National College of Ireland
Higher Diploma in Science in Data Analytics
2013/2014

Michele Groarke
X13120654
Michele.Groarke@student.ncirl.ie

Analysis of Million Song Dataset to Create an Accurate Recommender System

Dissertation

National
College of
Ireland

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: Michele Groarke
Student ID: X13120654
Supervisor: Dr. Ioana Ghergulescu

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature:____

Date:____________

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to your being suspended or expelled from college.

Executive Summary

This study proposes an alternative music recommender system which utilises users' feelings about songs in tailoring a music recommender system. As an alternative to current recommender systems which use a combination of collaborative and content-based filtering to predict a user's enjoyment of a song/album/artist, this project proposes that the contextual element of a users' feelings about a sample of songs taken from the Million Song Dataset (MSD) can be used to make more accurate predictions regarding a user's enjoyment of proposed tracks. In addition to 'feelings', other factors were taken into consideration , namely the users' age, gender and whether they are influenced by lyrics, rhythm or both.

In current systems, there is a 'cold-start' problem which occurs when a new user has not provided enough input to tailor a profile for them. In addition, some current music recommender systems will recommend according to artist, genre or album which can be unreliable due to the non-homogeneity of artists and albums and the difficulty of assigning a genre to every song.

The methods used in this study were to perform clustering analyses (k-Means) on the survey results and on the attributes of the MSD Subset. Then the k-NN classification technique was applied to the data.

Results of initial analysis indicate that further development is necessary to obtain the desired level of accuracy in recommendations.

Post dissertation work to be carried out includes testing and finalisation of resulting recommendations to the survey participants. The results will be reported to participants upon completion of testing and analysis.

Keywords: Data Mining, Collaborative filtering, content-based filtering, Contextual Filtering, User Profiles.

Definitions/Acronyms: RS – Recommender System; HDF5 – portable and extensible file format for high volume, complex data; MSD – Million Song Dataset; CF – Collaborative filtering; ETL – extract, transform and load.

Contents

1. Introduction

With the ubiquity of the internet as a source of entertainment especially in the world of music, systems that recommend content tailored to individual tastes are increasingly valuable to service providers and retailers in attracting and retaining customers' engagement and encouraging purchases on-line. There are various ways to tailor these recommender systems according to different attributes – those of the content (music tracks, in this case) and/or the preferences of the user.

Recommender systems (RS) enable the user to discover new music, access preferred music, and they enable music retailers to sell to the users. This study aims to combine analysis of the features of the tracks, loudness and tempo, for example, with user sentiment regarding the tracks (gleaned from a user survey) to predict user enjoyment of new tracks.

'Contextual' elements which might be suitable to use in predicting a user's enjoyment of a song could be where they are located, the weather conditions, who they are with, how they are feeling, etc. This study used the contextual element of the user's feeling or reaction to each song in the survey and combined these results with the 'content-based' method of using each song's attributes to create a more effective recommender system than the alternative collaborative, content-based or hybrid systems.

The Million Song Dataset was developed in collaboration between Columbia University's LabROSA department (Laboratory for the Recognition and Organization of Speech and Audio) (labrosa.ee, 2011) and The Echonest (Echonest, 2011), an industry leader in music intelligence. It is a freely available collection of over 1 million songs and their features to enable large-scale analysis on Music Information Retrieval (MIR) with a set of attributes for each song which includes tempo, loudness, artist hotness, danceability, etc.

The information included in this introductory section will describe current filtering methods, the motivation and aims of the project, the research questions being addresses and an overview of the solution.

1.1. Filtering methods

The main methods of Recommender Systems are Collaborative filtering, Content-based filtering, contextual filtering or hybrid systems.

Collaborative filtering (CF)

Collaborative filtering is a technology which uses recommendations by other users with similar tastes to recommend similar items. Also, the user can build a profile and the system 'learns' the user's preferences over time. The advantage of a collaborative system is that the recommendations are more personally tailored for the user.

The major disadvantage of CF is that it takes time to build a user profile, so there is a 'cold start' issue.

The combined contextual and content based technique used in this project addresses this issue as the content based information is available immediately. This information has been improved through the addition of 'emotion' and demographic information about each track to make the recommendations more accurate.

Content-based filtering

Content-based filtering techniques analyse attributes of items and identify similarities between relevant attributes in order to make recommendations based on groupings of items with similar attributes.

The disadvantage of this technique is that the recommendations are based on predominantly technical similarities (tempo, genre, etc.). This system takes little or no cognisance of a user's emotional engagement (and resulting desire to purchase) with the items of music.

This problem is addressed by the proposed system through the contextual or 'emotional' aspect of the system. This information enables the resulting recommendations to be more personal than those offered by the traditional content-based system alone.

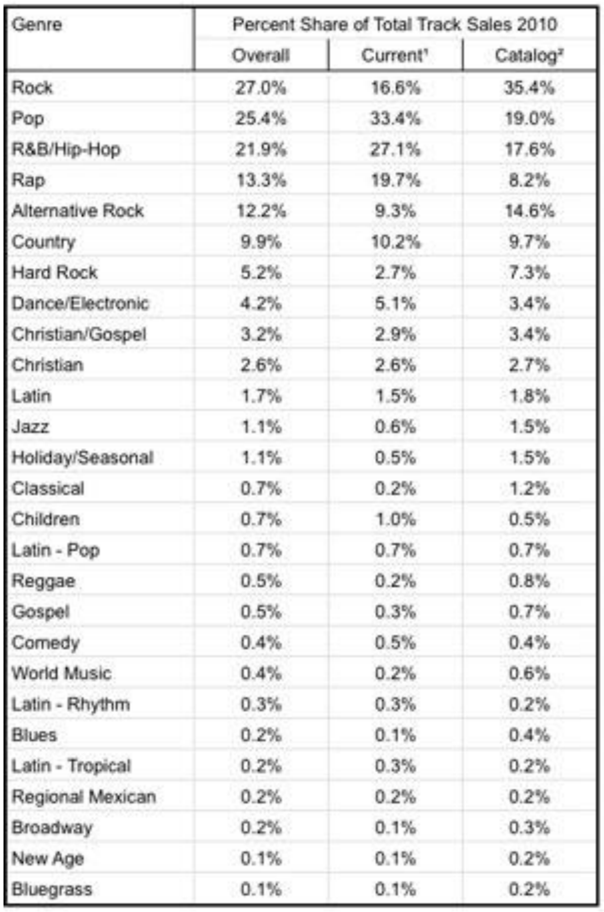
Hybrid Systems

Hybrid systems have been developed that incorporate elements of both, and in some cases also use statistical methods to create a more optimum system. In most cases these systems are more successful than either the collaborative or content based methods.

However, often recommender systems will often compare artists by album or genre, which represents a problem in that artists and albums are rarely homogeneous. It is impossible to assign every song to a genre – despite the large number of genres listed in Table 1 – as many songs span genres and genres themselves can evolve and subdivide as fashions and tastes change.

This problem is addressed by the proposed system since it does not group the songs by artist or genre. Rather, the groupings are formed naturally, using the k-Means clustering technique. This technique identifies natural groupings according to the attributes of the tracks (including the contextual attributes added as a result of the user survey)

Table 1: Music Genres



Genre	Percent Share of Total Track Sales 2010		
	Overall	Current[1]	Catalog[2]
Rock	27.0%	16.6%	35.4%
Pop	25.4%	33.4%	19.0%
R&B/Hip-Hop	21.9%	27.1%	17.6%
Rap	13.3%	19.7%	8.2%
Alternative Rock	12.2%	9.3%	14.6%
Country	9.9%	10.2%	9.7%
Hard Rock	5.2%	2.7%	7.3%
Dance/Electronic	4.2%	5.1%	3.4%
Christian/Gospel	3.2%	2.9%	3.4%
Christian	2.6%	2.6%	2.7%
Latin	1.7%	1.5%	1.8%
Jazz	1.1%	0.6%	1.5%
Holiday/Seasonal	1.1%	0.5%	1.5%
Classical	0.7%	0.2%	1.2%
Children	0.7%	1.0%	0.5%
Latin - Pop	0.7%	0.7%	0.7%
Reggae	0.5%	0.2%	0.8%
Gospel	0.5%	0.3%	0.7%
Comedy	0.4%	0.5%	0.4%
World Music	0.4%	0.2%	0.6%
Latin - Rhythm	0.3%	0.3%	0.2%
Blues	0.2%	0.1%	0.4%
Latin - Tropical	0.2%	0.3%	0.2%
Regional Mexican	0.2%	0.2%	0.2%
Broadway	0.2%	0.1%	0.3%
New Age	0.1%	0.1%	0.2%
Bluegrass	0.1%	0.1%	0.2%

(MusicRow, 2014)

1.2. Motivation

One of the main issues with existing collaborative filtering systems is the ‘cold-start’ problem from which these systems suffer. This is where the user has not input any or enough information for the system to make a prediction or recommendation based on user preferences.

Music RS are complicated by the many factors which influencing a user’s choice in music, including social and geographical context, personal preference, and the traits of the artists and songs themselves. The mood or feeling of a user is another factor that this paper will take into consideration.

Systems which group songs based on artist/album are rarely entirely effective as one artist can produce many different styles of song, some liked and some disliked by an individual user. Therefore recommending via this method is not optimal.

1.3. Aims

The Objective of this project is to build a model that will predict songs based on an individual's feelings about individual tracks. It will endeavour to facilitate more relevant predictions than are currently being offered on music streaming sites like Spotify as described below.

Another objective is an exploration of Clustering, Classification, Data reduction, Association, and other Data Analysis and Data Visualisation tasks.

1.3.1. Main Problems being addressed

The main problems being addressed in this project are:

- The introduction to new material that will be enjoyed by the user.
- The 'cold start' problem in collaborative filtering in which there is not yet enough input from the user to accurately predict material they might enjoy.
- Employing the user's feelings about previous songs into consideration when recommending new songs. Similarly, factors such as age and gender are included in the mix, as are narrower areas of focus such as the user's priority on lyrics, rhythm or a combination of these two.
- It is also hoped that the scope can be expanded to include a recommender system for films, books, art, dating, etc.

1.3.2. Research Questions

The main research questions proposed by this study are:

Can this system predict a user's enjoyment of songs more accurately, taking into account the user's initial classification of a 100 song subset of the MSD?

Do the lyrics or Rhythm of a song have more of an effect on the type of music a user enjoys?

1.4. Solution Overview

This project endeavoured to offer an accurate recommendation to users, based on their responses to tracks and the underlying attributes. It is hoped to predict tracks with similar attributes from the test set.

Resulting from "a collaboration between The Echo Nest and Columbia University's LabROSA department (Laboratory for the Recognition and Organization of Speech and Audio), (labrosa.ee, 2011) The Million Song Dataset has four main objectives:

- To encourage research on algorithms that scale to commercial sizes
- To provide a reference dataset for evaluating research
- As a shortcut alternative to creating a large dataset with The Echo Nest's API
- To help new researchers get started in the MIR field.

The Million Song Dataset offers researchers, engineers and commercial developers detailed sonic and cultural attributes for each song, as well as extensive metadata, both provided by The Echo Nest." (Echonest, 2011)

A 10,000 song subset is available for download and this is the subset that was used for this study. This subset of songs and their attributes was downloaded into a MySQL database called ‘msd’.

A survey concerning 100 of the tracks selected from the Million Song Dataset (MSD) was completed by a sample population of 18 respondents. These results were initially prepared in Excel and stored and transformed in aSQL database, called ‘survey’ before again being analysed in Excel and Tableau.

From the Tableau analysis, the most frequently occurring ‘feeling’ for each song was identifiable and became an extra attribute for each song, in addition to the attributes already provided in the MSD dataset.

The results of the survey were ‘blended’ with the attributes from the songs in the ‘msd’ database which created a rich source of information about each song.

Then Weka and R were used to perform analyses on the MSD attributes both independently and in conjunction with the results from the user survey.

1.5. Structure

The rest of this paper is organized as follows:

Section 2 introduces related work in the area of Recommender Systems. Examples of some existing Recommender systems are presented. In Section 3, the system and datasets are described. Testing and evaluation of the results is described in Section 4. Conclusions are presented in Section 5 where the advantages, disadvantages, opportunities and limits of the project are discussed. Finally, Section 6 proposes further development and research possibilities.

2. Related Work

This section provides a background to the study and gives a description of research in the field of Recommender Systems and some current examples of music recommender systems.

2.1. Review of relevant research

The first Recommender System is widely recognised in the literature as Tapestry (Goldberg, Nichols, Oki, & Terry, 1992) which was an experimental mail system developed at the Xerox Research Centre in Palo Alto over 20 years ago. From this starting point, the literature discusses collaborative and content-based filtering methods, with some modifications, for example with statistical elements also included.

Examples of such systems are a system based on music data grouping and user interests (Hung-Chen & Chen, 2014) which utilises three approaches – Collaborative, Content-based and Statistical – to predict recommendations for users. Another system based on Deep Content is proposed by Van den Oord et al (Van Den Oord, Dieleman, & Schrauwen, 2013) which utilises deep convolutional neural networks and compares results to a more traditional "Bag-of-words" approach. Jayalakshmi et al (Jayalakshmi, Shruthi, Sneha, & Uttarika Ratnakar, 2014) combine collaborative and content-based filtering for their Hybrid Music Recommender System which they found performed better than either of the combined systems when used alone. This paper will use a similar approach in that there is a combination of collaborative (survey responses) and content-based (MSD attributes) input to the system. However, the user input is unique in that it takes the user's emotions or feelings into consideration in the collaborative filtering stage.

2.2. Current Recommender System Examples

Recommender systems are vital to online music streaming services. Many offer sophisticated methods of matching suggestions with a user's preferences. Examples are;

Beats Music

Beats Music evolved from Beats by Dre headphones which resulted from the partnership between the rapper Dr. Dre, and Monster headphones. The recommendations from Beats Music are based on a personal profile created by the user when they log in and select favourite artists and genres. (Beats Music, 2014)

Last.fm

Last fm offers music recommendations for the user based on artists they already like. This music streaming service encourages the user to download a 'scrobbler' app which automatically populates the user's library with whatever they've been listening to on their computer. Last.fm has a library of over 54 million artists, 200 million albums and 640 million tracks. (Last.fm, 2014)

Pandora

Pandora is a service by which users can create up to 100 online 'radio stations' with their own personal preference in music. Pandora is supported by the Music Genome Project and in house musicians identify up to 400 'genes' or traits of each song in the Pandora corpus. Unfortunately, due to licensing agreements, Pandora is only available in the U.S. and New Zealand at the current time. (Pandora)

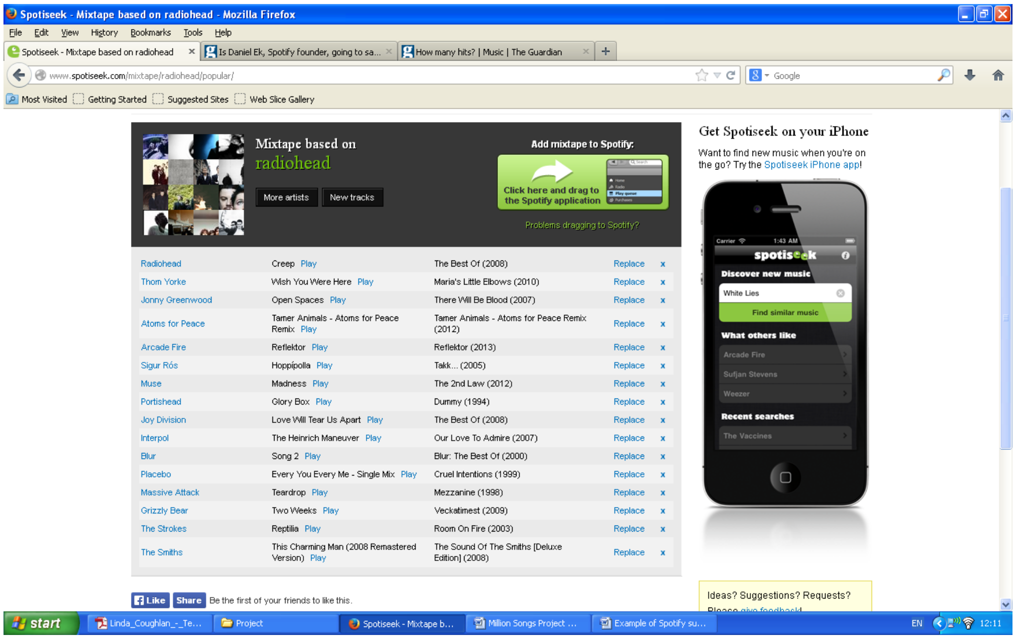
Spotify

Spotify was started 8 years ago in Sweden by Daniel Ek, a music fan and programming genius. He approached record labels with the radical/novel idea to provide music to rent and for free instead of buying it! Despite being understandably wary, he managed to persuade them that Spotify would be the music industry's saviour rather than completely destroying it. As of October 2013, on the 5th anniversary of its launch, more than 20million songs were made available to 24 million users in 52 territories. One billion playlists have been created so far. (Lynskey, 2013)

Spotiseek

There are many Spotify playlist generators available, for example, filtr.com, Spotibot.com, lazify.nl, and many more. Spotiseek helps users to find music similar to their favourite artists and creates "mixtape" playlists for Spotify.

Although the results achieved (see Figure 1) from several attempts on the Spotiseek playlist generator were ok for the song chosen by Spotiseek from the input artist, there was no option to input an individual song. So, the song chosen, and therefore the suggested playlist, could be completely different to another song by the same artist that the user might like. (Sommestad, 2009)


Spotiseek - Mixtape based on radiohead - Mozilla Firefox
File Edit View History Bookmarks Tools Help
www.spotiseek.com/mixtape/radiohead/popular/
Mixtape based on
radiohead
More artists
New tracks
Add mixtape to Spotify:
Click here and drag to the Spotify application
Problems dragging to Spotify?
Get Spotiseek on your iPhone
spotiseek
Discover new music
White Lies
Find similar music
What others like
Arcade Fire
Sufjan Stevens
Weezer
Recent searches
The Vaccines
Like
Share
start

Want to find new music when you're on the go? Try the Spotiseek iPhone app!

Radiohead	Creep Play	The Best Of (2008)	Replace	x
Thom Yorke	Wish You Were Here Play	Maria's Little Elbows (2010)	Replace	x
Jonny Greenwood	Open Spaces Play	There Will Be Blood (2007)	Replace	x
Atoms for Peace	Tamer Animals - Atoms for Peace Remix Play	Tamer Animals - Atoms for Peace Remix (2012)	Replace	x
Arcade Fire	Reflektor Play	Reflektor (2013)	Replace	x
Sigur Rós	Hoppípolla Play	Takk... (2005)	Replace	x
Muse	Madness Play	The 2nd Law (2012)	Replace	x
Portishead	Glory Box Play	Dummy (1994)	Replace	x
Joy Division	Love Will Tear Us Apart Play	The Best Of (2008)	Replace	x
Interpol	The Heinrich Maneuver Play	Our Love To Admire (2007)	Replace	x
Blur	Song 2 Play	Blur: The Best Of (2000)	Replace	x
Placebo	Every You Every Me - Single Mix Play	Cruel Intentions (1999)	Replace	x
Massive Attack	Teardrop Play	Mezzanine (1998)	Replace	x
Grizzly Bear	Two Weeks Play	Veckatimest (2009)	Replace	x
The Strokes	Reptilia Play	Room On Fire (2003)	Replace	x
The Smiths	This Charming Man (2008 Remastered Version) Play	The Sound Of The Smiths [Deluxe Edition] (2008)	Replace	x

Be the first of your friends to like this.

Ideas? Suggestions? Requests?

Figure 1: Example of playlist recommended by Spotiseek

3. System and Datasets

This section of the dissertation describes the construction of the survey and resulting survey dataset in addition to the MSD dataset. It also describes the system architecture and the extraction, transformation and loading (ETL) of the various data.

3.1. Survey

The survey consisted of 100 songs taken from the 10,000 song subset of the MSD. These were chosen from a shortlist of recognisable artists (as there were artists included on the original dataset that were particular to a certain geographical location and not recognised in Ireland). These were listed on an excel sheet with Hyperlinks included so the respondents could listen to any songs they weren’t familiar with. The survey was emailed to 133 recipients, randomly selected from the analysts email contact list and class members.

There were 18 respondents, ranging in age from 18 – 52 years, with a 50:50 male:female ratio.

A survey database was created in MySQL and the survey results were prepared and loaded into the database. The structure of the survey database can be found in Figure 3 below.

The required ethics document relating to the survey participants can be found in the appendix.

3.2. Design and Architecture

The Survey Database is a MySQL database which stores the results from the user survey

The Million Song Database is a MySQL database which stores the 10,000 song subset of songs and attributes extracted from the Million Song Dataset.

The Initial Analysis and **pre-processing** steps were carried out in Excel, Tableau and R

The Recommender System utilises kNN and kMeans to classify and group songs via their attributes and user’s feelings regarding them.

Test Results will be communicated to the survey respondents and the system will be evaluated accordingly.

Final results will be produced after **adjustments** are applied to the system.

The most basic process of the system is illustrated in Figure 2**:**

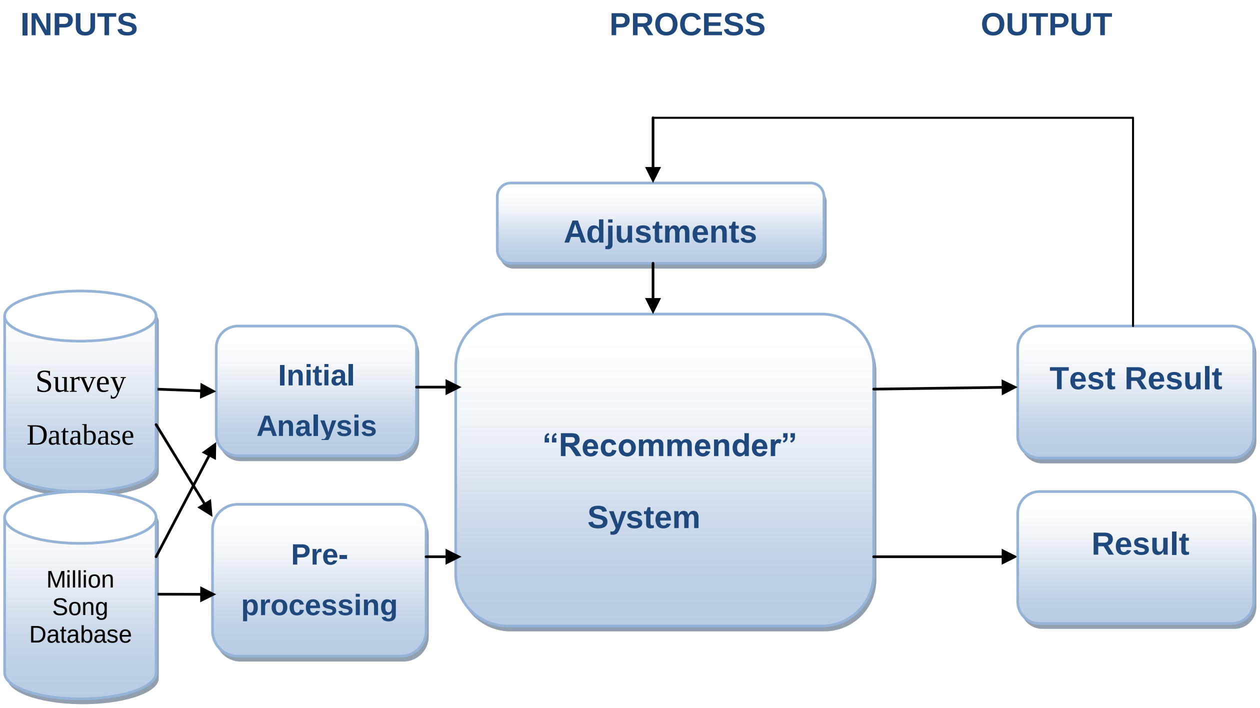


Figure 2: System Process

The survey results (survey and anonymised results in appendix, and on supplied disc) were received in an Excel format. These were prepared in excel before being exported to the 'survey' MySQL database.

The MSD 10,000 song subset was downloaded from labrosa (labrosa.ee, 2011) and initially stored in the laptop hard drive. They were then downloaded into the 'msd' MySQL database.

3.2 System Architecture

This section includes the survey database architecture, the fields and field types included for each song in the MSD and the storage and memory requirements of the computer used to work on this project.

Survey ERD

This Entity Relationship Diagram (ERD), shown in Figure 3: Survey ERDFigure 3, of the survey response database shows the attributes associated with the four entities – ‘artist’, ‘track’, ‘respondent’ and ‘feeling’. The ‘artist’ table became unnecessary as the system evolved; hence it is not joined to the other tables.

