Table of content

Table of figures

Dedication

I dedicate this work to my family, my parents Karim Khoufi and Monia Bouchlaghem, for giving birth to me at the first place and supporting me spiritually throughout my life, without them I wouldn't be able to be where I am today.

Thank you for always believing in me.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Mr. Sebastien Doubey for the continuous support of my work, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of my internship.

Besides my advisor, I also would like like to thank my college supervisor Mrs. Afef Kacem for her advice the rest and encouragement, insightful comments.

Besides, I would like to thank the rest of the committee professors, Mrs. Ahlem BEN YOUNES my reviewer and Mrs. Meriem Riahi the committee president.

General introduction

In an era where knowledge is key, one must possess the information so he cannot only single out his work but also gain the maximum benefit from it.

The problem is we are now overwhelmed with data, the amount of data collected is increasing by the day, especially, now that computers makes it more and more easy to save things that we would have trashed if we did not get access to those inexpensive disks and online storages. We are buried underneath a huge amount of information, which we, simple human beings, fail to explore at it full extent and extract the most valuable information we can use.

Our mission is to fill in our music database with the maximum of data we can find by comparing two solutions the data mining developing solution and the API consuming solution.

So in this report, we will present the company and its system in the first chapter, next we will explain our data mining notions. In the third chapter however, we will be detailing our specification diagrams, and finally we will set up our results and achievements.

Chapter 1 Company and project overview

Introduction

In this chapter, we will present the company where the internship was held: Soundytics, a services provider to music platforms and an approach distributor for shops, stationed in Luxembourg.

We will start by a brief presentation of the company, its goals and then we will detail the services that it offers to its clients. We will also have an overview about how the system works technically by use-cases diagrams and database schemas.

1 Soundytics presentation

Soundytics is a startup founded in 2013 by two co-founders Sébastien Doubey as CTO and Hugo Bon as CEO.

Soundytics is a music smart recommendation and search solution based on deep audio signal analysis and listening context detection.

Soundytics is the solution to create accurate and cross-devices experiences from large music catalogs in web, mobile and hardware environments.

Soundytics analyzes the audio signal, extracts the songs musical DNA and enriches it with editorial, semantic and social data mining. We provide an automatic indexation solution of music catalogs to create enhanced user experiences and ultra-customizable recommendation services.

1.1 Goals

Personalized recommendation is the key to catch users and develop the best music applications. The automated analysis of metadata from recordings is crucial to face the current inflation of multimedia data and to classify music. It is a challenge for music and entertainment industry.

Streaming platforms are looking for solutions to describe their catalog and recommend music: The famous Echonest music intelligence API has been purchased in March 2014 for 100M$ by Spotify, Songza by Google for 40M$, Aupeo by Panasonic (music embedded in connected cars), Playlists.net by Warner, TastemakerX by Radio, etc.

Cars manufacturers are looking for technologies to build their own strong ecosystems facing to new tech companies and interact with their customers.

Inflight Entertainment systems have to offer great onboard experiences to passengers to compete low-cost and keep first class customers.

Retailers groups are looking for new solutions to embed the customer in-store experience, especially in adapting the music to the time of the day and clients type.

Soundytics is the link between catalog owners and music daily life broadcasting.

1.2 Products and services

The Soundytics technology automatically detects the musical DNA of songs (mood, genre, rhythm, voice presence, tempo, instruments…) and generates metadata from massive semantic web analysis (country, decade, artist influences, language, lyrics theme, popularity…). This feature is implemented but the data based on is not sufficient, which makes the mission of my internship is to collect the maximum of information.

It is based on two main products:

1.2.1 Soundytics Prelude API

The Soundytics Prelude API is the application-programming interface that provides to music services the data and services to enhance their recommendation and search systems.

This service is destined to for music services and catalog owners.

Its features:

- **Auto tagging**: Automatic indexation of music catalogs and optimization of the management of musical databases.
- **Advanced search**: Playlist suggestion automatically generated through acoustic, semantic and social criteria selected by users.

- **Similarity**: Suggestion of similar tracks to the one currently played through recognition of common acoustic properties.
- **Themed playlist**: Create smart radios by matching acoustic criteria with external properties.
- **Personalized recommendation**: Suggestions by learning users tastes with Like, Dislike, Skip and Favorite features, suggestions according to the context and moment of the day, but that functionality in yet to be developed, as soon as we be collect enough information about users.

1.2.2 Soundytics Prelude UI

The Soundytics Prelude application enables to deploy a musical experience on music catalogs. With analysis and automatic classification of the tracks in a dedicated interface, publishers smartly recommend their content, increase visibility and value their offer to professional customers.

Prelude UI is a web interface for retailers and hotels with several features:

- Playlist creation from criteria like mood, tempo, genre, percussivity, voice, etc.
- Several pre-featured atmospheres
- Create playlist from an external track or MP3 (ex: all songs similar to Riana's one)
- Scheduling and edition of the playlists
- Playlists download for offline listening
- A sound box is available in option if the device is not connected to speakers.

1.3 Business model

The Soundytics business plan is detailed in the Figure 1 below:

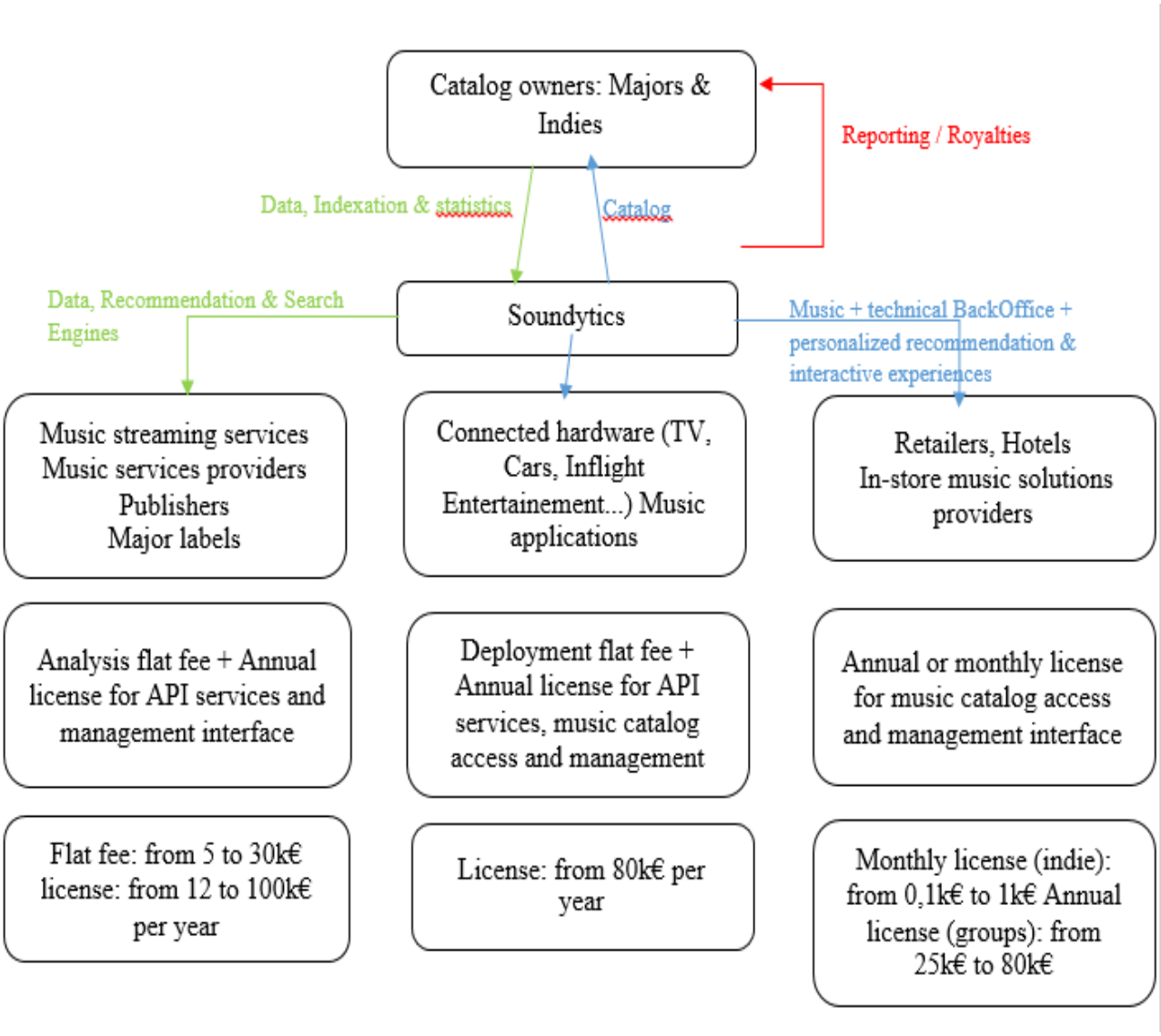


Figure 1 Business model

2 Existent work

The introduction to the company made above is still maybe somewhat vague technically. We need to explain further, what it is all about since our work is an extension to what already exists. Therefore, in this part we are going to detail our vision of the existing code.

2.1 Functionalities

To insure some visibility we will represent the functionalities in a general use-case diagram (see Figure 2) and then explain in detail each one.

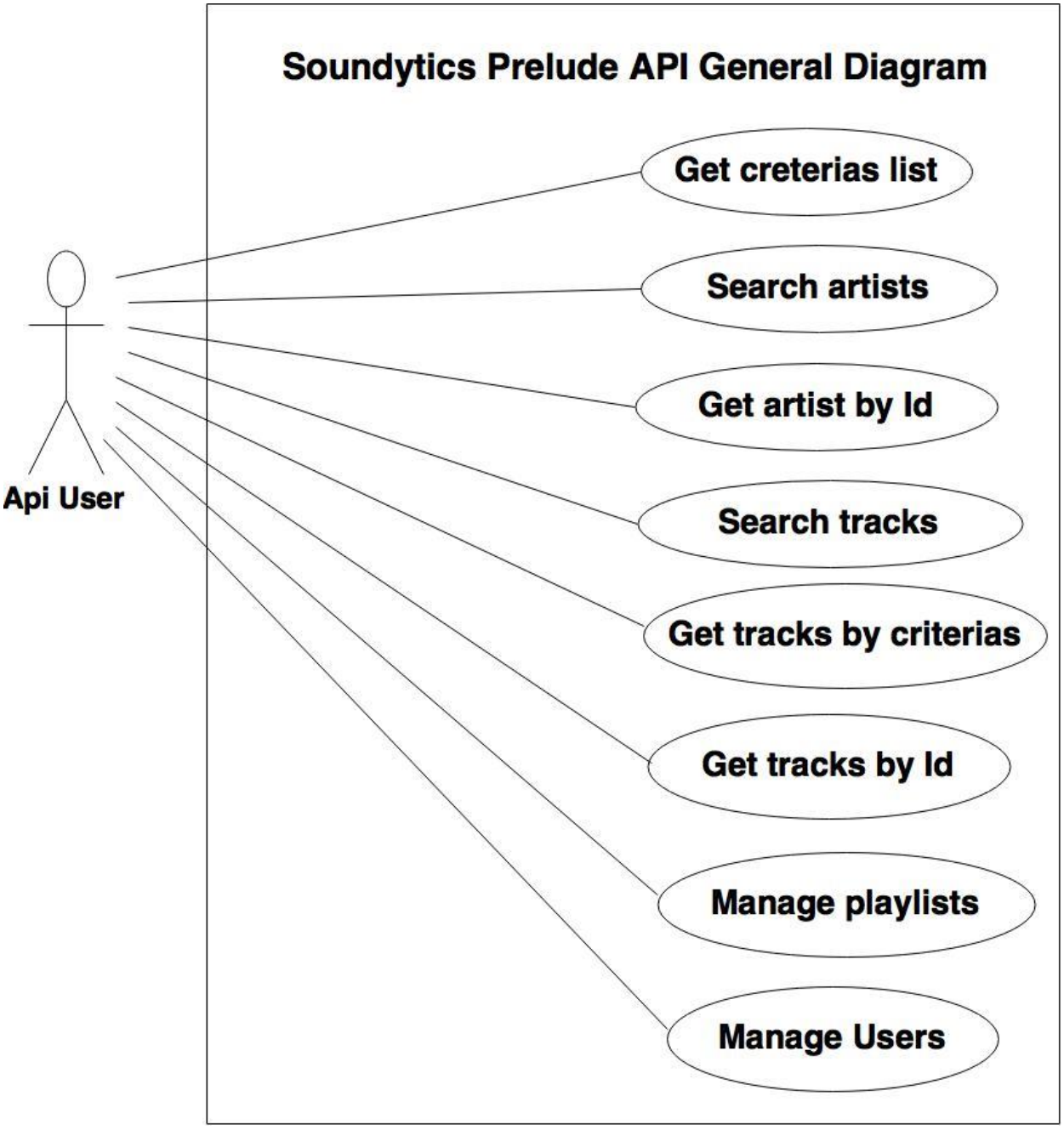


Figure 2 General use-case API diagram

2.1.1 Get criteria list

The criteria request is a request meant to be helpful for users, so they can get the list of all criteria in English or French to be used as entries in further functionalities (see Figure 3).

