

Analiza “Million Songs” skupa podataka

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Milana Kovačević

Ivan Ristović

jun 2018.

Sažetak

U ovom radu su dati rezultati istraživanja skupa podataka *Million Songs Dataset*. Najpre je ukratko opisana struktura samog skupa, a zatim je i primer sloga koji opisuje jednu pesmu. Vizuelnim metodama se pruža bolje upoznavanje sa samim skupom. Za obradu podataka, korišćen je alat *Knime*, a alat *SPSS Modeler* je korišćen za njihovu vizuelizaciju. Detaljno je opisan postupak preprocesiranja i analize podataka. Informacije o podacima su dobijene korišćenjem različitih algoritama, a najbolje rezultate je dao algoritam *Apriori*.

Sadržaj

1	Opis skupa podataka	2
2	Korišćeni alati	2
3	Preprocesiranje podataka	2
4	Vizuelizacija	5
5	Pravila pridruživanja	9
5.1	Zavisnosti između žanrova	9
5.2	Zavisnost žanra od decenije	9
5.3	Zavisnost žanra od lokacije	9
6	Zaključak	9
	Literatura	10
A	Konverzija iz HDF5 u CSV format	11

1 Opis skupa podataka

The Million Song Dataset [7] 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 izvođaču, segmentima, tempu kao i ID pesme na raznim online servisima (*Echo Nest* [5], *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 [8]. Mi ćemo izdvojiti informacije iz datog modela i podatke organizovati u CSV format, zarad lakšeg ubacivanja u alate koje ćemo opisati kasnije. Ova transformacija je izvršena korišćenjem Python skripti iz MSongsDB repozitorijuma [2], modifikovanih za naše potrebe. Kompletne skripte se mogu naći u A.

2 Korišćeni alati

Za obradu podataka, korišćeni su alati *Knime Analytics Platform* [10] i *IBM SPSS Modeler* [9]. *IBM SPSS Modeler* je pretežno korišćen pretežno za vizuelizaciju, dok je *KNIME AP* korišćen za manipulisanje podacima, vizuelizaciju i primenu algoritama.

3 Preprocesiranje podataka

// TODO STATISTICS

Jedna od opštih transformacija je nad atributom koji sadrži informacije o žanru. Žanr je podatak koji je originalno dat kao niz niski. Medjutim, sadržaj ovog niza nije tačno definisan, već on ponekad u sebi sadrži čitavu rečenicu koja opisuje žanr. Jednostavna, a neophodna transformacija je bila da iz ovog niza izbacimo pojavljivanje reči *and*, čije je često pojavljivanje remetilo rezultate.

Za upotrebu različitih algoritama potrebne su drugačije transformacije polaznog skupa podataka.

