Second meeting

24/10/2024


PLAYLIST
- Title
- Title song 1
- Title song 2
...
CORPUS

- Should we use Json file or CSV file ?

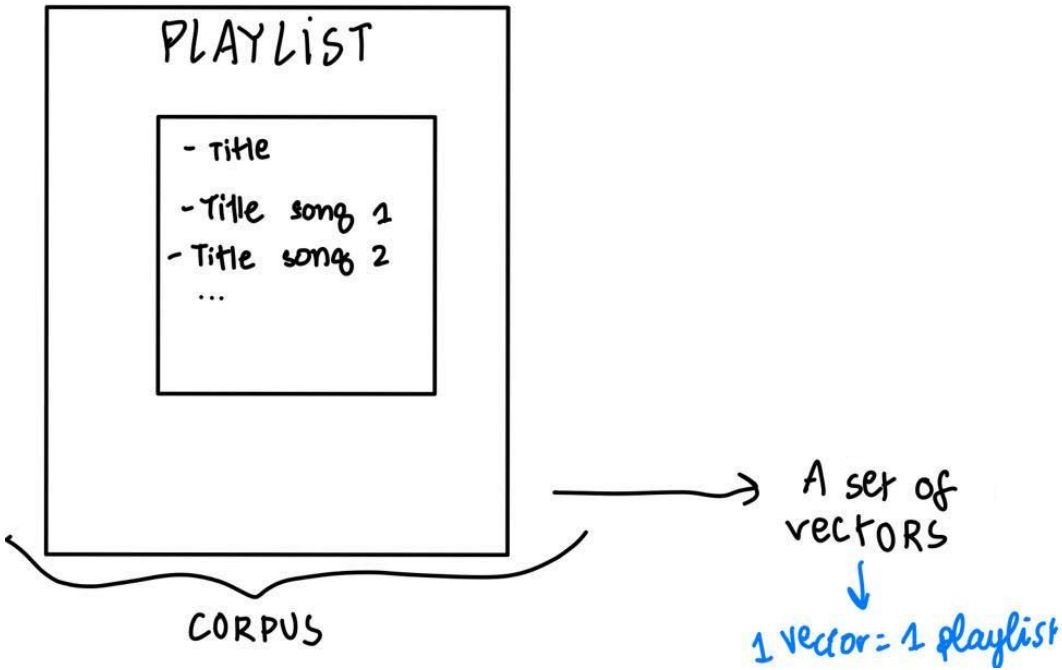
Dataset

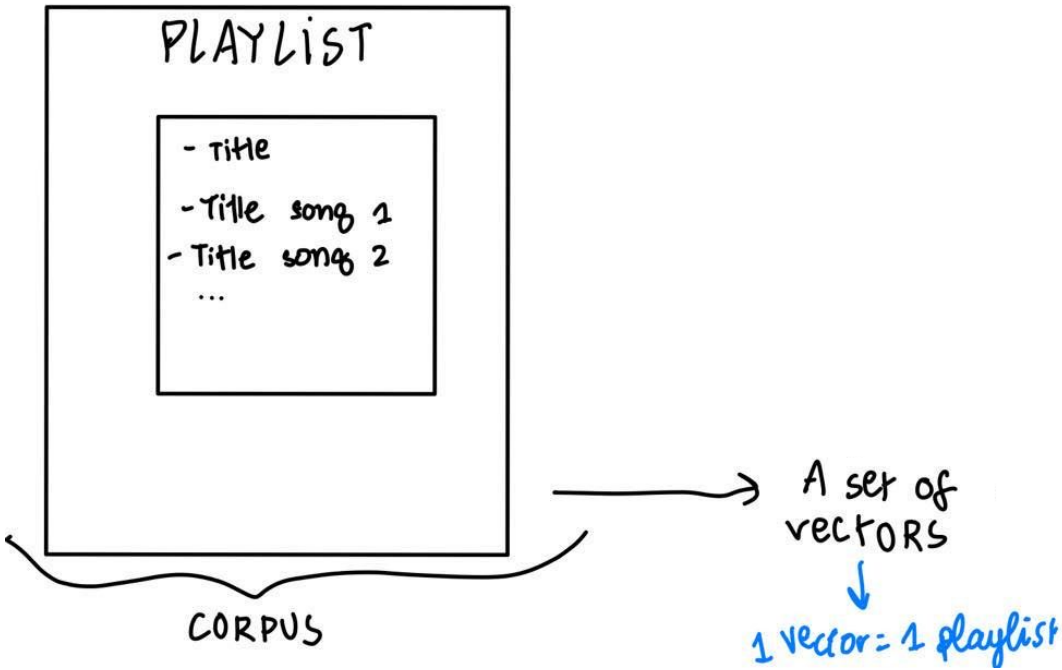
We converted the original JSON files in an equivalent CSV version.

```
python evaluation/mpd2csv.py --mpd_path /path/to/mpd --out_path dataset
python evaluation/challenge2csv.py --challenge_path /path/to/challenge.json --out_path dataset
```

