Analiziranje podskupa Million Songs Dataset-a

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Milana Kovačevoć
Ivan Ristović

jun 2018.

Sažetak

U ovom radu ćemo istraživati skup podataka *Million Songs Dataset*. Milana molim te pomozi Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sadržaj

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 izvodjaču, segmentima, tempu kao i ID pesme na raznim online servisima (*Echo Nest* [2], *7digital* [1], *MusicBrainz* [4] i *PlayMe* [5]). 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 [3], modifikovanih za naše potrebe. Kompletne skripte se mogu naći u A.

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	nig 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: Atributi prisutni u *The Million Song Dataset* skupu podataka

```
analysis_sample_rate: 22050
artist_7digitalid: 61424
artist_familiarity: 0.5467275539627645
artist_hotttnesss: 0.3861804160792181
artist_id: ARE26EG1187B990AEF
artist_latitude: 51.77045
artist_location: Essex, England
artist_longitude: 0.64255
artist_mbid: de212b3a-2f54-4def-a13d-5a877bfaeaf7
artist_mbtags: shape = (6,)
artist_mbtags_count: shape = (6,)
artist_name: Sunscreem
artist_playmeid: 19156
artist_terms: shape = (44,)
artist_terms_freq: shape = (44,)
artist_terms_weight: shape = (44,)
audio_md5: c2f7f92e66d18e86af3752478d3be966
bars_confidence: shape = (123,)
bars_start: shape = (123,)
beats_confidence: shape = (497,)
beats_start: shape = (497,)
danceability: 0.0
duration: 232.4371
end_of_fade_in: 0.0
energy: 0.0
key: 11
key_confidence: 0.625
loudness: -8.955
mode: 0
mode_confidence: 0.558
release: Looking At You: The Club Anthems
release_7digitalid: 196929
sections_confidence: shape = (6,)
sections_start: shape = (6,)
segments_confidence: shape = (1045,)
segments_loudness_max: shape = (1045,)
segments_loudness_max_time: shape = (1045,)
segments_loudness_start: shape = (1045,)
segments_pitches: shape = (1045, 12)
segments_start: shape = (1045,)
segments_timbre: shape = (1045, 12)
similar_artists: shape = (100,)
song_hotttnesss: nan
song_id: SOICLQB12A8C13637C
start_of_fade_out: 232.437
tatums_confidence: shape = (993,)
tatums_start: shape = (993,)
tempo: 130.201
time_signature: 4
time_signature_confidence: 0.0
title: Exodus
track_7digitalid: 2140010
track_id: TRBBBLA128F424E963
year: 1995
```

Slika 2: Primer sloga iz skupa podataka.

2 Korišćeni alati

Za obradu podataka, korišćeni su alati *Knime Analytics Platform* [6] i *IBM SPSS Modeler* [9]. *IBM SPSS Modeler* je pretežno koriš´¸en za vizuelizaciju, dok je *KNIME AP* koriś´¸en za manipulisanje podacima, vizuelizaciju i primenu algoritama.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

3 Preprocesiranje i analiza podataka

Za upotrebu različitih algoritama potrebne su drugačije transformacije.

3.1 Apriori algoritam

- *Zavisnost žanra od lokacije*
- *Zavisnost žanra od decenije*
- *Zavisnost žanra od lokacije*