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

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

```

1      analysis_sample_rate: 22050
2      artist_7digitalid: 61424
3      artist_familiarity: 0.5467275539627645
4      artist_hottnesss: 0.3861804160792181
5      artist_id: ARE26EG1187B990AEF
6      artist_latitude: 51.77045
7      artist_location: Essex, England
8      artist_longitude: 0.64255
9      artist_mbid: de212b3a-2f54-4def-a13d-5a877bfaef7
10     artist_mbtags: shape = (6,)
11     artist_mbtags_count: shape = (6,)
12     artist_name: Sunscreen
13     artist_playmeid: 19156
14     artist_terms: shape = (44,)
15     artist_terms_freq: shape = (44,)
16     artist_terms_weight: shape = (44,)
17     audio_md5: c2f7f92e66d18e86af3752478d3be966
18     bars_confidence: shape = (123,)
19     bars_start: shape = (123,)
20     beats_confidence: shape = (497,)
21     beats_start: shape = (497,)
22     danceability: 0.0
23     duration: 232.4371
24     end_of_fade_in: 0.0
25     energy: 0.0
26     key: 11
27     key_confidence: 0.625
28     loudness: -8.955
29     mode: 0
30     mode_confidence: 0.558
31     release: Looking At You: The Club Anthems
32     release_7digitalid: 196929
33     sections_confidence: shape = (6,)
34     sections_start: shape = (6,)
35     segments_confidence: shape = (1045,)
36     segments_loudness_max: shape = (1045,)
37     segments_loudness_max_time: shape = (1045,)
38     segments_loudness_start: shape = (1045,)
39     segments_pitches: shape = (1045, 12)
40     segments_start: shape = (1045,)
41     segments_timbre: shape = (1045, 12)
42     similar_artists: shape = (100,)
43     song_hottnesss: nan
44     song_id: SOICLQB12A8C13637C
45     start_of_fade_out: 232.437
46     tatums_confidence: shape = (993,)
47     tatums_start: shape = (993,)
48     tempo: 130.201
49     time_signature: 4
50     time_signature_confidence: 0.0
51     title: Exodus
52     track_7digitalid: 2140010
53     track_id: TRBBBLA128F424E963
54     year: 1995
55

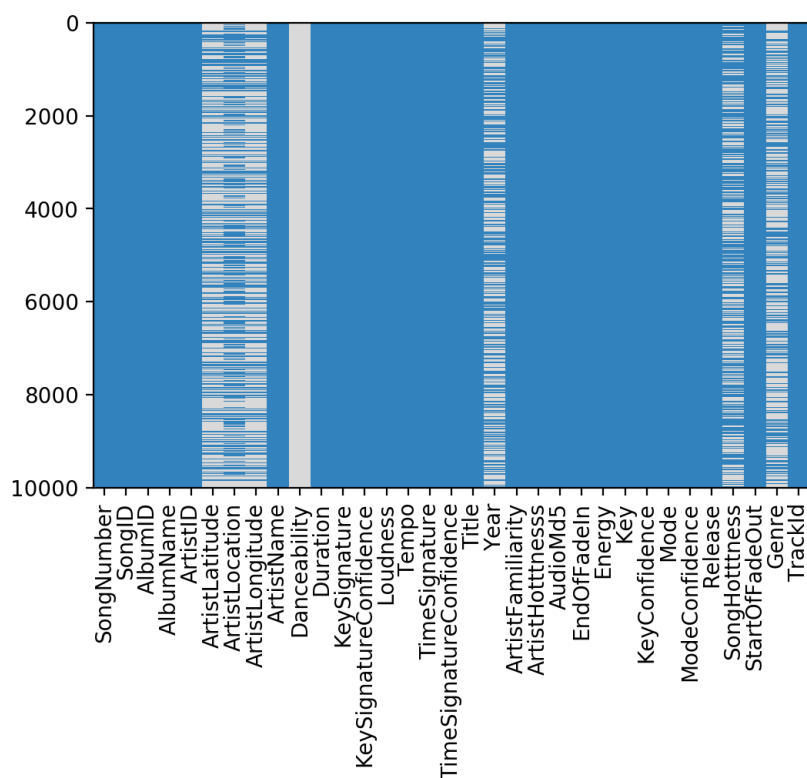
```

Slika 2: Primer jednog sloga

4 Vizuelizacija

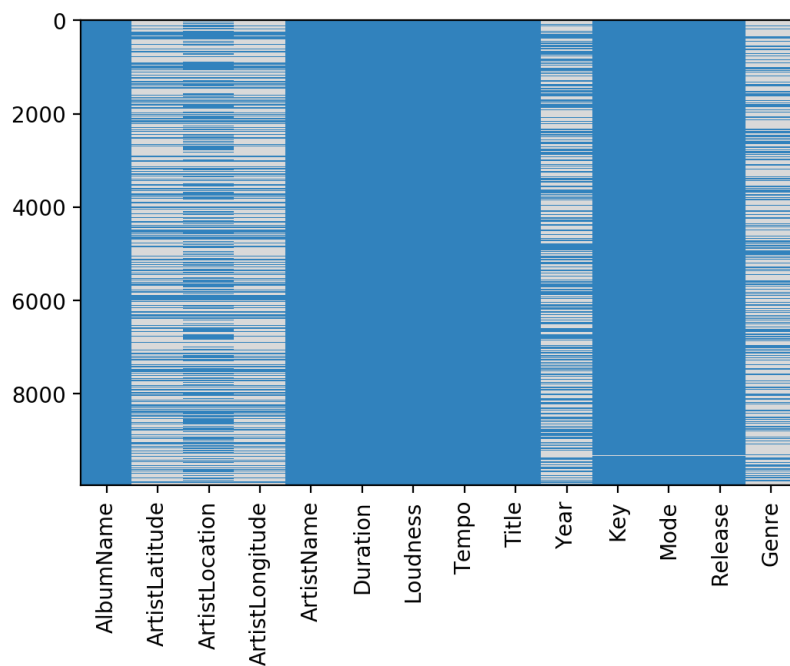
U ovom odeljku ćemo pokušati da čitaocu damo vizuelni prikaz raznovrsnosti skupa podataka, uz osvrt na neke zanimljive zaključke. Neki od atributa koji su vizuelizovani u ovom odeljku nisu u potpunosti prisutni u skupu, tako da je analiza takvih atributa radjena samo nad slogovima gde nema nedostajućih vrednosti za te attribute.