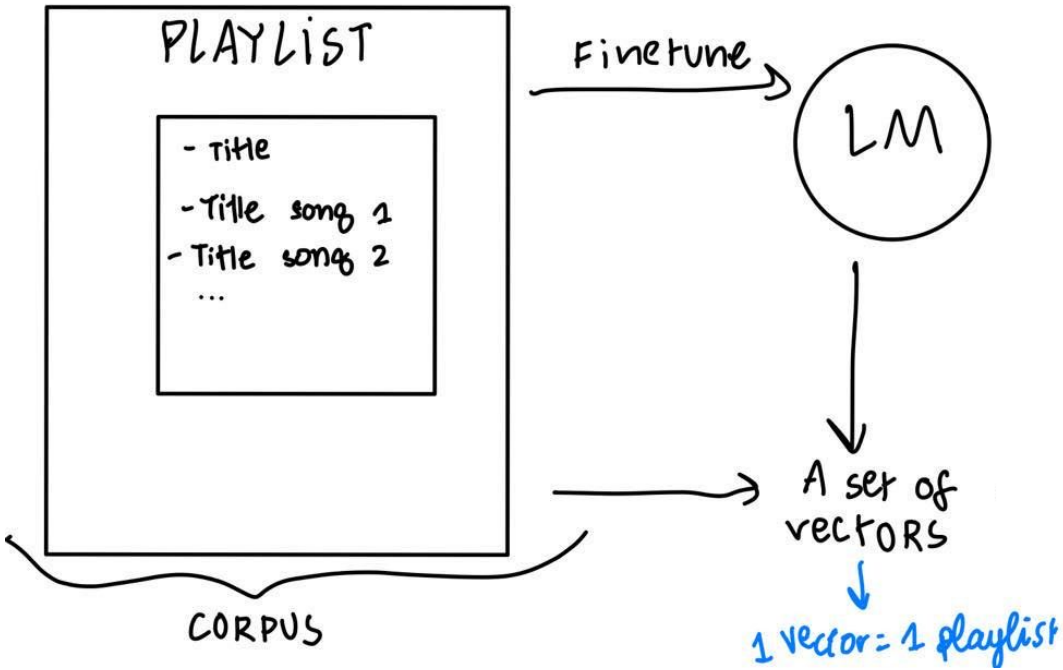
- Cleaning data
  - lowercasing;
  - detecting and separating emoji from words;
  - separating the skin code from the emoji;
  - detecting and separating emoticons from words;
  - transforming space-separated single letters into words (e.g. “*w o r k o u t*” becomes “*workout*”);
  - remove ‘#’ from hashtags.