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

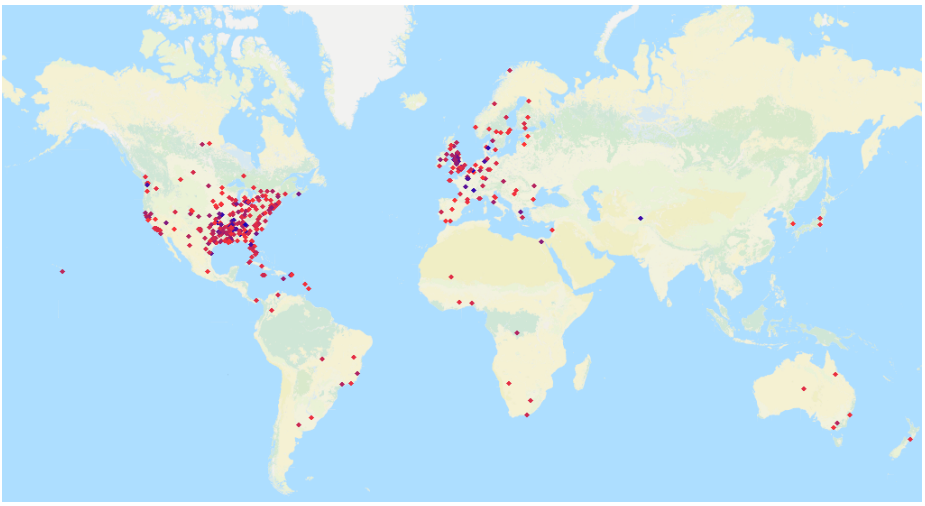
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4 Vizuelizacija

Kako bi se razumeli rezultati istraživanja, izvršena je njihova vizuelizacija. Na slici 4 prikazana je zastupljenost autora na različitim lokacijama. Ova informacija je korišćena u odeljku **??**.

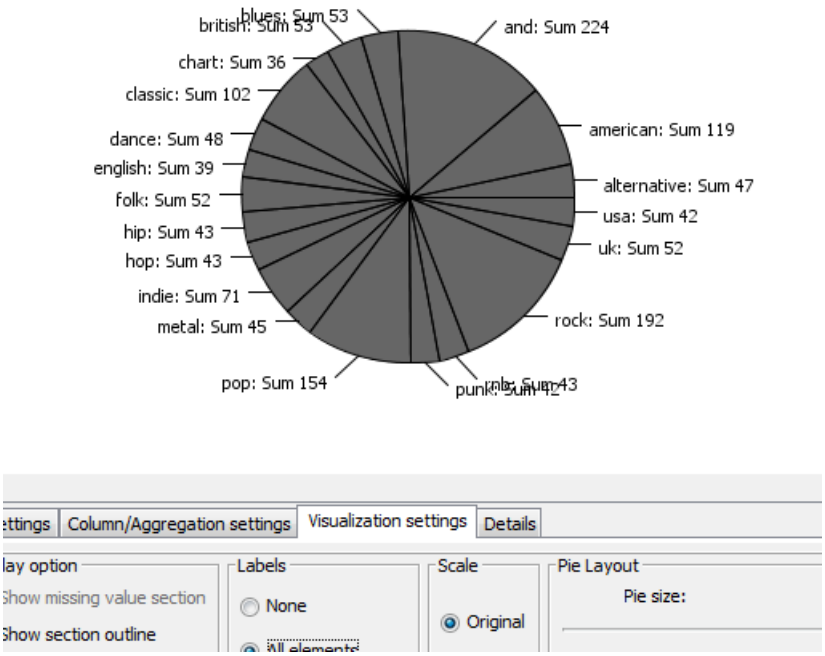
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



Slika 3: Geografksa rasprostranjenost autora uz osvrt na godine - gradijentni prelaz od plave (1950) do crvene (2010)

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

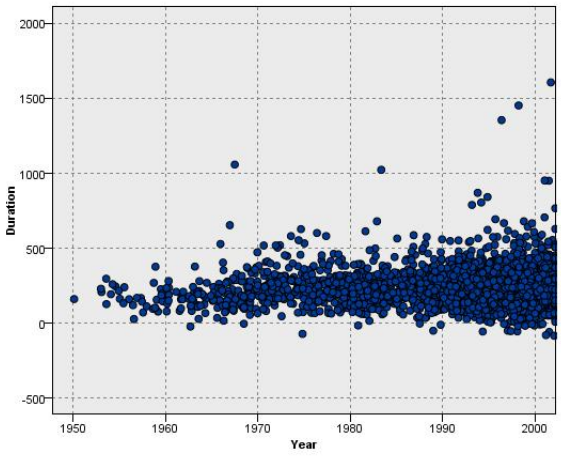
Zastupljenost žanrova je prikazana na slici 4.



