🕵️ Summary

The Spotify Million Playlist Dataset Challenge consists of a dataset and evaluation to enable research in music recommendations. It is a continuation of the RecSys Challenge 2018, which ran from January to July 2018. The dataset contains 1,000,000 playlists, including playlist titles and track titles, created by users on the Spotify platform between January 2010 and October 2017. The evaluation task is automatic playlist continuation: given a seed playlist title and/or initial set of tracks in a playlist, to predict the subsequent tracks in that playlist. This is an open-ended challenge intended to encourage research in music recommendations, and no prizes will be awarded (other than bragging rights).

Please note: The dataset associated with this challenge is not available for download anymore. We request you to directly reach out to Spotify Research for access to this dataset.

🎵 Background

Here at Spotify, we love playlists. Playlists like Today's Top Hits and RapCaviar have millions of loyal followers, while Discover Weekly and Daily Mix are just a couple of our personalized playlists made especially to match your unique musical tastes.

Our users love playlists too. In fact, the Digital Music Alliance, in their 2018 Annual Music Report, state that 54% of consumers say that playlists are replacing albums in their listening habits.

But our users don't love just listening to playlists, they also love creating them. To date, over 4 billion playlists have been created and shared by Spotify users. People create playlists for all sorts of reasons: some playlists group together music categorically (e.g., by genre, artist, year, or city), by mood, theme, or occasion (e.g., romantic, sad, holiday), or for a particular purpose (e.g., focus, workout). Some playlists are even made to land a dream job, or to send a message to someone special.

The other thing we love here at Spotify is playlist research. By learning from the playlists that people create, we can learn all sorts of things about the deep relationship between people and music. Why do certain songs go together? What is the difference between “Beach Vibes” and “Forest Vibes”? And what words do people use to describe which playlists?

By learning more about nature of playlists, we may also be able to suggest other tracks that a listener would enjoy in the context of a given playlist. This can make playlist creation easier, and ultimately help people find more of the music they love.

💾 Dataset

To enable this type of research at scale, in 2018 we sponsored the RecSys Challenge 2018, which introduced the Million Playlist Dataset (MPD) to the research community. Sampled from the over 4 billion public playlists on Spotify, this dataset of 1 million playlists consist of over 2 million unique tracks by nearly 300,000 artists, and represents the largest public dataset of music playlists in the world. The dataset includes public playlists created by US Spotify users between January 2010 and November 2017. The challenge ran from January to July 2018, and received 1,467 submissions from 410 teams. A summary of the challenge and the top scoring submissions was published in the ACM Transactions on Intelligent Systems and Technology.

In September 2020, we re-released the dataset as an open-ended challenge on AIcrowd.com. The dataset can now be downloaded by registered participants from the Resources page.

Each playlist in the MPD contains a playlist title, the track list (including track IDs and metadata), and other metadata fields (last edit time, number of playlist edits, and more). All data is anonymized to protect user privacy. Playlists are sampled with some randomization, are manually filtered for playlist quality and to remove offensive content, and have some dithering and fictitious tracks added to them. As such, the dataset is not representative of the true distribution of playlists on the Spotify platform, and must not be interpreted as such in any research or analysis performed on the dataset.

