BLUE MANGOES

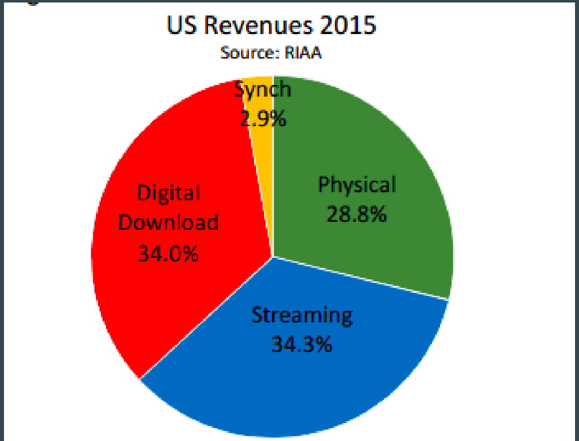
Piyush Khemka, Prachi Poddar, Snigdha Kamal, Sharang Bhat

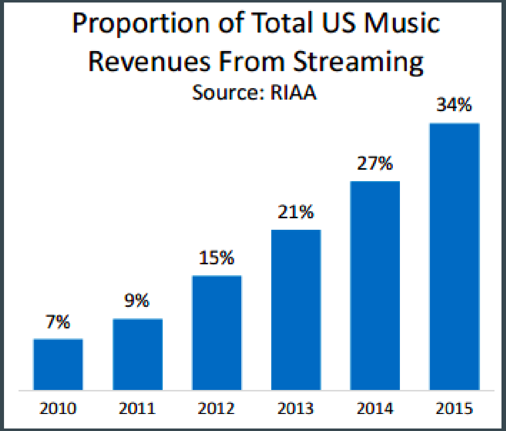
MUSIC RECOMMENDATION SYSTEM

Based on Million Song Dataset*

*labrosa.ee.columbia.edu/millionsong/

Background and Motivation





Growth in Music streaming consumption among consumers

Plethora of options available - Spotify, Pandora, 8tracks

*http://www.riaa.com/reports/riaa-2015-year-end-sales-shipments-data-report-riaa/
RIAA - Recording Industry Association of America

Background and Motivation

Music Recommendations - excellent feature for any music application.

Better Recommendations - Better Conversions, More engagement

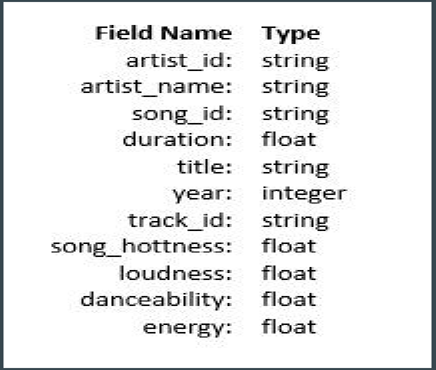
Develop a music recommendation system based on the Million Song Dataset using various recommendation methodologies and draw a comparative analysis between them

Dataset Description

The Million Song Dataset Dataset

Freely available collection of audio features and metadata for a million contemporary popular music tracks (280 GB) 1 M songs.

Subset: 2.8 GB (compressed) -> 10 GB (CSV)

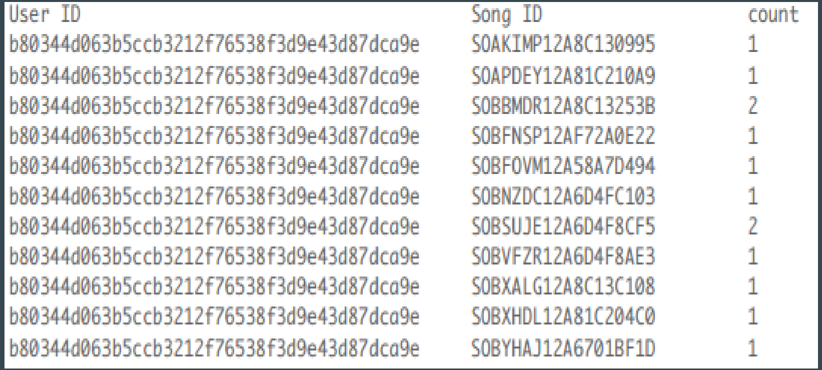


Field Name	Type
artist_id:	string
artist_name:	string
song_id:	string
duration:	float
title:	string
year:	integer
track_id:	string
song_hottness:	float
loudness:	float
danceability:	float
energy:	float