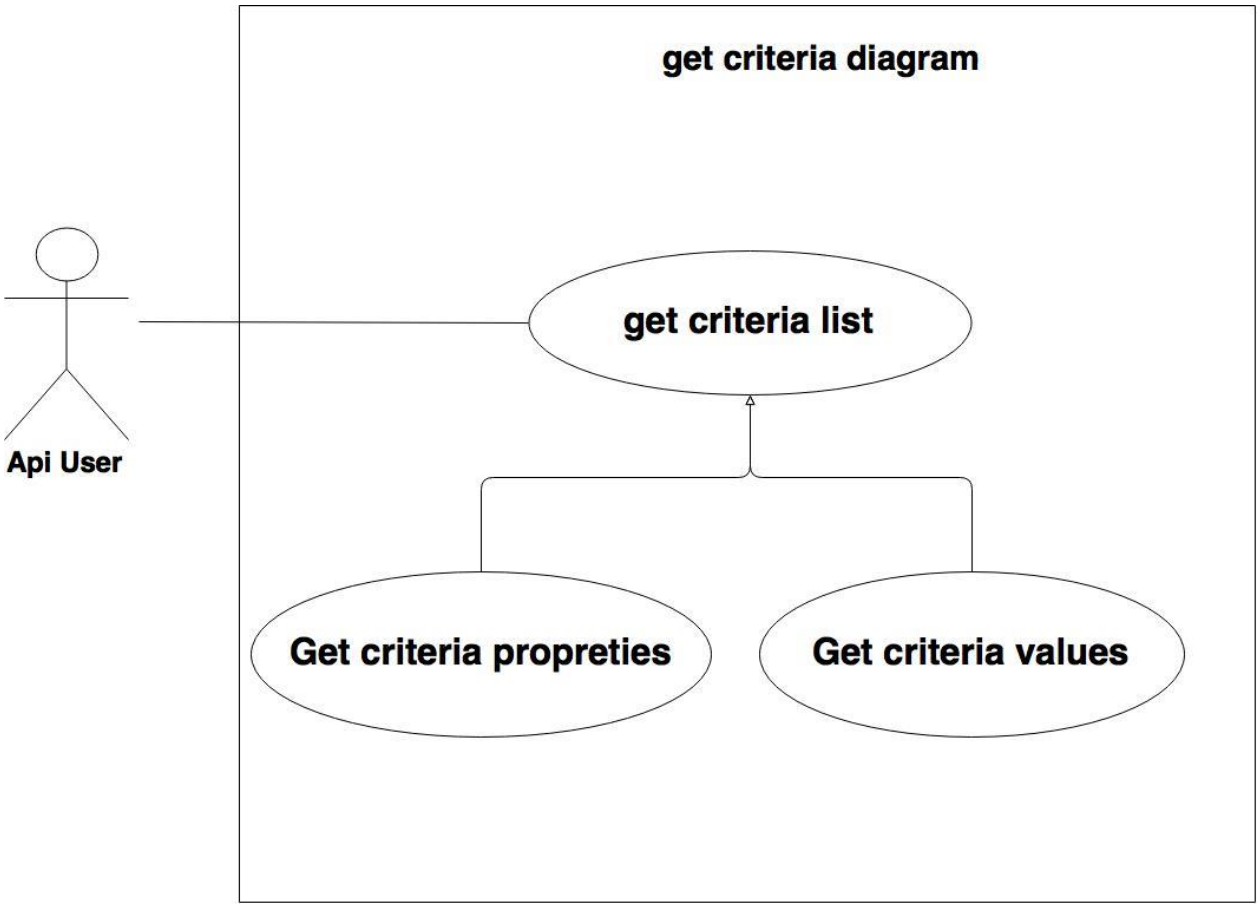


Figure 3 Get criteria use-case diagram

This is a list of the criteria properties and their values:

- **Style***:* classic, country, hip-hop, electronica, soul_rnb, rock, jazz
- **Mood:** melancholic, dramatic, energetic, stressful, aggressive, relaxing, calm, happy, danceable, fun
- **Tempo:** fast, midfast, midtempo, slowmid, slow
- **Rhythmic**: percussive, nonpercussive

- **Sound texture**: hard, soft

2.1.2 Search artists

With this request we can get a list of artists by country, by type (solo, duo or group), by gender (male or female, if it is a solo artist of course), or simply just by name (see Figure 4).

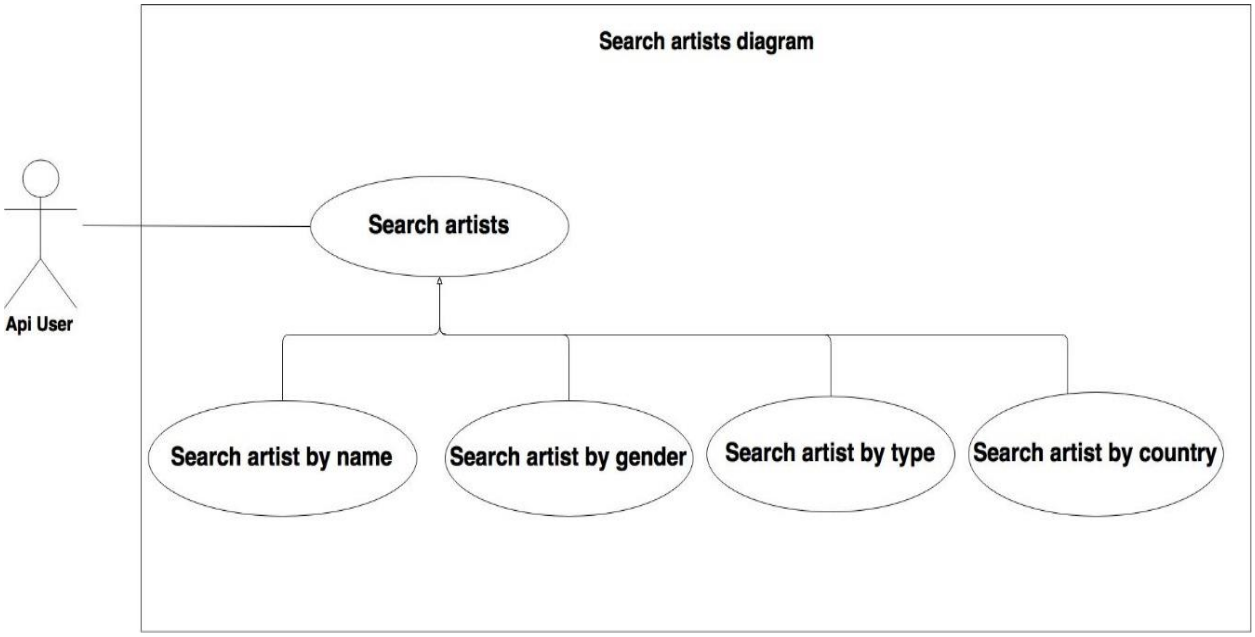


Figure 4 Search artists use-case diagram

It returns a list of artists corresponding to the research with the basic information available on each one, including an id, which can be used in the following functionality.

2.1.3 Get artist by id

This functionality returns all the information available about a specific artist represented by his id, we can of course ask for a specific information only like the external ids (the artist id in well-known databases like Wikipedia or MusicBrainz…), the biography, videos, images and official websites related to this artist (see Figure 5).

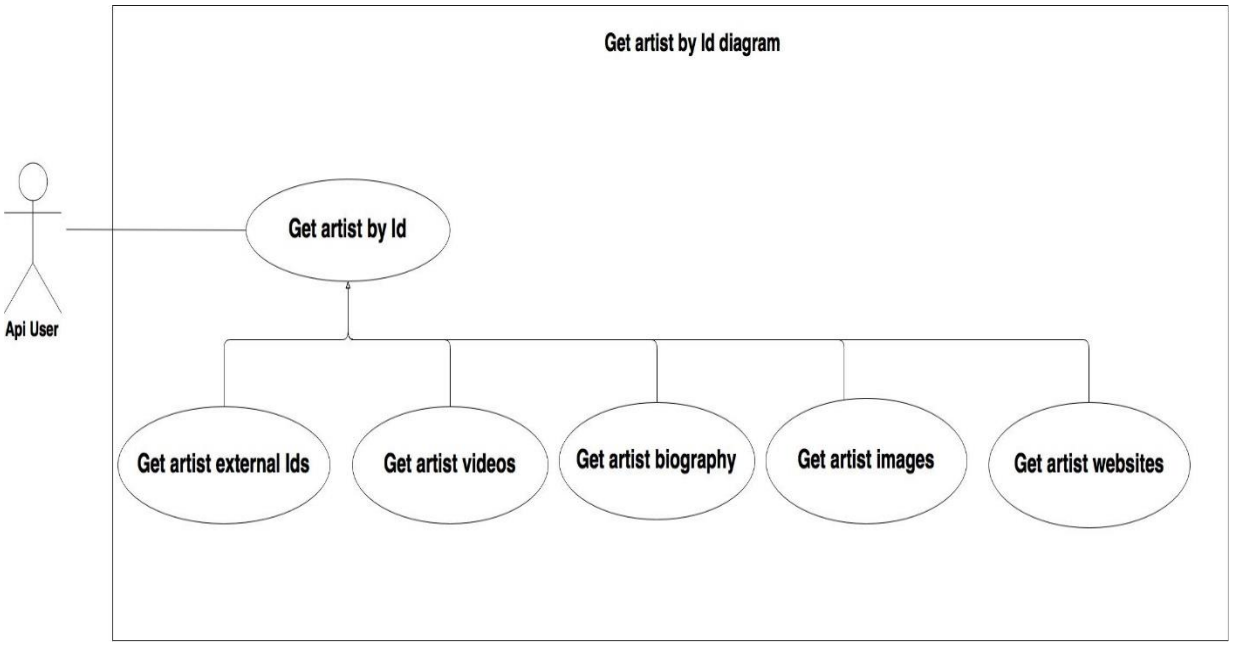


Figure 5 Get artist by id use-case diagram

2.1.4 Search tracks by criteria

This case does not need to be represented by more than what was noted in the general use-case diagram, and yet it is a very important functionality, maybe even the most important one alongside with the similarity search. This functionality is based on very effective learning algorithms to extract every song's criteria (see Figure 6).