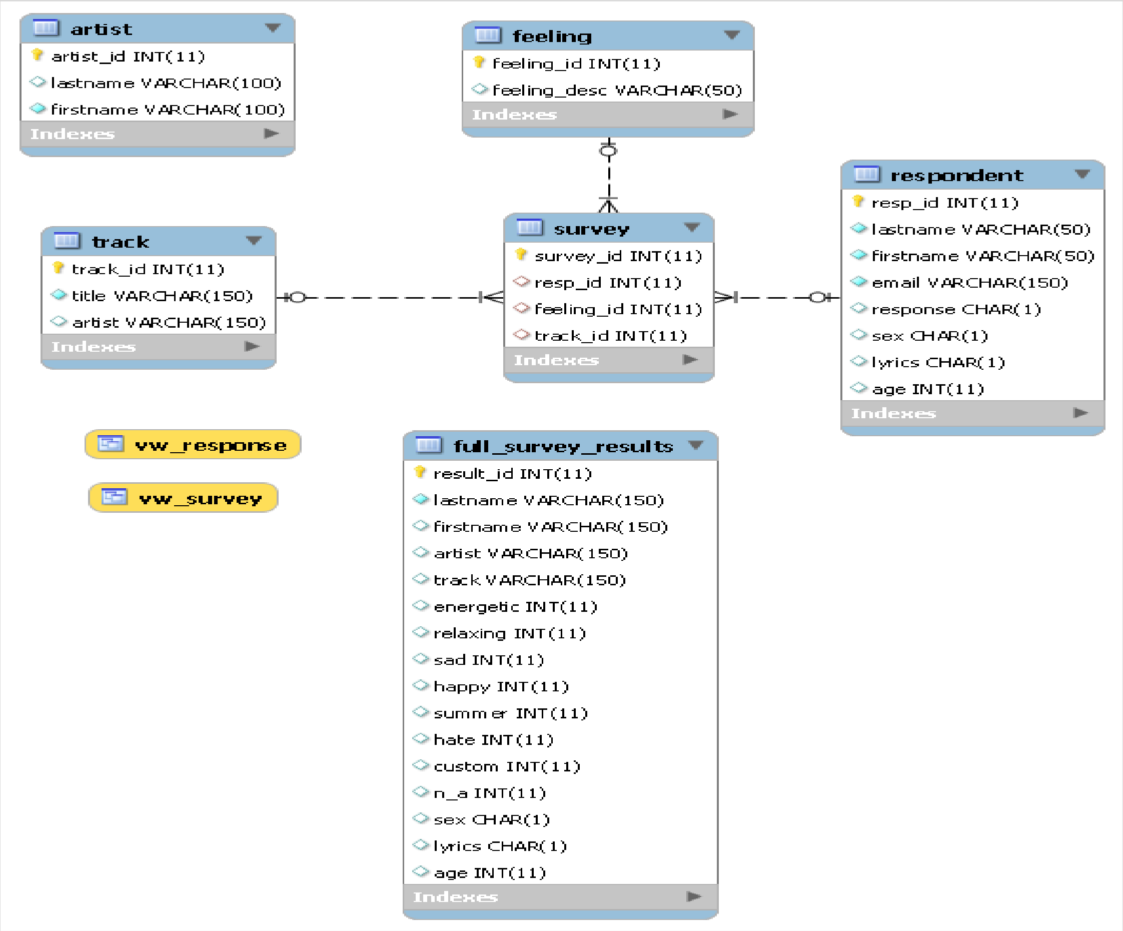


Figure 3: Survey ERD

The ‘full_survey_results’ table was the initial ‘staging’ table as part of the Extraction process of the project. The survey results arrived back in 18 separate excel sheets. These were combined in Excel into one sheet and copied as a .csv file (anonymised version in appendix) which was then uploaded into the ‘full_survey_results’ table in the MySQL ‘survey’ database with the following code:

```
LOAD DATA LOCAL INFILE 'C:/Scrap/Full_Survey_Results_1.csv'
INTO TABLE Full_Survey_Results
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'
(firstname,lastname,artist,track,energetic,
relaxing,sad,happy,summer,hate,custom,n_a, sex,lyrics, age);
```

This table has all the results from the respondents in their original format. This means a lot of redundancy but is necessary as a midpoint for viewing the data.

The next step was to populate the static tables – ‘track’, ‘feeling’ and ‘respondent’ in the ‘survey’ database using the following codes:

Populate ‘tracks’ table with 100 songs taken from the MSD subset:

```
LOAD DATA LOCAL INFILE 'C:/Scrap/tracks.csv'
INTO TABLE track
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'
(title, artist);
```

(The following code was necessary to remove a rogue ascii character, from the end of each row after loading the data)

```
UPDATE survey.track SET artist = TRIM(TRAILING '\r' FROM artist)
WHERE track_id>0;
```

Populate ‘feelings’ table:

```
LOAD DATA LOCAL INFILE 'C:/Scrap/feeling.csv'

INTO TABLE feeling

FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'

(feeling_id, feeling_desc);
```

(The following code was necessary to remove a rogue ascii character, from the end of each row after loading the data)

```
UPDATE feeling SET feeling_desc = TRIM(TRAILING '\r' FROM
feeling_desc) WHERE feeling_id>0;
```

Populate ‘respondent’ table with all survey recipients:

```
LOAD DATA LOCAL INFILE 'C:/Scrap/Survey recipients.csv'

INTO TABLE respondent

FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n'

(firstname, lastname, email, resp_code, undelivered, response,
sex, lyrics);
```

artist:

This entity includes the fields:

- artist_id – this is the primary key which is an auto-incremented integer type and is unique.
- lastname – this is a VARCHAR of no more than 100 characters
- firstname – this is a VARCHAR of no more than 100 characters

track:

This entity includes the fields:

- track_id - this is the primary key which is an auto-incremented integer type and is unique.
- title – this is the title of the song and is a VARCHAR of no more than 150 characters.
- artist_id – this is a foreign key linking to the artist table. It is an INTEGER.
- resp_id - this is a foreign key linking to the artist table. It is an INTEGER.

respondent:

This entity includes the fields:

- resp_id - this is the primary key which is an auto-incremented INTEGER and is unique.
- lastname – this is the respondent's last name and is a VARCHAR of no more than 50 characters.
- firstname – this is the respondent's first name and is a VARCHAR of no more than 50 characters
- email -. this is the respondent's email address and is a VARCHAR of no more than 150 characters
- feeling_id - this is a foreign key linking to the feeling table. It is an INTEGER.
- Sex – this is the respondent's gender and is a CHAR ('m', 'f')

- Lyrics – This indicates if the respondent reacts more to a song's lyrics, rhythm or both and is a CHAR ('l', 'r', or 'b')
- Age – the age of the respondent and is an INTEGER

feeling:

This entity includes the fields:

- feeling_id - this is the primary key which is an auto-incremented integer and is unique.
- feeling_desc – this is the title of the song and is a VARCHAR of no more than 150 characters.
- resp_id – this is a foreign key linking to the feeling table. It is an INTEGER.

3.3 Extraction of MSD 10,000 Song Subset

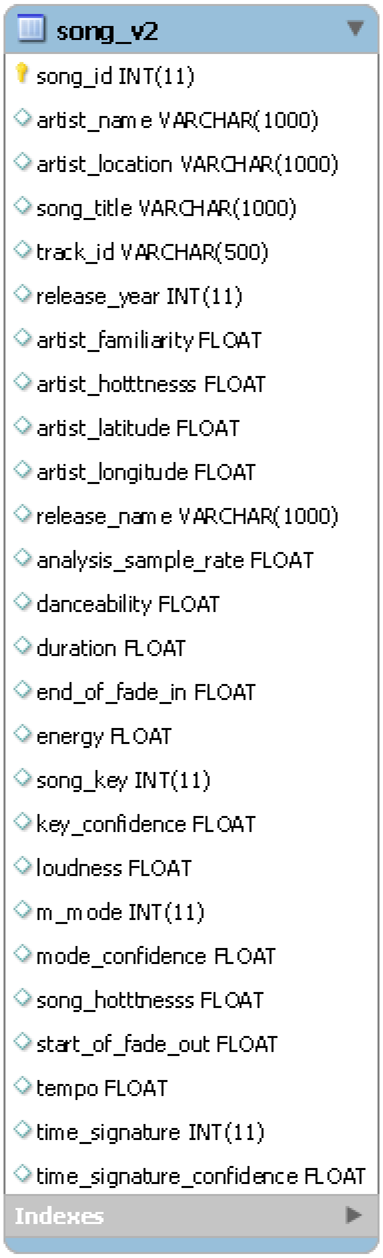
In order to read the hdf5 files in which the songs and their attributes were stored it was necessary to download some elements as follows:

Firstly, WinPython had to be downloaded from Sourceforge (Dice Holdings Inc., 2014) in order to provide the necessary libraries for importing the data into MySQL and to access the hdf5 files via hdf5 'getters'. The libraries required were Numpy, Math and Pytables which were needed to operate the hdf5 getters code, which was downloaded from Github (GitHub Inc., 2014)

The Python libraries for connecting to MySQL were in Numpy

The MSD database was created (code on accompanying disc). See Figure 4 below for details regarding how the table is populated. The initial subset of attributes was chosen from those available as they were not arrays or IDs. This resulted in 26 variables. Then it was populated using the 'Populate Table' script which used the hdf5 getters to specify which of the attributes were required to populate the table.

.



song_id INT(11)
artist_name VARCHAR(1000)
artist_location VARCHAR(1000)
song_title VARCHAR(1000)
track_id VARCHAR(500)
release_year INT(11)
artist_familiarity FLOAT
artist_hotttnesss FLOAT
artist_latitude FLOAT
artist_longitude FLOAT
release_name VARCHAR(1000)
analysis_sample_rate FLOAT
danceability FLOAT
duration FLOAT
end_of_fade_in FLOAT
energy FLOAT
song_key INT(11)
key_confidence FLOAT
loudness FLOAT
m_mode INT(11)
mode_confidence FLOAT
song_hotttnesss FLOAT
start_of_fade_out FLOAT
tempo FLOAT
time_signature INT(11)
time_signature_confidence FLOAT

song_id: unique identifier NOT NULL
artist_name: name of artist NOT NULL
artist_location: where artist is based (NULL)
song_title: song title
track_id: Echo Nest track ID
release_year: year when this song was released
artist_familiarity:
artist_hotttnesss:
artist_latitude: latitude
artist_longitude: longitude
release_name: album name
analysis_sample_rate:
danceability: (not currently available)
duration: in seconds
end_of_fade_in: seconds at the beginning of the song
energy: energy from listener point of view
song_key: key the song is in
key_confidence: confidence measure
loudness: overall loudness in dB
m_mode: major or minor
mode_confidence: confidence measure
song_hotttnesss: algorithmic estimation
start_of_fade_out: start time of the fade out, in seconds
tempo: estimated tempo in BPM
time_signature: usual number of beats per bar
time_signature_confidence: confidence of time signature estimation

Figure 4: MSD database

For the purposes of ‘blending’ the MSD data with the survey data, the attributes selected from the msd database were as follows:

artist_familiarity (AF)

artist_hotttnesss (AH)

duration (D)

end_of_fade_in (FI)

key_confidence (KC)

loudness (L)

mode_confidence (MC)

song_hotttnesss (SH)

tempo (T)

time_signature_confidence (TSC)

The feature selected from the survey database was the ‘feeling’ (F) assigned to each song by a majority vote of the survey respondents.

3.4 Technical Approach

The Million Song Data Set subset of 10,000 songs was downloaded (1% of the full dataset, 1.8GB compressed) to obtain the 100 songs for the survey set. The dataset is in *.gz format so conversion software had to be downloaded.

As there were potentially 53 fields associated with each song, some attribute reduction was applied, initially excluding array type attributes and then removing most of the id columns as they are not descriptive attributes.

Once the results of the survey were returned, the data had to be visualized, and similarity measures were identified and defined with regard to the demographics of the respondents.

It was necessary to use multiple strategies to generate multiple rankings on which to base possible user preferences for example, the preference of the respondent for lyrical content, or rhythm, or both, in a song might have an effect on their recommendations.

Backups were regularly performed on the data by saving to the cloud at least once a day.

3.5 Implementation

Because of difference in ranges of values between the different attributes it was necessary to standardise the data before carrying out any analysis. There are several methods of doing this including the standardize function in Excel and min-max normalisation. Min-Max normalisation was used in Excel for the kNN analysis and the 'scale' function in R was used for the kMeans analysis. Min-max normalisation involves subtracting the minimum value, $Xmin$ of an attribute, from X and dividing by the range of values of that attribute:

$$\boldsymbol{Normalised\ Xn = \frac{X - Xmin}{Xmax - Xmin}}$$

This returns a value between 0 and 1 for every value in the dataset.

The 'scale' function in R produces a z-score type scaling:

```
MSD_z <- as.data.frame(lapply(MSD1, scale))
```

Euclidean Distance

Both kNN and kMeans methods use the Euclidean distance formula to measure the distance between the data points:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

K Nearest Neighbour (kNN) Classifier

Combining the results of the survey with the MSD attributes resulted in each song being represented by an 11-tuple, (AF, AH, D, FI, KC, MC, SH, FO, T, TSC, F). The kNN classifier is a 'lazy learning' classifier which means that it is an instance-based technique and no model is produced. The algorithm kNN (D,d,k) has three steps:

1. Compute the distance between d and every example in D.
2. Choose k examples in D that are nearest to d, denote set by P($\subseteq$ D)
3. Assign d the class that is most frequent in P (or the majority class)

Figure 5 shows an example of how the algorithm works. When a new, unclassified item (in this case, a tomato) is introduced into the feature space, the distance is measured between it and every classified item in the feature space. Then it is assigned to the most frequently occurring class within the k nearest items (in this case, fruit).

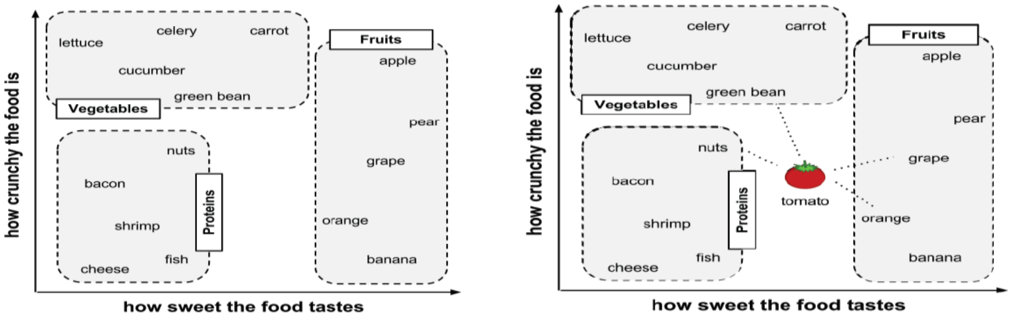


Figure 5: kNN example

(DWM - Lecture 4 , 2014)

kMeans Clustering Algorithm

To identify natural groupings within the MSD subset, a kMeans clustering algorithm was applied. kMeans is an unsupervised data mining technique, which means no training set is required and no model is produced as a result – it is often used for knowledge discovery about the data.

The kMeans algorithm k-Means (k, D) has 8 steps:

Choose k data points as the initial centroids (Figure 6)

```
1. Repeat
        a. For each datapoint , x €D do
        b. Compute distance from x to each centroids
        c. Assign x to the closest centroids
        d. Endfor
2. Recompute centroids for current cluster memberships (Figure 67&8)
    3. Until some stability criterion is met. (Figure 89)
```

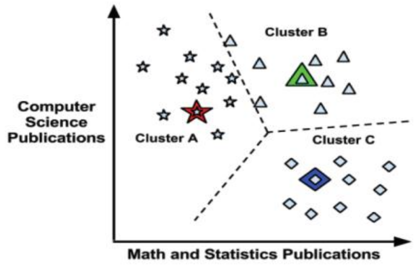

Cluster B
Computer Science Publications
Cluster A
Cluster C
Math and Statistics Publications

Figure 6: Initial clusters

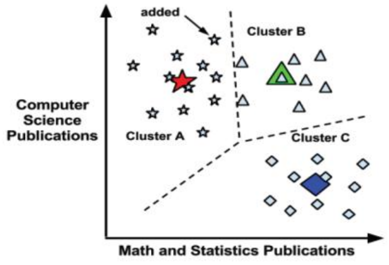

added
Cluster B
Computer Science Publications
Cluster A
Cluster C
Math and Statistics Publications

Figure 7: First update of centroids

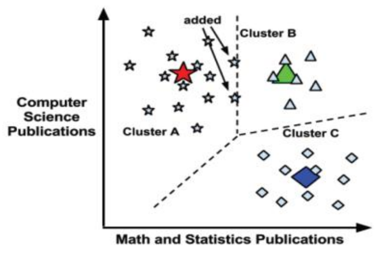

added
Cluster B
Computer Science Publications
Cluster A
Cluster C
Math and Statistics Publications

Figure 8: Second update of centroids

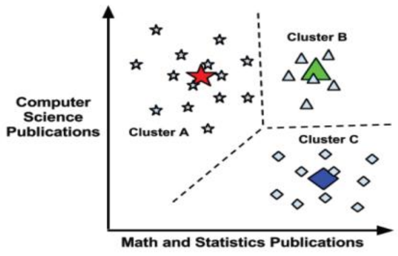

Cluster B
Computer Science Publications
Cluster A
Cluster C
Math and Statistics Publications

Figure 9: Final Clusters

(DWM - Lecture 6, 2014)

3.6 Requirements

This purpose of this section is to give a brief description of the original requirements (full document located in appendix) and explains any changes which may have taken place during the duration of the study.

3.6.1 Conduct Survey

Extract subset of the Million Song Dataset. Extract results from music survey.

- **Extract the Data**

  Download of the Million Song Subset.

  In addition to the original tasks specified in the original requirements specification (refer to appendix for full document), several more applications had to be downloaded in order to extract the songs and their attributes into the SQL database for analysis.
  In addition, the csv formatted results from the survey will have to be downloaded.
  The 18 respondent's results from the survey were returned in excel format, manually compiled into one excel file "Full_Survey_Results.csv", and loaded into the SQL database.

- **Compose Survey**

  The survey was composed in Excel and sent to sample population via Hotmail.

3.6.2 Analyse the Data

- **Prepare the Data.**

  Once the data was successfully extracted, it was necessary to clean and combine the two sets of data.

- **Preparing the Data**

  The csv files were uploaded into the MySQL database, addressing any omissions or errors.

- **Analysis of the Data**

  This use case describes the analysis of the prepared data in SQL Database by the Data Analyst (DA) in conjunction with a statistical expert to identify correlations between the songs attributes and the mood they are associated with by the survey respondents.

Report the Results

- **Report the Results.**

  Once the analysis is complete, the results will be reported.

- **Reporting the Analysis results.**

  The results will be communicated to the stakeholders via a written report and Powerpoint presentation.

3.6.3 Non-Functional Requirements

This area deals with any other particular non-functional attributes required by the system.

Performance/Response time requirement The response time, though not crucial, would ideally be within 1 minute of input.

Security requirement As the dataset is publicly available, there is no issue with security. Also, the survey respondents have been asked for permission to use their results within the scope of this project.

Portability requirement The system should be accessible from any portable device in addition to PC access.

Scalability requirement Extra features. It should be possible to upscale or downscale the system.

Resource utilization requirement A number of resources will be required for this project – Microsoft Excel, MySQL, Hotmail account, and access to SME. In addition to original requirements, Tableau was also required to be installed to the DA's laptop.

3.6.4 Interface requirements

In my project I have the following interfaces:

3.6.4.1 MySQL Workbench

This interface will be used to import the survey results, which are on excel worksheets, to the SQL database.

3.6.4.2 Hotmail

A Hotmail account will be required to send the survey out to the sample population.

3.6.4.3 WinPython

WinPython had to be downloaded from Sourceforge (Dice Holdings Inc., 2014) in order to provide the necessary libraries for importing the data into MySQL and to access the hdf5 files via hdf5 'getters'. The libraries required were Numpy, Math and Pytables which were needed to operate the hdf5 getters code, which was downloaded from Github (GitHub Inc., 2014) The Python libraries for connecting to MySQL were in Numpy.

3.6.4.4 Microsoft Excel

Microsoft Excel was used to create the survey. It was also required to load the responses into MySQL database. Microsoft Windows XP Professional Version 2002 Service Pack 3 is the system being used in this instance.

3.6.4.5 Tableau

Tableau is a powerful analytical and visualisation tool which is widely used in industry. Tableau can blend data from many different sources including SQL databases and Excel worksheets and as such was considered suitable for this analysis. It is capable of handling very large data sets. (Tableau, 2014)

3.6.4.6 Weka

Weka (Waikato Environment for Knowledge Analysis) is a software "workbench" for data mining, developed in the University of Waikato, New Zealand. It contains data mining algorithms and is suitable for very large datasets. It will be used to analyse the prepared data and to produce graphical representation of same. (University of Waikato)

Note:
Amazon Web Services (AWS) was initially considered to be necessary to analyse the main body of the MSD. However, after reducing the scope of the project to the 10,000 song subset and discovering the capabilities of Tableau, AWS was considered unnecessary to this analysis. (See appendix for original requirements specification).

3.7 Datasets

MSD dataset

The MSD is an open source dataset which due to its' size, enables realistic study and research and development of RSs than was previously possible. A subset of this dataset was downloaded into a MySQL database and prepared for analysis.

Survey dataset

The survey dataset was created from the individual survey responses from the sample population. These responses, initially in .xlsx format, were uploaded to a MySQL database, aggregated, and output as a csv file for initial analysis.

4 Testing and Evaluation

This section describes the methods for systems testing and data analysis. It includes initial analyses of the MSD data and the survey results and kMeans clustering and kNN classification. Tableau, R and Weka were used for these analyses.

4.1 Initial inspection of ‘blended survey and MSD data in Tableau

The data from the MSD database was joined with the data from the survey database and output as a ‘blended’ csv file which was then uploaded to Tableau for analysis.

4.1.1 Sample Demographic

Figure 1010 shows the demographic of the 18 survey respondents. As there were more respondents aged 43 than any other age, this can be seen by the ‘spikes’ in the histogram representing the ‘feelings’

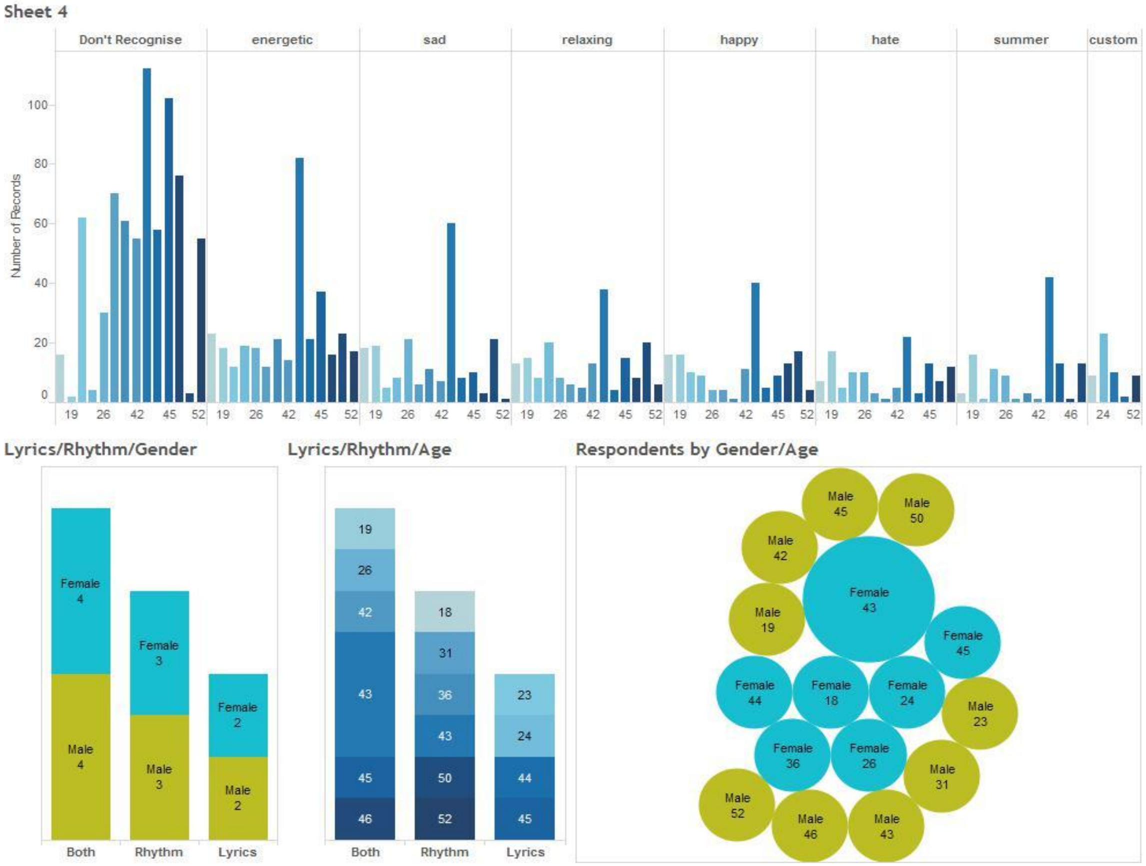


Figure 10: Sample User Demographic

Gender was equally represented in the responses with the respondents’ preferences for lyrics, rhythm, and both, being equally represented as illustrated in the bar chart in the bottom left hand side of Figure 10.

4.2 Initial inspection of MSD data in Tableau and R

The 10,000 songs and their attributes were exported from the MySQL ‘msd’ database and uploaded to Tableau for initial analysis and to R to check for correlation.

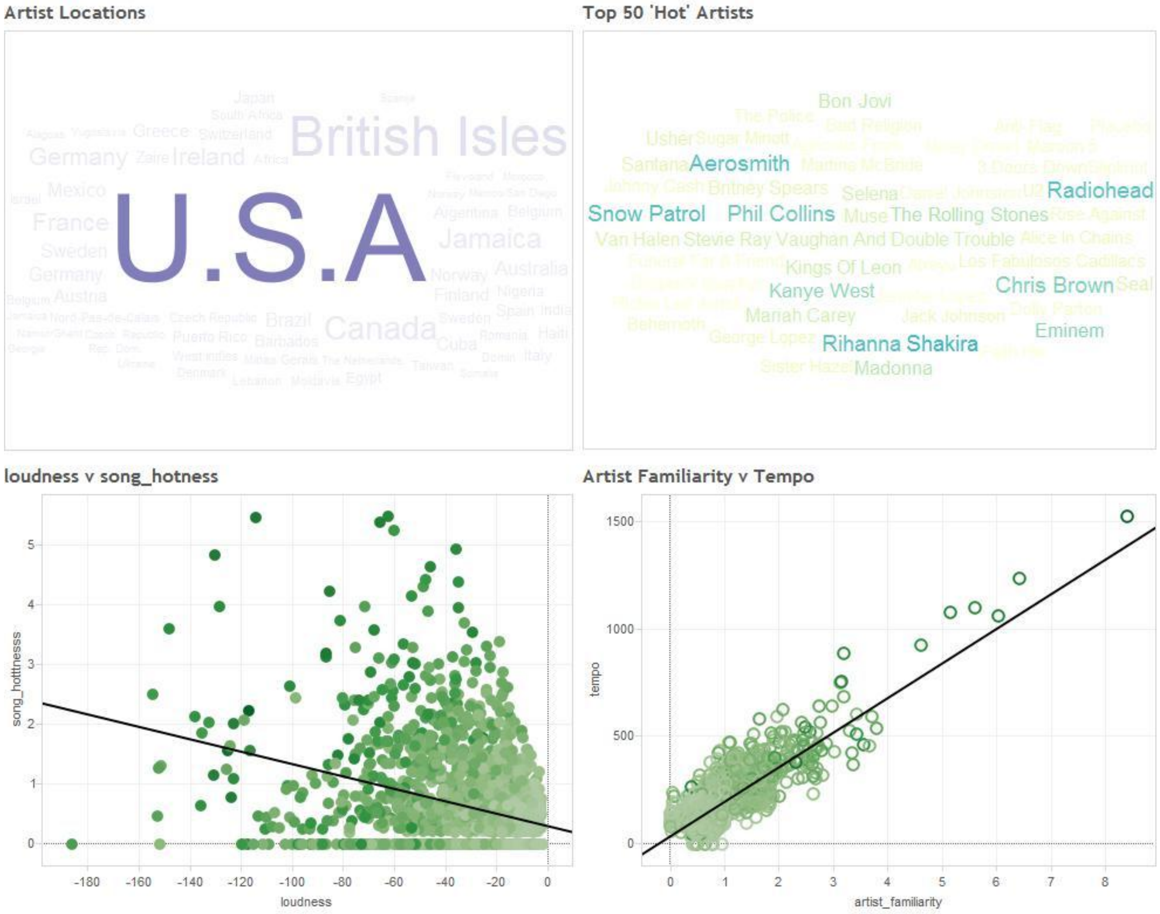


Figure 11: Tableau Visualisations of MSD 10,000 song subset

Tableau Analysis

From the Wordcloud in Figure 1111 it can be seen that the most frequently occurring artist location is the U.S.A., with the British Isles (including Wales and Scotland) a clear second place, and Ireland, Jamaica, Canada, Germany and

France following behind. Size and depth of colour correspond to the number of songs in the dataset by an artist from each location.