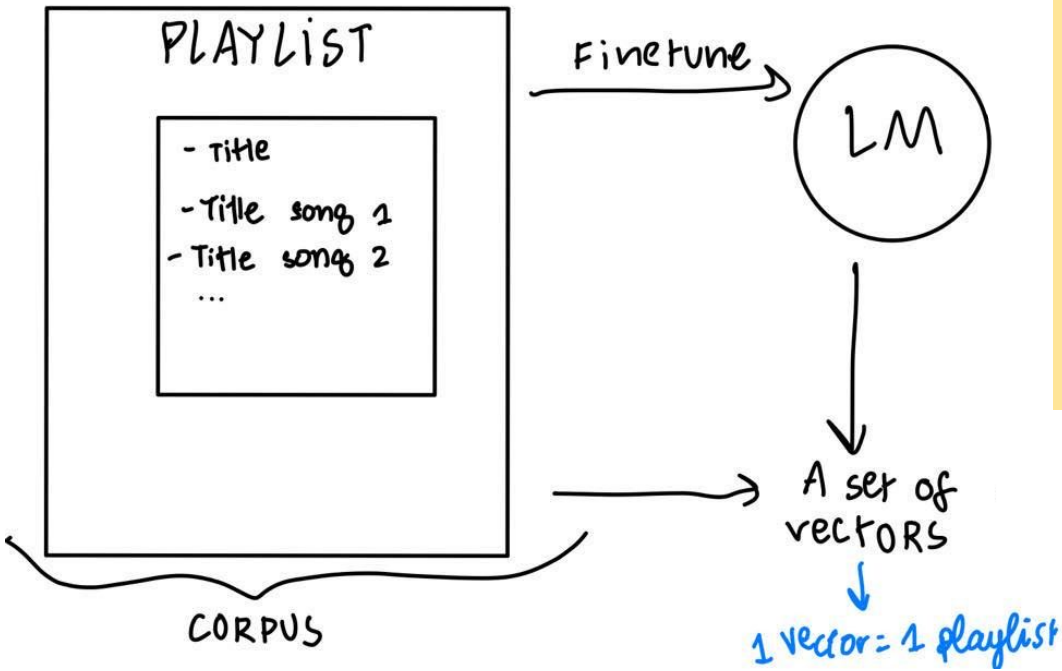
using “BERT un_cased “ ???


PLAYLIST
- Title
- Title song 1
- Title song 2
...
CORPUS
A set of vectors
1 vector = 1 playlist