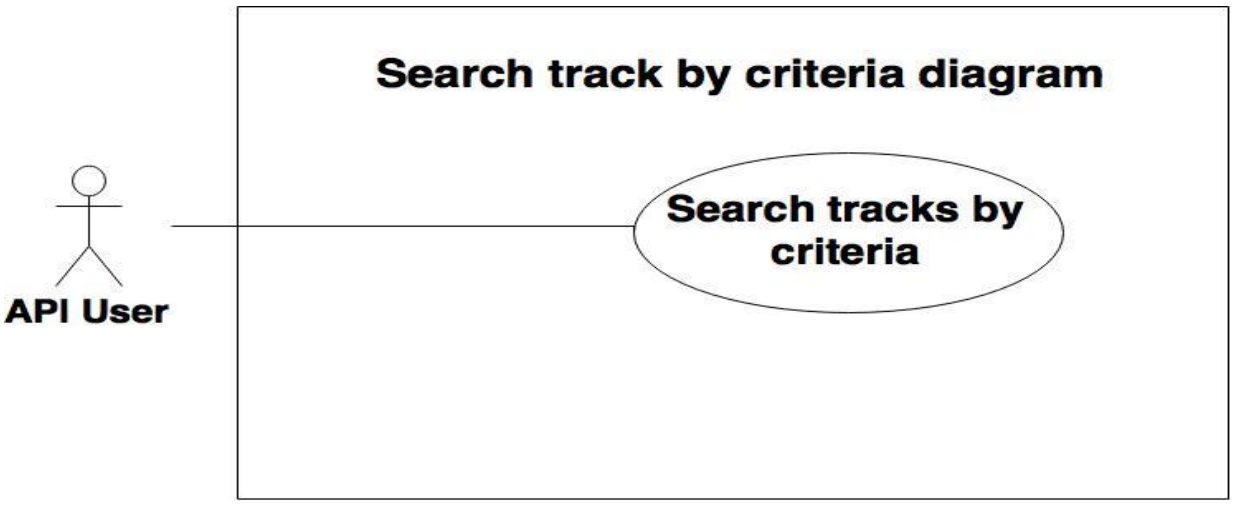


Figure 6 Search tracks by criteria use-case diagram

It is the functionality that can help to “create playlists” based on the user’s mood and preferences.

2.1.5 Search tracks

Track searching is very similar to search artist, only for tracks. We can browse for a list of tracks by its name or by the name of its singer but we can also combine both for more specificity (see Figure 7).

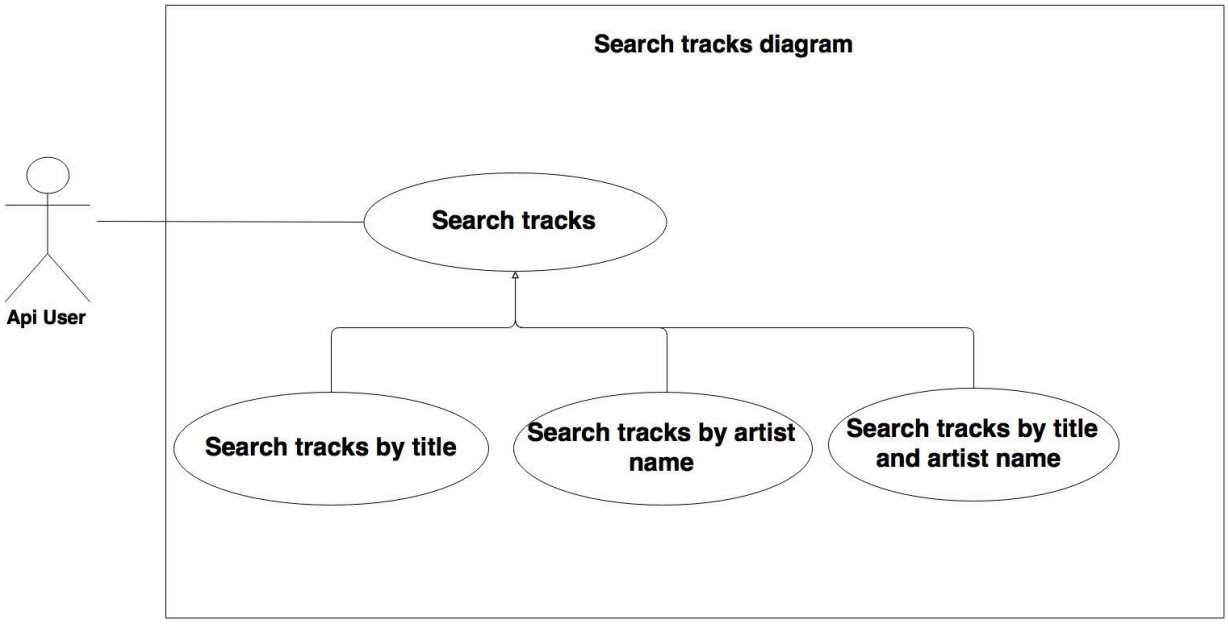


Figure 7 Search tracks use-case diagram

2.1.6 Search track by id

This one is also similar to the search artist by id functionality. We use the identifiers that we got from the search tracks request to get more details on a track, or a specific property like the track own criteria, its images, external ids or more importantly the stream of the track that we use to read it on a player (see Figure 8).

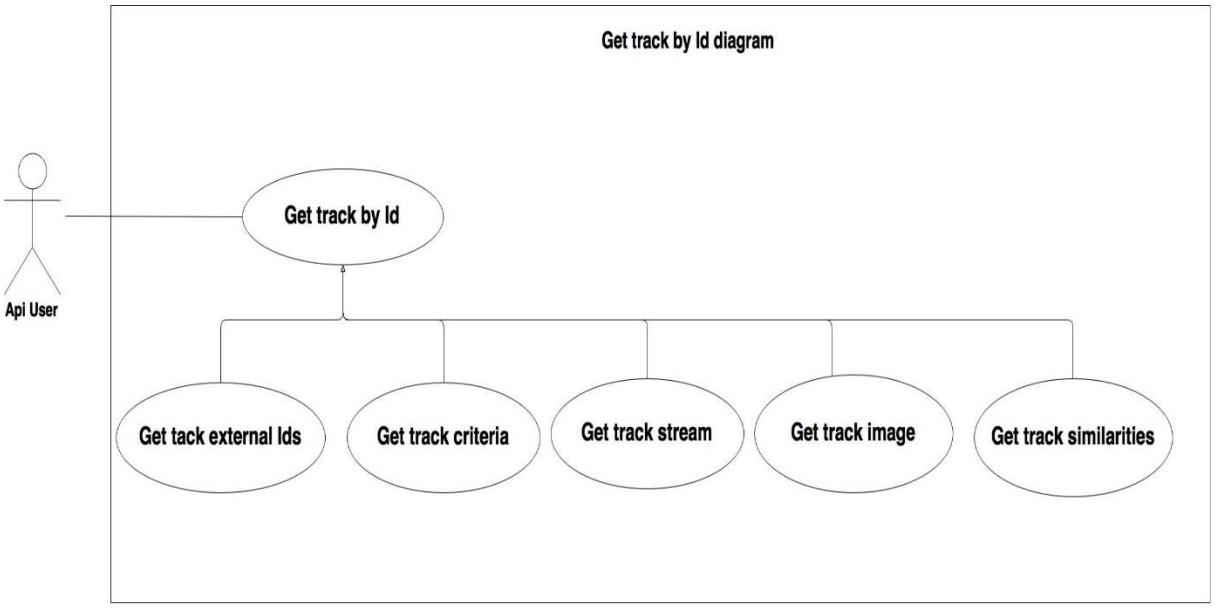


Figure 8 Search track by id use-case diagram

This feature also holds an important functionality, which is get similar tracks. It is used to procure the tracks that are similar to the specific track in criteria.

2.1.7 Manage playlists

It is a very basic functionally used to manage playlists like creating an empty playlist, adding songs to it, deleting songs or deleting the whole playlist and getting images as well as the playlist criteria (see Figure 9).

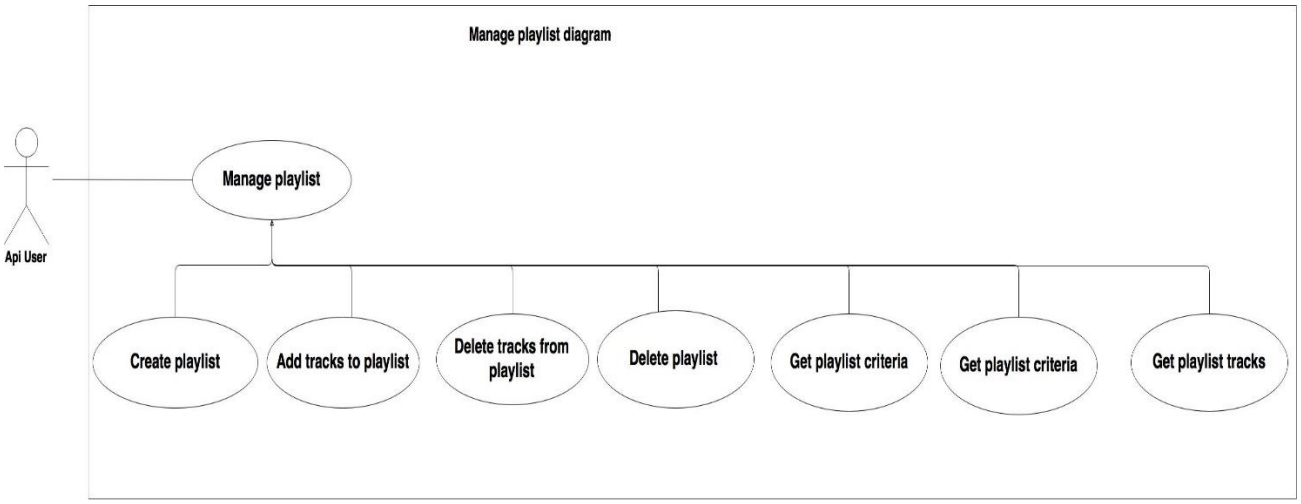


Figure 9 Manage playlists use-case diagram

2.1.8 Manage users

Like the previous case, this one is about managing user accounts, with basic features like adding, deleting, modifying and verifying if users credentials are correct and getting the user's playlists (see Figure 10).

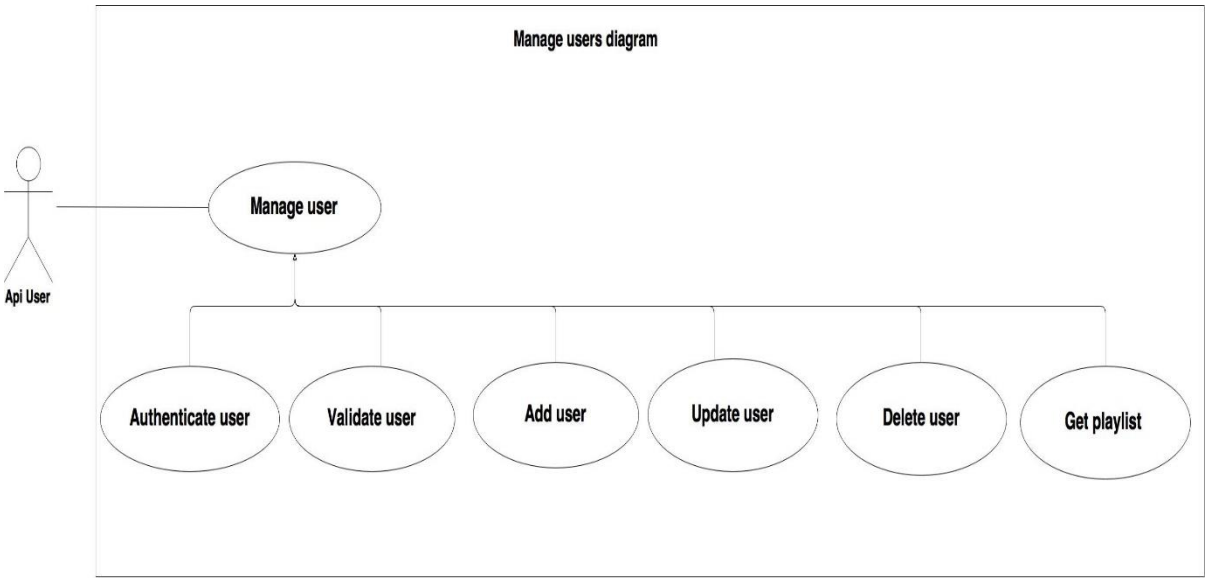


Figure 10 Manage users use-case diagram

2.2 Database architecture

The database architecture is quite complicated to be completely shown in this report in Figure 11. So, we will focus on what is important for our work: the tracks and especially the artists.