Vizuelizacija originalnog skupa podataka je prikazana na slici ???. Nedostajuće vrednosti su prikazane belom bojom, dok su plavom bojom predstavljene postojuće vrednosti atributa.



Slika 3: Originalni skup podataka

Sa prikaza skupa podataka se vidi da atribut koji opisuje plesnu moć pesme *eng. danceability*, ni u jednom slučaju nema postavljenu vrednost, te da je on neupotrebljiv. Takodje, godina, lokacija autora, geografska širina, geografska dužina i žanr su atributi koji u velikom broju slučajeva nemaju vrednost. Međutim, zbog njihove važnosti, mi ćemo svoje istraživanje vršiti nad onim slogovima za koje su ove vrednosti poznate. Nakon izdvajanja relevantnih atributa za naše istraživanje, pripremljen skup podataka je prikazan na slici 4.



Slika 4: Izdvojeni atributi korišćeni u istraživanju

// TODO outliers

Geografska rasprostranjenost autora čije su se pesme našle u skupu podataka se može videti na slici 5. Različite boje predstavljaju vizuelizaciju godine izdavanja pesme - gradijentni prelaz od plave (1950) do crvene (2010).

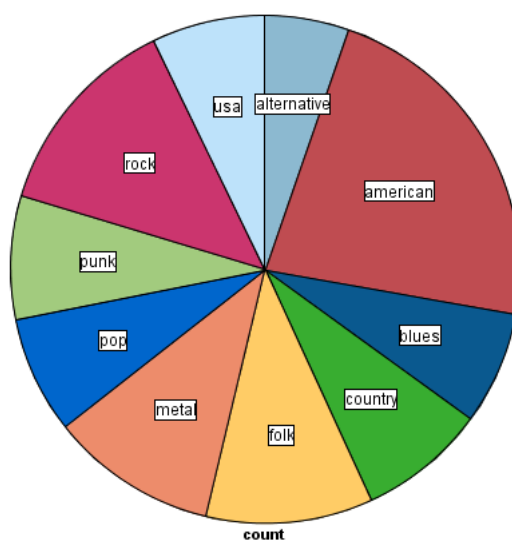


Slika 5: Geografska rasprostranjenost autora

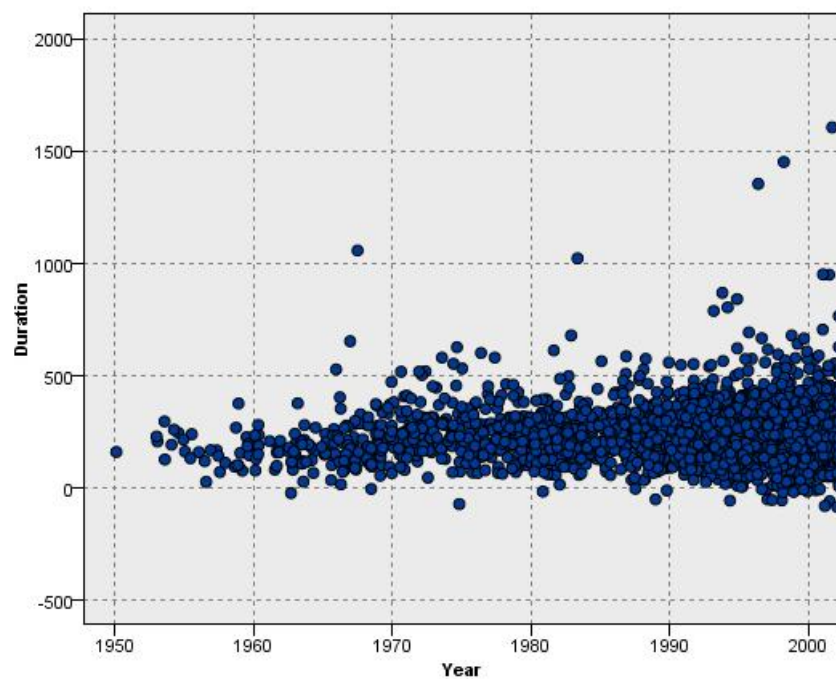
Spisak najzastupljenijih zanrova u skupu se može videti na slici 6. Žanr je atribut sa velikom stopom nedostajućih vrednosti, uz dodatni

problem rečenica prisutnih u nizu (podsećamo na problem naveden prilikom preprocesiranja, poglavlje 3), tako da je analiza vršena nad veoma ograničenim skupom od oko tri hiljade slogova. Smatramo, da su ovi rezultati u velikoj meri slični rezultatima koji bi se dobili da je potpuni skup analiziran - ukoliko ne bi bilo nedostajućih vrednosti za atribut žanr.

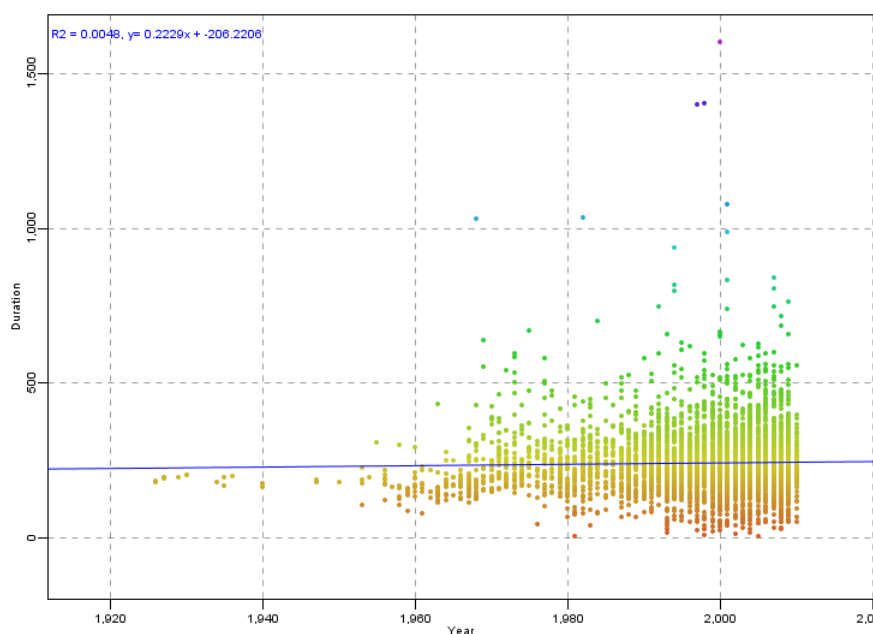
Grafik zavisnosti godine i trajanja pesme se može videti na slici 7. Jedan zanimljiv zaključak koji se nameće, je da se prosečno trajanje pesama povećava kroz vreme, sa razlikom od oko 20 sekundi u odnosu na 50-te godine prošlog veka. Detaljniji prikaz promene proseka trajanja se može videti na slici 8. Takodje, jasno je da se i raznovrsnost pesama mnogo veća danas - prisutne su i veoma kratke ali i veoma dugačke pesme.