User Taste Profile

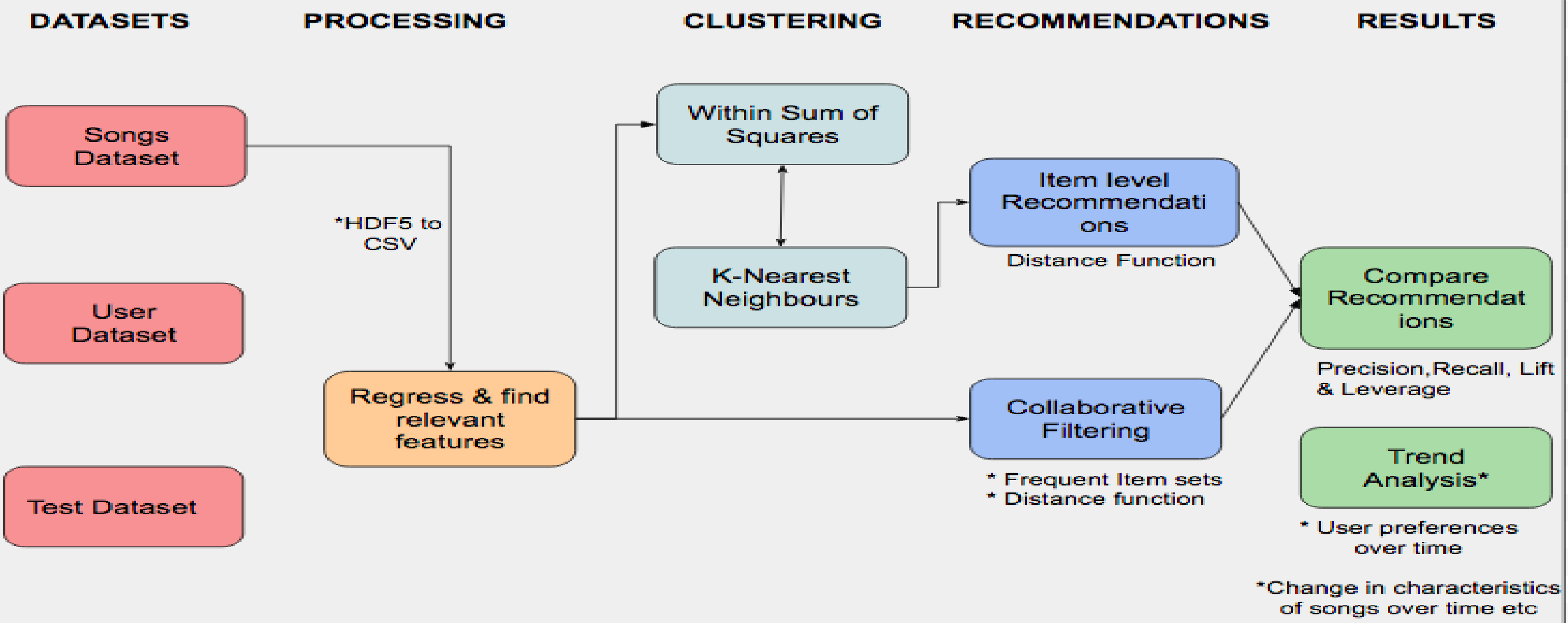
48 million triplets(User Id, Song ID, count)
Gathered from 1 million users

Size: 500 MB (compressed) -> 3 GB (.txt file)



User ID	Song ID	count
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAKIMP12A8C130995	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOAPDEY12A81C210A9	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBBMDR12A8C13253B	2
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBFNSP12AF72A0E22	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBFOVM12A58A7D494	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBNZDC12A6D4FC103	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBSUJE12A6D4F8CF5	2
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBVFZR12A6D4F8AE3	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBXALG12A8C13C108	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBXHDL12A81C204C0	1
b80344d063b5ccb3212f76538f3d9e43d87dca9e	SOBYHAJ12A6701BF1D	1

Project Pipeline



Collaborative Filtering

The Collaborative Filtering method uses previous user choices, and choices of similar users to predict the possible future song selections

Data was first fed through a data cleaning module, which removed erroneous entries, such as missing values for both Song and User ID’s.