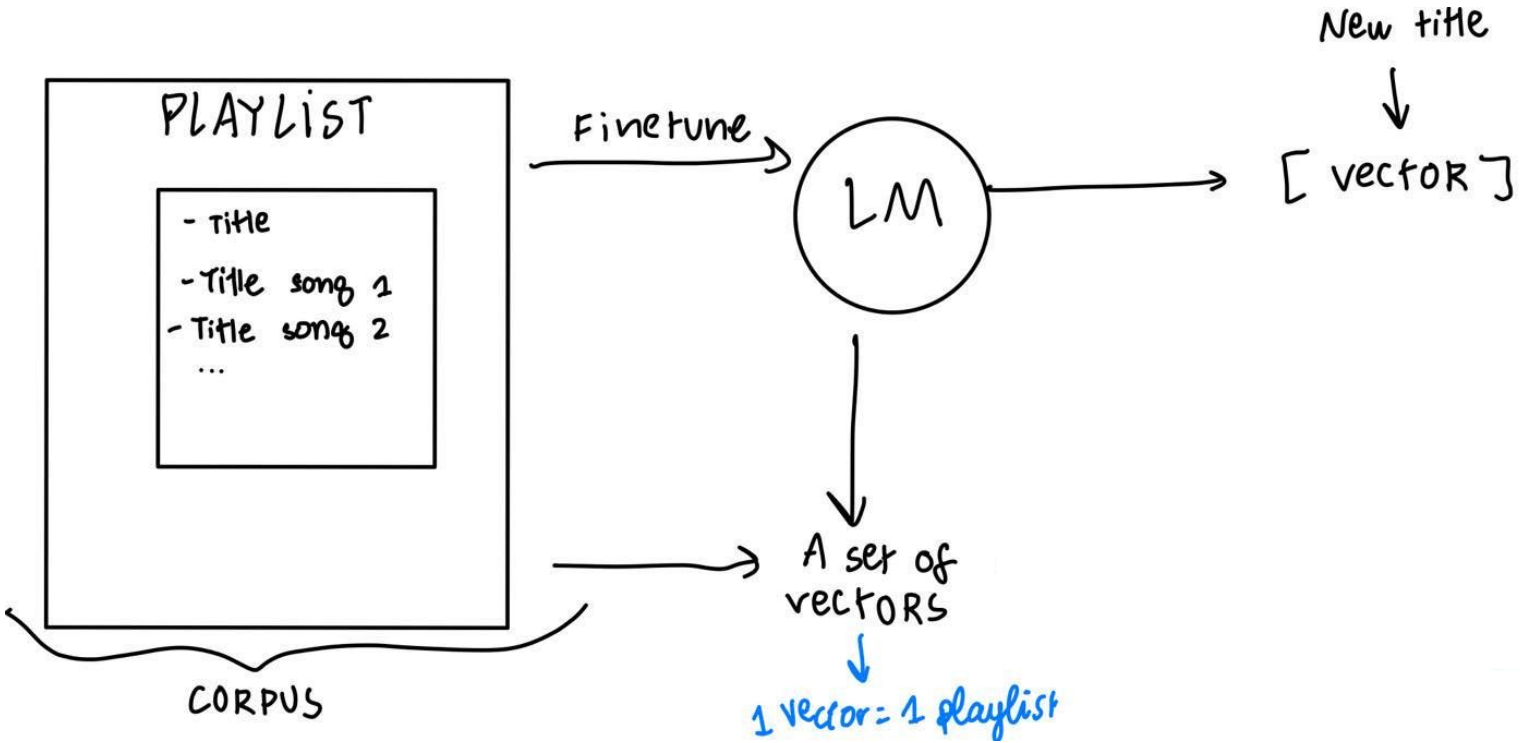


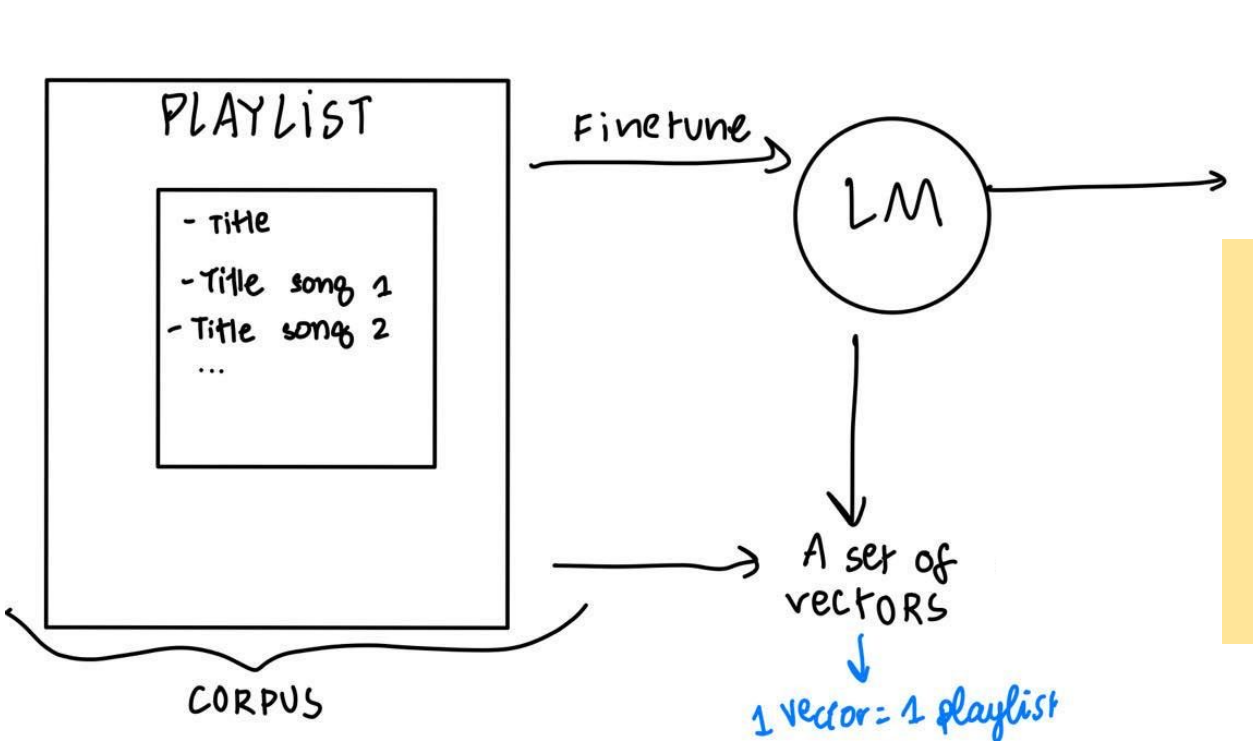
- Embedding using fastText?
- Word2Rec vs Title2Rec
- Title2Rec: K-means algorithm


