in Italia (Pacoda 1996; Filippone and Papini, 2002; Attolino, 2003; Scholz 2005)

outline ricerca

- web-scraping
- text-mining | analisi linguistica
- data visualization

webscraping

- fonte: raptxt.it
- addestramento web-crawler e estrazione: import.io e python beautifoulsoup module
- campi estratti: Autore; Coautori; Titolo Canzone; Titolo Album, Testo Canzone
- > 15.000 testi di canzoni estratte



A Cui!

Turi

Mi piace 0

♥ 2567

Rit.

A cui!
Cu sugnu eu!
Cu cazzu siti vui, cunta 1 e 2
e vidi poi cu simu nui, cotrari
ditincillu vui chi 'ndannu a fari (aundi vai,aundi vai,si c'è stu cumpari)

Chi 'nda a sapiri, aundi 'nda a jiri,
pigghja pili, avogghja u spari palli,
cca non po trasiri!
Dassati stari cumpari, ca scorcia chi ssi 'nduri ora
i sordi fannu comu o sceccu 'nte lenzola!
--
(Al momento sono a scuola, nn ho il lettore cd dietro e non posso finirlo..PROMETTO DI FARLO ENTRO 2 GIORNI)

text-mining e analisi

L’analisi è stata realizzata su un campione composto dai testi di 100 autori tra i più significativi per un totale di circa 3900 canzoni

Per l’analisi è stato usato il linguaggio Python con NLTK e alcuni moduli sviluppati ad-hoc

Preprocessing

- Data cleaning
- Tokenizzazione
- PoS Tagging
- Lemmatitazzione

Analisi statistica-linguistica

- Statistiche, distribuzioni, ranking
- Bigrammi e Trigrammi
- TF/IDF

text-mining e analisi

Output

- Corpus annotato di 1.866.089 tokens (3900 canzoni)
- Dizionario di frequenze del rap italiano
- 1milione di types circa
- Part Of Speech
- Bigrammi e trigrammi
- Mappa della forza di associazione tra parole e autori (tf-idf)

data visualization

Rendere il corpus facilmente consultabile, navigabile, osservabile nel suo complesso

“Distant Reading”
(Moretti 2005)

Affrontare le problematiche poste dalla visualizzazione di grandi corpora testuali
(Wise et al., 1995; Fortuna et al., 2005; Alencar et al., 2012; Sinclair et al., 2013; Kucher, 2014; Brath and Banissi, 2015)

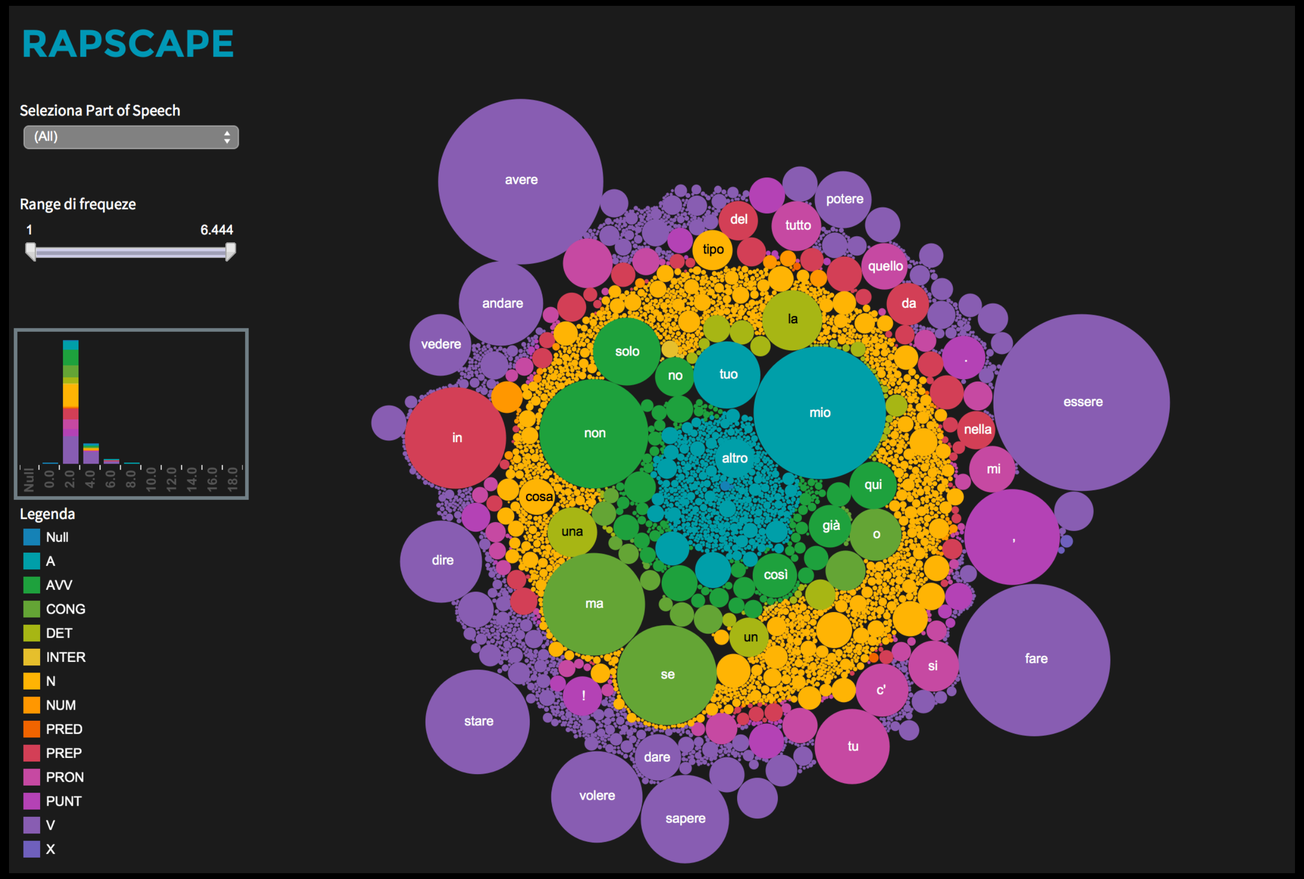
data visualization

Prototipo: Tableau Public; Javascript

Viste e Filtri Multipli

Overview first, zoom and filter, then details-on-demand (Shneidermann, 1996)

data visualization



https://public.tableau.com/profile/publish/rapscape/Overview#!/publish-confirm

data visualization

conclusioni e prossimi passi

- Estensione analisi all’intero dataset
- Arricchimento dataset con metadati via API pubbliche (anno, etichetta, geografia, ecc…)

- Topic Modeling
- Analisi stilometrica (rime, authorship attribution)

- Finalizzazione del tool e pubblicazione sul web

- Analisi socio-linguistica con ricercatori esperti di dominio


grazie
LENGTHOB
ICE
CUBE