ID’s were then mapped to a unique integer via a dictionary as Spark functions require numeric values

Collaborative Filtering

Used Matrix Factorization instead of the conventional distance function

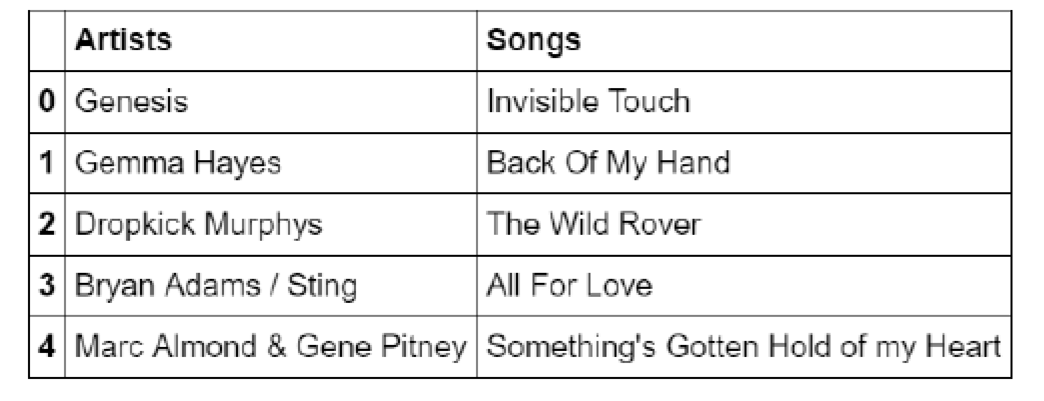
Latent factor methods were used to train on some known data

K User Features (latent) were extracted , which in this case was the song count

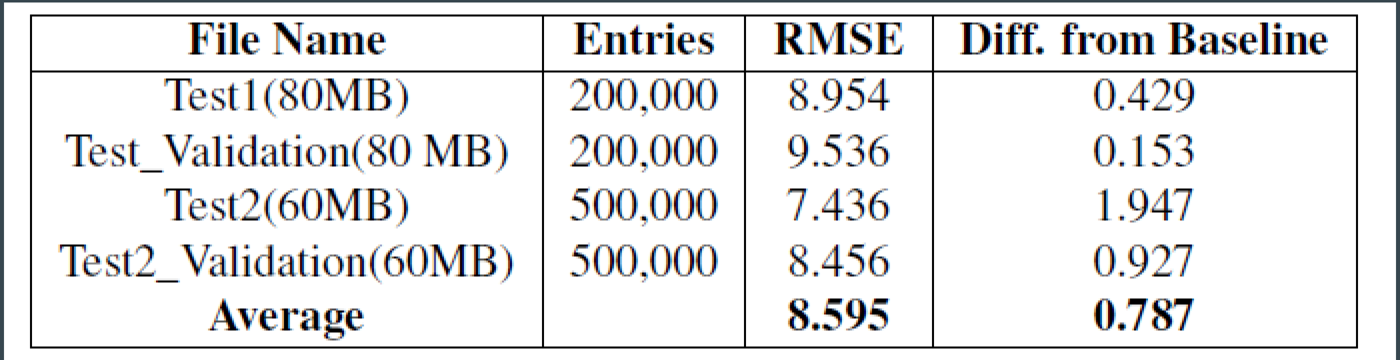
Assumption-Song Count represents all the factors that could have contributed to a user choosing to listen to a song

Trained over different values of K (ranks) and selected best rank to

Results - Collaborative Filtering



	Artists	Songs
0	Genesis	Invisible Touch
1	Gemma Hayes	Back Of My Hand
2	Dropkick Murphys	The Wild Rover
3	Bryan Adams / Sting	All For Love
4	Marc Almond & Gene Pitney	Something's Gotten Hold of my Heart