PLAYLIST
- Title
- Title song 1
- Title song 2
...
Finetune
LM
A set of vectors
1 vector = 1 playlist
CORPUS

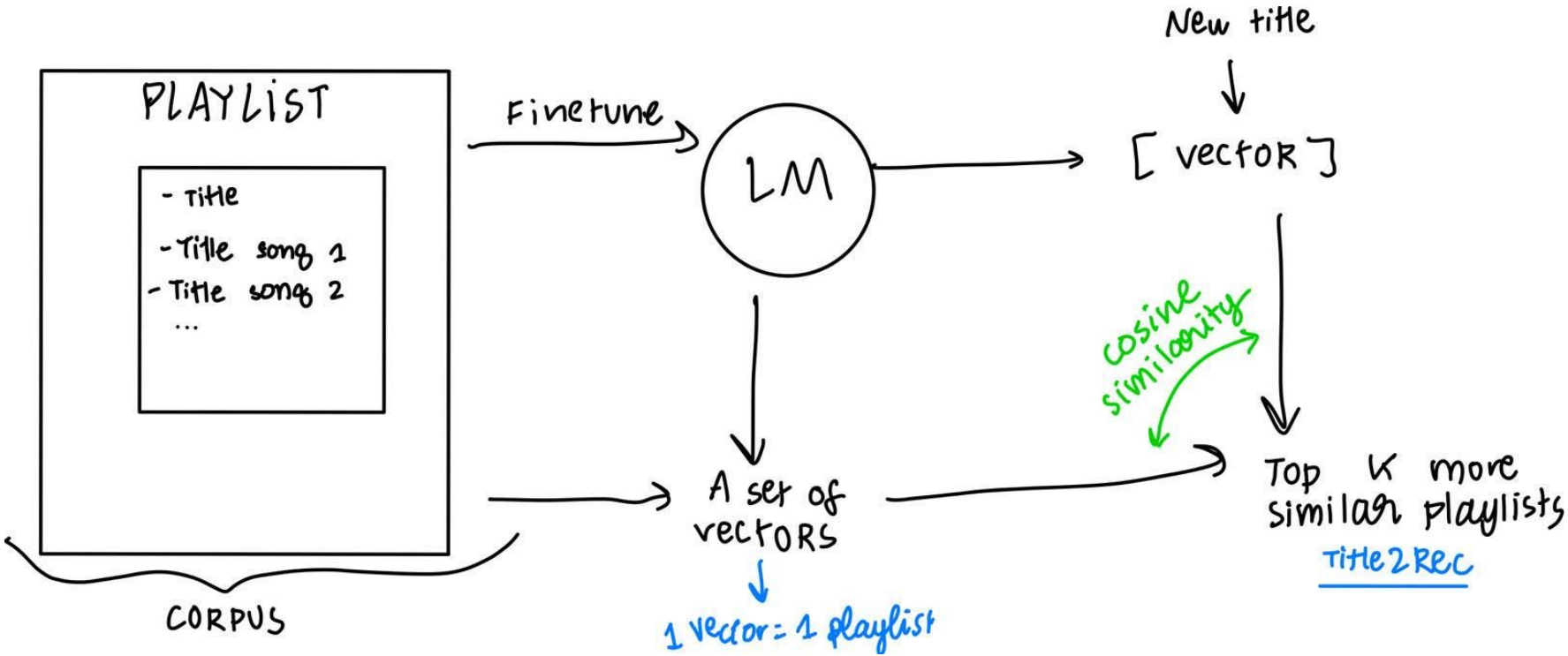


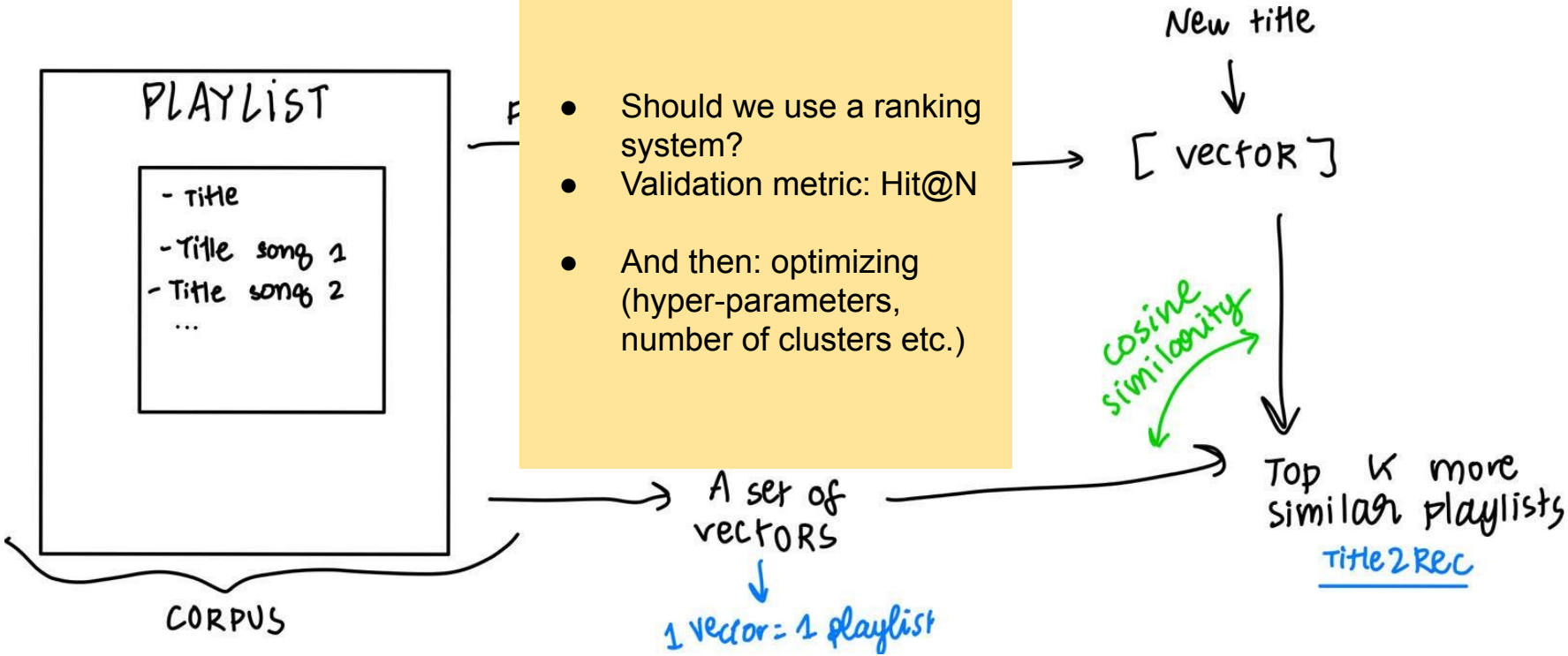
- Bert: bidirectional context
- Pre-training: mask words and sentences
- Easy to fine-tune (few additional layers needed)


PLAYLIST
- Title
- Title song 1
- Title song 2
...
CORPUS
Finetune
LM
New title
[vector]
A set of vectors
1 vector = 1 playlist



- Data cleaned
- Prompt engineering?


PLAYLIST
- Title
- Title song 1
- Title song 2
...
CORPUS
Finetune
LM
New title
[vector]
cosine similarity
A set of vectors
1 vector = 1 playlist
Top K more similar playlists
Title2Rec



- Should we use a ranking system?
- Validation metric: Hit@N
- And then: optimizing (hyper-parameters, number of clusters etc.)