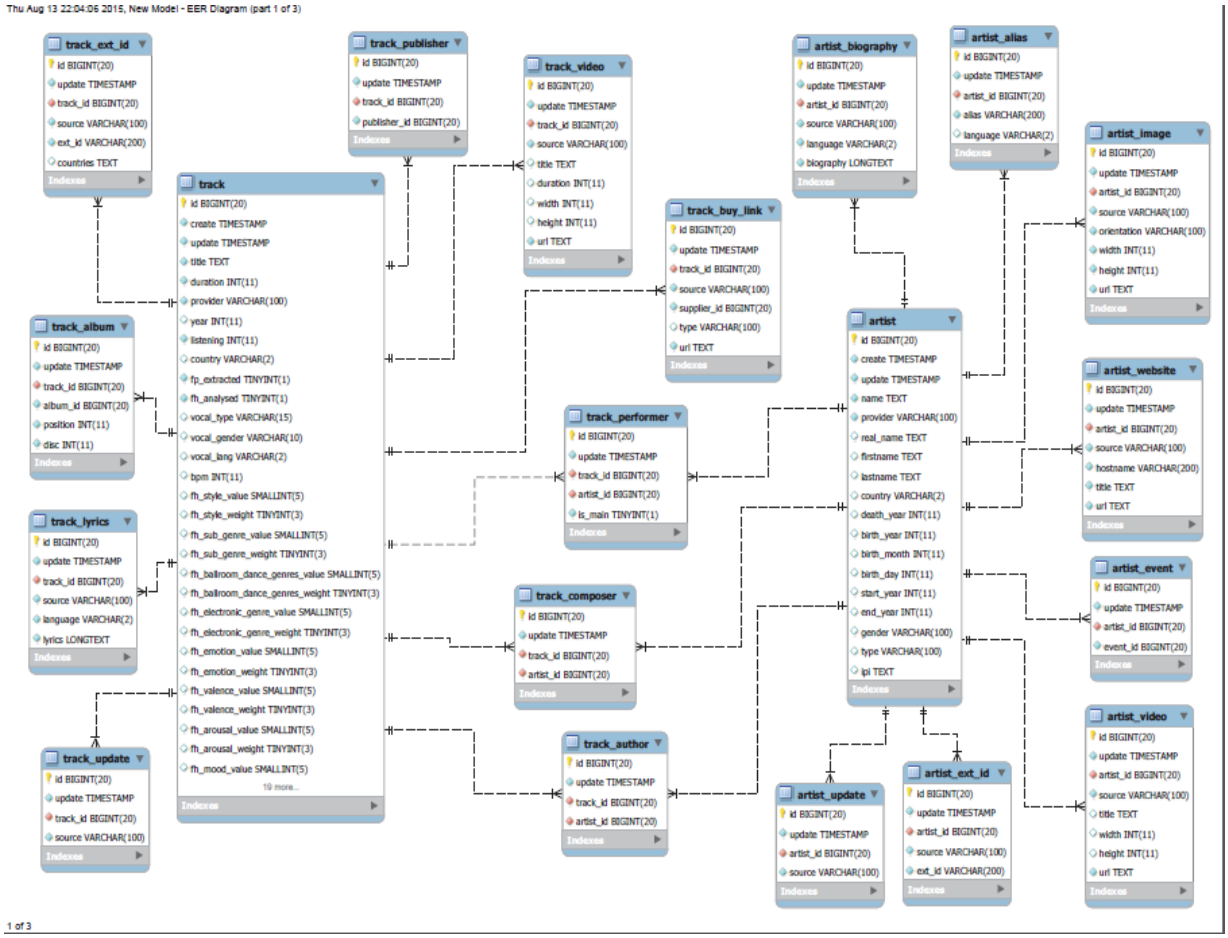


Figure 11 Database diagram

In the use-case explanation, we talked about what we can procure as track information like the external ids, or the videos, the images, but there is also the album to which this track belongs, the publisher, the composer, the performer or the author whom are artists, or the link to buy the track. An artist has biographies, alias, events to come, websites, videos and images.

Conclusion

In this chapter, we tried to introduce as best as we can how the Soundytics system is running by presenting the company, its goal and features as a whole. We finally detailed the functionalities and the system architecture.

Chapter 2 Data mining notions overview

Introduction

After reading the previous chapter, which covers the base project of the society, the API that allows the users to get all the information about artists and tracks. In fact, the existent system is fully finished in terms of operations that can handle the database components but the database itself lucks of data.

The system uses a mainstream or an indie catalog to fill the tracks and artists tables, and then it processes the tracks into knowing their criteria. However, both the tables are still missing quite a lot of information like the gender of the artist or the language of a song.

In order to fulfill that mission we are going to compare two techniques of comparing data. One is an internal, self-implemented technique based on data mining algorithms. The other is external API based.

In this chapter, we will explain notions like data mining, document classification and defining the algorithms, we are working with, like Naïve Bayes and the one class classification algorithms one and the data mining process they fit into.

1 Data mining

Data mining is the process of extracting and exploiting patterns in data. That process can be automatic or a semi-automatic in a large quantity of data.

The patterns extracted needs to be meaningful and advantageous to make non-trivial prediction on new data. In essence, data mining is the acquisition of knowledge and the ability to use it [1].

1.1 Application fields

1.1.1 Web mining

It is commonly used in search engines and page ranking. Search engines mine the web content and the content of the queries circulating the web to select advertisement that the user might be interested in and to improve the search results next time.

1.1.2 Screening images

Since we have sent satellites in orbit and we received images from them, environmental scientists have been trying to use them to give an early warning of ecological disasters and predict daily weather.

1.1.3 Load forecasting

To make significant economies in the electricity supply industry it is important to be able to determine future demands for power as far in advance as possible. That is where machine learning became handy and automated load forecasting assistant has been operating for years to generate hourly forecast two days in advance.

1.1.4 Diagnosis

When we talk diagnosis in this field, the first thing that comes to our minds is expert systems which are based on rules that are handcrafted at first which can become quite laborious, this is why we added a learning process to ease the work, but we still check afterword since it is not fully achieved.

1.1.5 Marketing and sales

This field is where data mining is the most active. The companies usually process massive volumes of recorded data, which, with the help of data mining, are used to help predict costumers' behavior.

1.1.6 Other fields

Machine leaning is used in various and innumerable fields like biology, biomedicine, astronomy, chemistry and even everyday use for recommendation in entertainment for example the Videos on demand platform Netflix or the music platform Spotify.

1.2 Data mining and ethic

The data mining approach has serious ethical implications, depending on the applications it is used in. Using sexual and racial information for medical diagnosis for example is ethical but using zip codes and ethnic identities to mine loan payment behavior is not.

2 Machine learning

Things learn when they change their behavior in a way that makes them better in the future. One can test learning by comparing present behavior and past one. However, learning is a philosophical concept and there by a slippery concept, because not all that can change and become more

performant is learning. In this field of study we also use the term training, the difference is that learning implies thinking and a purpose; it is intentional but if it does not implies a purpose it is merely training.

Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Machine learning is therefore a study field of artificial intelligence.

2.1 Types of learning

There are three common types of machine learning, according to the context and the data we have:

2.1.1 Supervised learning

We talk about supervised learning, or what is more commonly called classification, when classes (labels, categories) and testing examples are known in advance.

First, an expert needs to label each example instance according to what class they belong. Then our chosen algorithm figures out a pattern in this example set, a pattern needed in the second phase, which consists on predicting every new entering data label.

It is usually preferable to determine a probability of affiliation to each class.

This type of learning has different approaches like:

- Support Vector Machines or simply SVM
- K-Nearest Neighbors algorithm
- Artificial neural network
- Bayesian statistics
- Decision tree learning
- Gaussian process regression
- Kernel estimators
- Naïve Bayes classifier
- Random Forests
- Ensembles of Classifiers
- Ordinal classification…

2.1.2 Unsupervised learning

When the system or the operator holds a set of examples but cannot figure out their label or the number or nature of classes, we call unsupervised learning; because the algorithm will have to find by itself a pattern and not based on a learned one.

The most known approach for this type of learning is the data-clustering algorithm.

Approaches to unsupervised learning include:

- Clustering (e.g. k-means, mixture models, hierarchical clustering) [2]
- Approaches for learning latent variable models such as
- Expectation–maximization algorithm (EM)
- Method of moments
- Blind signal separation techniques,
- Principal component analysis,
- Independent component analysis,
- Non-negative matrix factorization,
- Singular value decomposition

2.1.3 Semi-supervised learning

This method can be probabilistic or not. It starts with a semi-labeled set of examples. We mention semi-supervised learning when we to which class the instance does not belongs to but we do not know to which one it actually belongs.

Some methods for semi-supervised learning:

- Generative models
- Low density separation
- Graph based methods
- Heuristic approaches

We are not going to detail the other types such as reinforcement learning or inductive transfer. In the rest of our work, we will concentrate on the supervised approach, more specifically the naïve Bayes algorithm, and the one class classification one, in terms of document classification.

3 Data mining process

When data mining analysis is conducted, a general process is usually followed like the CRISP-DM process [2], which stands for Cross-Industry standard Process for Data Mining. It consists of six phases like stipulated in Figure 12:

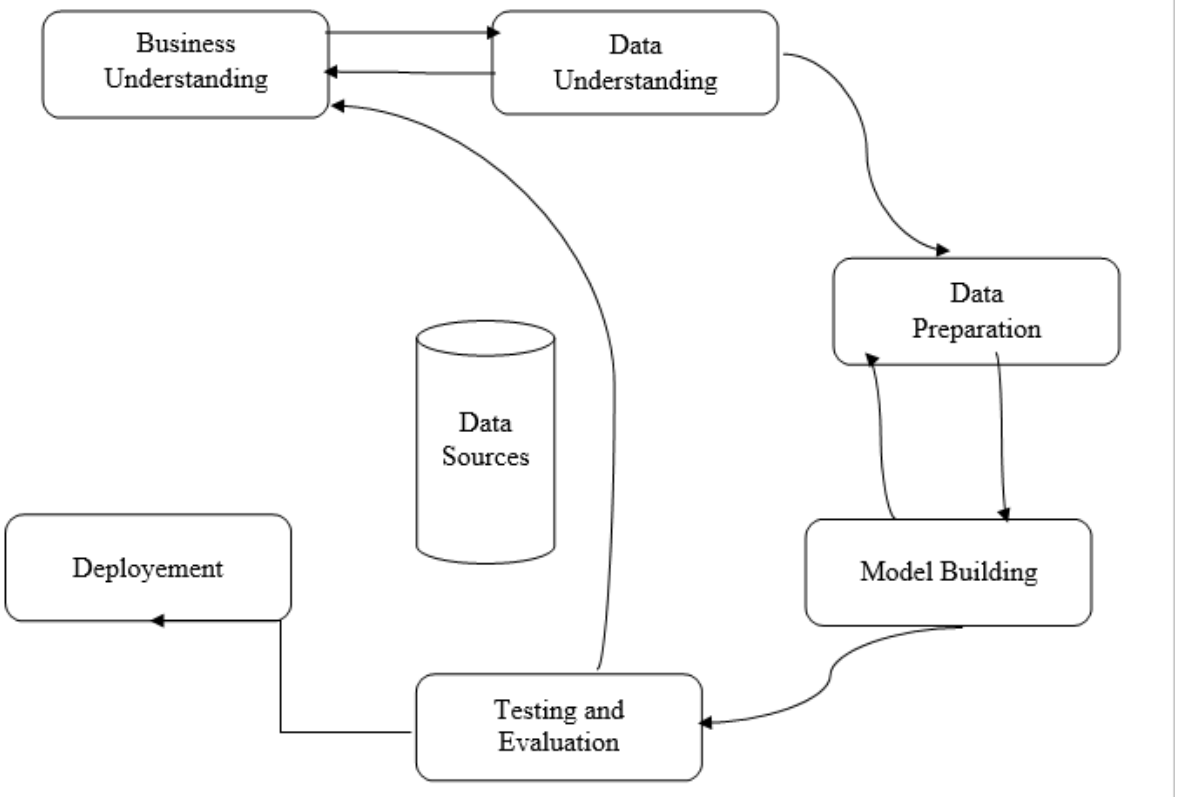


Figure 12 CRISP-DM process model

3.1 Business Understanding

In this phase, we establish the objectives from our data mining and developing a project plan. In our case, the project plan and our requirements will be detailed in chapter 3.