The top 5 in the 50 ‘Hot’ artists are Shakira, Snow Patrol, Aerosmith, Rihanna and Radiohead according to the ‘artist_hotttnesss’ attribute in the MSD and the corresponding Wordcloud can be seen in Figure 1111.

Trend Lines Models

Two examples of graphs from the MSD attributes can also be seen in Figure 1111 with loudness having a weak inverse relationship with song hotness as described by the following result:

P-value:	< 0.0001
Equation:	song_hotttnesss = -0.0103744*loudness + 0.284561

Coefficients

Term	Value	StdErr	t-value	p-value
loudness	-0.0103744	0.0006099	-17.0102	< 0.0001
intercept	0.284561	0.0213342	13.3382	< 0.0001

This graph also represents tempo by depth of colour, with a higher value being represented by a deeper colour.

The linear trend model in the bottom r.h.s. of Figure 1111 was computed in Tableau for sum of tempo given sum of artist_familiarity. Song duration is represented by depth of colour with longer durations having a deeper colour. The model may be significant at p <= 0.05.

Model formula:	(artist_familiarity + intercept)
Number of modeled observations:	4501
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	4499
SSE (sum squared error):	1.19324e+007
MSE (mean squared error):	2652.23
R-Squared:	0.656453

Standard error: 51.4998

p-value (significance): < 0.0001

Individual trend lines:

Table 2: Correlation Analysis results

Interpretation of the Result:

It is clear from the correlation analysis in Table 2, that there is a moderate to strong positive relationship between tempo and artist familiarity as the R-Squared value is 0.656453.

4.3 Building a Linear Model in R

Correlation Analysis

Firstly, the attributes were checked for multi-collinearity:

```
> cor(msdNumeric)
                          artist_familiarity artist_hotttnesss     duration
song_id                         0.0013007403       0.012101662 -0.003293714
artist_name                     0.0055724635      -0.005447061 -0.002960727
song_title                      0.0109645010       0.010001812  0.003605418
artist_familiarity              1.0000000000       0.852876831  0.070734909
artist_hotttnesss               0.8528768310       1.000000000  0.058613852
duration                        0.0707349086       0.058613852  1.000000000
end_of_fade_in                 -0.0387813119      -0.017478457  0.039255559
song_key                        0.0112516393       0.013900940 -0.012213488
key_confidence                 -0.0189938532      -0.010219757 -0.033824765
loudness                        0.1041693158       0.077019675  0.058099309
m_mode                         -0.0266400612      -0.019407881 -0.067726753
mode_confidence                -0.0212047527      -0.006076516 -0.049092219
song_hotttnesss                 0.4305003320       0.424557881  0.045483136
start_of_fade_out               0.0729363457       0.059557043  0.966980989
tempo                          -0.0004867632      -0.005932106 -0.014707993
time_signature                  0.0603201300       0.053408423  0.072252390
time_signature_confidence       0.0296306825       0.030791341  0.098220003
```

Table 3: R Correlation results

From the extract of results obtained and shown in Table 3, it is clear that 'start_of_fade_out' and 'duration' are highly correlated, with a value of 0.97, so it was not necessary to use both in the analysis and 'start_of_fade_out' was removed.

Similarly, 'artist_hotttnesss' and 'artist_familiarity' were highly correlated with a value of 0.853, so 'artist_hotttnesss' was removed.

'mode_confidence' and 'key_confidence' were also correlated with a value of 0.78, so 'mode_confidence' was removed.

Convert to dataframe from matrix:

In order to construct a linear model it was necessary to first convert the matrix 'msdNumeric' Into a dataframe:

msdDF <- data.frame(msdNumeric)

Build Linear Model

Then the linear model was built with the dependent variable defined as 'song_hotttnesss' and the independent variables of 'artist_familiarity', 'song_key', 'time_signature', 'tempo' and 'duration'.

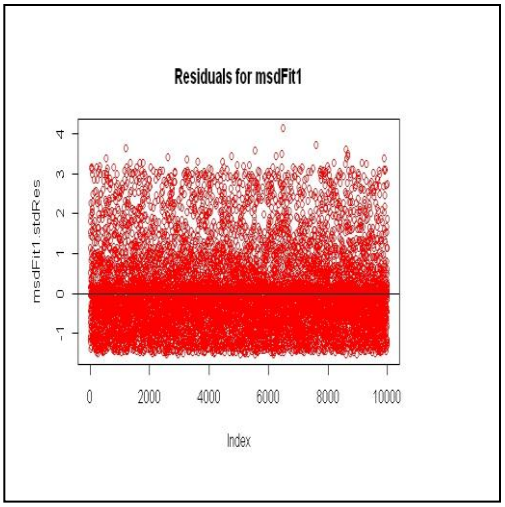
```
msdFit <-
lm(song_hotttnesss~artist_familiarity+song_key+time_signature+tempo+duration,
data=msdDF)

Call:
lm(formula = song_hotttnesss ~ artist_familiarity + song_key +
   time_signature + tempo + duration, data = msdDF)

Coefficients:
      (Intercept)  artist_familiarity           song_key
       -1.343e+02           1.739e-01          5.769e-01
```

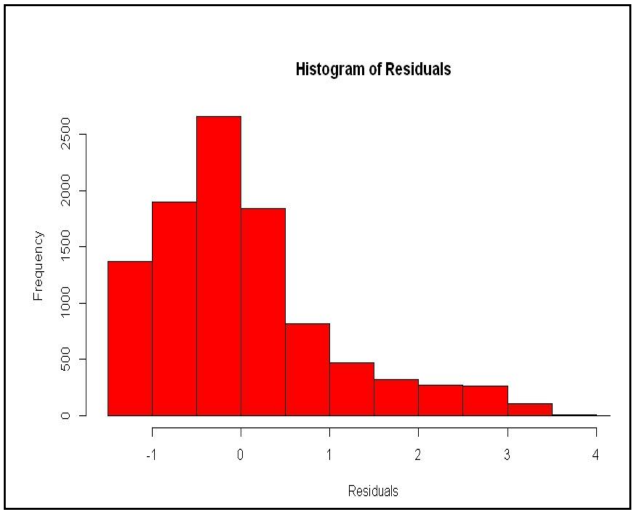
Table 4: Linear Model Results

The resulting linear model, shown in Table 4 is not effective as the independent variables have negligible effect on the dependent variable.

Check the Residuals

```
msdFit.stdRes <- rstandard(msdFit)
> plot(msdFit.stdRes)
> plot(msdFit.stdRes, col='red')
> abline(0,0)
```





The model is a good fit if the residuals fall within 2 standard deviations of the mean. From the Residuals plot shown above we can see that although most of the residuals are close to zero, there are more residuals than desired.

To gain a better view of the amount of residuals that fall outside 2 standard deviations, the Histogram above shows that although these are in the minority and therefore the model is an acceptable fit.

Normality Plot

A common test for normality is the qq plot.

```
qqnorm(msdFit.stdRes, ylab="Standardised Residuals", xlab = "Normal
Scores", main="Normality Plot", col="blue")qqline(msdFit.stdRes)
```

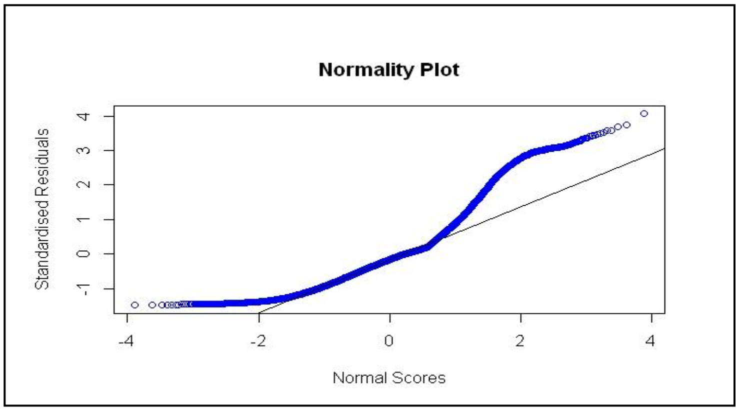


Figure 12: Normality Plot

If the residuals fall in a straight line, that means the normality condition is met – not strictly met here, as illustrated in Figure 12, for the chosen variables, although a slight banana shape may be acceptable.

New Model with 6 variables, chosen from normality plots. Table 5 shows the results from the new correlation analysis.

```
msdFit1 <- lm(song_hotttnesss~artist_familiarity+time_signature+tempo+duration+
loudness+key_confidence,data=msdDF)
>
> msdFit1

Call:
lm(formula = song_hotttnesss ~ artist_familiarity + time_signature +
   tempo + duration + loudness + key_confidence, data = msdDF)

Coefficients:
     (Intercept)  artist_familiarity      time_signature
      -1.757e+02          1.722e-01          9.767e+00
          tempo            duration          loudness
      -2.393e-04          3.262e-03          9.783e-03
   key_confidence
       3.028e-02
```

Table 5: Correlation with 6 variables

Test conclusion: There was no difference in residual or normality plots using these variables

4.4 kMeans Analysis in R

A kMeans clustering analysis was performed on the 10,000 song subset in R to investigate natural groupings that might occur within the data. Initially, R was reading the file's attributes as factors so they had to be converted to numeric:

```
MSDNumeric <- data.matrix(MSD, rownames.force=NA)
```

Then, it had to be converted to a data frame (Table 65) before the clustering could proceed:

```
MSD1<-data.frame(MSDNumeric)
```

```
> str(MSD1)
'data.frame':  10001 obs. of  15 variables:
song_id                   : int  1 1113 2224 3335 4446 5557 6668 7779
artist_familiarity        : int  3717 2482 2984 1453 2982 3153 1999
artist_hotttnesss         : int  3481 2339 2560 1266 2933 2334 2077 311
duration                  : int  4244 2619 718 1444 3028 2342 3955 151
end_of_fade_in            : int  176 184 86 217 1 8 847 37 253 1259
song_key                  : int  8 2 9 11 1 5 8 2 7 7 ...
key_confidence            : int  555 736 169 643 751 93 635 1 1 717 ...
loudness                  : int  3791 602 7347 7255 6890 4280 7057
m_mode                    : int  2 1 1 2 2 2 2 2 1 2 ...
mode_confidence           : int  471 607 401 536 720 342 528 1 131 623
song_hotttnesss           : int  1068 1094 1 1 1 1108 1 1 1 279 ...
start_of_fade_out         : int  6546 4286 741 2019 4191 3272 5847 216
tempo                     : int  5831 8521 2117 34 1852 3032 4528$
time_signature            : int  5 4 4 2 4 4 3 2 3 4 ...
time_signature_confidence: int  118 773 379 1 1 557 449 1 403 482 ...
```

Table 6: Dataframe Conversion

Because there were a wide range of differences in the attribute values, as shown in Table 6, the numeric data had to be normalized via the 'scale' function in R:

```
MSD_z <- as.data.frame(lapply(MSD1, scale))
```

```
> str(MSD_z)
```

Then kMeans clustering was performed:

Initially with k=6:

```
MSD_clusters<-kmeans(MSD_z, 6)
```

Check size of clusters:

```
> MSD_clusters$size
[1] 1670 1823 1634 1797 1153 1924
16.69%, 18.23%, 16.34%, 17.97%, 11.53%, 19.24%
```

These were the songs at the centre of each of the 6 clusters:

1	Delroy Wilson	Half Way Up The Stairs
2	Bill Perry	Man On The Side
3	Galactic	Cafe deClouet
4	The Charms	Ascolta mio dio
5	Hevia	Albo
6	A Static Lullaby	The Everlasting Gaze (A Static Lullaby)

Specifying 7 clusters resulted in the following songs which were different again to those produced by k=6 and other values of k:

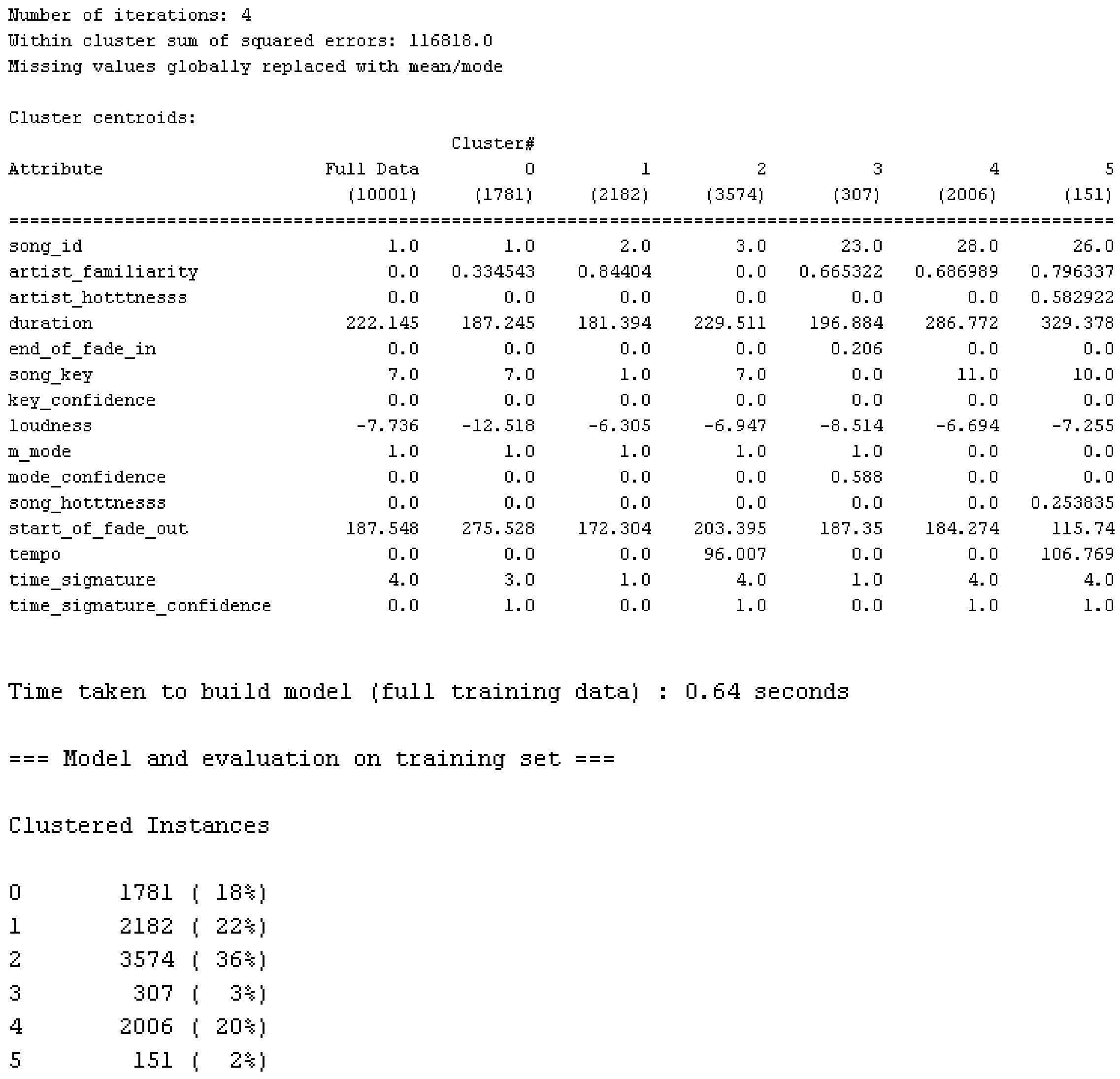
```
> MSD_clusters<-kmeans(MSD_z, 7)
> MSD_clusters$size
[1] 1645 1449 1576 1489 1070 1393 1379
```

1	Azukx	124 Stomp
2	Casiotone For The Painfully Alone	Town Topic (Instrumental)
3	Bizzy Bone Presents	Intro
4	Leon Russell	Out In The Woods
5	R.E.M.	Final Straw (Album Version)
6	Agnostic Front	Come Alive
7	Richie McDonald	If Every Day Could Be Christmas

4.5 KMeans in Weka

The kMeans algorithm was then applied in Weka to compare and test the results achieved in R.

K=6



```
Number of iterations: 4
Within cluster sum of squared errors: 116818.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                         Cluster#
Attribute                    Full Data          0          1          2          3          4          5
                               (10001)     (1781)     (2182)     (3574)      (307)     (2006)      (151)
=========================================================================================================
song_id                            1.0        1.0        2.0        3.0       23.0       28.0       26.0
artist_familiarity                 0.0   0.334543    0.84404        0.0   0.665322   0.686989   0.796337
artist_hotttnesss                  0.0        0.0        0.0        0.0        0.0        0.0   0.582922
duration                       222.145    187.245    181.394    229.511    196.884    286.772    329.378
end_of_fade_in                     0.0        0.0        0.0        0.0      0.206        0.0        0.0
song_key                           7.0        7.0        1.0        7.0        0.0       11.0       10.0
key_confidence                     0.0        0.0        0.0        0.0        0.0        0.0        0.0
loudness                        -7.736    -12.518     -6.305     -6.947     -8.514     -6.694     -7.255
m_mode                             1.0        1.0        1.0        1.0        1.0        0.0        0.0
mode_confidence                    0.0        0.0        0.0        0.0      0.588        0.0        0.0
song_hotttnesss                    0.0        0.0        0.0        0.0        0.0        0.0   0.253835
start_of_fade_out              187.548    275.528    172.304    203.395     187.35    184.274     115.74
tempo                              0.0        0.0        0.0     96.007        0.0        0.0    106.769
time_signature                     4.0        3.0        1.0        4.0        1.0        4.0        4.0
time_signature_confidence          0.0        1.0        0.0        1.0        0.0        1.0        1.0


Time taken to build model (full training data) : 0.64 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       1781 ( 18%)
1       2182 ( 22%)
2       3574 ( 36%)
3        307 (  3%)
4       2006 ( 20%)
5        151 (  2%)
```

Table 7: Cluster Centroids and sizes for k=6

As shown in Table 7, there are 4 main clusters into which 96% of the data fall. The remaining clusters are very small, although with more than one instance, which can mean they are possible meaningful clusters.

The corresponding cluster centroids are listed below:

0	Mastodon	Deep Sea Creature
1	Casual	I Didn't Mean To
2	The Box Tops	Soul Deep
3	Alice Stuart	Kassie Jones
4	Clp	Superconfidential
5	SUE THOMPSON	James (Hold The Ladder Steady)

K=7

The centroids are:

0	Mastodon	Deep Sea Creature
1	Sonora Santanera	Amor De Cabaret
2	The Box Tops	Soul Deep
3	Alice Stuart	Kassie Jones
4	Clp	Superconfidential
5	SUE THOMPSON	James (Hold The Ladder Steady)
6	Casual	I Didn't Mean To

Result:

The same centroids are showing up in Weka for every value of k, whereas the centroids resulting from kMeans in Weka are different every time. The cluster sizes resulting from the R analysis are more evenly distributed, whereas the cluster sizes in Weka have a wide variance as shown in Figure 13 and Figure 14 below.

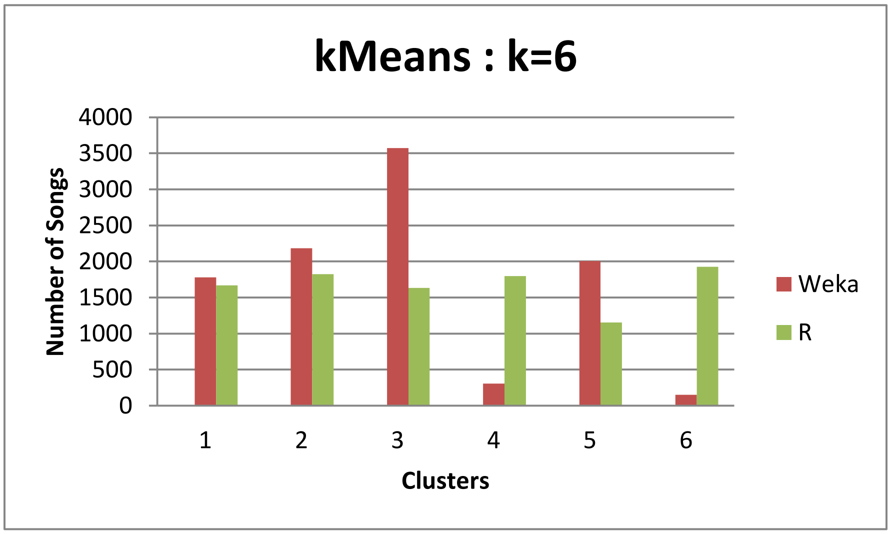


Figure 13: kMeans Results, k=6

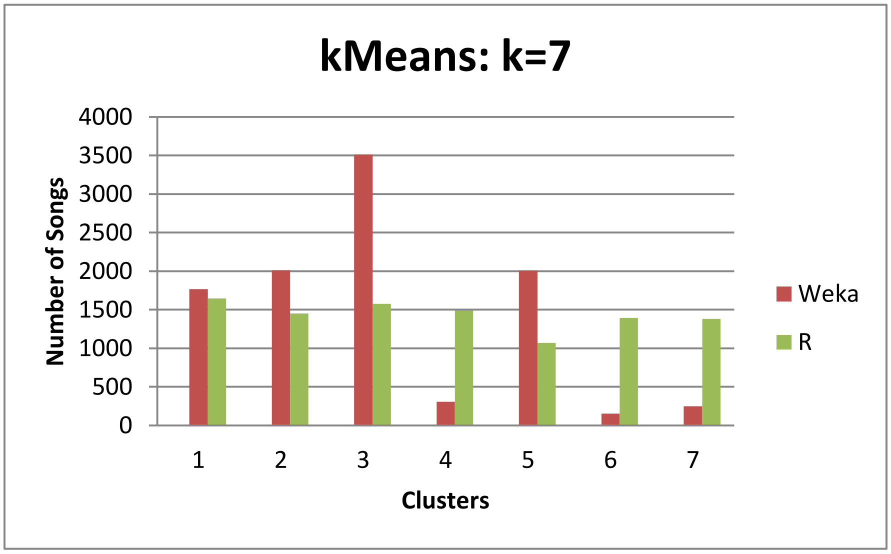


Figure 14: kMeans Results, k=7

kNN Process

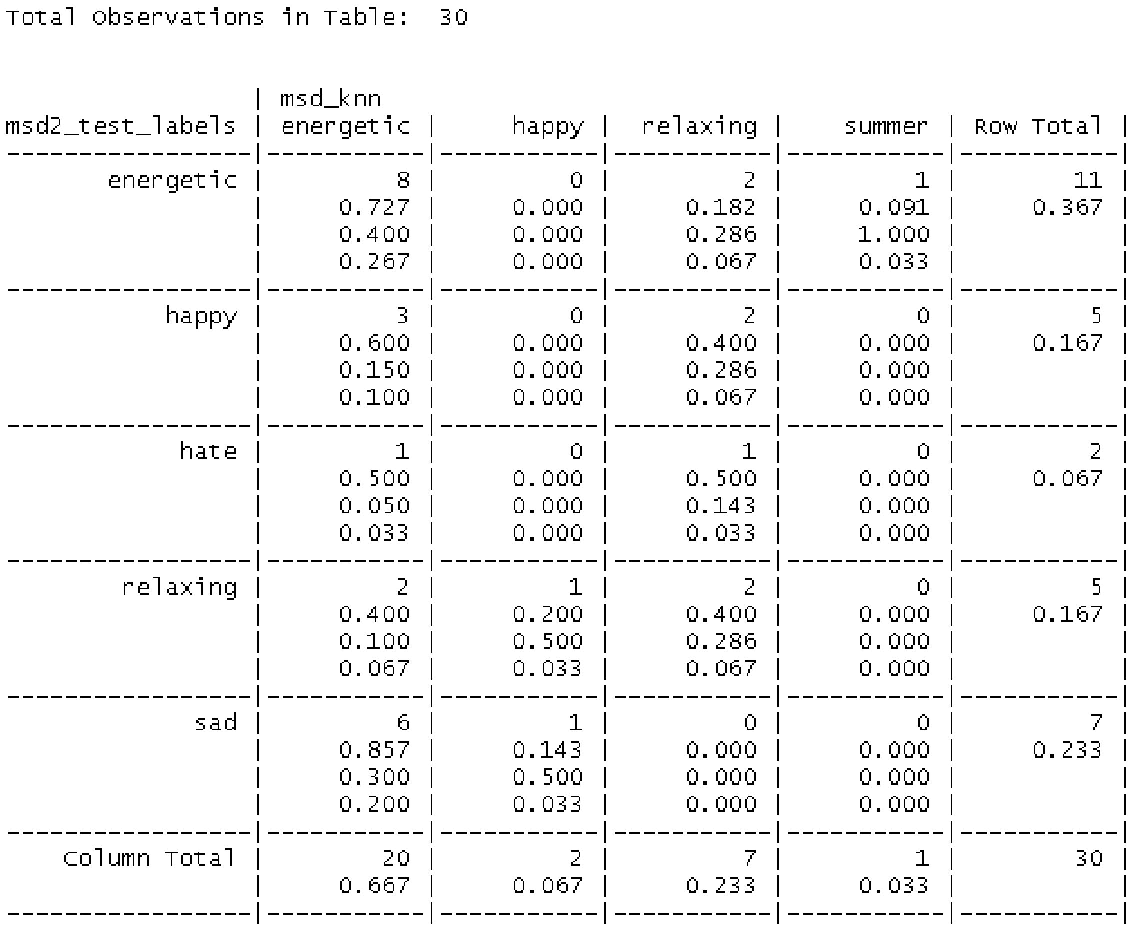
Before applying kNN, it was necessary for the data to be randomised as the rows had originally been in order. This was done with the ‘sample’ function in R:

```
> msd2 <- msd1 [sample (nrow(msd1)),]
```

Then the data had to be split into a classified training set and a test set with the ‘class’ or target ‘value’ unlabelled (in this case, the ‘feeling’ variable).