File Name	Entries	RMSE	Diff. from Baseline
Test1(80MB)	200,000	8.954	0.429
Test_Validation(80 MB)	200,000	9.536	0.153
Test2(60MB)	500,000	7.436	1.947
Test2_Validation(60MB)	500,000	8.456	0.927
Average		**8.595**	**0.787**

Content Based Recommendation

We extracted certain features from the dataset, which describes features of a song

Normalized those features by taking its product with its confidence to get a final value, such as mode and mode confidence to get a final mode estimate

Removed features not related to audio features

Clustered songs in a higher dimensional space, and found similar songs within each cluster

Content Based Recommendation

Training Method- Cross Validation

User-triples dataset was split in a ratio of 80:20.

80% of the dataset was used to train a clustering model

Created a profile for each user by merging the user-song dataset

Each user profile consisted of a “mean” of all songs heard by a user in his lifetime (as per the dataset)

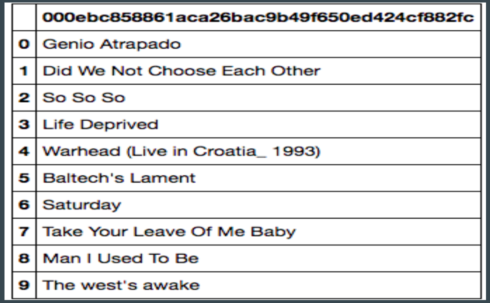
Clustered using K-means

Content Based Recommendation

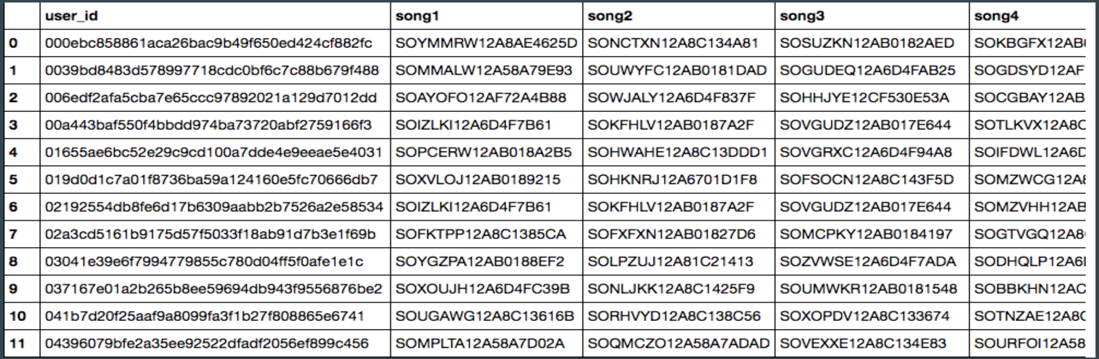
Testing

Generated 10 nearest neighbours for each user

Evaluated by comparing recommendations to actual values present in the dataset



	000ebc858861aca26bac9b49f650ed424cf882fc
0	Genio Atrapado
1	Did We Not Choose Each Other
2	So So So
3	Life Deprived
4	Warhead (Live in Croatia_ 1993)
5	Baltech's Lament
6	Saturday
7	Take Your Leave Of Me Baby
8	Man I Used To Be
9	The west's awake



