Challenge Set

Version 1, February 13, 2018 (Documentation updated August 5, 2020)

This is the challenge set for the Spotify Million Playlist Dataset Challenge.

This challenge set contains 10,000 incomplete playlists. The challenge is to recommend tracks for each of these playlists. See https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge for challenge details.

Format

The challenge set consists of a single JSON dictionary with three fields:

- **date** - the date the challenge set was generated. This should be "2018-01-16 08:47:28.198015"
- **version** - the version of the challenge set. This should be "v1"
- **playlists** - an array of 10,000 incomplete playlists. Each element in this array contains the following fields:
    - **pid** - the playlist ID
    - **name** - (optional) - the name of the playlist. For some challenge playlists, the name will be missing.
    - **num_holdouts** - the number of tracks that have been omitted from the playlist
    - **tracks** - a (possibly empty) array of tracks that are in the playlist. Each element of this array contains the following fields:
        - **pos** - the position of the track in the playlist (zero offset)
        - **track_name** - the name of the track
        - **track_uri** - the Spotify URI of the track
        - **artist_name** - the name of the primary artist of the track
        - **artist_uri** - the Spotify URI of the primary artist of the track
        - **album_name** - the name of the album that the track is on
        - **album_uri** -- the Spotify URI of the album that the track is on
        - **duration_ms** - the duration of the track in milliseconds
    - **num_samples** the number of tracks included in the playlist
    - **num_tracks** - the total number of tracks in the playlist.

    Note that len(tracks) == num_samples and num_samples + num_holdouts == num_tracks

Playlist Challenge Categories

The 10,000 playlists are made up of 10 different challenge categories, with 1,000 playlists in each category:

1. Predict tracks for a playlist given its title only
2. Predict tracks for a playlist given its title and the first track
3. Predict tracks for a playlist given its title and the first 5 tracks
4. Predict tracks for a playlist given its first 5 tracks (no title)
5. Predict tracks for a playlist given its title and the first 10 tracks
6. Predict tracks for a playlist given its first ten tracks (no title)

7. Predict tracks for a playlist given its title and the first 25 tracks
8. Predict tracks for a playlist given its title and 25 random tracks
9. Predict tracks for a playlist given its title and the first 100 tracks
10. Predict tracks for a playlist given its title and 100 random tracks

How the challenge set was built

The playlists in the challenge set are selected using the same criteria used to select playlists for the full Million Playlist Dataset (MPD). See the README.md file in the MPD distribution for more details on how the playlists were selected. Additionally, playlists in the challenge set meet the following constraints:

- All tracks in the challenge set appear in the MPD
- All holdout tracks appear in the MPD

Tools

Scripts for checking and verifying submissions

- check.py - checks the challenge set to make sure that it is internally consistent and properly formatted.
- verify_submission.py - verifies that a given challenge submission is properly formatted

Verifying the challenge set

To verify that you have an uncorrupted challenge set you can check its md5. E.g.

```
  % md5sum --check md5

  challenge_set.json: OK
```

Use check.py to verify that the challenge set is internally consistent.

```
% python check.py

stats:
   tests: 4634003
   errors: 0

  challenge_set.json is OK
```

Sample Submission

Included in the challenge set is a sample challenge submission:

```
 sample_submission.csv
```

This sample shows the expected format for your submission to the challenge. Your submssion should follow the following rules:

- All fields are comma separated. It is ok, but optional to have whitespace before and after the comma.
- Comments are allowed with a '#' at the beginning of a line.

- Empty lines are ok (they are ignored).
- The first non-commented/blank line must start with "team_info" and then include the team name, and contact email address.
- For each challenge playlist there must be a line of the form: pid, trackuri_1, trackuri_2, trackuri_3, ..., trackuri_499, trackuri_500 with exactly 500 tracks.
- The seed tracks, provided as part of the challenge set for any particular playlist, must *not* be included in the submission for that playlist.
- The submission for a particular playlist must *not* contain duplicated tracks.
- The submission for a particular playlist must have exactly 500 tracks.
- Any submission violating one of the rules will be rejected by the scoring system.

'pid' is the playlist id of the challenge playlist

Before submission, the csv should be gzipped.

You can verify that your submission is in the proper format as follows:

```
python verify_submission.py challenge_set.json sample_submission.csv
```