3.2 Data understanding

In this step, we must establish our data resources and assess its quality.

In this work, our input data will come from different sources depending on its nature:

- Our artist and track database are self-stored but comes originally from two types of databases, one mainstream and the other one indie from the Soundcloud music platform.

- Our mining resources will be from the Wikipedia database
- The lyrics needed for determining the language of our tracks comes from the Lyric Find database.

3.3 Data preparation

3.3.1 Data representation

In text classification, which is the type of classification we are doing, documents are categorized by the words that appear in them. One way to apply machine learning to document classification is to treat the presence or absence of each word as a Boolean attribute. However, a document can also be seen as a bag of words, so we can take account of the occurrence of each word. Words frequencies can be accommodates by applying a modified form of Naïve Bayes called multinomial Naïve Bayes.

One other way to represent data is by segmentation the words into grams, or n consecutive letters (for some small value such as n=3).

We can also represent our data into TF-IDF TF.IDF (Term Frequency times Inverse Document Frequency). It is normally computed as follows. Suppose we have a collection of N documents. Define f_{ij} to be the frequency (number of occurrences) of term (word) i in document j. Then, define the term frequency TF_{ij} to be: $\mathrm{TF}_{ij} = \frac{f_{ij}}{max_k f_{kj}}$.

That is, the term frequency of term i in document j is f_{ij} normalized by dividing it by the maximum number of occurrences of any term (perhaps excluding stop words) in the same document. Thus, the most frequent term in document j gets a TF of 1, and other terms get fractions as their term frequency for this document. The IDF for a term is defined as follows. Suppose term i appears in ni of the N documents in the collection. Then, $IDF_i = log_2 ({}^{N}/_{n_i})$ .The TF.IDF score for term i in document j is then defined to be $\mathrm{TF}_{ij} \times IDF_i$. The terms with the highest TF.IDF score are often the terms that best characterize the topic of the document [3].

3.3.2 Data transformation

There are of course an immense number of different words in a document, and most of them are not very useful for document classification. Words like stop-words can be eliminated in most cases; otherwise, it can mislead your classifier. In fact, your classifier is much more performing when the

chosen parameter value during the tuning process is reliable. This is why this pre-treatment step is very important. There are many ways to treat the input data to make it more amenable for learning methods, including: attribute selection, attribute discretization, data projection, sampling and data cleansing…

In our work, we needed the attribute selection technique, which eliminates all but the most relevant attributes. However, automatic methods exist, like reducing the dimensionality of the data by deleting unsuitable attributes.

3.4 Model building

This step is where we process our input data to mine information, as it is a document classification problem and more specifically a text mining one, we will detail two methods that we will work by, the naïve Bayes classifier and the one class classification.

3.4.1 Supervised document classification

An important domain of machine learning is document classification, which is the science of assigning a document to a category, or a class determined by the document certain attribute like in our case, for example, the gender of the artist that the text is about. The documents to be classified may be texts, images, music etc. Each kind of document possesses its special classification problems. When not otherwise specified, text classification is implied.

The performance of four document classification methods has been measured: the naïve Bayes classifier, the nearest classifier, decision trees and a subspace method [4]. The naïve Bayes classifier and the subspace method outperform the others.

3.4.2 Naïve Bayes classifier

Naïve Bayes gives a simple approach, with clear semantics to representing, using and learning probabilistic knowledge. It can achieve impressive results. [1]

Naïve Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. A naïve Bayes classifier considers each of these features to contribute independently to the probability that this

fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

For some types of probability models, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood; in other words, one can work with the naïve Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naïve Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naïve Bayes classifiers [5]. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests [6].

An advantage of naïve Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

3.4.3 One class classifier

Classification is usually about classifying items between two or more classes but in some classification problems, only a single class of instances in available during training, but in prediction time some instances with unknown label may belong to this target class or another class that is unknown. [1]

This kind of classification problems is called one class classification. In many cases the one class issues can be resolved by reformulating them into two class ones because we have access to data that do not belong to our target class and we can build our training set from them, but it is not always as obvious to get hold on such data.

One class classification is also called outlier or novelty detection because it is usually used to detect abnormal data from normal one.

3.5 Evaluation

Model results should be evaluated in the context of the business objectives established in the first phase (business understanding). This will lead to the identification of other needs (often through

pattern recognition), frequently reverting to prior phases of CRISP-DM. Gaining business understanding is an iterative procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations.

However, it is usually difficult to provide general and meaningful evaluations because the mining task is highly sensitive to the particular text under consideration.

Some evaluation methods, the ones that we will be using for comparing our results, will be detailed in chapter 4.

3.6 Deployement

This phase is about choosing the results issued from the best model and put into use. In our case, it will be added to the database to be accessible from the Prelude Soundytics API.

Conclusion

This chapter was all about notions, definitions and settlements. We explained the data mining and machine learning notions in general then we chose a data mining process adapted to our need to detail and introduce the techniques and methods to be used.

Chapter 3 Requirements specification and design

Introduction

In this chapter, we will describe the process that we adopted to work with our project. One may think that we already mentioned the CRISP-DM process, but this one is only for the data mining part, so it will be included in the project plan process, with the API solution.

The process we will talk about is an R&D (research and development) process proper to our mission and it is called the TRIZ process. Next to that, we will detail our work design using UML language (the **U**nified **M**odeling **L**anguage) diagrams.

1 TRIZ process

TRIZ stands for a Russian word “teoriya resheniya izobretatelskikh zadach”, which means literally: "theory of the resolution of invention-related tasks". It is "a problem-solving, analysis and forecasting tool derived from the study of patterns of invention in the global patent literature".

Maybe what we are doing it is not really an invention but its purpose is to develop our own data-mining module to add value to the company [7]. If the self-developed module is not reliable enough, we will work with API services to implement our databases. It is actually, what this process is all about, as it demands a generic solution in our case the API solution and an inventive one in our case the data mining system development. The Figure 13 below explains it further more.

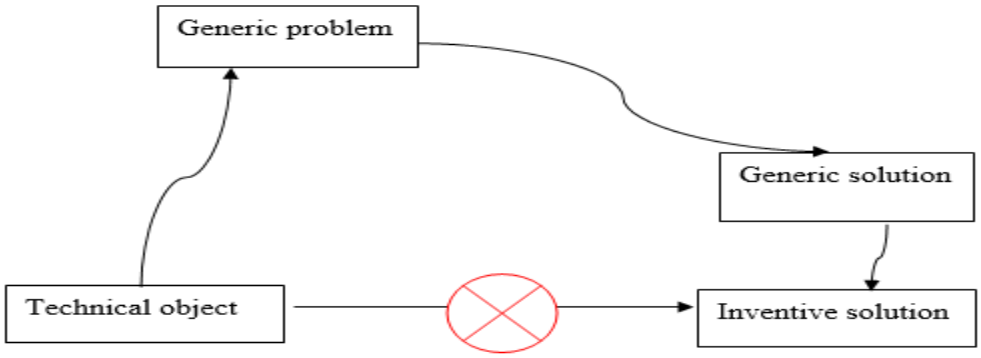


Figure 13 Prism of TRIZ

2 Requirements specification

2.1 Functional requirements

Our requirements are:

- Determining the gender of an artist (male or female) and its type (solo artist, group or duo)
- The artist biography, date of birth, and other information.
- In order to do that we need to be sure that the profile of the name we look for is an actual artist, and more specifically a music artist. We call this step the disambiguation step.
- Specifying the language of a track if it is a lyrical song and not mare instrumental.

2.2 Nonfunctional requirements

In this project, the nonfunctional requirements are very important because it will directly affect the performance of the outer system, and this system is for users to work with.

- Efficiency
    - ✓ Short response time
    - ✓ High throughput (rate of processing work)
    - ✓ Low utilization of computing memory and processors.
    - ✓ High availability
    - ✓ Short data transmission time
    - ✓ Scalability
- Reliability
    - ✓ Resiliency

- ✓ Structural solidity
- ✓ Low potential to failure

- Security
  - ✓ Low potential security breaches
  - ✓ Low vulnerability
- Maintainability
  - ✓ Adaptability
  - ✓ Responsive to business driven changes (portability/transferability)

3 Design

As we detailed in the first chapter, the use-case diagram and the database structure diagram, which is very similar to our class diagram, we will concentrate on the diagrams that describe the most our work. First, we will summarize the classification process into an activity diagram then we will detail each process in a sequence diagram, what we will do for the API process too.

3.1 Classification activity diagram

This diagram in Figure 14 is all about the steps we followed in each of our classification works. It balances between the internal system that we are developing and the Weka module as we see in Figure 14. Each of those activities will be detailed further in the following sequence diagrams.