Initially the training set had 70 instances and the test set had 30 instances.

K=6:



```
Total Observations in Table:  30 

 
                  | msd_knn 
msd2_test_labels  | energetic |     happy |  relaxing |    summer | Row Total | 
------------------|-----------|-----------|-----------|-----------|-----------|
        energetic |         8 |         0 |         2 |         1 |        11 | 
                  |     0.727 |     0.000 |     0.182 |     0.091 |     0.367 | 
                  |     0.400 |     0.000 |     0.286 |     1.000 |           | 
                  |     0.267 |     0.000 |     0.067 |     0.033 |           | 
------------------|-----------|-----------|-----------|-----------|-----------|
            happy |         3 |         0 |         2 |         0 |         5 | 
                  |     0.600 |     0.000 |     0.400 |     0.000 |     0.167 | 
                  |     0.150 |     0.000 |     0.286 |     0.000 |           | 
                  |     0.100 |     0.000 |     0.067 |     0.000 |           | 
------------------|-----------|-----------|-----------|-----------|-----------|
             hate |         1 |         0 |         1 |         0 |         2 | 
                  |     0.500 |     0.000 |     0.500 |     0.000 |     0.067 | 
                  |     0.050 |     0.000 |     0.143 |     0.000 |           | 
                  |     0.033 |     0.000 |     0.033 |     0.000 |           | 
------------------|-----------|-----------|-----------|-----------|-----------|
         relaxing |         2 |         1 |         2 |         0 |         5 | 
                  |     0.400 |     0.200 |     0.400 |     0.000 |     0.167 | 
                  |     0.100 |     0.500 |     0.286 |     0.000 |           | 
                  |     0.067 |     0.033 |     0.067 |     0.000 |           | 
------------------|-----------|-----------|-----------|-----------|-----------|
              sad |         6 |         1 |         0 |         0 |         7 | 
                  |     0.857 |     0.143 |     0.000 |     0.000 |     0.233 | 
                  |     0.300 |     0.500 |     0.000 |     0.000 |           | 
                  |     0.200 |     0.033 |     0.000 |     0.000 |           | 
------------------|-----------|-----------|-----------|-----------|-----------|
     Column Total |        20 |         2 |         7 |         1 |        30 | 
                  |     0.667 |     0.067 |     0.233 |     0.033 |           | 
------------------|-----------|-----------|-----------|-----------|-----------|
```

Table 8: kNN Result, k=6

The mis-classification rate was calculated as 67%:

```
mean(msd2_test_labels != msd_knn)
[1] 0.6666667
```

K=9 resulted in:

```
Mis-classification rate:

> mean(msd2_test_labels != msd_knn)
[1] 0.6
```

A 60/40 train/test split yielded no better results:

K=3:

```
table(msd2_test_labels, msd_knn)
                msd_knn
msd2_test_labels energetic happy hate relaxing sad summer
       energetic         9     1    1        3   0      1
       happy             2     0    0        2   0      1
       hate              3     0    0        0   1      0
       relaxing          2     2    0        1   1      2
       sad               5     0    1        1   1      0
       summer            0     0    0        0   0      0

mean(msd2_test_labels != msd_knn)
[1] 0.725
```

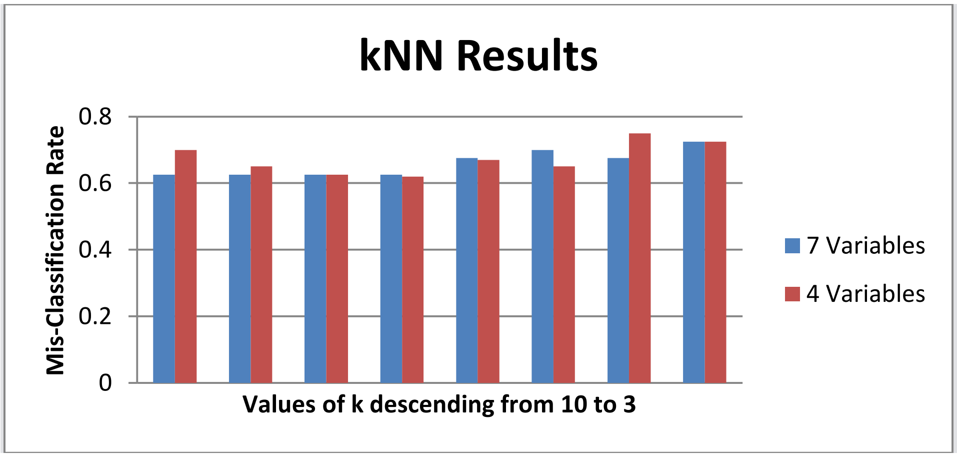


Figure 15: Graph of kNN results

Result:

The kNN algorithm has classified the ‘energetic’ category correctly most of the time for different values of k. This would indicate that the MSD attributes are good for classifying the energy of a song and also that the ‘energetic’ category in the survey data was easy to define from the participants’ perspectives

4.3 Customer/User Testing

4.3.1 Survey Feedback

Feedback was supplied by some survey respondents which will be helpful in shaping future work:

Quote from someone who started/but couldn't finish the questionnaire.

"I'm very sorry Michele, while I was very interested in the survey, I found that it was somewhat daunting in scale. Maybe a shorter survey could be developed based on the findings of this survey. The shorter survey might focus on some key indicator?”

Quote from respondent who did the survey but did not listen to all songs:

“Did as many as I could but there are lots of oddball songs in here so even when I know a singer I don’t always recognise the song. Listened to 40 or so anyway.”

Quote from respondent indicating that the categories need to be better defined:

“I have completed the survey to the best of my ability. Where I did not know the songs I clicked on the link and listened to it. I found myself changing between Happy and Energetic a lot. Happy and Summer music also a challenge. I must say that I was probably biased in terms of Rock & Roll Us bands as I tend to think about these as Summer Music probably because of the better weather they have on the videos and the Slane etc. I hate those Rap songs and this comes across in the survey.”

Notes regarding survey:

The ‘Don’t Recognise’ option wasn’t very helpful and despite the hyperlinks, many respondents took this option. Therefore this category will not be an option in future iterations of the study.

Some respondents had neglected to fill in some of the fields, so these were populated with an ‘8’for ‘Don’t Recognise’.

Four respondents used the ‘put your own category here if you require’ field with:

- ‘”Not my type of music”
- “Hard to define”
- “Indifferent/boring/wouldn’t listen to it again”
- “Like this song”

These fields were not used in this study but will be considered in combination with feedback from the survey respondents for future work.

4.3.2 Results

The results from the kNN analysis lead to the conclusion that a new survey needs to contain more clearly defined ‘feelings’ or emotional response choices to the tracks. The kMeans analysis shows that there are definitely 6 or 7 clusters which the songs naturally form. Further tasks to complete include:

- Creating and distributing new survey to participants
- Researching and sourcing alternative datasets to the MSD, with different attributes
- Performing a combination of analyses on the resulting data
- Evaluating and reporting the new resulting system

5 Conclusions

There are no conclusive results as yet from this preliminary analysis. There is a clear connection between some of the attributes in the MSD and there is also an accurate classification for the ‘energetic’ category. Further investigation is required to deliver more accurate classifications for other categories to be decided in conjunction with the users (participants in the survey). The scope of the project has changed somewhat from the initial project plan (see appendix for complete document) due to a more thorough analysis being required to develop more accurate classifications for the ‘feelings’ category.

6 Further development or research

Including contextual elements has been a more recent development in the realm of Recommender Systems and includes considerations such as where the user is located, who they are with, even heart rate and temperature with the introduction of smartphones that measure heart rate and temperature.
(Reference Dr. Dre and Spotify)

There are many challenges to the perfect recommender system, not least of which is the system’s current inability to identify if there are several users on a single account for example in a family, perhaps a father and a child or 2 use the mother ‘s account or vice versa. Currently this skews a recommenders ability to predict songs according to a user’s tastes, whereas if the system could identify that there are actually 3 users on the account, it would be able to provide 3 different relevant playlists rather than a ‘composite’ (of course, it could also recommend that the extra users open their own accounts!)

From a prioritisation viewpoint this project concentrated on reporting and presenting the results to date in the analysis. The plan is to complete the remaining tasks after the dissertation presentation and this will include reporting the results to the survey participants.

7 References

Bibliography

Adomavicius, G. M., & Tuzhilin, A. M. (2005, June). *IEEE Computer Society.* Retrieved Feb 7, 2014

Adomavicius, G., & Jannach, D. (2013). Preface to the Special Issue on Context-Aware Recommender Systems. *AI Magazine* , pp. 1-6.

Aiolli, F. (2013). A Preliminary Study on a Recommender System for the Million Song Dataset Challenge. *Roberto Basili; Fabrizio Sebastiani 0001 & Giovanni Semeraro, ed., 'IIR'* , 73-83.

Bertin-Mahieux, T., P.W.Ellis, D., Whitman, B., & Lamere, P. (2011). *Million Song Dataset.* Retrieved February 19, 2014, from http://labrosa.ee.columbia.edu/millionsong/

Boulkrinat, S., Hadjali, A., & Mokhtari, A. (2013). *Enhancing Recommender Systems Prediction Through Qualitative Preference Relations.* Algiers: Programming and Systems (ISPS) 2013, 11th International symposium on.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). *Using Collaborative Filtering to weave an information Tapestry.* Xerox Palo Alto Research Center. San Diego: Association for Computing Machinery.

Hornung, T., Ziegler, C.-N., Franz, S., Przyjaciel-Zablocki, M., Schatzle, A., & LAusen, G. (2014). *Evaluating Hybrid Music Recommender Systems.* Munich: 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT).

Hung-Chen, C., & Chen, A. L. (2014). *Association for Computing Machinery.* Retrieved April 25, 2014, from dl.acm.org: http://dl.acm.org/citation.cfm?id=502625

Hyun-Tae, K., Eungyeong, K., Jong-Hyun, L., & Chang, W. A. (2010). *A Recommender System Based on GeneticAlgorithm for Music Data.* Suwon, South Korea: 2010 2nd International Conference on Computer Engineering and Technology.

International Institute of Business Analysts. (2009). *A Guide to the Business Analysis Body of Knowledge (BABOK Guide) Version 2.* Ontario: IIBA.

Jayalakshmi, D., Shruthi, J., Sneha, S., & Uttarika Ratnakar, S. (2014). *A Hybrid Music Recommender System.* Retrieved April 25, 2014, from www.academia.edu: http://www.academia.edu/4697281/A_Hybrid_Music_Recommender_System#

Jones, N., & Pu, P. (2009). *User Acceptance Issues in Music Recommender Systems.* The Human Computer Interaction Group. Lausanne: Swiss Federal Institute of Technology.

labrosa.ee. (2011). Retrieved February 19, 2014, from http://labrosa.ee.columbia.edu/millionsong/

Liu, D.-R., & Shih, Y.-Y. (2004). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information And Management , I* (42), 387-400.

Lynskey, D. (2013, November 10). *www.theguardian.com.* Retrieved February 19, 2014, from http://www.theguardian.com/technology/2013/nov/10/daniel-ek-spotify-streaming-music

MindGenius. (2014). *www.mindgenius.com*. Retrieved March 8, 2014, from http://www.mindgenius.com/?_kk=mindgenius&_kt=d38e0e9d-4e25-4730-a2eb-6f625a811669&gclid=CJ378oHNl70CFaKx2woddbUAYQ

MusicRow. (2014). *www.musicrow.com*. Retrieved April 23, 2014, from http://www.musicrow.com/2011/01/ss-debuts-digital-song-genre-rpts/

O'Loughlin, E. (2011). *Youtube*. Retrieved Feb 6, 2014, from http://www.youtube.com/watch?v=sA67g6zaKOE

Resnick, P., & Varian, H. R. (1997). Recommender Systems. *Communications of the ACM , XL* (3), 3.

Sommestad, K. (2009). *www.spotiseek.com*. Retrieved February 19, 2014, from http://www.spotiseek.com/mixtape/moby/popular/

University of Waikato. (n.d.). *weka*. Retrieved February 27, 2014, from waikato.ac.nz: http://www.cs.waikato.ac.nz/ml/weka/

Van Den Oord, A., Dieleman, S., & Schrauwen, B. (2013). *Neural Information Processing Systems Foundation.* Retrieved April 25, 2014, from www.nips.cc: https://nips.cc/Conferences/2013/Program/event.php?ID=4028

Yoshii, K., Goto, M., & Ogata, T. (2008). An Efficient Hybrid Music Recommender System Using an Incrementally Trainable Probabilistic Generative Model. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING* , 435-447.

Acknowledgements:

Thank you to all the survey respondents for taking the time to complete and return the survey. Thank you to friends and family for support, advice and patience. Thank you to Podge for your expert guidance.

8 Appendix

8.1 Project Proposal

Objective

The Objective of this project is to build a model that will predict songs based on an individual’s previous choices. It will endeavour to make better predictions than are currently being performed on sites like Spotify and its’ offshoots as described below.

Another objective is an exploration of Clustering, Classification, Link prediction, Data reduction, Association, Multiple linear regression, and other Data Mining, Data Analysis and Data Visualisation tasks.

It is also hoped that the scope can be expanded to include a recommender system for films, books, art, etc.

Background

Million Song dataset

“The Million Song Dataset is a freely available collection of audio features and metadata for a million contemporary popular music tracks.”

This dataset enables large-scale analysis on Music Information Retrieval (MIR) with a set of attributes for each song which includes tempo, loudness, artist hotness, danceability, etc.

Platinum Blue

Though not the primary objective of this analysis, a possible by-product could be the identification of what makes a hit. Many attempts have been made to predict what makes one song a hit and one not. One company that seems to be making headway in this area is Platinum Blue.

Using data for every hit song since the 60’s, they were all plotted three-dimensionally, hence leading to the discovery of “clusters”. This led to the discovery that 80% of all top hits shared a relatively small number of common elements. Another interesting discovery was that songs in the same cluster didn’t necessarily sound the same. This leads to the possibility that the link between the songs is mathematical.

Though Platinum Blue’s technology can’t yet predict a hit, they can predict with some confidence what will not be a hit.

http://www.theguardian.com/music/2006/nov/11/popandrock.news [accessed 19/2/14]

Oliver Burkeman, The Guardian, 11 November 2006

Spotify

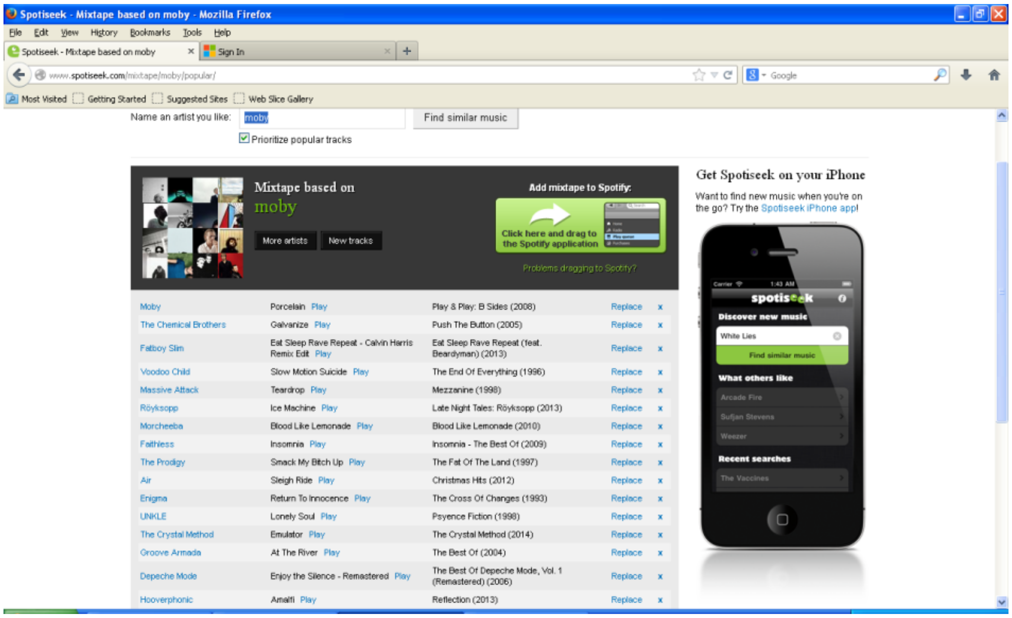
Spotify was started 8 years ago in Sweden by Daniel Ek, a music fan and programming genius. He approached record labels with the radical/novel idea to provide music to rent and for free instead of buying it! Despite being understandably wary, he managed to persuade them that Spotify would be the music industry's saviour rather than completely destroying it. As of October 2013, on the 5th anniversary of its launch, more than 20million songs were made available to 24 million users in 52 territories. One billion playlists have been created so far.

(Lynskey, 2013)

Spotiseek

There are many Spotify playlist generators available, for example, filtr.com, Spotibot.com, lazify.nl, and many more. I have tested some to see how relevant the returned suggestions were, and although the results were relevant enough, I think they can be improved upon. Spotiseek helps users to find music similar to their favourite artists and creates "mixtape" playlists for Spotify.

(Sommestad, 2009)



Name an artist you like: moby

☑ Prioritize popular tracks

Mixtape based on moby

Get Spotiseek on your iPhone

Want to find new music when you're on the go? Try the Spotiseek iPhone app!

White Lies

Moby	Porcelain Play	Play & Play: B Sides (2008)	Replace	x
The Chemical Brothers	Galvanize Play	Push The Button (2005)	Replace	x
Fatboy Slim	Eat Sleep Rave Repeat - Calvin Harris Remix Edit Play	Eat Sleep Rave Repeat (feat. Beardyman) (2013)	Replace	x
Voodoo Child	Slow Motion Suicide Play	The End Of Everything (1996)	Replace	x
Massive Attack	Teardrop Play	Mezzanine (1998)	Replace	x
Röyksopp	Ice Machine Play	Late Night Tales: Röyksopp (2013)	Replace	x
Morcheeba	Blood Like Lemonade Play	Blood Like Lemonade (2010)	Replace	x
Faithless	Insomnia Play	Insomnia - The Best Of (2009)	Replace	x
The Prodigy	Smack My Bitch Up Play	The Fat Of The Land (1997)	Replace	x
Air	Sleigh Ride Play	Christmas Hits (2012)	Replace	x
Enigma	Return To Innocence Play	The Cross Of Changes (1993)	Replace	x
UNKLE	Lonely Soul Play	Psyence Fiction (1998)	Replace	x
The Crystal Method	Emulator Play	The Crystal Method (2014)	Replace	x
Groove Armada	At The River Play	The Best Of (2004)	Replace	x
Depeche Mode	Enjoy the Silence - Remastered Play	The Best Of Depeche Mode, Vol. 1 (Remastered) (2006)	Replace	x
Hooverphonic	Amalfi Play	Reflection (2013)	Replace	x

Although the results achieved from several attempts on the Spotiseek playlist generator were ok for the song chosen by Spotiseek from the input artist, there was no option to input an individual song. So, the song chosen, and therefore the suggested playlist, could be completely different to another song by the same artist that the user might like.

Technical Approach

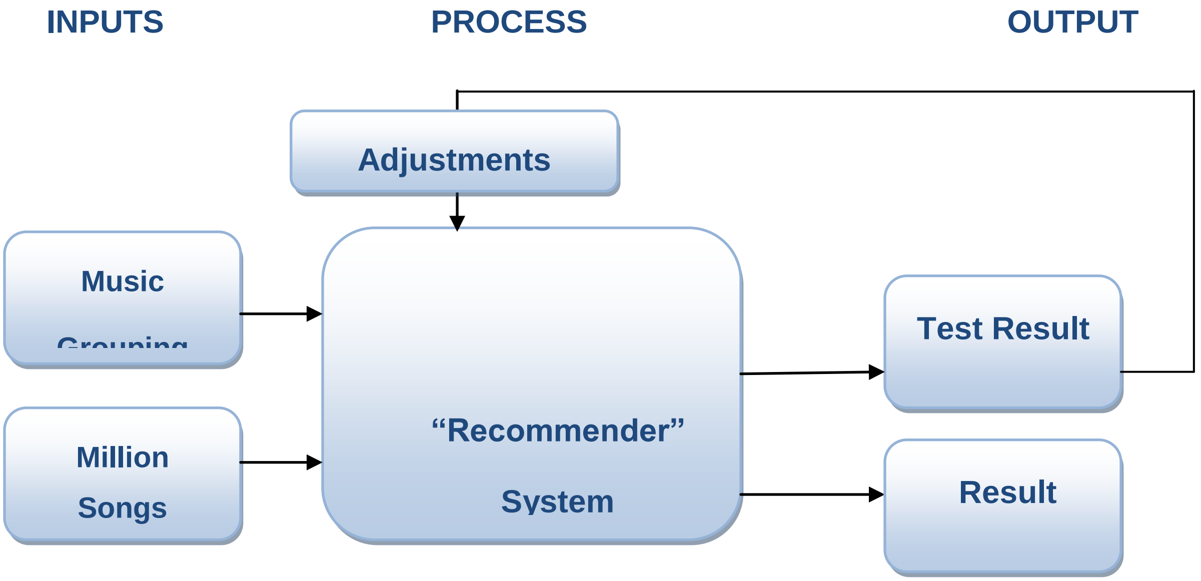
It is intended that the above objectives will be obtained as follows: The Million Song Data Set will be downloaded, with an initial download of a subset of 10,000 songs (1%, 1.8GB compressed) to obtain the 100 songs for the training set. The dataset is in *.gz format so conversion software will have to be downloaded.

As there is a possibility of 53 fields associated with each song, there will have to be some data reduction applied.

Once the results of the survey are returned, the data will have to be visualized, similarity measures will need to be identified and defined

It will probably be necessary to use multiple strategies to generate multiple rankings on which to base possible user preferences

The most basic process of the system is demonstrated in the following diagram:



Identifying Customer Needs

In order to identify user needs, a sample of the user population will be surveyed regarding their natural logical groupings of songs. The ideal sample size will be more than 30 to achieve a strong representation of the total population. However we accept that given time constraints and the necessity of the sample to be re-surveyed regarding the output of the model, this ideal sample size might not be achieved.

Identifying Target Specifications

In the course of my research, I intend to also survey the sample population on what they would expect from the proposed system compared to what is already available as mentioned in the "Background" section of this proposal. I also intend to conduct research via the internet sites mentioned above and previous research documents on the same or similar subjects.

Project Management

In order to meet the strict project deadlines it is necessary to plan, schedule and time- manage all the tasks and activities involved in this project. Each task and activity has been identified and allocated a start and completion date as indicated in the Gantt chart below:

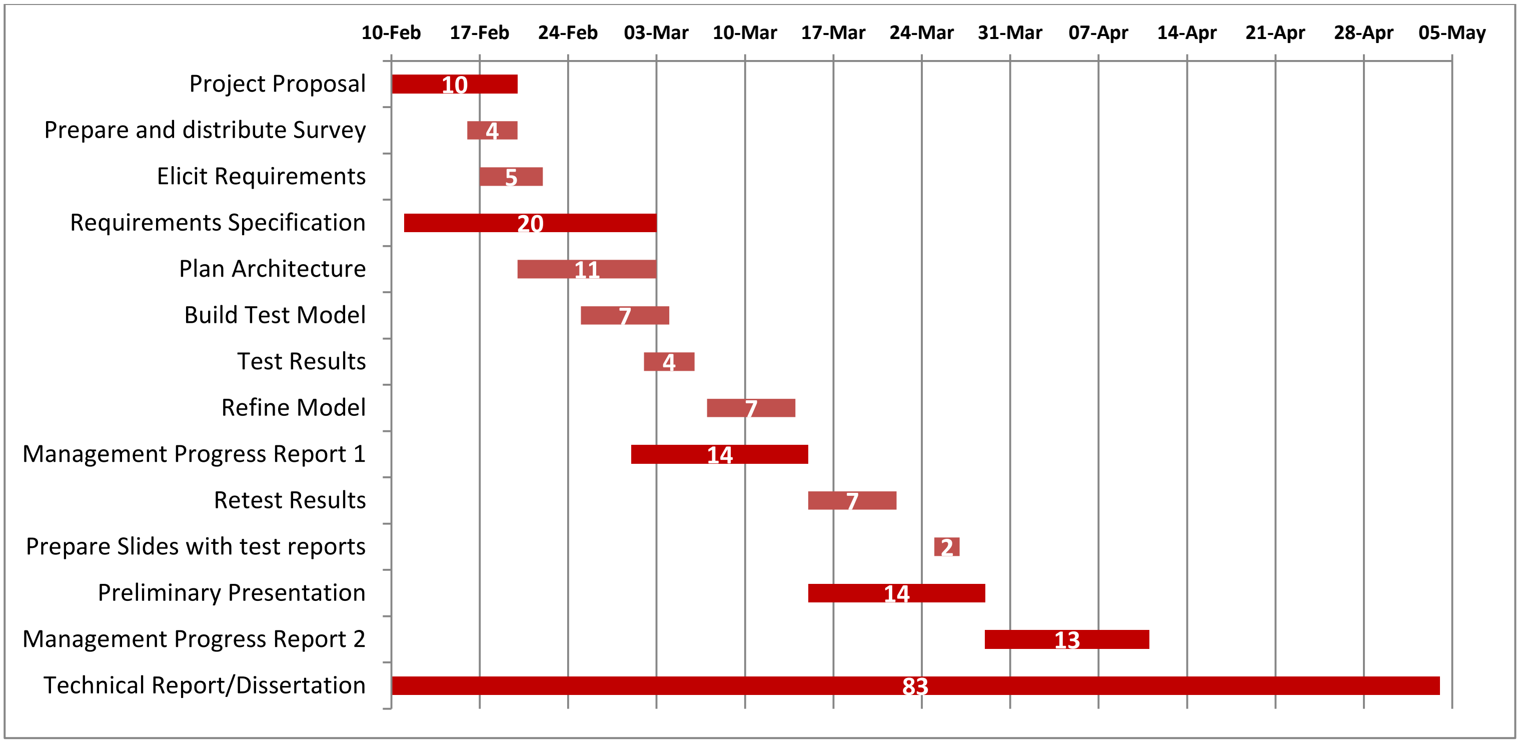


Figure 1: Gantt chart for the project. The bars indicate the tasks to accomplish. The darker coloured bars indicate the deliverables required for the project. (O'Loughlin, 2011)

Deliverables

The deliverable resulting from this project will be as follows:

- **Project Proposal** – this document describes the background to the project; A brief description of the approach to be followed in implementing the project; The major implementation steps and timelines; Names of academic staff members consulted.
- **Requirements Specification –** The requirements prioritized, organized, specified, verified and validated using, for example, a use case model and use case descriptions.
- **Management Progress Report –** A monthly management report that outlines highlights, issues and progress. This report is completed in collaboration with the Academic supervisor.