Slika 4: Zastupljenost žanrova

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Zavisnost dužine pesama u odnosu na godinu nastanka prikazana je na slici 4.



Slika 5: Odnos godine i dužine pesama

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5 Zaključak

Zakljucak... TODO

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Literatura

[1] 7digital. on-line at https://www.7digital.com/.

[2] The echo nest. on-line at http://the.echonest.com/.

[3] Million song database git repository. on-line at https://github.com/tbertinmahieux/MSongsDB.

[4] Musicbrainz. on-line at https://www.musicbrainz.org/.

[5] Play.me. on-line at https://www.playme.com/.

[6] KNIME AG. Knime analytics platform. on-line at https://www.knime.com/knime-analytics-platform.

[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.

Dodatak A Konverzija iz HDF5 u CSV format

```
"""
Alexis Greenstreet (October 4, 2015) University of Wisconsin-
    Madison
"""
class Song:
    songCount = 0

    def __init__(self, songID):
        self.id = songID
        Song.songCount += 1
        self.albumName = None
        self.albumID = None
        self.artistID = None
        self.artistLatitude = None
        self.artistLocation = None
        self.artistLongitude = None
        self.artistName = None
        self.danceability = None
        self.duration = None
        self.genreList = []
        self.keySignature = None
        self.keySignatureConfidence = None
        self.lyrics = None
        self.loudness = None
        self.popularity = None
        self.tempo = None
        self.timeSignature = None
        self.timeSignatureConfidence = None
        self.title = None
        self.year = None
        self.artistFamiliarity = None #float
        self.artistHottnesss = None #float
        self.audioMd5 = None # string
        self.endOfFadeIn = None #float
        self.energy = None #float
        self.key = None #int
        self.keyConfidence = None #float
        self.mode = None # int
        self.modeConfidence = None #float
        self.release = None #string
        self.songHotttness = None #float
        self.songId = None #string
        self.startOfFadeOut = None #float
        self.trackId = None #string
        self.genre = None # list of strings
```

Slika 6: Klasa korišć]ena za deserijalizaciju podataka.

```
"""
Alexis Greenstreet (October 4, 2015) University of Wisconsin-
    Madison
"""
outputFile1 = open('SongCSV.csv', 'w')
csvRowString = ""

csvRowString = ("SongID,AlbumID, ...")
    csvAttributeList = re.split('\W+', csvRowString)
for i, v in enumerate(csvAttributeList):
    csvAttributeList[i] = csvAttributeList[i].lower()
outputFile1.write("SongNumber,");
outputFile1.write(csvRowString + "\n");
csvRowString = ""

####################################################
#Set the basedir here, the root directory from which the
    search
basedir = "/home/m/Documents/MillionSongSubset/data"
ext = ".h5"
####################################################

csvRowStringTotal = ""

for root, dirs, files in os.walk(basedir):
    files = glob.glob(os.path.join(root,'*'+ext))
    for f in files:
        print f

        songH5File = hdf5_getters.open_h5_file_read(f)
        song = Song(str(hdf5_getters.get_song_id(songH5File)))

        song.artistID = str(hdf5_getters.get_artist_id(
    songH5File))
        # Isto za ostala polja

        artistMbtags = np.array(hdf5_getters.get_artist_mbtags
    (songH5File))
        song.genre = ' | '.join(artistMbtags)

        csvRowString += str(song.songCount) + ","
        csvRowString += song.id + ","
        # Isto za ostala polja

        csvRowString += song.trackId + ","
        csvRowString += song.genre + "\n"
        csvRowStringTotal += csvRowString
        csvRowString = ""

        songH5File.close()

outputFile1.write(csvRowStringTotal)
outputFile1.close()
```

Slika 7: Uprošćena verzija programa korišćenog za konvertovanje iz HDF5 u CSV format.