Slika 6: Zastupljenost žanrova



Slika 7: Odnos godine izdavanja i dužine pesme



Slika 8: Promena proseka dužine pesama kroz vreme

5 Pravila pridruživanja

// TODO uvodni tekst

5.1 Zavisnosti izmedju žanrova

Na slici 9 prikazano je 10 najizraženijih pravila sortirana prvo po podršci, a zatim po lift meri, opadajuće.

Podrška	Pouzdanost	Lift	Glava pravila	Telo pravila
0.2646	0.7404	1.5567	rock	pop
0.2646	0.5563	1.5567	pop	rock
0.2248	0.9930	2.0878	rock	classic
0.2163	0.8176	3.6114	classic	pop, rock
0.2163	0.9623	2.6926	pop	rock, classic
0.2163	1.0000	2.1026	rock	pop, classic
0.1612	0.9967	2.0957	rock	indie
0.1405	0.9170	5.9225	british	uk
0.1405	0.9075	5.9225	uk	british
0.1373	0.5582	1.1736	rock	american, classic

Slika 9: Rezultati Apriori algoritma koji pokazuju zavisnost medju žanrovima

5.2 Zavisnost žanra od decenije

Na tabeli 10 prikazana su dobijena pravila sortirana prvo po podršci, a zatim po lift meri, opadajuće. Kako se kasnije vidi na slici 7, najveći deo skupa obradjenih podataka pripada dvehiljaditim godinama. Zbog ovoga se dobijeni rezultati koncentrišu na ovoj deceniji. Korišćenje većeg i raznovrnijeg skupa podataka bi doneo drugačije rezultate.

Podrška	Pouzdanost	Lift	Glava pravila	Telo pravila
0.0716	0.8133	1.6369	00s	pop, chart
0.0525	0.7857	1.5815	00s	rnb, hop, hall, dance, hip
0.0764	0.7385	1.4864	00s	dance
0.0615	0.6824	1.3734	00s	rnb
0.0615	0.6667	1.3419	00s	hop hip
0.0541	0.6000	1.2077	00s	metal
0.0530	0.5917	1.1910	00s	rock, alternative
0.0578	0.5619	1.1309	00s	alternative
0.0896	0.5541	1.1153	00s	indie
0.0891	0.5526	1.1123	00s	rock, indie

Slika 10: Rezultati Apriori algoritma koji pokazuju zavisnost izmedju žanrova i decenija tokom koje su bili popularni

5.3 Zavisnost žanra od lokacije

Lokacija autora je atribut čije ja vrednost često nepostojuća. A ukoliko jeste, iste lokacije su na različitim pesmama drugačije napisane. Na primer, jedna pesma ima lokaciju *London*, a druga *London, UK*. Ovakvi podaci su doveli do loših rezultata istraživanja.

6 Zaključak

Dobijeni rezultati se oslanjaju na podskup skupa podataka koji u sebi sadrži milion pesama. Korišćenje celog skupa podataka bi donelo još

pouzdanije rezultate. Međutim, i sam podskup podataka je bio dovoljan da se uoče prethodno navedeni zaključci.

Rezultati imaju praktičnu primenu. Uočavanje zavisnosti muzičkih žanrova i godina njihove popularnosti, ili njihove popularnosti na različitim podnevljima bi se moglo iskoristiti za automatsko generisanje lista pesama. Ovo bi predstavljao zanimljiv i koristan dodatak za muzički plejer.

Sajtovi kao što su YouTube [6] koriste prethodno prikupljene podatke o pesmama. Njih, u kombinaciji sa saznanjem o korisnikovim prethodno preslušanim pesmama, koristi kako bi korisniku pružio što bolje preporuke za sledeću pesmu, a samim tim ga i zadržao na sajtu. Implementacija sličnog, ali dosta jednostavnijeg, sistema bila bi moguća dobijenim rezultatima koji su predstavljeni u ovom radu.

