

TODO

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Milana Kovačević
Ivan Ristović

jun 2018.

Sadržaj

1	Opis skupa podataka	2
2	Korišćeni alati	2
3	Preprocesiranje	2
4	Zaključak	2
	Literatura	2
A	Dodatak	3

1 Opis skupa podataka

The Million Song Dataset [6] je skup od milion slogova koji sadrže informacije o popularnim pesmama. S obzirom da ovaj skup preveliki za udoban rad, ograničićemo se na podskup od deset hiljada slogova izdvojen od strane autora originalnog skupa.

Svaki slog pomenutog skupa podataka sadrži informacije o jednoj pesmi: detalje o izvodjaču, segmentima, tempu kao i ID pesme na raznim online servisima (*Echo Nest* [2], *7digital* [1], *MusicBrainz* [3] i *PlayMe* [4]). Detaljne informacije o atributima se mogu videti na slici 1.

Skup u svojoj originalnoj formi je organizovan u *HDF5* format [7]. Mi ćemo izdvojiti informacije iz datog modela i podatke organizovati u CSV format, zarad lakšeg ubacivanja u alate koje ćemo koje ćemo opisati kasnije. Izdvajanje podataka i njihova konverzija u CSV format je izvršena korišćenjem Python programa koji se mogu naći u dodatku A.

2 Korišćeni alati

Za obradu podataka ćemo koristiti alate *Knime Analytics Platform* [5] i *IBM SPSS Modeler* [8].

3 Preprocesiranje

TODO Preprocesiranje

4 Zaključak

Zaključak... TODO

Literatura

- [1] 7digital. on-line at <https://www.7digital.com/>.
- [2] The echo nest. on-line at <http://the.echonest.com/>.
- [3] Musicbrainz. on-line at <https://www.musicbrainz.org/>.
- [4] Play.me. on-line at <https://www.playme.com/>.
- [5] KNIME AG. Knime analytics platform. on-line at <https://www.knime.com/knime-analytics-platform>.
- [6] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset, 2011. on-line at <https://labrosa.ee.columbia.edu/millionsong/>.
- [7] The HDF Group. The hdf5 data model, 2011. on-line at <https://support.hdfgroup.org/HDF5/>.
- [8] IBM. Ibm spss modeler. on-line at <https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software>.

A Dodatak

python kod ovde

Atribut	Tip podatka	Kratki opis
analysis sample rate	float	učestalost uzorkovanja
artist 7digitalid	int	7digital ID izvođača ili -1
artist familiarity	float	algoritamska aproksimacija
artist hotttnesss	float	algoritamska aproksimacija
artist id	string	Echo Nest ID izvođača
artist latitude	float	geografska širina
artist location	string	lokacija autora
artist longitude	float	geografska dužina
artist mbid	string	MusicBrainz ID izvođača
artist mbtags	array string	niz MusicBrainz tagova
artist mbtags count	array int	broj MusicBrainz tagova
artist name	string	ime autora
artist playmeid	int	PlayMe ID izvođača ili -1
artist terms	array string	niz Echo Nest tagova
artist terms freq	array float	frekvencije Echo Nest tagova
artist terms weight	array float	težina Echo Nest tagova
audio md5	string	MD5 heš kod audio zapisa
bars confidence	array float	pouzdanost takta
bars start	array float	niz početaka taktova
beats confidence	array float	pouzdanost ritma
beats start	array float	niz početaka ritmova
danceability	float	algoritamska aproksimacija
duration	float	trajanje audio zapisa (u sekundama)
end of fade in	float	vreme u odnosu na pocetak u kom prestaje
energy	float	fade-in efekat (u sekundama)
key	int	algoritamska aproksimacija energije
key confidence	float	pesme od strane slušaoca
loudness	float	tonalitet u kojem je audio zapis
mode	int	pouzdanost tonaliteta
mode confidence	float	prosečna jačina (u dB)
release	string	mod - dur ili mol
release 7digitalid	int	pouzdanost moda
sections confidence	array float	ime albuma
sections start	array float	7digital ID albuma ili -1
segments confidence	array float	niz pouzdanosti stihova
segments loudness max	array float	počeci stihova
segments loudness max time	array float	niz pouzdanosti segmenata
segments loudness max start	array float	niz maksimalnih jačina unutar
segments pitches	2D array float	segmenata (u dB)
segments start	array float	niz vremena dostizanja maksimalne jačine
segments timbre	2D array float	unutar segmenata
similar artists	array string	niz jačina na počecima segmenata
song hotttnesss	float	niz jačina po segmentima, jedna
song id	string	vrednost za svaku notu
start of fade out	float	počeci segmenata
tatums confidence	array float	informacije o teksturi (MFCC + PCA)
tatums start	array float	niz Echo Nest sličnih izvođača
tempo	float	algoritamska aproksimacija
time signature	int	Echo Nest ID pesme
time signature confidence	float	vreme u odnosu na pocetak u kom počinje
title	string	fade-out efekat (u sekundama)
track id	string	pouzdanost najmanjih elemenata ritma
track 7digitalid	int	niz najmanjih elemenata ritma
year	int	procenjen tempo (u BPM)
		procenjen broj ritmova u taktu, npr. 4
		pouzdanost procene broja ritmova u taktu
		naziv pesme
		Echo Nest ID pesme
		ID 7digital ID pesme ili -1
		godina izdavanja uzeta sa MusicBrainz ili 0

Slika 1: Atributi prisutni u *The Million Song Dataset* skupu podataka