- **Analytics Artefacts –** Analytics artefacts generated and used will be provided by the learners in such a way as to be easily accessible by supervisors and examiners. Such artefacts may consist of databases, dashboards, web-sites, models etc.
- **Preliminary Report** – A statement of the progress on work carried out so far to a panel of experts. Feedback from the panel is a vital component to this report.
- **Dissertation** – Includes Executive Summary, Introduction, Background, Technologies, Background, System/Data Sets, Conclusions, Further development or research, Bibliography, Appendix.
- **Final Presentation** – A presentation of the project followed by a demonstration of all data analysis conducted to a panel of data analytics experts.
- **Showcase** – Presentation of research and data analysis to venture capitalists and the public.

Consultations

Michael Bradford – I discussed several project options with Michael. He suggested some options regarding three initial proposals I had been considering.

In a further consultation Michael confirmed that the chosen project would be a suitable choice and also suggested that I could “architect the solution in such a way so that the same process can be used for data other than songs (eg. If people were to give responses based on viewing objects of art, pieces of literature, Netflix movies). The same process may be able to be applied in more general settings – so that you are essentially creating a recommender / sentiment analysis system based on mood classifications. You could use the songs database as a testbed but it would be interesting to consider applications in more general settings (even as an indication of future work).”

In a final consultation with Michael, we trashed out how the analysis of both the survey results and the Million Song Subset might proceed. Michael suggested initially “eyeballing” the survey results to identify any obvious groupings, then performing a number of methods including Clustering, K nearest neighbour, association techniques.

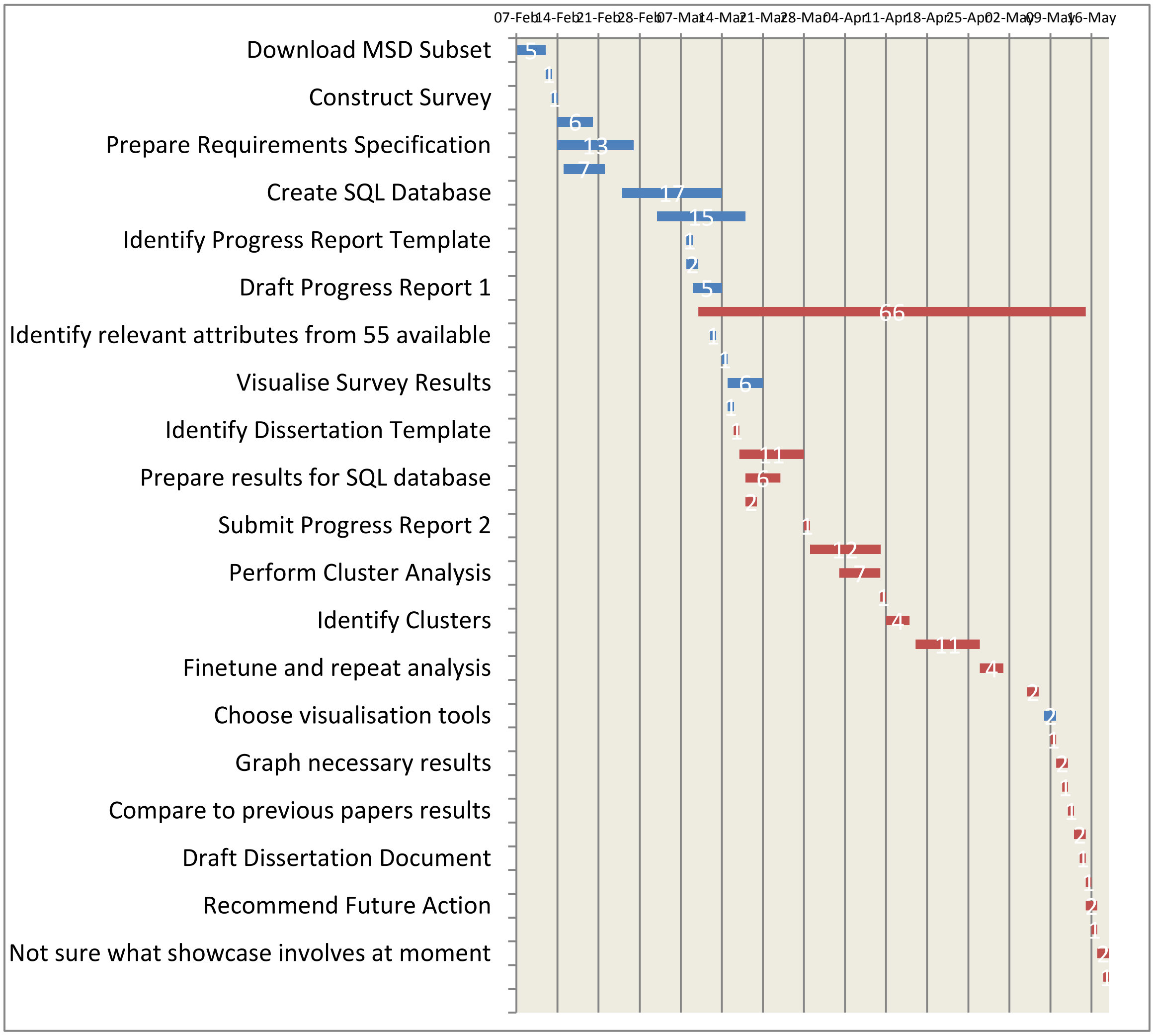
Jonathan Lambert – Having discussed another project proposal at length with Jonathan, he helped me identify the current project as the best choice to learn the skills identified in the project specification.

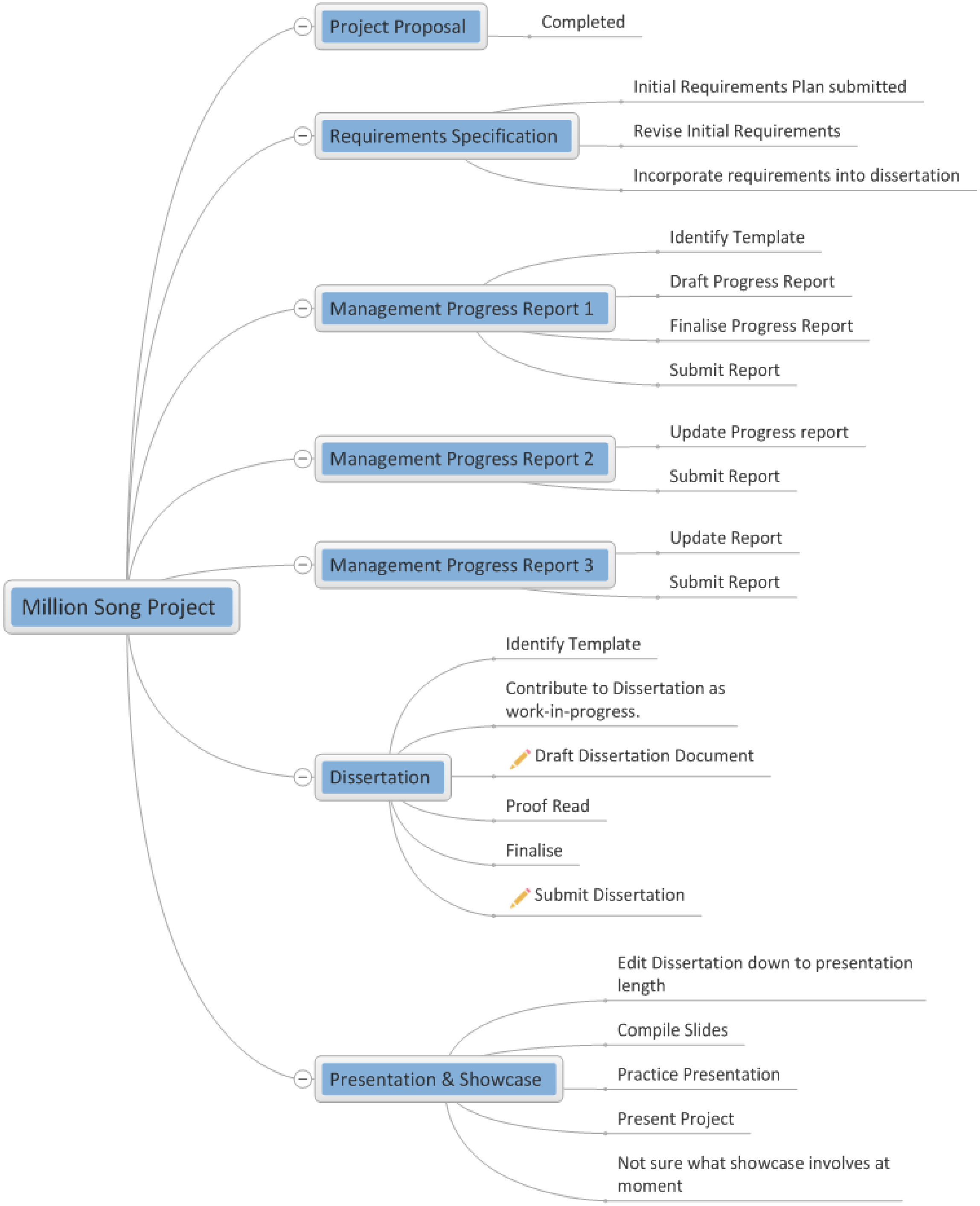
Ioana Ghergulescu – Ioana advised on potential issues and risks relating to initial project under consideration – “Paddy Power Payroll and Scheduling”. I decided against this project as I didn’t think it covered the skill and learning areas required for the course and for future career directions.

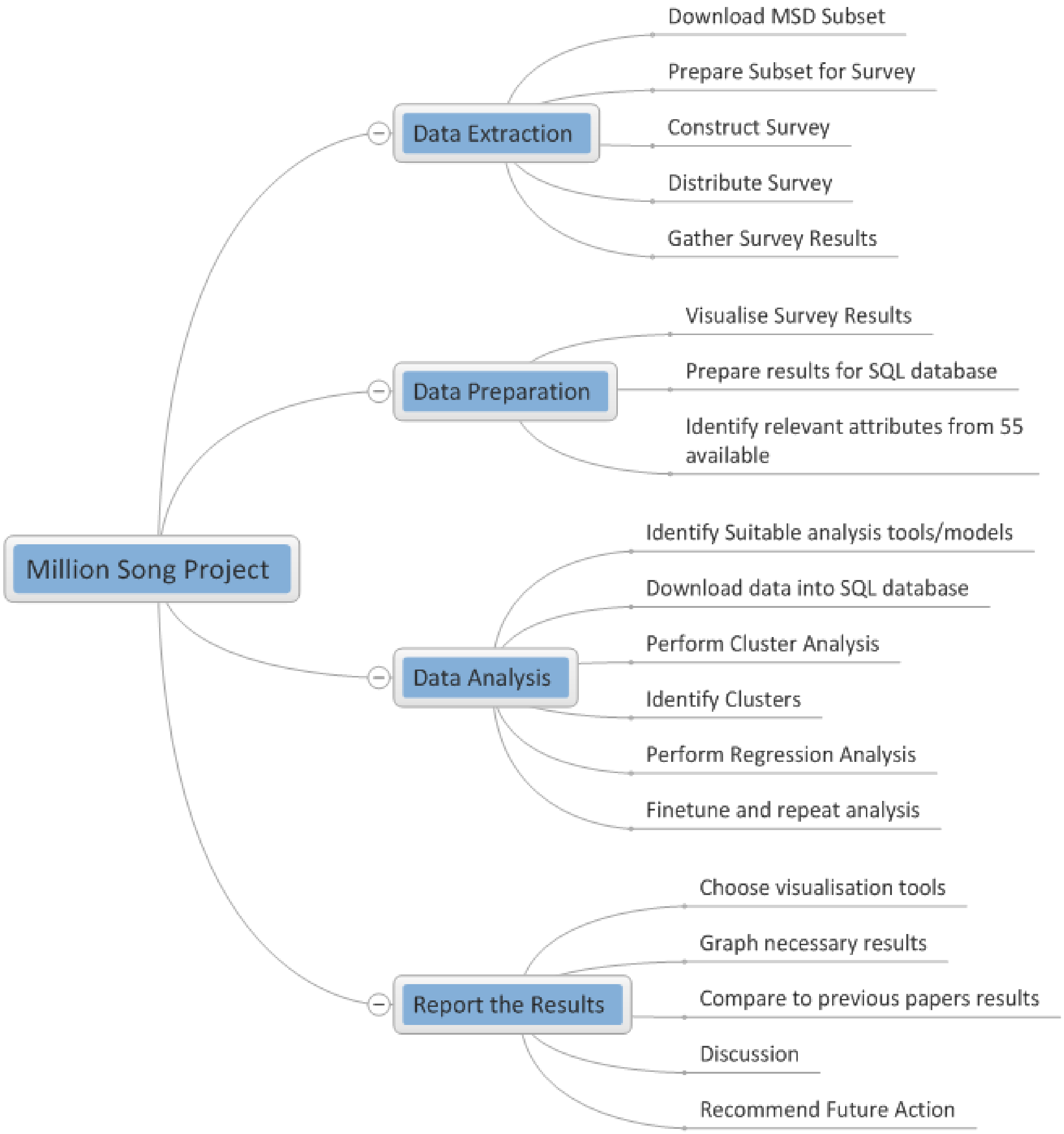
Qualifications

Michele Groarke has a Degree in Electronic Engineering obtained from Kevin Street College of Technology in 1991.

Initial Project Plan




Million Song Project
Project Proposal
Completed
Requirements Specification
Initial Requirements Plan submitted
Revise Initial Requirements
Incorporate requirements into dissertation
Management Progress Report 1
Identify Template
Draft Progress Report
Finalise Progress Report
Submit Report
Management Progress Report 2
Update Progress report
Submit Report
Management Progress Report 3
Update Report
Submit Report
Dissertation
Identify Template
Contribute to Dissertation as work-in-progress.
Draft Dissertation Document
Proof Read
Finalise
Submit Dissertation
Presentation & Showcase
Edit Dissertation down to presentation length
Compile Slides
Practice Presentation
Present Project
Not sure what showcase involves at moment


Million Song Project
Data Extraction
Download MSD Subset
Prepare Subset for Survey
Construct Survey
Distribute Survey
Gather Survey Results
Data Preparation
Visualise Survey Results
Prepare results for SQL database
Identify relevant attributes from 55 available
Data Analysis
Identify Suitable analysis tools/models
Download data into SQL database
Perform Cluster Analysis
Identify Clusters
Perform Regression Analysis
Finetune and repeat analysis
Report the Results
Choose visualisation tools
Graph necessary results
Compare to previous papers results
Discussion
Recommend Future Action

8.2 Initial Requirement Specification

Document Control

Revision History

Date	Version	Scope of Activity	Prepared	Reviewed	Approved
24/2/14	1	Create	MG		

Distribution List

Name	Title	Version
Dr. Ioana Ghergulescu	Lecturer	1

Related Documents

Title	Comments

Do I remove the document control above?

Introduction

Purpose

The purpose of this Requirements Specification document is to give a breakdown of the requirements for the development of this system. It lists the functional and non-functional requirements required for each step of the project and also lists the use cases for the system. In addition, the Requirements Specification helps to clarify exactly what is required to complete this project. The intended customers are initially music streaming companies, extending to any sales or service provider. The requirements have been prioritised, organised, specified, assessed, verified and validated as recommended in the Guide to Business Analysis Body of Knowledge. (International Institute of Business Analysts, 2009)

Project Scope

The scope of the project is to create an accurate song recommender system. The system shall have a method of extracting the data from the sources, i.e. labrosa (labrosa.ee, 2011) and survey results.

To elicit the following requirements, several methods were used in accordance with the Guide to the Business Analysis Body of Knowledge (International Institute of Business Analysts, 2009)reference book studied last semester. Document analysis of sorts was also carried out on playlist recommender platforms such as Spotify[i], Spotiseek[ii] Last.fm[iii] and, more broadly, Amazon[iv], which recommends a whole range of items (books, cds, Dvds, etc.) based on previous purchases.

The requirements elicitation process was somewhat constrained by the short time period between the project proposal and the due date for the Requirements Specification.

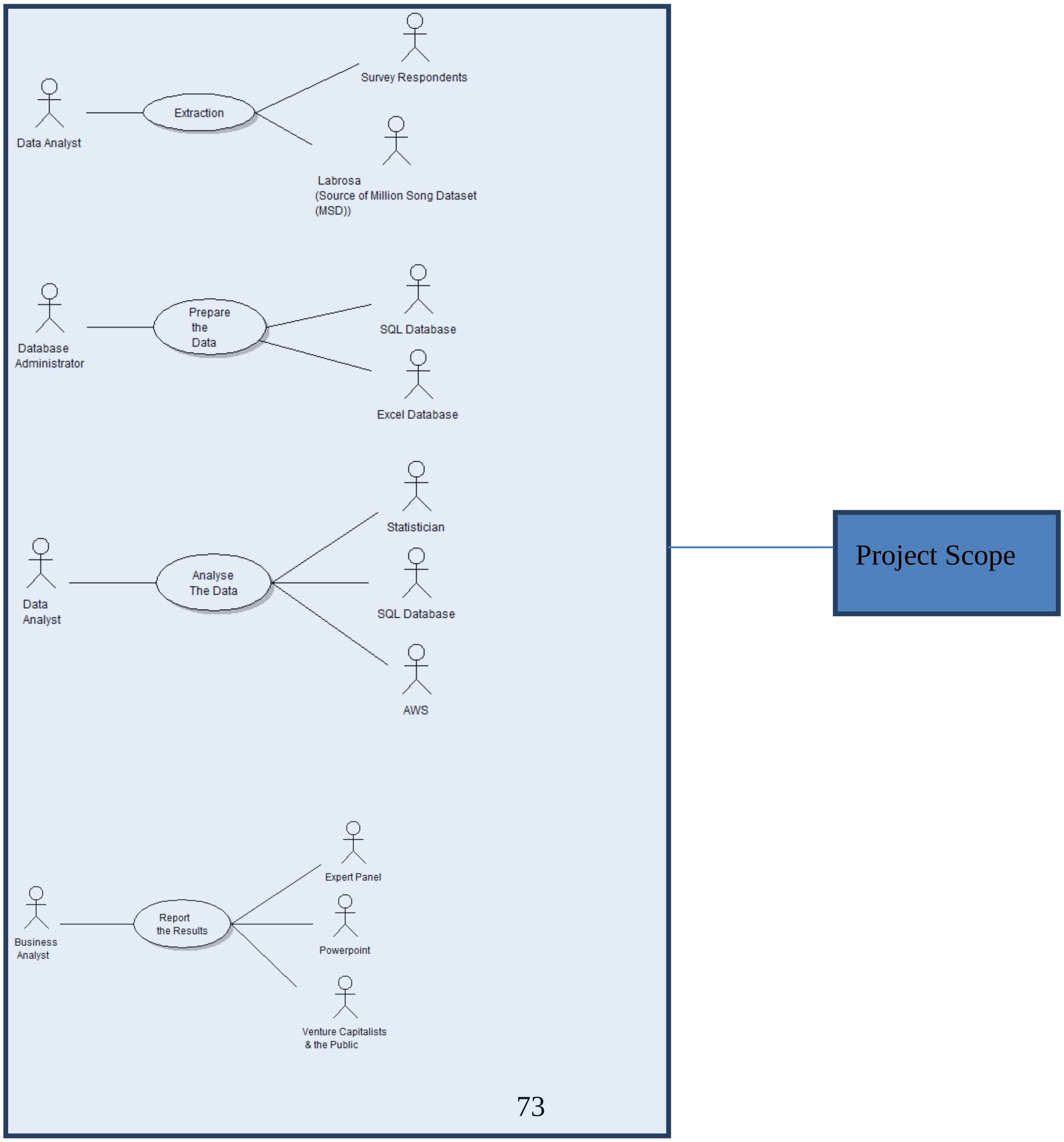
Definitions, Acronyms, and Abbreviations

MSD	Million Songs Dataset
MIR	Music Information Retrieval
GB	Gigabyte
tar.gz	Compressed **T**ape**Ar**chive format
SQL	Structured Query Language
SME	Subject Matter Expert

User Requirements Definition

From interviews with a sample of the end user's customer base, it has emerged that they require a more accurate recommender system with songs recommended not according to genre, artist or location, but more according to mood or sentiment. They are eager to be exposed to new music and the options are limited to a decreasing number of radio and television music programmes, word-of-mouth, and online music platforms. The latter source is the area to be focused on in this analysis.

Use Case Diagram



Requirements Specification

Users shall be able to use the system with no training apart from normal internet and computer usage.

Functional requirements

The functional requirements necessary for the analysis are described below:

Requirement 1 – Conduct Survey

Extract subset of the Million Song Dataset. Extract results from music survey.

Extract the Data

Download of the Million Song Subset. This is a subset consisting of 10,000 songs (1% of the full dataset, 1.8GB in size).This is essential for the compilation of the music survey to be sent out to the sample population. This is in .gz archive compressed format so will have to install 7-Zip File Manager and Winpython[v] to read the hdf5 files contained within.

In addition, the csv formatted results from the survey will have to be downloaded.

Compose Survey

Compose survey in Excel and send it to sample population via Hotmail.

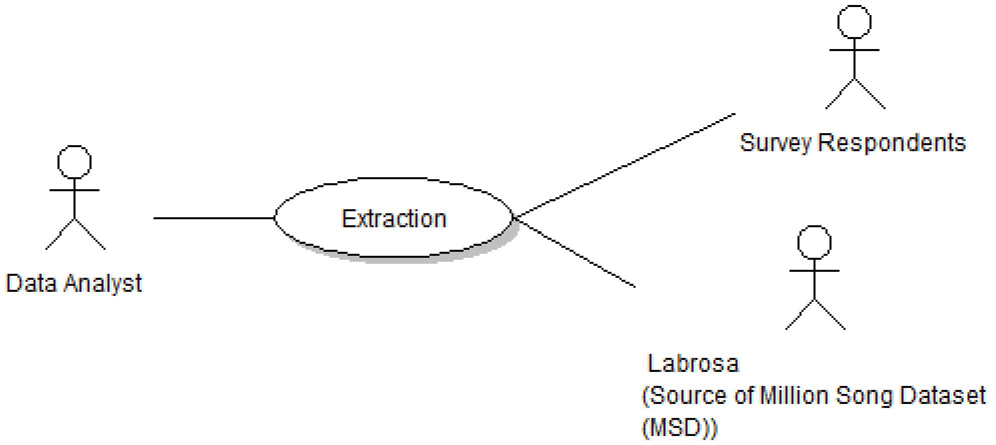
Scope

The scope of this use case is to extract the million song subset from the website labrosa.com and to extract the data resulting from the responses to the survey.

Description

This use case describes the creation of the survey from tracks extracted from the MSD dataset and the extraction of the data from the survey responses and the Million Song Subset by the Data Analyst in a complete and relevant format to have it ready for preparation.

Use Case Diagram 1



Flow

Description

Precondition

The system is in initialisation mode when the survey results are returned.

Activation

This use case starts when the Data Analyst downloads the Million Song Subset.

Main flow

The Million Song Subset is downloaded.

Tracks for survey chosen from Subset.

The survey is created.

The survey is emailed.

The survey database is created.

The survey responses are saved as csv files.

Termination

This Use Case terminates when all the data has been extracted.

Post condition

The data is ready for preparation.

Requirement 2 - Prepare the Data

Prepare the Data.

Once the data has been successfully extracted, it is necessary to clean and combine the two sets of data.

Preparing the Data

The csv files are uploaded into SQL database, addressing any omissions or errors.

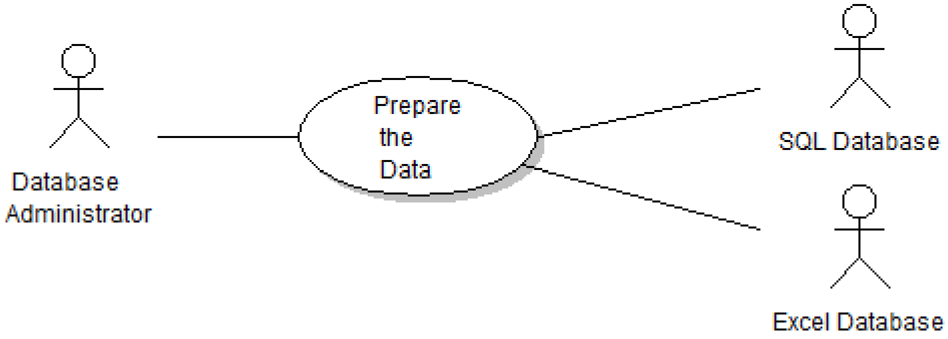
Scope

The scope of this use case is to prepare the extracted data.

Description

This use case describes the preparation of the extracted data. It needs to be formatted, errors corrected and omissions addressed.

Use Case Diagram 2



Flow Description

Precondition

The system is in initialisation mode when all the data has been extracted.

Activation

This use case starts when the Database Administrator (DBA) commences the preparation of the data.

Main flow

1. Format the data
2. The DBA identifies any errors
3. The DBA identifies omissions
4. The DBA corrects errors
5. The DBA deals with omissions

Termination

This use case terminates when the data has been suitably prepared.

Post condition

The data is ready for analysis.

Requirement 3 – Analyse the Data

Analyse the Data.

Once the data has been prepared, the analysis of the data can commence.

Analysis of the Data

This use case describes the analysis of the prepared data in SQL Database by the Data Analyst (DA) in conjunction with a statistical expert to identify correlations between the songs attributes and the mood they are associated with by the survey respondents.