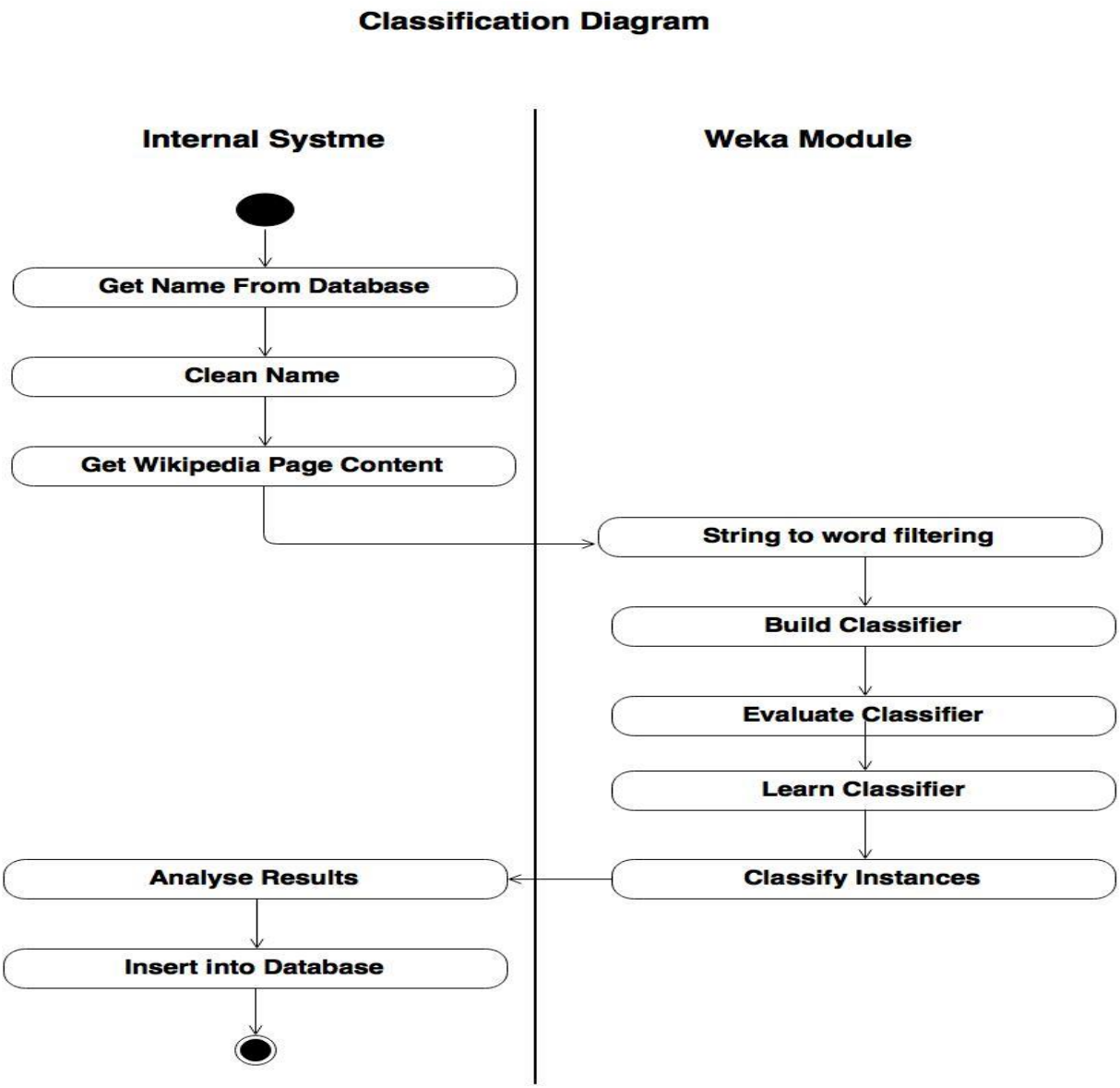


Figure 14 Classification activity diagram

3.2 Initialization sequence diagrams

The initialization step is what differentiate our classification processes. The disambiguation and the gender classification have some sequence in common like the fact that they create a StringToWordVector filter, which is going to fragment the strings into alphabetical words (setTokenizer), it will also transform the results into TFIDF representation. We add a normalization to the document length.

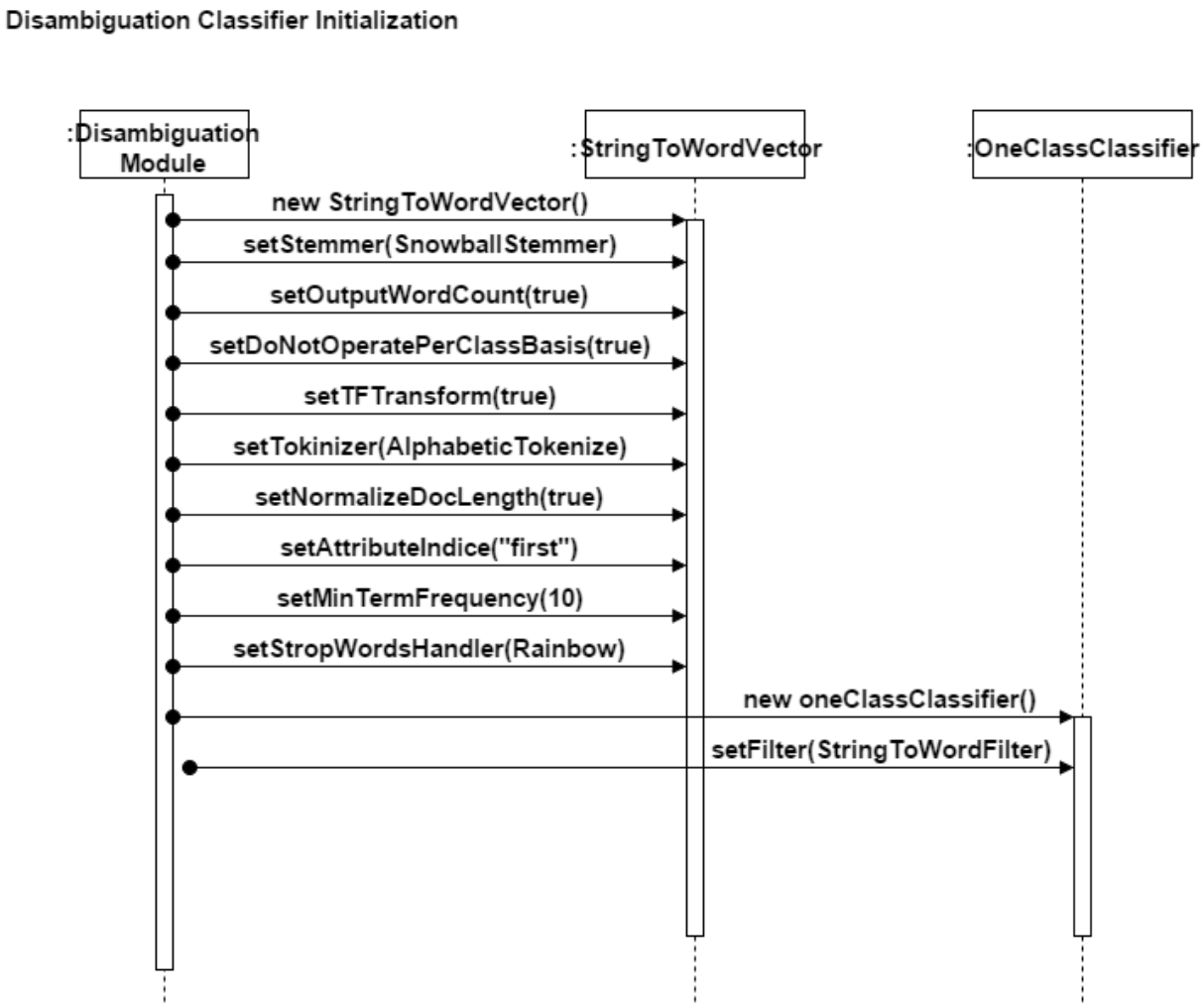


Figure 15 Disambiguation classifier initialization sequence diagram

About the Disambiguation classifier (see Figure 15), we will set the minimum term frequency to 10, which means we will delete every term with the frequency of 10. Besides, we will add a stop word handler, to eliminate the stop words.

For the disambiguation, we will use a one-class classifier, as it is a one-class classification problem: the text content is about an artist or it is not.

Figure 16 Gender classifier initialization sequence diagram

However, in the gender classifier (see Figure 16) we will not use a stop word handler because there are important to differentiate between genders like the "he" and the "she". We will add an attribute selection to improve the outcome. The classifier we will be using is the naïve Bayes classifier, as it is one of the best classifiers for text mining.

3.3 Wikimedia importer sequence diagram

The Wikimedia importer function (see Figure 17) is a very important function in the data mining solution as it returns the raw texts from the Wikipedia pages to operate on, an extract from them.

We start by cleaning the name from all the unwanted characters and make it lowercase. We send our artist name to be searched by the API, and it returns all the titles that corresponds for disambiguation.

We compare each title with the name according to the Levenshtein distance, which is the minimum number of single character insertion, deletion or substitution [8]. If it is at 95% resembling we ask to get the content page and parse it from HTML to text.

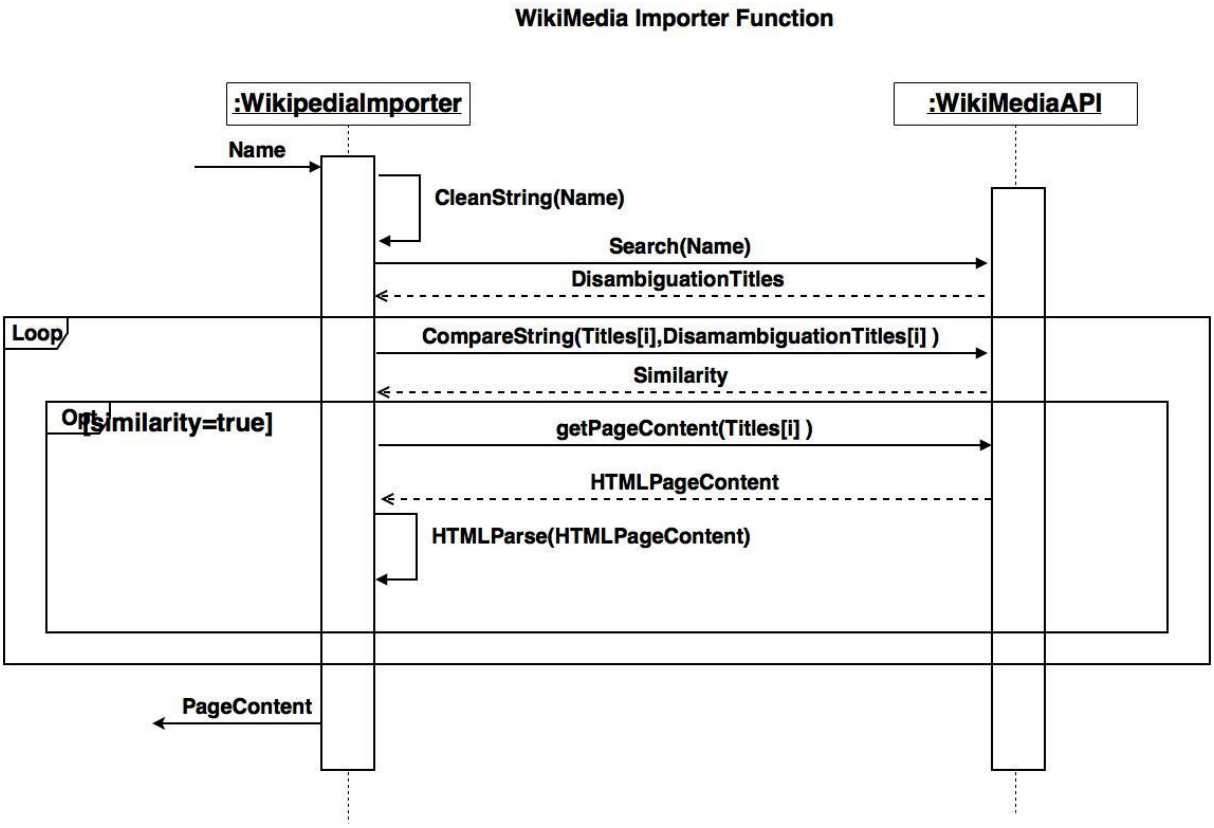


Figure 17 Wikimedia importer sequence diagram

3.4 Add instances sequence diagram

The add instances function (see Figure 19) is a simple method to add the lines to our model file which looks like this model in Figure 18 below (it is not our working model, because ours is quite complicated to capture)

```
@relation SMS_Spam_Collection

@attribute class_target {spam,ham}
@attribute text String

@data
ham,'Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...'
ham,'Ok lar... Joking wif u oni...'
spam,'Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C\'s apply 08452810075over18\'
ham,'U dun say so early hor... U c already then say...'
ham,'Nah I don\'t think he goes to usf, he lives around here though'
spam,'FreeMsg Hey there darling it\'s been 3 week\'s now and no word back! I\'d like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv'
ham,'Even my brother is not like to speak with me. They treat me like aids patent.'
ham,'As per your request \'Melle Melle (Oru Minnaminunginte Nurungu Vettam)\' has been set as your callertune for all Callers. Press *9 to copy your friends Calle
spam,'WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only
spam,'Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030'
ham,'I\'m gonna be home soon and i don\'t want to talk about this stuff anymore tonight, k? I\'ve cried enough today.'
spam,'SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info'
spam,'URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18'
ham,'I\'ve been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have bee
ham,'I HAVE A DATE ON SUNDAY WITH WILL!!'
spam,'XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL'
ham,'Oh k...i\'m watching here:)'
ham,'Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.'
ham,'Fine if that\'s the way u feel. That\'s the way its gota b'
spam,'England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ú1.20 POBOXox36504W45WQ 16+'
ham,'Is that seriously how you spell his name?'
ham,'I\'m going to try for 2 months ha ha only joking'
ham,'So ü pay first lar... Then when is da stock comin...'
ham,'Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?'
```

Figure 18 Model example

There is the class attribute “ham” or “spam” then the value attribute “the message”. All the line is called an instance.