Here’s an example of a typical playlist entry:

```
{
        "name": "musical",
        "collaborative": "false",
        "pid": 5,
        "modified_at": 1493424000,
        "num_albums": 7,
        "num_tracks": 12,
        "num_followers": 1,
        "num_edits": 2,
        "duration_ms": 2657366,
        "num_artists": 6,
        "tracks": [
            {
                "pos": 0,
                "artist_name": "Degiheugi",
                "track_uri":
"spotify:track:7vqa3sDmtEaVJ2gcvxtRID",
                "artist_uri":
"spotify:artist:3V2paBXEoZIAhfZRJmo2jL",
                "track_name": "Finalement",
                "album_uri":
"spotify:album:2KrRMJ9z7Xjoz1Az4O6UML",
                "duration_ms": 166264,
                "album_name": "Dancing Chords and Fireflies"
            },
            {
                "pos": 1,
                "artist_name": "Degiheugi",
                "track_uri":
"spotify:track:23EOmJivOZ88WJPUbIPjh6",
                "artist_uri":
"spotify:artist:3V2paBXEoZIAhfZRJmo2jL",
                "track_name": "Betty",
                "album_uri":
"spotify:album:3lUSlvjUoHNA8IkNTqURqd",
                "duration_ms": 235534,
                "album_name": "Endless Smile"
            },
            {
                "pos": 2,
                "artist_name": "Degiheugi",
                "track_uri":
"spotify:track:1vaffTCJxkyqeJY7zF9a55",
                "artist_uri":
"spotify:artist:3V2paBXEoZIAhfZRJmo2jL",
                "track_name": "Some Beat in My Head",
                "album_uri":
"spotify:album:2KrRMJ9z7Xjoz1Az4O6UML",
                "duration_ms": 268050,
                "album_name": "Dancing Chords and Fireflies"
            },
            // 8 tracks omitted
            {
                "pos": 11,
                "artist_name": "Mo' Horizons",
                "track_uri":
"spotify:track:7iwx00eBzeSSSy6xfESyWN",
                "artist_uri":
"spotify:artist:3tuX54dqgS8LsGUvNzgrpP",
                "track_name": "Fever 99\u00b0",
                "album_uri":
"spotify:album:2Fg1t2tyOSGWkVYHlFfXVf",
                "duration_ms": 364320,
                "album_name": "Come Touch The Sun"
            }
        ],

    }
```

More details on how the data is stored in files, and on the individual metadata fields can be found in the README file included in the dataset distribution.

🎧 Task

The goal of the challenge is to develop a system for the task of automatic playlist continuation. Given a set of playlist features, participants' systems shall generate a list of recommended tracks that can be added to that playlist, thereby "continuing" the playlist. We define the task formally as follows:

Input

- A user-created playlist, represented by:
    - Playlist metadata (see the dataset README)
    - K seed tracks: a list of K tracks in the playlist, where K can equal 0, 1, 5, 10, 25, or 100.

Output

- A list of 500 recommended candidate tracks, ordered by relevance in decreasing order.

Note that the system should also be able to cope with playlists for which no initial seed tracks are given! To assess the performance of a submission, the output track predictions are compared to the ground truth tracks ("reference set") from the original playlist.

🖊️ Evaluation

Submissions will be evaluated using the following metrics. All metrics will be evaluated at both the track level (exact track match) and the artist level (any track by the same artist is a match).

In the following, we denote the ground truth set of tracks by and the ordered list of recommended tracks by . The size of a set or list is denoted by , and we use from:to-subscripts to index a list. In the case of ties on individual metrics, earlier submissions are ranked higher.

R-precision

R-precision is the number of retrieved relevant tracks divided by the number of known relevant tracks (i.e., the number of withheld tracks):

The metric is averaged across all playlists in the challenge set. This metric rewards total number of retrieved relevant tracks (regardless of order).

Normalized Discounted Cumulative Gain (NDCG)

Discounted Cumulative Gain (DCG) measures the ranking quality of the recommended tracks, increasing when relevant tracks are placed higher in the list. Normalized DCG (NDCG) is determined by calculating the DCG and dividing it by the ideal DCG in which the recommended tracks are perfectly ranked:

The ideal DCG or IDCG is, in our case, equal to:

If the size of the set intersection of and is empty, then the IDCG is equal to 0. The NDCG metric is now calculated as:

Recommended SongS Clicks

Recommended Songs is a Spotify feature that, given a set of tracks in a playlist, recommends 10 tracks to add to the playlist. The list can be refreshed to produce 10 more tracks.

Recommended Songs clicks is the number of refreshes needed before a relevant track is encountered. It is calculated as follows:

If the metric does not exist (i.e. if there are no relevant tracks in , a value of 51 is picked (which is 1 greater than the maximum number of clicks possible).

Rank Aggregation

Final rankings will be computed by using the Borda Count election strategy. For each of the rankings of *p* participants according to R-precision, NDCG, and Recommended Songs Clicks, the top ranked system receives *p* points, the second system received *p*-1 points, and so on. The participant with the most total points wins. In the case of ties, we use top-down comparison: compare the number of 1st place positions between the systems, then 2nd place positions, and so on.