Scope

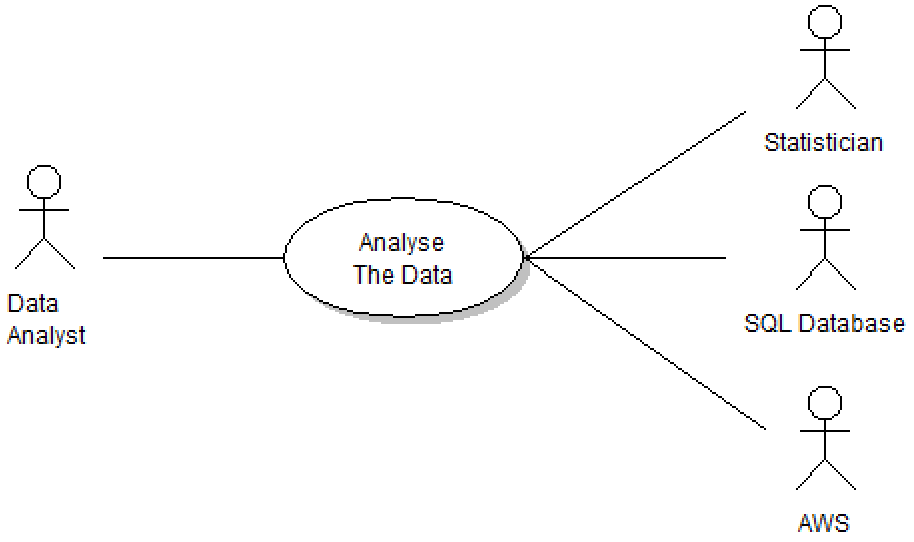
The scope of this use case is to analyse the prepared data.

Description

This use case describes the analysis of the prepared data. Using statistical methods the data will be analysed, first of all "Clustering" the raw data from the survey to observe any natural groupings within the MSD attributes of the songs included in the survey.

Then an appropriate analysis (probably regression) method will be used to identify patterns between individual songs and the "mood" they produce in the respondents.

Use Case Diagram 3



Flow Description

Precondition

The system is in initialisation mode when all the data has been suitably prepared.

Activation

This use case starts when the Data Analyst (DA) commences analysis of the prepared data.

Main flow

6. Initial Analysis
7. Secondary Analysis(See A1)
8. Test (See E1)
9. Adjust Analysis as necessary
10. Test
11. Result

Termination

This use case terminates when the DA is satisfied with the results.

Post condition

The results of the analysis are ready for reporting.

Requirement 4 – Report the Results

Report the Results.

Once the analysis is complete, the results will be reported.

Reporting the Analysis results.

The results will be communicated to the stakeholders via a written report and Powerpoint presentation.

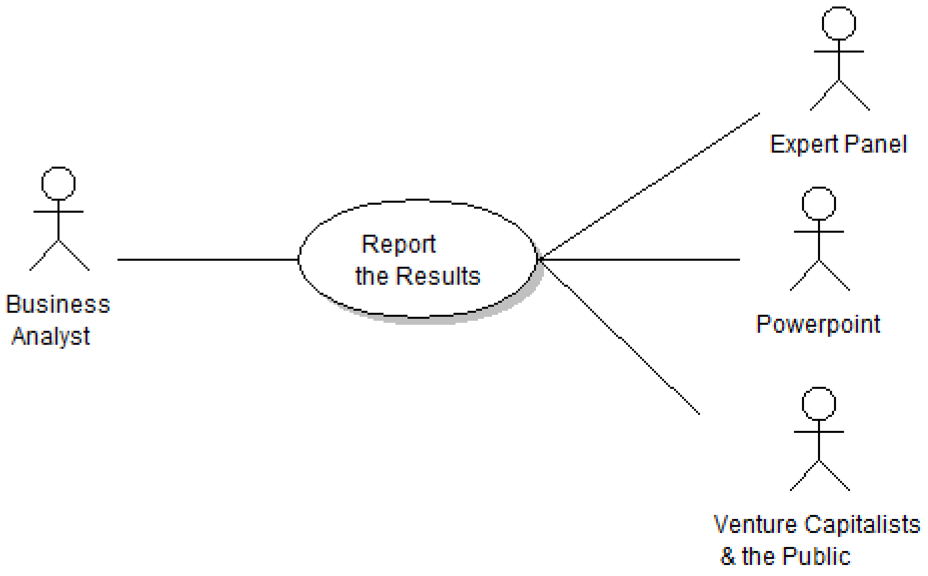
Scope

The scope of this use case is to report the results of the analysis.

Description

Once the results have been obtained from the analysis/analyses, they need to be communicated to the various stakeholders. Due to the varying levels of technical knowledge amongst the stakeholder group, two separate presentations will be prepared. These presentations will be prepared using various visualisation tools. These tools are at this time anticipated to be R, Excel, Tableau and Powerpoint.

Use Case Diagram 4



Flow Description

Precondition

The system is in initialisation mode when all the analyses have been conducted and the results are obtained.

Activation

This use case starts when the Business analyst (BA) receives the results from the analyses.

Flow

12. BA receives results of analysis.
13. BA compiles results

14. BA prepares presentations
15. BA communicates results

Termination

This use case terminates when the BA has prepared the report and presentations.

Post condition

The results have been communicated to the stakeholders.

Non-Functional Requirements

This area deals with any other particular non-functional attributes required by the system.

Performance/Response time requirement

The response time, though not crucial, would ideally be within 1 minute of input.

Security requirement

As the dataset is publicly available, there is no issue with security. Also, the survey respondents have been asked for permission to use their results within the scope of this project.

Portability requirement

The system should be accessible from any portable device in addition to PC access.

Scalability requirement

Extra features. It should be possible to upscale or downscale the system.

Resource utilization requirement

A number of resources will be required for this project – Microsoft Excel, MySQL, Hotmail account, and access to SME.

Interface requirements

In my project I have the following interfaces:

MySQL Workbench

This interface will be used to import the survey results, which are on excel worksheets, to the SQL database.

AWS

Amazon Web Services will be required to analyse the main body of the MSD

Login details are required to access this service. (These login details are still active from the Programming For Big Data module last semester.)

Hotmail

A Hotmail account will be required to send the survey out to the sample population.

Microsoft

Microsoft Excel will be used to create the survey. It will also be required to load the responses into MySQL database. Microsoft Windows XP Professional Version 2002 Service Pack 3 is the system being used in this instance.

Weka

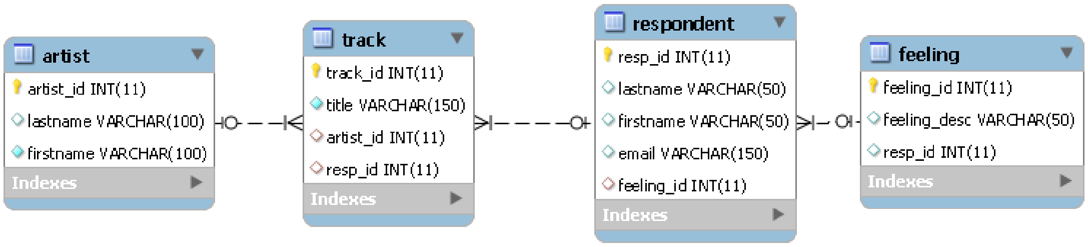
Weka (Waikato Environment for Knowledge Analysis) is a software “workbench” for data mining, developed in the University of Waikato, New Zealand. It contains data mining algorithms and is suitable for very large datasets. It will be used to analyse the prepared data and to produce graphical representation of same. (University of Waikato)

System Architecture

This section includes the survey database architecture, the fields and field types included for each song in the MSD and the storage and memory requirements of the computer used to work on this project.

Survey ERD

This ERD of the survey response database shows the attributes associated with the four entities – ‘artist’, ‘track’, ‘respondent’ and ‘feeling’.



artist:

This entity includes the fields:

- artist_id – this is the primary key which is an auto-incremented integer type and is unique.
- lastname – this is a VARCHAR of no more than 100 characters
- firstname – this is a VARCHAR of no more than 100 characters

track:

This entity includes the fields:

- track_id - this is the primary key which is an auto-incremented integer type and is unique.
- title – this is the title of the song and is a VARCHAR of no more than 150 characters.
- artist_id – this is a foreign key linking to the artist table. It is an INTEGER.
- resp_id - this is a foreign key linking to the artist table. It is an INTEGER.

respondent:

This entity includes the fields:

- resp_id - this is the primary key which is an auto-incremented INTEGER and is unique.

- lastname – this is the respondent’s last name and is a VARCHAR of no more than 50 characters.
- firstname – this is the respondent’s first name and is a VARCHAR of no more than 50 characters
- email -. this is the respondent’s email address and is a VARCHAR of no more than 150 characters
- feeling_id - this is a foreign key linking to the feeling table. It is an INTEGER.

feeling:

This entity includes the fields:

- feeling_id - this is the primary key which is an auto-incremented integer and is unique.
- feeling_desc – this is the title of the song and is a VARCHAR of no more than 150 characters.
- resp_id – this is a foreign key linking to the feeling table. It is an INTEGER.

List of Fields and field types in the MSD

	Type	Description	Link
analysis sample rate	float	sample rate of the audio used	url
artist 7digitalid	int	ID from 7digital.com or -1	url
artist familiarity	float	algorithmic estimation	url
artist hotttnesss	float	algorithmic estimation	url
artist id	string	Echo Nest ID	url
artist latitude	float	latitude	
artist location	string	location name	
artist longitude	float	longitude	
artist mbid	string	ID from musicbrainz.org	url
artist mbtags	array string	tags from musicbrainz.org	url
artist mbtags count	array int	tag counts for musicbrainz tags	url
artist name	string	artist name	url
artist playmeid	int	ID from playme.com, or -1	url
artist terms	array string	Echo Nest tags	url

artist terms freq	array float	Echo Nest tags freqs	url
artist terms weight	array float	Echo Nest tags weight	url
audio md5	string	audio hash code	
bars confidence	array float	confidence measure	url
bars start	array float	beginning of bars, usually on a beat	url
beats confidence	array float	confidence measure	url
beats start	array float	result of beat tracking	url
danceability	float	algorithmic estimation	
duration	float	in seconds	
end of fade in	float	seconds at the beginning of the song	url
energy	float	energy from listener point of view	
key	int	key the song is in	url
key confidence	float	confidence measure	url
loudness	float	overall loudness in dB	url
mode	int	major or minor	url
mode confidence	float	confidence measure	url
release	string	album name	

release 7digitalid	int	ID from 7digital.com or -1	url
sections confidence	array float	confidence measure	url
sections start	array float	largest grouping in a song, e.g. verse	url
segments confidence	array float	confidence measure	url
segments loudness max	array float	max dB value	url
segments loudness max time	array float	time of max dB value, i.e. end of attack	url
segments loudness max start	array float	dB value at onset	url
segments pitches	2D array float	chroma feature, one value per note	url
segments start	array float	musical events, ~ note onsets	url
segments timbre	2D array float	texture features (MFCC+PCA-like)	url
similar artists	array string	Echo Nest artist IDs (sim. algo. unpublished)	url
song hotttnesss	float	algorithmic estimation	
song id	string	Echo Nest song ID	
start of fade out	float	time in sec	url

tatums confidence	array float	confidence measure	url
tatums start	array float	smallest rythmic element	url
tempo	float	estimated tempo in BPM	url
time signature	int	estimate of number of beats per bar, e.g. 4	url
time signature confidence	float	confidence measure	url
title	string	song title	
track id	string	Echo Nest track ID	
track 7digitalid	int	ID from 7digital.com or -1	url
year	int	song release year from MusicBrainz or 0	url

(labrosa.ee, 2011)

The full complement of fields should not be required for this project and will be edited as necessary.

The computer used for this project is a Dell LatitudeD630 with Intel Core Duo CPU, a T7300 chip operating at 2GHz with 3GB of RAM and 150GB of memory on the hard drive.

System Evolution

In time it might be possible to apply this system to literature, films, art and possibly even the world of online dating. In this case the database will need to scale up and down as necessary. Storage will also need to be scalable.

Risks

There is a risk that the project proposal will be rejected.

Issues

- So far issues have arisen with regard to the downloading of the Million Songs Subset. It is not as simple as it initially seemed to gain access to the actual data. This has led to a slight delay in producing the survey.
- Also, though there are so many songs available in this dataset, judging from the artists and songs available on the Subset, there are many obscure artists, and many of the songs from well-known artists are not so well-known so the survey has been restricted somewhat by what is available. I have included hyperlinks on the survey to enable respondents to listen to any songs they might not recognise.
-

Assumptions

I have made the following assumptions in order to produce this document:

- I am assuming that the project proposal has been accepted.
- I am assuming I will receive an acceptable response to my survey
- I am assuming that I will be able to learn all the skills required for this project that I do not yet possess, whether through formal classes or personal research
- I am assuming that I will have appropriate access to expert advice

Stakeholder List

- The Business Analyst
- National College of Ireland (NCI)
- Music Industry

8.3 Management Progress Reports

8.3.1 Management Progress Report 1

1 Report History

1.1 Document Location

This document is only valid on the day it was printed.

The source of the document will be found in the ‘project’ folder on Project Managers laptop.

1.2 Revision History

Date of this revision: 15/3/14

Date of Next revision: 29/3/14

Revision date	**Previous revision date**	**Summary of Changes**	**Changes marked**
N/A	N/A	First issue	0

1.3 Approvals

This document requires the following approvals.

Signed approval forms are filed in the Management section of the project files.

Name	**Signature**	**Title**	**Date of Issue**	**Version**
Dr. Ioana Ghergulescu		Project Supervisor	15/3/14	1

1.4 Distribution

This document has been distributed to:

Name	**Title**	**Date of Issue**	**Version**
Dr. Ioana Ghergulescu	Project Supervisor	15/3/14	1

2 Highlight Report from 03 Feb 2014 – 16 Mar 2014

3 Purpose of Document

To provide the Project Supervisor with a summary of the stage status at intervals as set out in project proposal.

4 Activities During This Period

- Consultations with subject matter experts (SMEs)
- Submission of Project Proposal Document
- Submission of Requirements Specification Document
- Downloading MSD subset
- Constructing Survey
- Building SQL database
- Distribution of survey

5 Risks, Assumptions, Issues & Dependencies (RAID)

The current and closed risks, assumptions, issues and dependencies are indicated in the following tables.

Open Risks Date last reviewed 16/03/2014

Risk Ref	Risk Category	Risk Description	Raised by	Date Identified	Priority	Impact	Prob	Mitigation Category	Mitigation	Owner	Update	Date updated
R01	technology	Might not get enough survey responses	M.Groarke	14-Feb-14	H	H	L	prevention	Send Reminder email	MG		

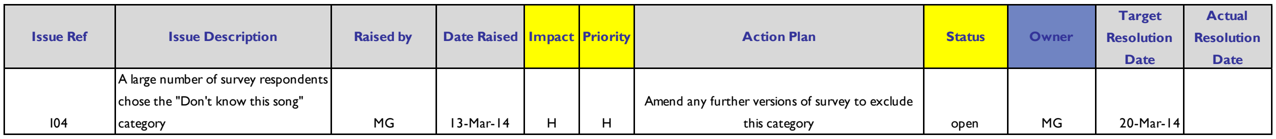
Closed Risks

Risk Ref	Risk Category	Risk Description	Raised by	Date Identified	Priority	Impact	Prob	Mitigation Category	Mitigation	Owner	Update	Date updated
Closed Risks	time	Might get job and will be constrained with regard to time available to work on project	M.Groarke	10-Feb-14	H	H	H	contingency	Will drop internship position	MG	Got Job, handing in notice on Monday	08-Mar-14

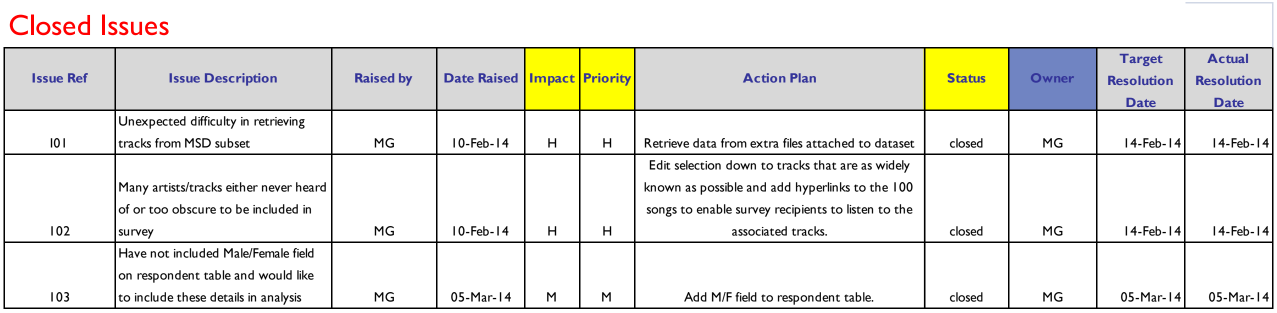
Assumptions

Ref #	Assumption	Importance	Certainty	Influence	Test	Test Date
A01	Lecturers will provide prompt feedback and guidance	4 - critical	3 - Probable	H	Send request to test level of response	07-Feb-14
A01	Lecturers will provide prompt feedback and guidance	4 - critical	1 - unknown	L	Send request to test level of response	07-Feb-14
A02	There are no confidentiality issues with the datasets I have chosen	1 - minor but worth noting	3 - Probable	H	Include note re confidentiality on survey	14-Feb-14
A03	Access to Expert Advice	4 - critical	3 - Probable	M	Request meetings with lecturers	18-Feb-14
A04	That I will learn the necessary skills to complete the project	4 - critical	3 - Probable	H	Feedback from lecturer/SME	

Issues



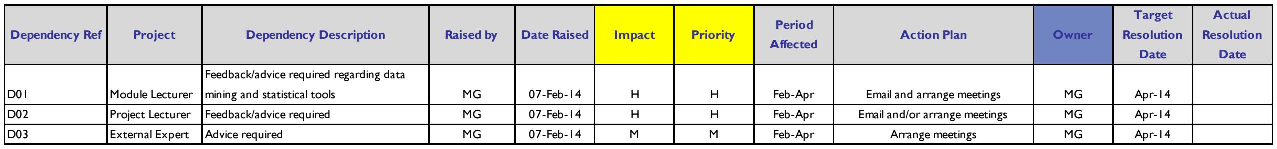
Issue Ref	Issue Description	Raised by	Date Raised	Impact	Priority	Action Plan	Status	Owner	Target Resolution Date	Actual Resolution Date
I04	A large number of survey respondents chose the "Don't know this song" category	MG	13-Mar-14	H	H	Amend any further versions of survey to exclude this category	open	MG	20-Mar-14	



Closed Issues

Issue Ref	Issue Description	Raised by	Date Raised	Impact	Priority	Action Plan	Status	Owner	Target Resolution Date	Actual Resolution Date
I01	Unexpected difficulty in retrieving tracks from MSD subset	MG	10-Feb-14	H	H	Retrieve data from extra files attached to dataset	closed	MG	14-Feb-14	14-Feb-14
I02	Many artists/tracks either never heard of or too obscure to be included in survey	MG	10-Feb-14	H	H	Edit selection down to tracks that are as widely known as possible and add hyperlinks to the 100 songs to enable survey recipients to listen to the associated tracks.	closed	MG	14-Feb-14	14-Feb-14
I03	Have not included Male/Female field on respondent table and would like to include these details in analysis	MG	05-Mar-14	M	M	Add M/F field to respondent table.	closed	MG	05-Mar-14	05-Mar-14

Dependencies



Dependency Ref	Project	Dependency Description	Raised by	Date Raised	Impact	Priority	Period Affected	Action Plan	Owner	Target Resolution Date	Actual Resolution Date
D01	Module Lecturer	Feedback/advice required regarding data mining and statistical tools	MG	07-Feb-14	H	H	Feb-Apr	Email and arrange meetings	MG	Apr-14	
D02	Project Lecturer	Feedback/advice required	MG	07-Feb-14	H	H	Feb-Apr	Email and/or arrange meetings	MG	Apr-14	
D03	External Expert	Advice required	MG	07-Feb-14	M	M	Feb-Apr	Arrange meetings	MG	Apr-14	

6 Products Completed during Period

6.1 Project planning

The work required for each stage of the project has been detailed. The Project plan has been agreed by the Project Supervisor.

6.2 Extraction of Data

The MSD subset of 10,000 songs has been downloaded. From this, the survey has been constructed, distributed and any results gathered so far are currently being prepared for download into SQL database.

6.3 Requirements Specification

The initial Requirements Specification document has been submitted to the project supervisor – currently awaiting feedback.

6.4 Data Analysis

The SQL database has been built and is ready for data gathered from survey.

6.5 Research

Several papers relating to the Million Song Dataset and Recommender Systems have been identified and printed (Bertin-Mahieux, et al., 2011).

6.6 Presentation

A possible visualization tool has been identified as Tableau software and training is currently underway.

7 Products due for Completion

By the next period the following products should be completed:

7.1 Dissertation

Template to be identified and draft dissertation in progress.

7.2 Data Analysis

Results from survey prepared and downloaded to SQL database. Some analysis may also be performed in Excel. Relevant attributes to be identified from the 55 available on the MSD tracks.

7.3 Survey Results

Initial visualization of results to be performed.

8 Impact of Changes

8.1 Schedule

On March 8th the Preliminary Presentation deliverable was removed from the project Specification, enabling more time to be spent on the analysis. However, this has the effect of adding an extra weighting onto the Dissertation deliverable, changing it from 60 percent to 80 percent. This results in more work to be focused on the dissertation document and presentation.

9 Variance from Plan

9.1 Survey

Upon consultation with one of the survey recipients it was deemed more insightful to include a Gender category in the survey. Gender was added to any results that were returned before the change.

10 Planned Work for next period to 30 Mar 2014

See Figure 1 below for overview of work completed and work due for next period and beyond.

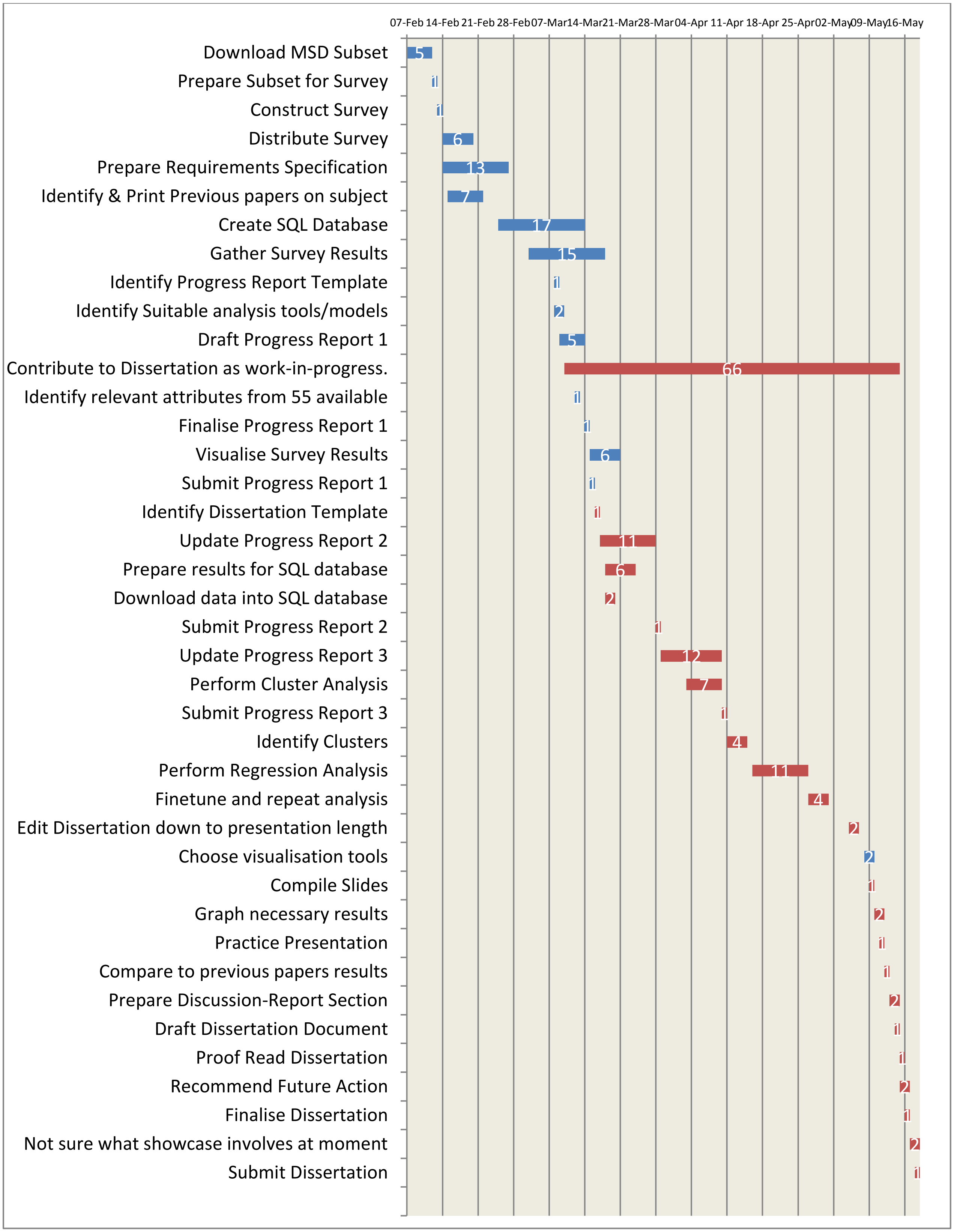

07-Feb
14-Feb
21-Feb
28-Feb
07-Mar
14-Mar
21-Mar
28-Mar
04-Apr
11-Apr
18-Apr
25-Apr
02-May
09-May
16-May
Download MSD Subset
5
Prepare Subset for Survey
1
Construct Survey
1
Distribute Survey
6
Prepare Requirements Specification
13
Identify & Print Previous papers on subject
7
Create SQL Database
17
Gather Survey Results
15
Identify Progress Report Template
1
Identify Suitable analysis tools/models
2
Draft Progress Report 1
5
Contribute to Dissertation as work-in-progress.
66
Identify relevant attributes from 55 available
1
Finalise Progress Report 1
1
Visualise Survey Results
6
Submit Progress Report 1
1
Identify Dissertation Template
1
Update Progress Report 2
11
Prepare results for SQL database
6
Download data into SQL database
2
Submit Progress Report 2
1
Update Progress Report 3
12
Perform Cluster Analysis
7
Submit Progress Report 3
1
Identify Clusters
4
Perform Regression Analysis
11
Finetune and repeat analysis
4
Edit Dissertation down to presentation length
2
Choose visualisation tools
2
Compile Slides
1
Graph necessary results
2
Practice Presentation
1
Compare to previous papers results
1
Prepare Discussion-Report Section
2
Draft Dissertation Document
1
Proof Read Dissertation
1
Recommend Future Action
2
Finalise Dissertation
1
Not sure what showcase involves at moment
2
Submit Dissertation
1

Figure 16: Gantt chart showing tasks completed (in blue) and tasks due (in red). This Gantt chart was created by combining the mind maps in Mind Genius. (MindGenius, 2014) (Mind maps located in Appendix)

8.3.2 Management Progress Report 2

1 Report History

1.1 Document Location

This document is only valid on the day it was printed.