So in the add instances function we create an Instance (Weka class) with its content, and if it is for the learning section we add the class value too. And then we insert it into the Instances class (Weka class).

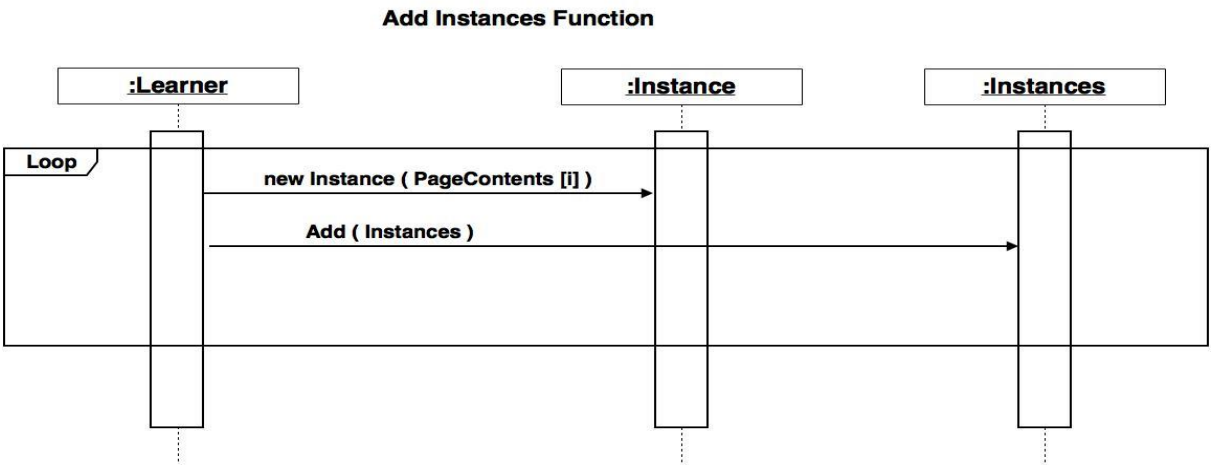


Figure 19 Add instances sequence diagram

3.5 Learning sequence diagram

In the learning diagram (see Figure 20), we call the disambiguation or the gender classifier initialization function, then we procure our data from the Wikimedia importer function and then we add the content as instances.

Then we build our classifier, that instance of classifier is the one we created in the initialization part (OneClassClassifier or Naïve BayesClassifier), we evaluate it and then we operate the learning to get hold on the model will be using for classification.

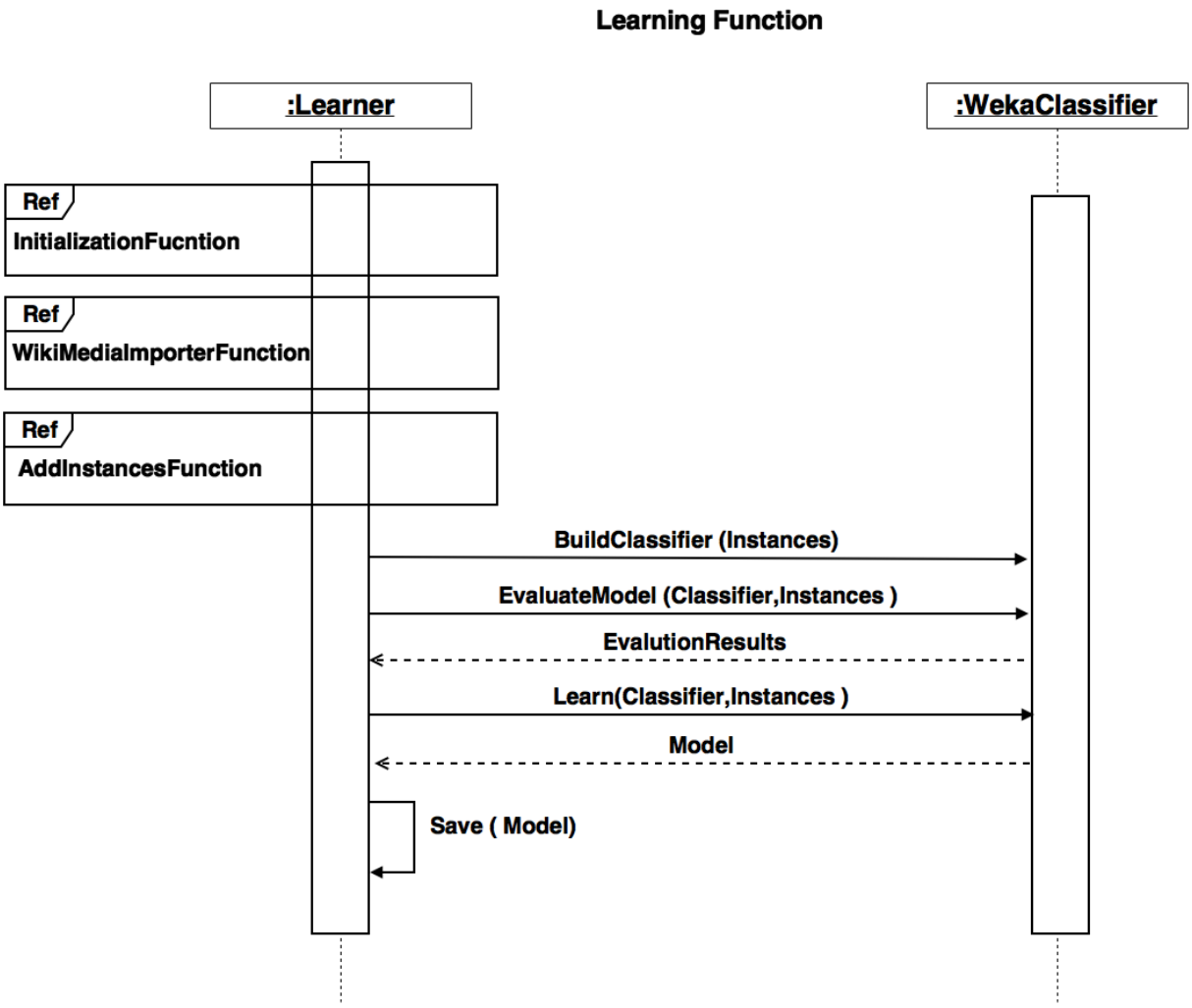


Figure 20 Learning sequence diagram

3.6 Classification sequence diagram

The first part of our classification function (see Figure 21) is very similar to our learning function. Nevertheless, instead of building, evaluating and learning, we are going to load our module, saved before, and classify each instance with the according module. Each classification loop returns a prediction number, if the prediction we can guess or decide for the membership of our instance.

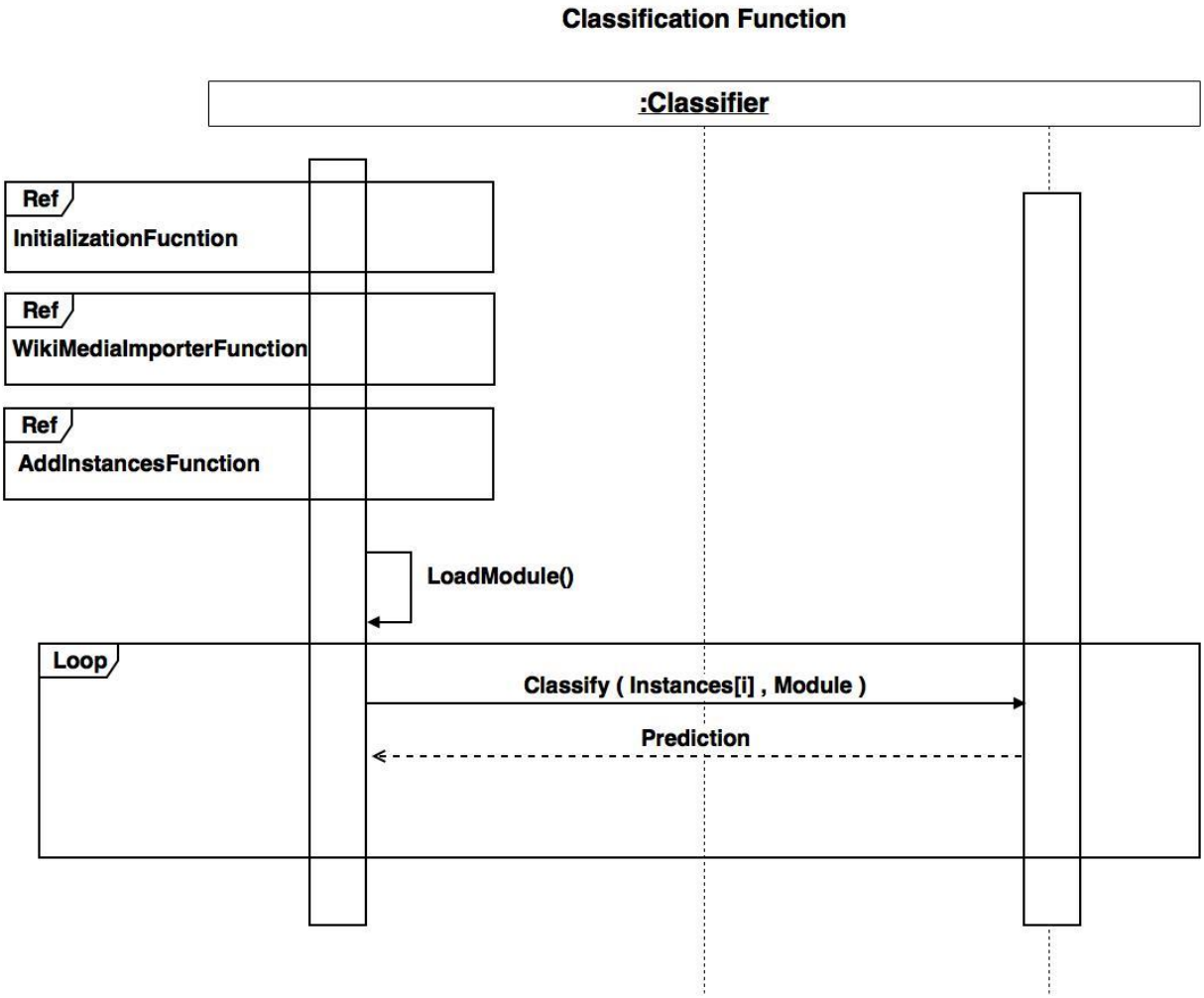


Figure 21 Classification sequence diagram

3.7 Wikidata API importer sequence diagram

This sequence diagram (see Figure 22) is related to the API solution part. It imports the structured data from the wikidata database.

We start by searching our already cleaned artist name into the wikidata API, which returns a list of item ids and “instance of” ids, the instance of ids are related to the type of the subject of the name

(human, band, animal, object….) and an occupation id if he is human. So we check if it is a human and it is in our manually collected list of occupation ids, if it is, then he is an artist and we can collect its data. If not, we check if it is a band or a duo, if not, sometimes it can say "rock band" or "soul band" for example and their id is different from our band id so we ask for the upper-class of our 'instance of' and we compare it to our band id. If our item is a band or an artist, we insert its information into the database.