Literatura

- [1] 7digital. on-line at:
<https://www.7digital.com/>.
- [2] Million Song Database GitHub repository. on-line at:
<https://github.com/tbertinmahieux/MSongsDB>.
- [3] MusicBrainz. on-line at:
<https://www.musicbrainz.org/>.
- [4] Play.me. on-line at:
<https://www.playme.com/>.
- [5] The Echo Nest. on-line at:
<http://the.echonest.com/>.
- [6] YouTube. on-line at:
<https://www.youtube.com/>.
- [7] 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/>.
- [8] The HDF Group. The HDF5 data model, 2011. on-line at: <https://support.hdfgroup.org/HDF5/>.
- [9] IBM. IBM SPSS Modeler. on-line at:
<https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software>.
- [10] KNIME AG. KNIME Analytics Platform. on-line at:
<https://www.knime.com/knime-analytics-platform>.

Dodatak A Konverzija iz HDF5 u CSV format

```
1  """
2  Alexis Greenstreet (October 4, 2015) University of Wisconsin-
   Madison
3  """
4  class Song:
5      songCount = 0
6
7      def __init__(self, songID):
8          self.id = songID
9          Song.songCount += 1
10         self.albumID = None           # string
11         self.albumName = None         # string
12         self.artistFamiliarity = None # float
13         self.artistHottnesss = None   # float
14         self.artistID = None          # string
15         self.artistLatitude = None    # float
16         self.artistLocation = None    # string
17         self.artistLongitude = None   # float
18         self.artistName = None        # string
19         self.audioMd5 = None          # string
20         self.danceability = None      # float
21         self.duration = None          # float
22         self.endOfFadeIn = None       # float
23         self.energy = None            # float
24         self.genre = None             # string
25         self.genreList = []           # list of
   strings
26         self.key = None               # int
27         self.keyConfidence = None     # float
28         self.keySignature = None      # float
29         self.keySignatureConfidence = None # float
30         self.loudness = None          # float
31         self.mode = None              # int
32         self.modeConfidence = None    # float
33         self.release = None           # string
34         self.songHottness = None      # float
35         self.songId = None            # string
36         self.startOfFadeOut = None    # float
37         self.tempo = None             # float
38         self.timeSignature = None     # int
39         self.timeSignatureConfidence = None # float
40         self.title = None             # string
41         self.trackId = None          # string
42         self.year = None              # int
```

Slika 11: Klasa korišćena za deserijalizaciju podataka.

```

1  """
2  Alexis Greenstreet (October 4, 2015) University of Wisconsin-
   Madison
3  """
4  outputFile1 = open('SongCSV.csv', 'w')
5  csvRowString = ""
6
7  csvRowString = ("SongID,AlbumID, ...")
8      csvAttributeList = re.split('\W+', csvRowString)
9  for i, v in enumerate(csvAttributeList):
10      csvAttributeList[i] = csvAttributeList[i].lower()
11  outputFile1.write("SongNumber,");
12  outputFile1.write(csvRowString + "\n");
13  csvRowString = ""
14
15  #####
16  #Set the basedir here, the root directory from which the
   search
17  basedir = "/home/m/Documents/MillionSongSubset/data"
18  ext = ".h5"
19  #####
20
21  csvRowStringTotal = ""
22
23  for root, dirs, files in os.walk(basedir):
24      files = glob.glob(os.path.join(root, '*' + ext))
25      for f in files:
26          print f
27
28          songH5File = hdf5_getters.open_h5_file_read(f)
29          song = Song(str(hdf5_getters.get_song_id(songH5File)))
30
31          song.artistID = str(hdf5_getters.get_artist_id(
songH5File))
32          # Isto za ostala polja
33
34          artistMbtags = np.array(hdf5_getters.get_artist_mbtags
(songH5File))
35          song.genre = ' | '.join(artistMbtags)
36
37          csvRowString += str(song.songCount) + ","
38          csvRowString += song.id + ","
39          # Isto za ostala polja
40
41          csvRowString += song.trackId + ","
42          csvRowString += song.genre + "\n"
43          csvRowStringTotal += csvRowString
44          csvRowString = ""
45
46          songH5File.close()
47
48  outputFile1.write(csvRowStringTotal)
49  outputFile1.close()

```

Slika 12: Uprošćena verzija programa korišćenog za konvertovanje iz HDF5 u CSV format.