Model Used : Vector Space Model

Please follow the steps to create a model, process it and perform information retrieval

Step 1: Creating a text document for the dataset in bag of words format

1. The first line should contain all the stemmed words (top 5000, preferably) from the complete corpus.
2. Each line should be in the bag of words format.
3. Pass each document (here, song) to the function **lyrics\_to\_bow(lyrics)** which is in the **stemmer.py** file.
4. Store the returned value in a text document as comma separated values in which first term would be the track\_id, second term being mxm\_id (can use custom id)
5. **Track\_id** , **mxm\_id** , **term\_index** (from first line) : **term freq (**in this song)
6. Example: TRAAAAV128F421A322,4623710,1:6,2:4,3:2,4:2,5:5,6:3,7:1,8:1,11:1

Step 2: Create a text document to store the information of the songs

1. Create a separate text file in which each line corresponds to a particular song.
2. The information should be in the following format
3. **track\_id**<SEP>**Artist name**<SEP>**Song name**<SEP>**custom\_id(here,** **mxm\_id)**<SEP>**Artist name for mxm**<SEP>**Song name for mxm**
4. Example: TRMMMKD128F425225D<SEP>Karkkiautomaatti<SEP>Tanssi vaan<SEP>4418550<SEP>Karkkiautomaatti<SEP>Tanssi vaan

Step 3: Preprocessing the dataset

1. Now use the text file created in steps 1 and 2 in **process\_data.py**
2. It process the data in the text files and stores them in a few data structures and then dumps them into the pickle file

Step 4: Query Processing

1. Now run the **main.py** file.
2. It loads the data from the pickfile generated in step 3 which contains all the relevant information.
3. It also implements the additional feature present in the **search\_yt.py** file and enables the user to choose to play the song.
4. Input the query and get the results.