The source of the document will be found in the ‘project’ folder on Project Managers laptop.

1.2 Revision History

Date of this revision: 29/3/14

Date of Next revision: 11/4/14

Revision date	**Previous revision date**	**Summary of Changes**	**Changes marked**
29/3/14	15/3/14	First issue	0

1.3 Approvals

This document requires the following approvals.

Signed approval forms are filed in the Management section of the project files.

Name	Signature	Title	Date of Issue	Version
Dr. Ioana Ghergulescu		Project Supervisor	15/3/14	1
Dr. Ioana Ghergulescu		Project Supervisor	29/3/14	2

1.4 Distribution

This document has been distributed to:

Name	Title	Date of Issue	Version
Dr. Ioana Ghergulescu	Project Supervisor	15/3/14	1
Dr. Ioana Ghergulescu	Project Supervisor	29/3/14	2

2 Highlight Report from 16 Mar 2014 – 11 Apr 2014

3 Purpose of Document

To provide the Project Supervisor with a summary of the stage status at intervals as set out in project proposal.

4 Activities During This Period

Consultations with subject matter experts (SMEs)

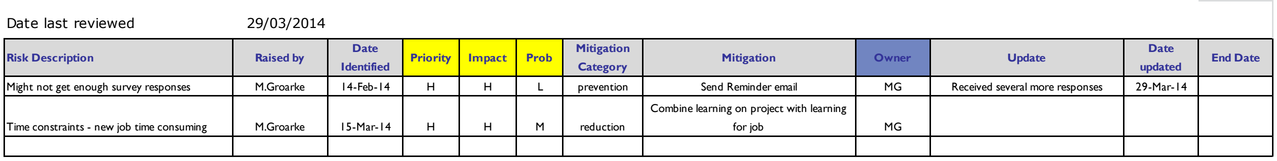
Training in Visualisation tool - Tableau

Collecting Survey results

Provisional analysis of survey results in SQL and Excel

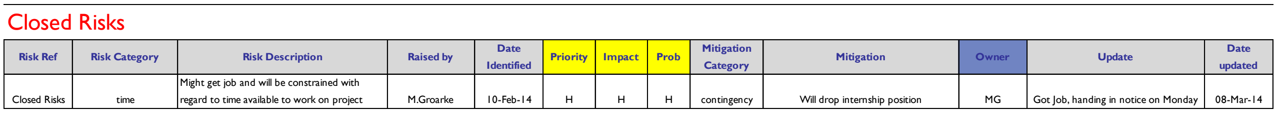
5 Risks, Assumptions, Issues & Dependencies (RAID)

The current and closed risks, assumptions, issues and dependencies are indicated in the following tables.



Date last reviewed 29/03/2014

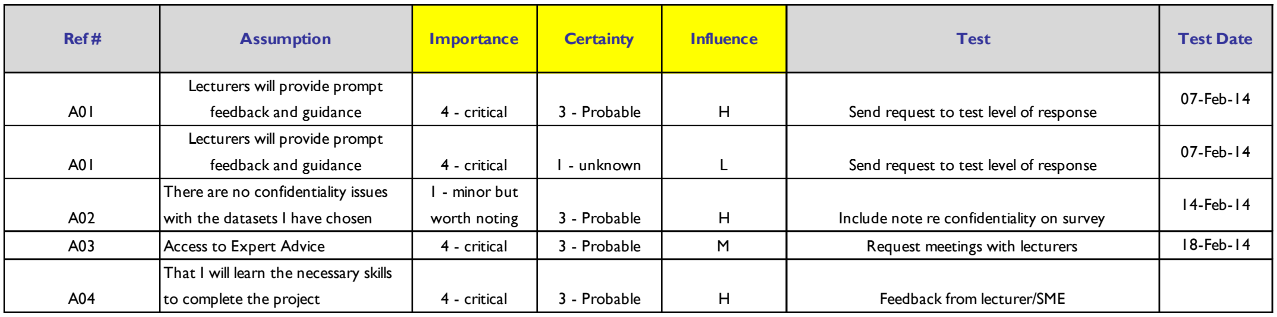
Risk Description	Raised by	Date Identified	Priority	Impact	Prob	Mitigation Category	Mitigation	Owner	Update	Date updated	End Date
Might not get enough survey responses	M.Groarke	14-Feb-14	H	H	L	prevention	Send Reminder email	MG	Received several more responses	29-Mar-14	
Time constraints - new job time consuming	M.Groarke	15-Mar-14	H	H	M	reduction	Combine learning on project with learning for job	MG			



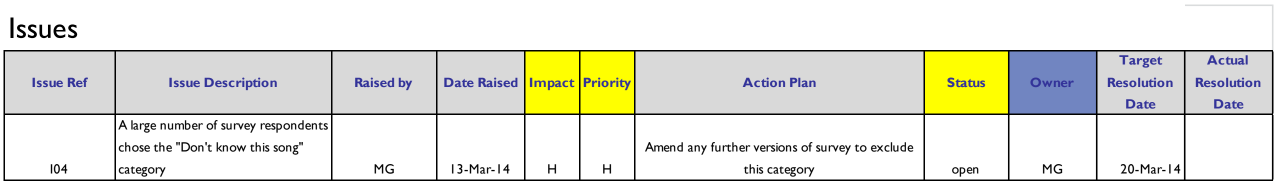
Closed Risks

Risk Ref	Risk Category	Risk Description	Raised by	Date Identified	Priority	Impact	Prob	Mitigation Category	Mitigation	Owner	Update	Date updated
Closed Risks	time	Might get job and will be constrained with regard to time available to work on project	M.Groarke	10-Feb-14	H	H	H	contingency	Will drop internship position	MG	Got Job, handing in notice on Monday	08-Mar-14

Assumptions

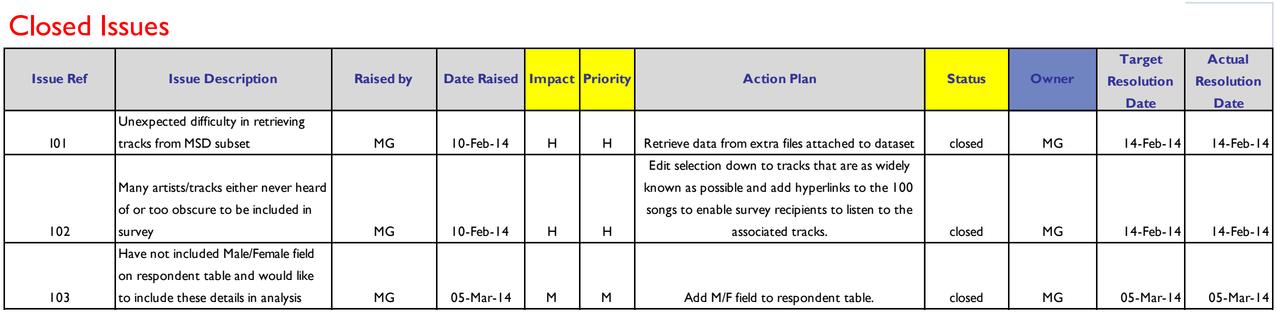


Ref #	Assumption	Importance	Certainty	Influence	Test	Test Date
A01	Lecturers will provide prompt feedback and guidance	4 - critical	3 - Probable	H	Send request to test level of response	07-Feb-14
A01	Lecturers will provide prompt feedback and guidance	4 - critical	1 - unknown	L	Send request to test level of response	07-Feb-14
A02	There are no confidentiality issues with the datasets I have chosen	1 - minor but worth noting	3 - Probable	H	Include note re confidentiality on survey	14-Feb-14
A03	Access to Expert Advice	4 - critical	3 - Probable	M	Request meetings with lecturers	18-Feb-14
A04	That I will learn the necessary skills to complete the project	4 - critical	3 - Probable	H	Feedback from lecturer/SME	



Issues

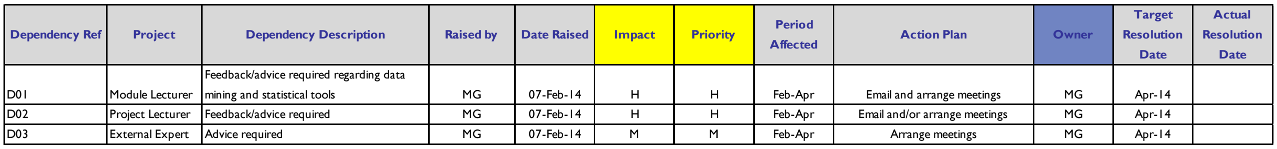
Issue Ref	Issue Description	Raised by	Date Raised	Impact	Priority	Action Plan	Status	Owner	Target Resolution Date	Actual Resolution Date
I04	A large number of survey respondents chose the "Don't know this song" category	MG	13-Mar-14	H	H	Amend any further versions of survey to exclude this category	open	MG	20-Mar-14	



Closed Issues

Issue Ref	Issue Description	Raised by	Date Raised	Impact	Priority	Action Plan	Status	Owner	Target Resolution Date	Actual Resolution Date
I01	Unexpected difficulty in retrieving tracks from MSD subset	MG	10-Feb-14	H	H	Retrieve data from extra files attached to dataset	closed	MG	14-Feb-14	14-Feb-14
I02	Many artists/tracks either never heard of or too obscure to be included in survey	MG	10-Feb-14	H	H	Edit selection down to tracks that are as widely known as possible and add hyperlinks to the 100 songs to enable survey recipients to listen to the associated tracks.	closed	MG	14-Feb-14	14-Feb-14
I03	Have not included Male/Female field on respondent table and would like to include these details in analysis	MG	05-Mar-14	M	M	Add M/F field to respondent table.	closed	MG	05-Mar-14	05-Mar-14

Dependencies



Dependency Ref	Project	Dependency Description	Raised by	Date Raised	Impact	Priority	Period Affected	Action Plan	Owner	Target Resolution Date	Actual Resolution Date
D01	Module Lecturer	Feedback/advice required regarding data mining and statistical tools	MG	07-Feb-14	H	H	Feb-Apr	Email and arrange meetings	MG	Apr-14	
D02	Project Lecturer	Feedback/advice required	MG	07-Feb-14	H	H	Feb-Apr	Email and/or arrange meetings	MG	Apr-14	
D03	External Expert	Advice required	MG	07-Feb-14	M	M	Feb-Apr	Arrange meetings	MG	Apr-14	

6 Products Completed during Period

6.1 Project planning

The work required for each stage of the project has been detailed. The Project plan has been agreed by the Project Supervisor.

6.2 Requirements Specification

Currently awaiting feedback from project supervisor.

6.3 Data Analysis

The results have been prepared and uploaded to SQL database. The data has been pre-processed in SQL and output as a csv file for initial analysis. The initial analysis has been conducted in Excel.

6.4 Research

Several papers relating to the Million Song Dataset and Recommender Systems are being read. (Bertin-Mahieux, et al., 2011).

6.5 Presentation

Preliminary results have been uploaded into Tableau but further preparation is necessary before analysis/preparation can be conducted with this particular tool.

6.6 Dissertation

Template identified and draft dissertation in progress.

6.7 Data Analysis

Preliminary analysis performed in Excel. Relevant attributes identified from the 55 available on the MSD tracks

6.8 Survey Results

Initial visualisation of survey results has been performed

7 Products due for Completion

By the next period the following products should be completed:

7.2 Data Analysis

Continue dissertation development.

7.3 Survey Results

Perform cluster analysis on survey results. Perform cluster analysis on same songs in MSD subset to identify common attributes within groupings.

8 Impact of Changes

8.1 Schedule

On March 28th the management report due dates were extended to 6/4/14 for management report 2 and 20/4/14 for management report 3. This project will continue according to the original timetable as there will be an impact on external projects should timings change.

9 Variance from Plan

9.1 Survey

Upon receipt of the survey results it was necessary to prepare the initial results in SQL before they could be visualized in Excel.

10 Planned Work for next period to 11 Apr 2014

See Figure 1 below for overview of work completed and work due for next period and beyond.

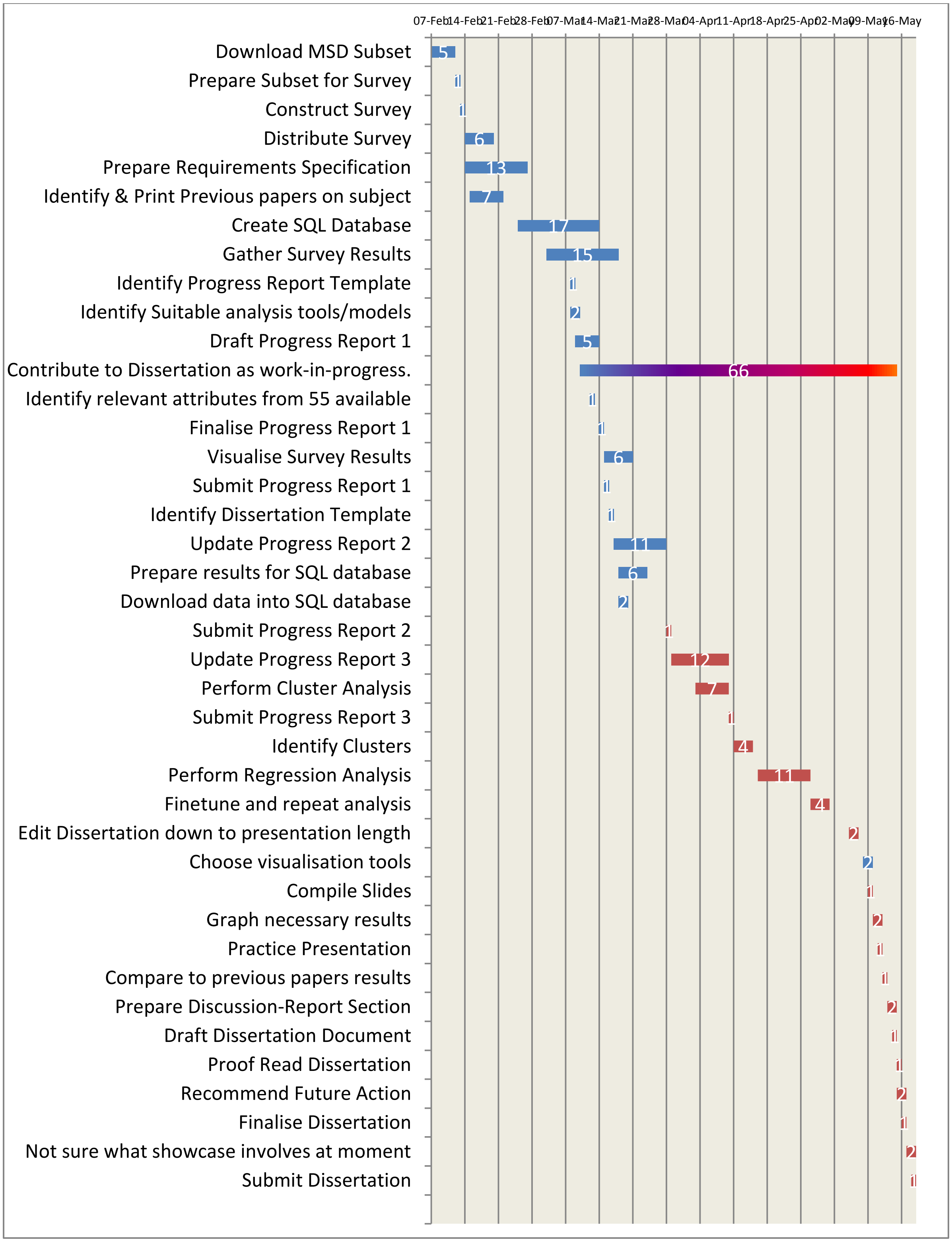


Figure 17: Gantt chart showing tasks completed (in blue) and tasks due (in red). This Gantt chart was created by combining the mind maps in Mind Genius. (MindGenius, 2014) (Mind maps located in Appendix)

8.3.3 Management Progress Report 3

1 Report History

1.1 Document Location

This document is only valid on the day it was printed.

The source of the document will be found in the ‘project’ folder on Project Managers laptop.

1.2 Revision History

Date of this revision: 18/4/14

Date of Next revision: 28/5/14 – Project Submission

Revision date	Previous revision date	Summary of Changes	Changes marked
29/3/14	15/3/14	First issue	0
18/4/14	29/3/14	Second Issue	

1.3 Approvals

This document requires the following approvals.

Signed approval forms are filed in the Management section of the project files.

Name	**Signature**	**Title**	**Date of Issue**	**Version**
Dr. Ioana Ghergulescu		Project Supervisor	15/3/14	1
Dr. Ioana Ghergulescu		Project Supervisor	29/3/14	2
Dr. Ioana Ghergulescu		Project Supervisor	18/4/14	3

1.4 Distribution

This document has been distributed to:

Name	**Title**	**Date of Issue**	**Version**
Dr. Ioana Ghergulescu	Project Supervisor	15/3/14	1
Dr. Ioana Ghergulescu	Project Supervisor	29/3/14	2

Dr. Ioana Ghergulescu	Project Supervisor	18/4/14	3

2 Highlight Report from 29 Mar 2014 – 18 Apr 2014

3 Purpose of Document

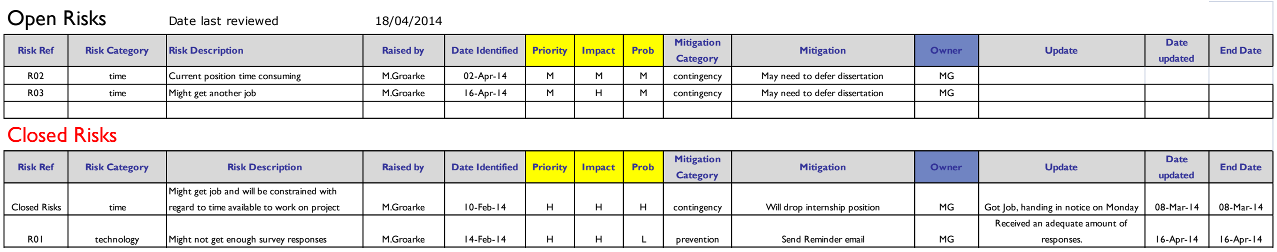
To provide the Project Supervisor with a summary of the stage status at intervals as set out in project proposal.

4 Activities During This Period

- Further consultations with subject matter experts (SMEs)
- Further training in Visualisation tool - Tableau
- Research subject papers and literature
- Preparation of MSD attribute data and provisional import into SQL database

5 Risks, Assumptions, Issues & Dependencies (RAID)

The current and closed risks, assumptions, issues and dependencies are indicated in the following tables.



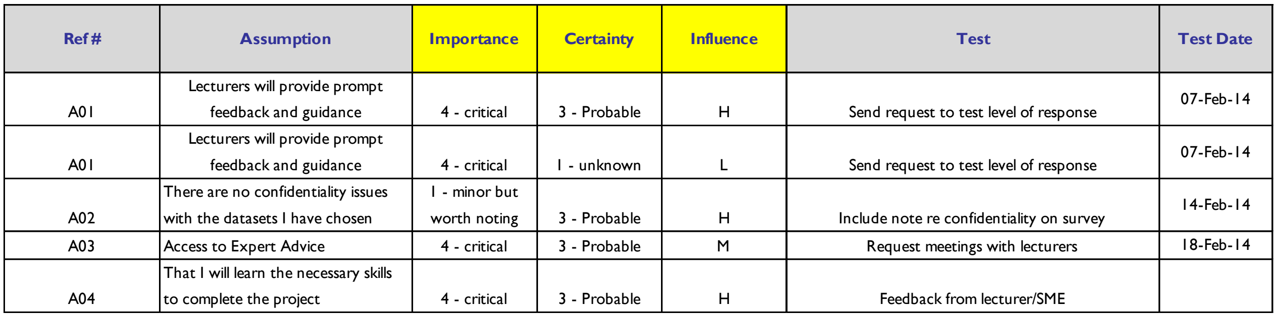
Open Risks Date last reviewed 18/04/2014

Risk Ref	Risk Category	Risk Description	Raised by	Date Identified	Priority	Impact	Prob	Mitigation Category	Mitigation	Owner	Update	Date updated	End Date
R02	time	Current position time consuming	M.Groarke	02-Apr-14	M	M	M	contingency	May need to defer dissertation	MG			
R03	time	Might get another job	M.Groarke	16-Apr-14	M	H	M	contingency	May need to defer dissertation	MG			

Closed Risks

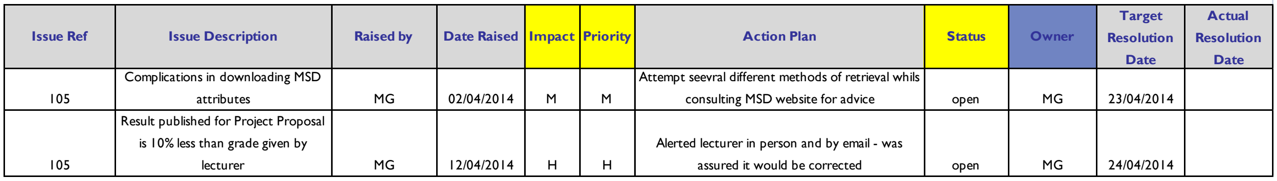
Risk Ref	Risk Category	Risk Description	Raised by	Date Identified	Priority	Impact	Prob	Mitigation Category	Mitigation	Owner	Update	Date updated	End Date
Closed Risks	time	Might get job and will be constrained with regard to time available to work on project	M.Groarke	10-Feb-14	H	H	H	contingency	Will drop internship position	MG	Got Job, handing in notice on Monday	08-Mar-14	08-Mar-14
R01	technology	Might not get enough survey responses	M.Groarke	14-Feb-14	H	H	L	prevention	Send Reminder email	MG	Received an adequate amount of responses.	16-Apr-14	16-Apr-14

Assumptions



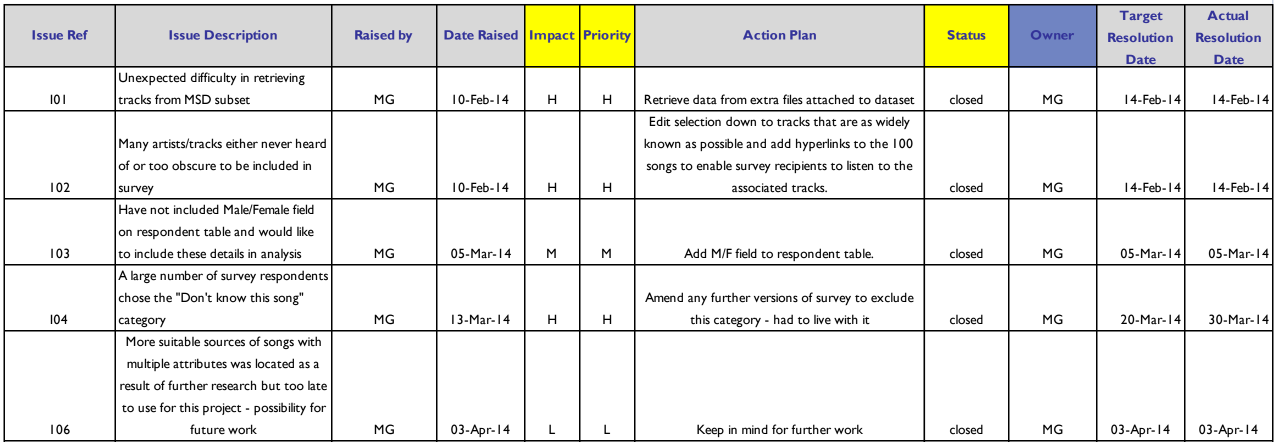
Ref #	Assumption	Importance	Certainty	Influence	Test	Test Date
A01	Lecturers will provide prompt feedback and guidance	4 - critical	3 - Probable	H	Send request to test level of response	07-Feb-14
A01	Lecturers will provide prompt feedback and guidance	4 - critical	1 - unknown	L	Send request to test level of response	07-Feb-14
A02	There are no confidentiality issues with the datasets I have chosen	1 - minor but worth noting	3 - Probable	H	Include note re confidentiality on survey	14-Feb-14
A03	Access to Expert Advice	4 - critical	3 - Probable	M	Request meetings with lecturers	18-Feb-14
A04	That I will learn the necessary skills to complete the project	4 - critical	3 - Probable	H	Feedback from lecturer/SME	

Issues



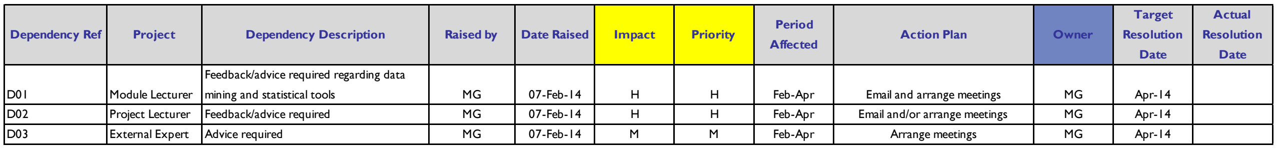
Issue Ref	Issue Description	Raised by	Date Raised	Impact	Priority	Action Plan	Status	Owner	Target Resolution Date	Actual Resolution Date
105	Complications in downloading MSD attributes	MG	02/04/2014	M	M	Attempt seevral different methods of retrieval whils consulting MSD website for advice	open	MG	23/04/2014	
105	Result published for Project Proposal is 10% less than grade given by lecturer	MG	12/04/2014	H	H	Alerted lecturer in person and by email - was assured it would be corrected	open	MG	24/04/2014	

Closed Issues



Issue Ref	Issue Description	Raised by	Date Raised	Impact	Priority	Action Plan	Status	Owner	Target Resolution Date	Actual Resolution Date
101	Unexpected difficulty in retrieving tracks from MSD subset	MG	10-Feb-14	H	H	Retrieve data from extra files attached to dataset	closed	MG	14-Feb-14	14-Feb-14
102	Many artists/tracks either never heard of or too obscure to be included in survey	MG	10-Feb-14	H	H	Edit selection down to tracks that are as widely known as possible and add hyperlinks to the 100 songs to enable survey recipients to listen to the associated tracks.	closed	MG	14-Feb-14	14-Feb-14
103	Have not included Male/Female field on respondent table and would like to include these details in analysis	MG	05-Mar-14	M	M	Add M/F field to respondent table.	closed	MG	05-Mar-14	05-Mar-14
104	A large number of survey respondents chose the "Don't know this song" category	MG	13-Mar-14	H	H	Amend any further versions of survey to exclude this category - had to live with it	closed	MG	20-Mar-14	30-Mar-14
106	More suitable sources of songs with multiple attributes was located as a result of further research but too late to use for this project - possibility for future work	MG	03-Apr-14	L	L	Keep in mind for further work	closed	MG	03-Apr-14	03-Apr-14

Dependencies



Dependency Ref	Project	Dependency Description	Raised by	Date Raised	Impact	Priority	Period Affected	Action Plan	Owner	Target Resolution Date	Actual Resolution Date
D01	Module Lecturer	Feedback/advice required regarding data mining and statistical tools	MG	07-Feb-14	H	H	Feb-Apr	Email and arrange meetings	MG	Apr-14	
D02	Project Lecturer	Feedback/advice required	MG	07-Feb-14	H	H	Feb-Apr	Email and/or arrange meetings	MG	Apr-14	
D03	External Expert	Advice required	MG	07-Feb-14	M	M	Feb-Apr	Arrange meetings	MG	Apr-14	

6 Products Completed during Period

6.2 Requirements Specification

Have received feedback from project supervisor and some small changes required.