	user_id	song1	song2	song3	song4
0	000ebc858861aca26bac9b49f650ed424cf882fc	SOYMMRW12A8AE4625D	SONCTXN12A8C134A81	SOSUZKN12AB0182AED	SOKBGFX12AB
1	0039bd8483d578997718cdc0bf6c7c88b679f488	SOMMALW12A58A79E93	SOUWYFC12AB0181DAD	SOGUDEQ12A6D4FAB25	SOGDSYD12AF
2	006edf2afa5cba7e65ccc97892021a129d7012dd	SOAYOFO12AF72A4B88	SOWJALY12A6D4F837F	SOHHJYE12CF530E53A	SOCGBAY12AB
3	00a443baf550f4bbdd974ba73720abf2759166f3	SOIZLKI12A6D4F7B61	SOKFHLV12AB0187A2F	SOVGUDZ12AB017E644	SOTLKVX12A8C
4	01655ae6bc52e29c9cd100a7dde4e9eeae5e4031	SOPCERW12AB018A2B5	SOHWAHE12A8C13DDD1	SOVGRXC12A6D4F94A8	SOIFDWL12A6D
5	019d0d1c7a01f8736ba59a124160e5fc70666db7	SOXVLOJ12AB0189215	SOHKNRJ12A6701D1F8	SOFSOCN12A8C143F5D	SOMZWCG12A
6	02192554db8fe6d17b6309aabb2b7526a2e58534	SOIZLKI12A6D4F7B61	SOKFHLV12AB0187A2F	SOVGUDZ12AB017E644	SOMZVHH12AB
7	02a3cd5161b9175d57f5033f18ab91d7b3e1f69b	SOFKTPP12A8C1385CA	SOFXFXN12AB01827D6	SOMCPKY12AB0184197	SOGTVGQ12A8
8	03041e39e6f7994779855c780d04ff5f0afe1e1c	SOYGZPA12AB0188EF2	SOLPZUJ12A81C21413	SOZVWSE12A6D4F7ADA	SODHQLP12A6
9	037167e01a2b265b8ee59694db943f9556876be2	SOXOUJH12A6D4FC39B	SONLJKK12A8C1425F9	SOUMWKR12AB0181548	SOBBKHN12AC
10	041b7d20f25aaf9a8099fa3f1b27f808865e6741	SOUGAWG12A8C13616B	SORHVYD12A8C138C56	SOXOPDV12A8C133674	SOTNZAE12A8C
11	04396079bfe2a35ee92522dfadf2056ef899c456	SOMPLTA12A58A7D02A	SOQMCZO12A58A7ADAD	SOVEXXE12A8C134E83	SOURFOI12A58

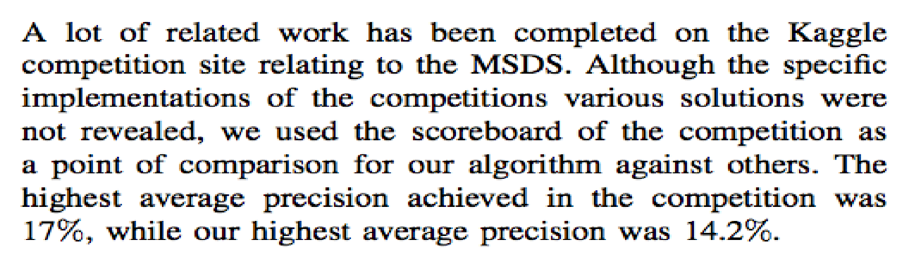
Results - Content Based Filtering

Precision = True Positives / (True Positive + False Positive)