💾 Challenge Dataset

As part of the challenge, we release a separate challenge dataset ("test set") that consists of 10,000 playlists with incomplete information. It has many of the same data fields and follows the same structure as the Million Playlist Dataset ("training set"), but the playlists may include incomplete metadata (no title), and only include *K* tracks. More specifically, the challenge dataset is divided into 10 scenarios, with 1000 examples of each scenario:

Title only (no tracks)
Title and first track
Title and first 5 tracks
First 5 tracks only
Title and first 10 tracks
First 10 tracks only
Title and first 25 tracks
Title and 25 random tracks
Title and first 100 tracks
Title and 100 random tracks

🚀 Submission Format

For each playlist in the challenge set, participants will submit a ranked list of 500 recommended track URIs. The file format should be a gzipped csv (.csv.gz) file. The order of the recommended tracks matters: more relevant recommendations should appear first in the list. Submissions should be made in the following comma-separated format:

- 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 a contact email address. (Note: If you previously participated in the RecSys Challenge 2018, there was an additional field specifying "main" or "creative" track. Since this challenge only has one track, that field has been removed from the first line.) Example:

  team_info, my awesome team name, my_awesome_team@email.com
- 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.

Important note about submissions:

- The seed tracks, provided as part of the challenge set, must not be included in the submission.
- The submission for any particular playlist must not contain duplicated tracks.
- A submission must contain exactly 500 tracks (post deduplication).
- Any submission violating one of the formatting rules will be rejected by the scoring system.

A sample submission (sample_submission.csv) is included with the challenge set. The sample shows the expected format for your submission to the challenge. Also included with the challenge set is a Python script called *verify_submission.py*. You can use this program to verify that your submission is properly formatted. See the challenge set README file for more information on how to verify and submit your challenge results.

📜 Rules

The dataset and challenge are available strictly for research and non-commercial use. You may not redistribute or make available any part or whole of this dataset. You may not use the dataset or challenge to reverse engineer any aspect of Spotify's technology, or intellectual property, nor attempt to identify any individuals from the data. As mentioned above, the dataset has been non-uniformly sampled, and is not representative of the true distribution of playlists on the Spotify platform, and must not be interpreted as such in any research or analysis performed on the dataset. Please read the full Terms and Conditions at https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge/challenge_rules carefully before participating in this challenge.

🧐 Citation

To use the Spotify Million Playlist Dataset and/or your challenge results in research publications, please cite the following paper:

C.W. Chen, P. Lamere, M. Schedl, and H. Zamani. Recsys Challenge 2018: Automatic Music Playlist Continuation. In Proceedings of the 12th ACM Conference on Recommender Systems (RecSys ’18), 2018.

References

For a summary of the submissions from the 2018 RecSys Challenge, read "An Analysis of Approaches Taken in the ACM RecSys Challenge 2018 for Automatic Music Playlist Continuation" by H. Zamani, M. Schedl, P. Lamere, C.W. Chen.

Details on each of the top submissions, including papers, slides, and code, can be found on the RecSys Challenge 2018 website, and in the Proceedings of the ACM Recommender Systems Challenge 2018.

📚 Acknowledgments

The Million Playlist Dataset was developed by the following researchers at Spotify:

- Cedric De Boom
- Paul Lamere
- Ching-Wei Chen
- Ben Carterette
- Christophe Charbuillet
- Jean Garcia-Gathright
- James Kirk
- James McInerney
- Vidhya Murali
- Hugh Rawlinson
- Sravana Reddy
- Marc Romejin
- Romain Yon
- Yu Zhao

The RecSys Challenge 2018 was organized by:

- Ching-Wei Chen, Spotify, New York, USA
- Markus Schedl, Johannes Kepler University, Linz, Austria
- Hamed Zamani, University of Massachusetts Amherst, MA, USA
- Paul Lamere, Spotify, USA

📱 Contact

For any queries, please create a post on Discourse or contact:

- Yoogottam Khandelwal
- Shivam Khandelwal
- Sharada Mohanty