6.3 Data Analysis

Results from survey have been observed with the Tableau visualisation tool.

6.4 Research

Several papers relating to the Million Song Dataset and Recommender Systems are being read. (Bertin-Mahieux, et al., 2011).

6.5 Presentation

Preliminary results have been uploaded into Tableau but further preparation is necessary before analysis/preparation can be conducted with this particular tool.

6.6 Dissertation

Template identified and draft dissertation in progress.

6.7 Data Analysis

Preparation of MSD files including attributes has commenced. More complicated to retrieve the attributed than originally anticipated.

7 Products due for Completion

By the next period the following products should be completed:

7.2 Data Analysis

Continue dissertation development.

7.3 Survey Results

Perform cluster analysis on survey results. Perform cluster analysis on same songs in MSD subset to identify common attributes within groupings.

8 Impact of Changes

8.1 Schedule

On April 9th the due dates on deliverables both related to this project and other deliverables were extended with the result that there is more time available to complete work on the project. However, it was also revealed that there will be two printed copies of the dissertation required to be delivered which will consume time and resources.

9 Variance from Plan

9.1 Survey

Upon receipt of the survey results it was necessary to prepare the initial results in SQL before they could be visualized in Excel.

10 Planned Work for next period to 3rd June 2014

See Figure 1 below for overview of work completed and work due for next period and beyond.

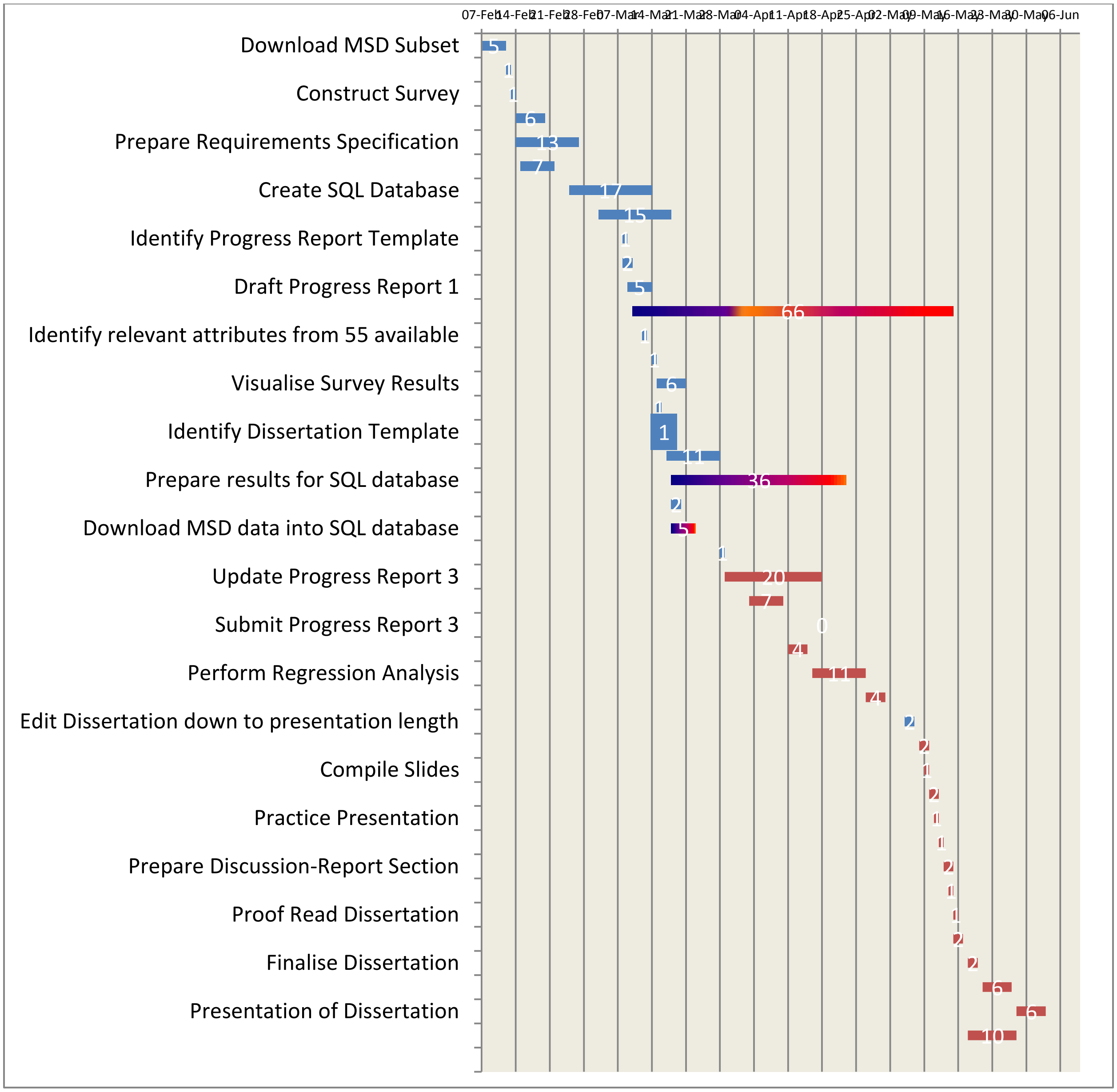


Figure 18: Gantt chart showing tasks completed (in blue) and tasks due (in red). This Gantt chart was created by combining the mind maps in Mind Genius. (MindGenius, 2014)

8.5 Other Material Used

(Any other reference material used in the project for example evaluation surveys etc.

CD containing code should be glued to the technical report.

)

Survey

Survey cover email :

In order to revolutionise the world of music recommender systems, I need to get some music fan input - this is where you come in!
(A "recommender" system is what Amazon uses to recommend books, Cds etc you might like based on previous purchases, or Spotify uses to predict what music you might like to listen to)

For merely the time it takes to complete the attached survey -estimated at around an hour to complete- you could be part of music history (i.e. You will be acknowledged in my dissertation).

- Simply fill a "1" into the cell that is most appropriate for each song on the attached Excel sheet, adding an extra heading if you require it. Below is an example of how to fill out the sheet.
- Then email the completed sheet back to michelegroarke@hotmail.com, before Friday 7th March if possible, and sit back to wait for the results! (I will email you a sample of the "recommenders" choices for your judgement at a later date.)
- If you're not familiar with a song you can either put a "1" in the "Don't Know" column **OR** you can click on the hyperlink to listen to the song and then make your choice (the more helpful and surprisingly enjoyable option)

Feel free to pass on to your friends and acquaintances too - the more results, the better!

Thanks so much for your help on this project.

Please Note: By filling out survey you agree to the results being used in this project.

	Lyrics	Rhythm	Both
Generally, do you listen to a songs' lyrics or are you more influenced by Rhythm?			

Artist	Track	Energetic	Relaxing	Sad	Happy	Summer Music	Hate this Song	Put your own category here if you require	Don't Recognise
The Shangri-Las	Twist and Shout								
The Rolling Stones	Angie (1993 Digital Remaster)								
Chris Rea	Driving Home For Christmas								
Black Eyed Peas	Let's Get It Started								
Public Image Ltd	(This Is Not A) Love Song (Live)								
Mastodon	Shadows That Move								
The Irish Tenors	Mountains Of Mourne								
Dean Martin	Until You Love Someone								
The Black Velvet Band	Dancing To A Standstill								
Gwen Stefani	Harajuku Girls								
Oasis	Morning Glory								
SNOWPATROL	We Wish You A Merry Christmas								
Amy Winehouse	Stronger Than Me								
Creedence Clearwater Revived	Have You Ever Seen The Rain								
Kings Of Leon	Knocked Up								
Mary Black	Turning Away								
Ms. Dynamite	Mr. Prime Minister								
Green Day	Wake Me Up When September Ends (Live at Foxboro)								
Jeff Beck	A Day In The Life (Album Version)								
Kelly Clarkson	My Life Would Suck Without You								
Christina Aguilera	Walk Away								
Michael Jackson	HISTORY								
Naughty By Nature	Swing Swang								
Red Hot Chili Peppers	Naked In The Rain (Album Version)								
Nirvana	Heart-Shaped Box								
Phil Spector	Spanish Harlem								
Avril Lavigne	Innocence								
Muse	Stockholm Syndrome								
David Bowie	Space Oddity (1997 Digital Remaster)								
Rihanna	Good Girl Gone Bad								
Kenny Rogers	The Son Of Hickory Holler's Tramp								
Aerosmith	Angel								
The Rolling Stones	Time Is On My Side								
Jean-Jacques Goldman	LÃ -Bas								
Jimi Hendrix	Fire								
Curtis Mayfield	This Year (ReMastered)								
U2	Vertigo								
Cake	Alpha Beta Parking Lot								
Ben E. King	Let It Be Me								
Backstreet Boys	PDA								
Righteous Brothers	Harlem Shuffle								
Nick Cave & The Bad Seeds	Rock Of Gibraltar								
Backstreet Boys	Shape Of My Heart								
Michael Jackson	Dirty Diana								
The Rolling Stones	Rocks Off								
Ms. Dynamite	Unbreakable								
Daryl Hall & John Oates	Everything Your Heart Desires								
Phil Collins	You Can't Hurry Love (LP Version)								
Janet Jackson	When I Think Of You								
Neil Sedaka	Superbird								
KT Tunstall	Boo Hoo								
Foo Fighters	Next Year								
Arctic Monkeys	From The Ritz To The Rubble								
Basement Jaxx	Sfm								
Dean Martin	Pennies From Heaven								
Blondie	Rapture (Us Disco Version) (1999 Digital Remaster)								
Bryan Ferry	Slave To Love (7" Version) (2009 Digital Remaster)								
Phil Collins	Two Hearts								
Rihanna	SOS								
Happy Mondays	Kinky Afro [Remastered Version]								
Coldplay	A Rush Of Blood To The Head (Live In Sydney)								
OutKast	Return Of The "G								
Iggy Pop	Lust for life (recorded during the us tour in 1986)								
Harry Connick_ Jr.	Here Comes The Big Parade								
Lily Allen	LDN (Warbox Original Cut Dub)								
Hot Chip	Ready For The Floor (Radio Edit)								
Jason Mraz	Song For A Friend (Live From Montalvo)								
Aerosmith	Livin' On The Edge								
The Strokes	Between Love & Hate								
Daft Punk	Da Funk								
Kanye West / Lupe Fiasco	Touch The Sky								
Everly Brothers	Oh What A Feeling								
Gene Pitney	For Me This Is Happy								
The Jam	The Modern World								
Nirvana	Polly								
Billy Bragg	Loving You Too Long								
Amy Winehouse	Valerie								
Muse	Supermassive Black Hole (Album Version)								
Creedence Clearwater Revisited	Proud mary								
ARRESTED DEVELOPMENT	Raining Revolution (Live) (Unplugged)								
Natasha Bedingfield	These Words								
Igor Stravinsky;Columbia Symphony Orchestra	Le Sacre du Printemps (The Rite of Spring)/Part One: The Adoration of the Earth - Introduction								
Aerosmith	Cryin'								
Franz Ferdinand	Do You Want To								
Aerosmith	Crazy								
Lady GaGa	Poker Face								
Chris Rea	Josephine								
Aerosmith	Janie's Got A Gun								
Bob Marley & The Wailers	Three Little Birds								
U2	October								

Ethics Document

National College of Ireland

Human Participants Ethical Review Application Form

All parts of the below form must be completed. However in certain cases where sections are not relevant to the proposed study, clearly mark NA in the box provided.

Part A: Title of Proiect and Contact Information

Name

Michele Groarke

Student Number (if applicable)

X13120654

Email

Michele.groarke@student.ncirl.ie

Status:

Undergraduate □
Postgraduate Y
Staff □

Title of Research Project

Analysis of Million Song Data Set to create an Accurate Recommender System

Have you read the NCI Ethical Guidelines for Research with Human Participants?

Yes Y

No □

Please indicate any other ethical guidelines or codes of conduct you have consulted

N/A

Has this research been submitted to any other research ethics committee?

Yes □

No N

If yes please provide details, and the outcomes of this process, if applicable:

N/A

Is this research supported by any form of research funding?

Yes □

No N

If yes please provide details, and indicate whether any restrictions exist on the freedom of the researcher to publish the results:

N/A

Part B: Research Proposal

Briefly outline the following information (not more than 200 words in any section).

Proposed starting date and duration of project

14th February 2014 – 12 weeks

The research aims and objectives

The aim of this research is to accurately predict which new songs a user will potentially like according to feelings and mood as opposed to classifying songs according to “artist” or “genre”

The rationale for the project

Recommender systems are increasingly replacing word of mouth referral and music recommender systems in particular, so this is a very topical subject.

The research design

The data retrieved from the survey will be analysed firstly on its own and then using the attributes from the Million Song Dataset, to create a prediction model that will be able to tailor recommendations for each individual user.

The methods of data collection

The data was collected via a survey in the form of an Excel sheet attached to an email.

The research sample and sample size

The participants were 20 volunteer participants who responded from a population of 132 personal contacts and classmates.

The nature of any proposed pilot study

The questions related to a 100 song subset of the Million song dataset. Participants were required to indicate their feeling about each track from 8 possibilities. This was to help identify the feelings each user had about the songs and also to identify songs that were common in feeling across all users. The study is questionnaire based.

The methods of data analysis

The data will be analysed via Excel, SQL, some Python in the preparation stages and there will be statistical analysis and clustering in R and Weka. Descriptive, inference statistics and modelling.

Part C: Ethical Risk

Please identify any ethical issues which will arise and how you will address them.

There will be no ethical issues as there is no personal information supplied that would identify the participants.

Please indicate any risk of harm or distress to participants.

There is no risk, harm or distress to participants.

Please indicate how you will address this risk (e.g. debriefing procedures, etc.).

N/A

Do the participants belong to any of the following vulnerable groups?

(Please tick all those involved).

□ Children;

□ The very elderly;

□ People with an intellectual or learning disability

□ Individuals or groups receiving help through the voluntary sector

□ Those in a subordinate position to the researchers such as employees

□ Other groups who might not understand the research and consent process

□ Other vulnerable groups

How will the research participants in this study be selected, approached and recruited?

The research participants were selected from my email contacts and student group and were emailed from my personal email address.

What inclusion or exclusion criteria will be used?

Will be excluded if under 18

How will participants be informed of the nature of the study and participation?

The participants were informed of the nature of the study and participation by the covering email.

What procedures will be used to document the participants' consent to participate?

By completing the survey they agreed to participate as was noted in the body of the covering email.

If vulnerable groups are participating, what special arrangements will be made to deal with issues of informed consent/assent?

N/A

Please include copies of any information letters and consent forms with the application.

Part D: Confidentiality and Data Protection

Please indicate the form in which the data will be collected.

Y Identified   □ Potentially Identifiable   □ De-Identified

What arrangements are in place to ensure that the identity of participants is protected?

I will be removing any identifiers from the data before processing and presenting/publishing.

Please indicate any recording devices being used to collect data (e.g. audio/video).

No recording devices will be used

Please describe the procedures for securing specific permission for the use of these recording devices in advance.

N/A

Please indicate the form in which the data will be stored.

□ Identified □ Potentially Identifiable Y De-Identified

Who will have responsibility for the data generated by the research?

Michele Groarke

Please describe the procedures of the storage and destruction of data.

The data is stored on my Dell Latitude D630 in the project file, also backed up on a USB key. Any identifiable data will be deleted directly after Project has been graded.

Dissemination and Reporting

Please describe how the participants will be informed of dissemination and reporting (e.g. submission for examination, reporting, publications, presentations)?

Upon submission, the participants will receive an email with a brief report of the results and possible publication.

If any dissemination entails the use of audio, video and/or photographic records (including direct quotes), please describe how participants will be informed of this in advance.

N/A

Part E: Signed Declaration

I confirm that I have read the NCI Ethical Guidelines for Research with Human Participants, and agree to abide by them in conducting this research. I also confirm that the information provided on this form is correct.

Signature of Applicant

Date 12/04/14

Signature of Supervisor (where appropriate)

Date ____________________________

National College of Ireland

Human Participants Ethical Review Exemption Form

All parts of the below form must be completed. However in certain cases where sections are not relevant to the proposed study, clearly mark NA in the box provided.

Part A: Title of Proiect and Contact Information

Name

Michele Groarke

Student Number (if applicable)

X13120654

Email

Michele.groarke@student.ncirl.ie

Status:

Undergraduate □
Postgraduate Y
Staff □

Title of Research Project

Analysis of Million Song Data Set to create an Accurate Recommender System

Have you read the NCI Ethical Guidelines for Research with Human Participants?

Yes Y
No □

Please indicate any other ethical guidelines or codes of conduct you have consulted

N/A

Has this research been submitted to any other research ethics committee?

Yes □
No N

If yes please provide details, and the outcomes of this process, if applicable:

N/A

Is this research supported by any form of research funding?

Yes □
No N

If yes please provide details, and indicate whether any restrictions exist on the freedom of the researcher to publish the results:

N/A

Part B: Research Proposal

Briefly outline the following information (not more than 200 words in any section).

Proposed starting date and duration of project

14th February 2014 – 12 weeks

The research aims and objectives

The aim of this research is to accurately predict which new songs a user will potentially like according to feelings and mood as opposed to classifying songs according to “artist” or “genre”

The rationale for the project

Recommender systems are increasingly replacing word of mouth referral and music recommender systems in particular, so this is a very topical subject.

The research design

The data retrieved from the survey will be analysed firstly on its own and then using the attributes from the Million Song Dataset, to create a prediction model that will be able to tailor recommendations for each individual user.

The methods of data collection

The data was collected via a survey in the form of an Excel sheet attached to an email.

The research sample and sample size

The participants were 20 volunteer participants who responded from a population of 132 personal contacts and classmates.

The nature of any proposed pilot study

The questions related to a 100 song subset of the Million song dataset. Participants were required to indicate their feeling about each track from 8 possibilities. This was to help identify the feelings each user had about the songs and also to identify songs that were common in feeling across all users. The study is questionnaire based.

The methods of data analysis

The data will be analysed via Excel, SQL, some Python in the preparation stages and there will be statistical analysis and clustering in R and Weka. Descriptive, inference statistics and modelling.

Part C: Grounds for Exemption

Please indicate the grounds on which you are applying for exemption from ethical review.

Please confirm that the research does NOT involve any of the following:

- □ Vulnerable groups
- □ Sensitive topics
- □ Risk of psychological or mental distress
- □ Risk of physical stress or discomfort
- □ Any other risk to participants
- □ Use of drugs or invasive procedures (e.g. blood sampling)
- □ Deception or withholding information from participants
- □ Conflict of interest issues
- □ Access to data by individuals/organisations other than the researchers
- □ Any other ethical dilemmas

Part D: Signed Declaration

I confirm that I have read the NCI Ethical Guidelines for Research with Human Participants, and agree to abide by them in conducting this research. I also confirm that the information provided on this form is correct.

Signature of Applicant _

Date **12/04/2014**

Signature of Supervisor (where appropriate)

____

Date ___________________________________

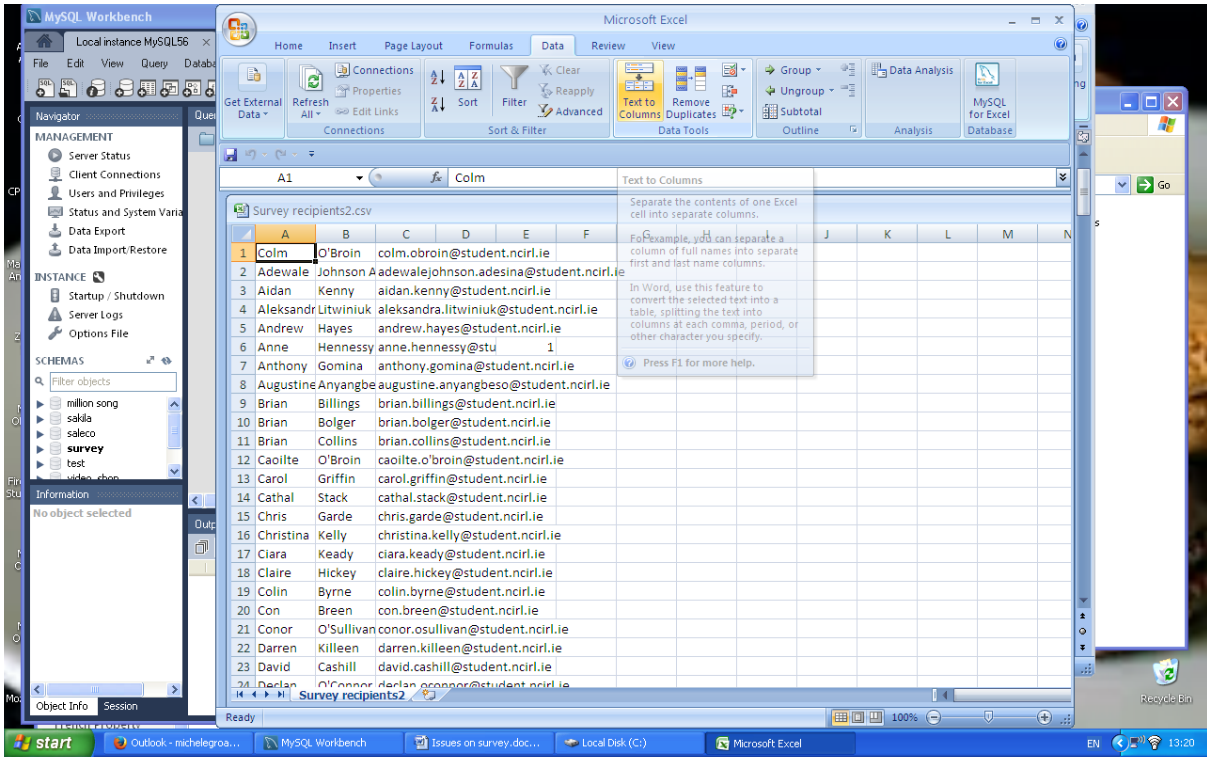
CSV File Creation

Copied the list of names from ‘sent’ email folder into notepad, saved, then opened from Excel, using ‘;’ as initial delimiters.

This opened the file with all the respondents and their emails in one row. I copied whole row and ‘special’ pasted, choosing the ‘transpose’ option.

Then I had one column with a separate row for each respondent.

To separate the names from the email addresses, I used the ‘Text to columns’ option in the ‘Data’ tab. This time I used ‘(‘ as delimiters. To get rid of the ‘)’ characters, I repeated, using ‘)’ as delimiters. Finally I separated Firstnames from lastnames by repeating with ‘ ‘, as delimiter.



Text to Columns

Separate the contents of one Excel cell into separate columns.

For example, you can separate a column of full names into separate first and last name columns.

In Word, use this feature to convert the selected text into a table, splitting the text into columns at each comma, period, or other character you specify.

Press F1 for more help.

	A	B	C	D	E	F
1	Colm	O'Broin	colm.obroin@student.ncirl.ie			
2	Adewale	Johnson A	adewalejohnson.adesina@student.ncirl.ie			
3	Aidan	Kenny	aidan.kenny@student.ncirl.ie			
4	Aleksandr	Litwiniuk	aleksandra.litwiniuk@student.ncirl.ie			
5	Andrew	Hayes	andrew.hayes@student.ncirl.ie			
6	Anne	Hennessy	anne.hennessy@stu		1	
7	Anthony	Gomina	anthony.gomina@student.ncirl.ie			
8	Augustine	Anyangbe	augustine.anyangbeso@student.ncirl.ie			
9	Brian	Billings	brian.billings@student.ncirl.ie			
10	Brian	Bolger	brian.bolger@student.ncirl.ie			
11	Brian	Collins	brian.collins@student.ncirl.ie			
12	Caoilte	O'Broin	caoilte.o'broin@student.ncirl.ie			
13	Carol	Griffin	carol.griffin@student.ncirl.ie			
14	Cathal	Stack	cathal.stack@student.ncirl.ie			
15	Chris	Garde	chris.garde@student.ncirl.ie			
16	Christina	Kelly	christina.kelly@student.ncirl.ie			
17	Ciara	Keady	ciara.keady@student.ncirl.ie			
18	Claire	Hickey	claire.hickey@student.ncirl.ie			
19	Colin	Byrne	colin.byrne@student.ncirl.ie			
20	Con	Breen	con.breen@student.ncirl.ie			
21	Conor	O'Sullivan	conor.osullivan@student.ncirl.ie			
22	Darren	Killeen	darren.killeen@student.ncirl.ie			
23	David	Cashill	david.cashill@student.ncirl.ie			
24	Declan	O'Connor	declan.oconnor@student.ncirl.ie			

[i] https://www.spotify.com

[ii] http://www.spotiseek.com

[iii] http://www.last.fm/

[iv] www.amazon.com

[v] http://winpython.sourceforge.net/