Final Precision- 6.1% on entire MSD subset

Performance for Content Based decreases with increase in size of data

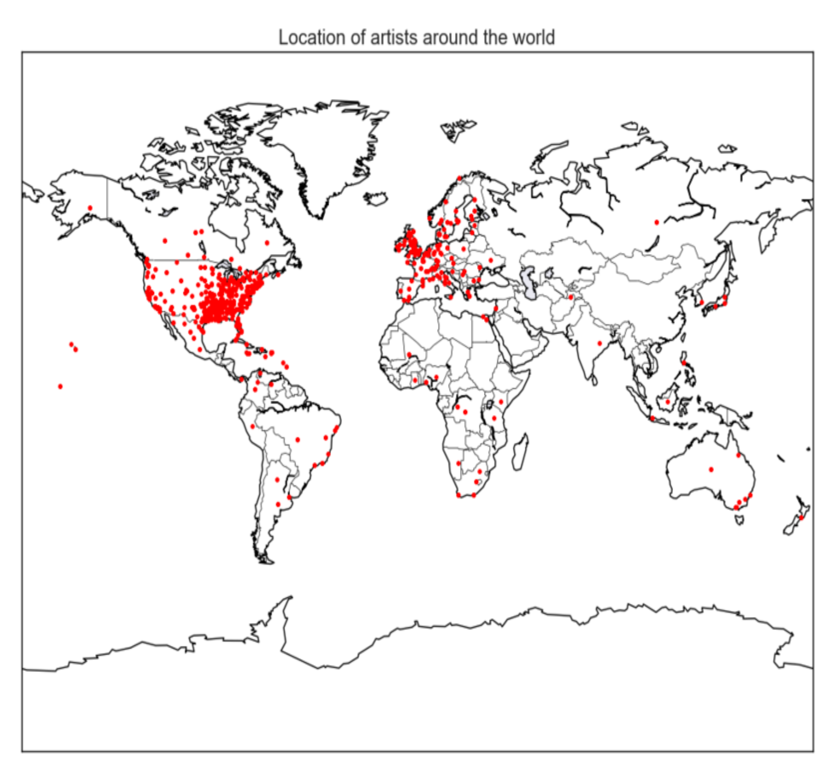
*

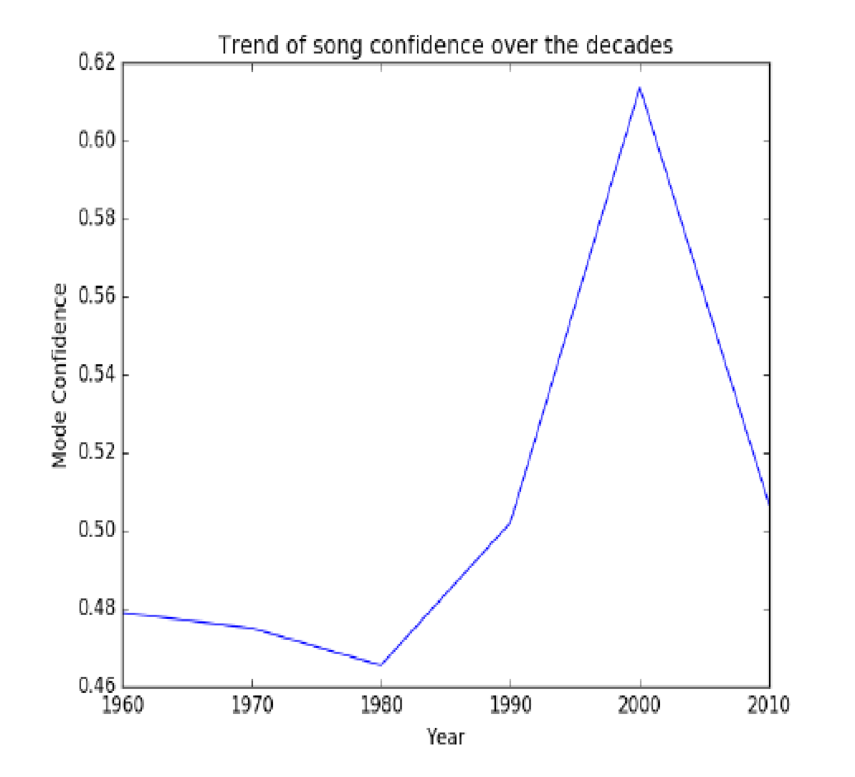


A lot of related work has been completed on the Kaggle competition site relating to the MSDS. Although the specific implementations of the competitions various solutions were not revealed, we used the scoreboard of the competition as a point of comparison for our algorithm against others. The highest average precision achieved in the competition was 17%, while our highest average precision was 14.2%.

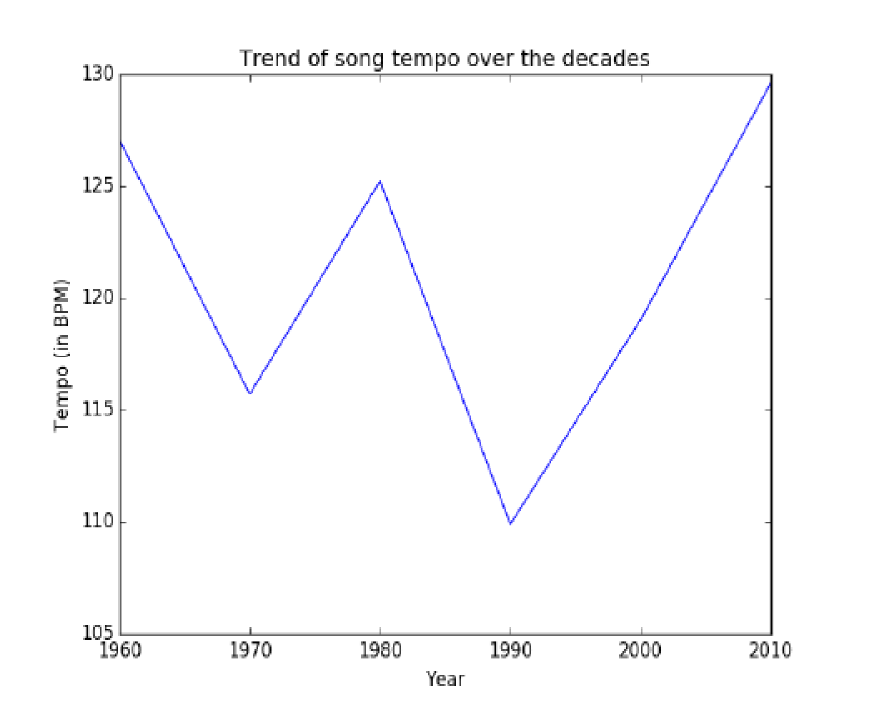
* http://www-personal.umich.edu/~yjli/content/projectreport.pdf

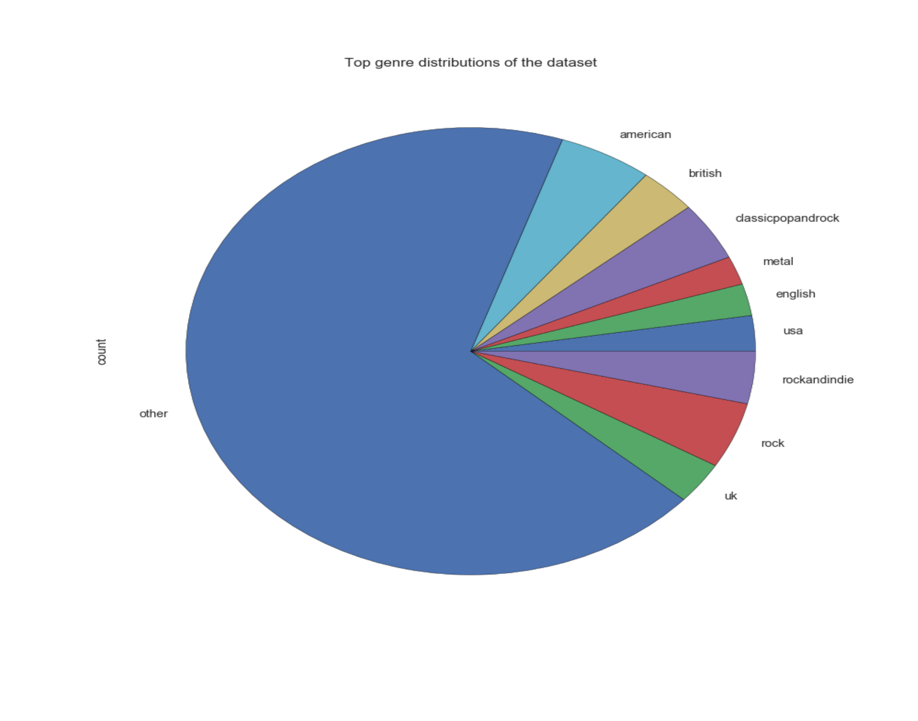
Visualizations and Trend Analysis





Visualizations and Trend Analysis





Team Work

- Scoping the project, data extraction, data cleaning - All

- Collaborative Filtering - Sharang

- Content Based Filtering - Piyush

- Trend Analysis & Visualizations - Prachi, Snigdha

- Conclusion, Report, Presentation - All

- Asking questions on Piazza - Anonymous

Conclusion

Eye opener on big data and its difficulties

We were fairly satisfied with our results, we managed to reduce the RMSE by almost 1, and virtually predicted how many times a user will play a song

We fairly satisfied with a precision of 6.1 % considering we tried a new recommendation methodology

Collaborative Filtering was easier to implement and evaluate

Unearthed some interesting music trends from across the years

THANK YOU