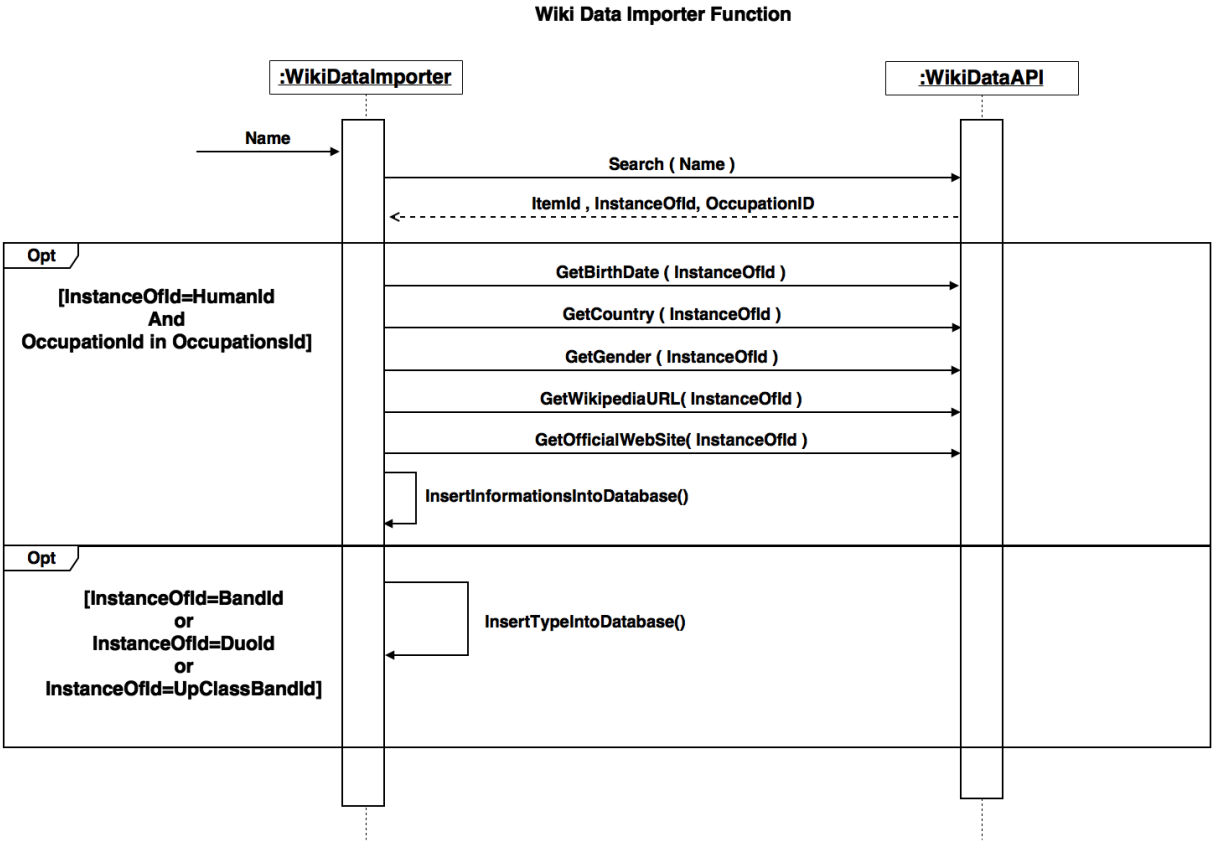


Figure 22 Wikidata importer sequence diagram

3.8 Lyric language detection sequence diagram

The lyric language detection function (see Figure 23) is a very simple function, as it only searches the right lyrics with LyricsFind API and checks for the language with the help of an external system, the Cybozu Lang Detection.

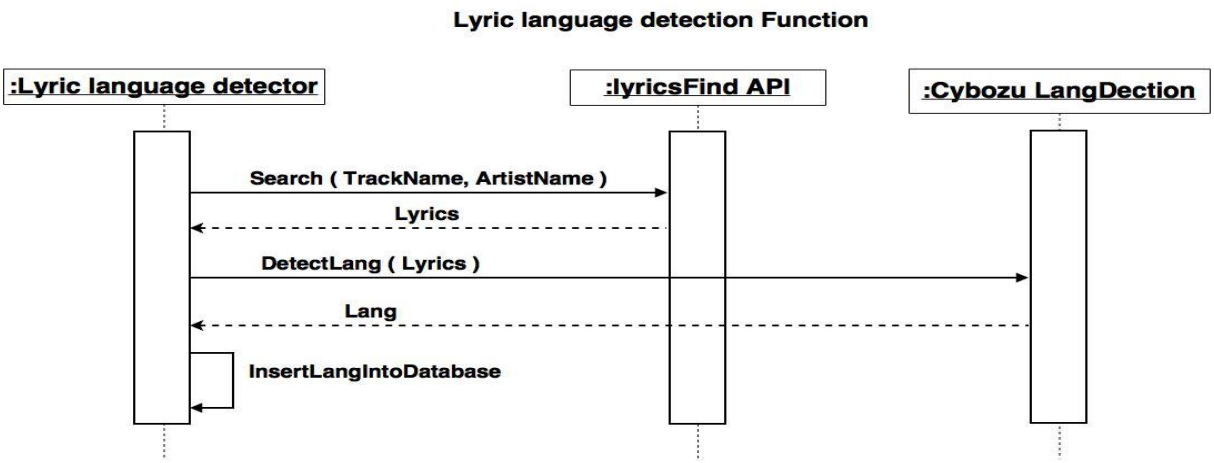


Figure 23 Lyric language detection sequence diagram

Conclusion

In this chapter, we tried as hard as we can to explain our work in detail and yet stay clear, by clarifying our activity and sequence diagrams. Afterward, we defined the work process of comparison we relied on, the TRIZ process.

Chapter 4 Evaluation and achievements

Introduction

After detailing every part of our work process and our business plan, now comes the time to assess our work by describing our achievements. Therefore, this chapter is going to be about our work environment, the hardware, software and the API we manipulated to get this work done. Then we will check the results of our work by comparing between the API solution results and the Data mining results.

1 Work environment

Our work environment is divided between hardware, software and API environments.

1.1 Hardware

The system architecture is made to be very fast and secure. Amazon (Amazon EC2 offer) in Europe zone and more specifically the «Ireland» and «Frankfurt» areas hosts the servers.

Storage of tracks is done via Amazon S3 in the "Ireland" region.

Every machines run on Ubuntu 14.04.2 LTS, the system is automatically updated with the use of the "unattended-upgrades" package. The following diagram in Figure 24 is quite self-explicatory about how requests travel between the different components.

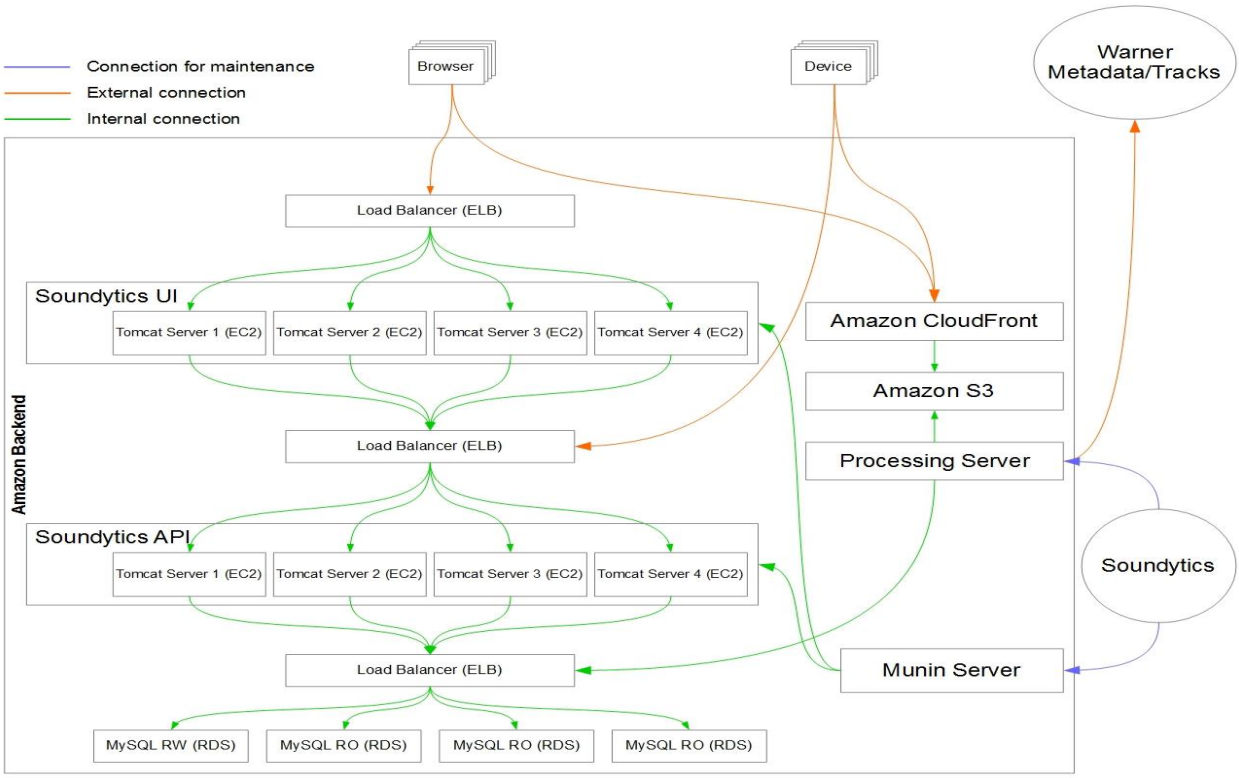


Figure 24 System architecture diagram

1.2 Software

Many tools were used during the accomplishment of this mission:

1.2.1 Eclipse Luna release

Eclipse is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Written mostly in Java, Eclipse can be used to develop applications. Released under the terms of the Eclipse Public License, Eclipse SDK is free and open source software (although it is incompatible with the GNU General Public License). It was one of the first IDEs to run under GNU Classpath.

The Luna release integrated Java 8 support because in the previous version this was possible via a "Java 8 patch" plugin.

1.2.2 Org.JSON

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. We used la lib org.json for java on eclipse.

1.2.3 JSOUP

JSOUP is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

1.2.4 Cybozu language detection

The language-detection library is a Java opensource library to detect languages in which texts are written. It presents 99% over precision for more than 40 languages including Arabic, French, Japanese, English…It detects language of a text using naive Bayesian filter.

1.2.5 Weka

Weka was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning.

The system is written in Java and distributed under the terms of the GNU General Public License. It also presents a plugin eclipse for java development.

1.3 APIs

We also relied a lot on external APIs like:

1.3.1 MediaWiki

The MediaWiki web API is a web service that provides convenient access to wiki features, data, and meta-data over HTTP.

The MediaWiki web API can be used to monitor a MediaWiki installation, or create a bot to automatically maintain one. It provides direct, high-level access to the data contained in MediaWiki databases. Client programs can log in to a wiki, get data, and post changes automatically by making HTTP requests to the web service.

This API was used to provide texts as input to our data mining solution.

1.3.2 Wikidata

Wikidata is a free linked database with 14,708,840 data items that can be read and edited by both humans and machines.

Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects. The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

As this API is more structured and oriented, we use it in our API solution, to furnish our database directly.

1.3.3 LyricFind

LyricFind provides a legal music lyrics service. LyricFind has licensing from over 2,000 music publishers, including the four majors EMI Music Publishing, Universal Music Publishing Group, Warner/Chappell Music Publishing, and Sony/ATV Music Publishing as well as a database of those lyrics available for licensing.

2 Data mining solution evaluation

For the data mining classifier evaluation, we used different evaluating methods as the holdout method and the cross validation one.

2.1 The holdout evaluation method

The holdout method reserves a certain amount for testing and uses the remainder for training. In practical terms, it is common to hold out one-third of the data for testing and use the remaining two-thirds for training. However, the sample used for training or testing might not be representative.

2.1.1 Disambiguation classifier evaluation

The results (see Figure 25) of our one class classifier were very disappointing, especially with the holdout method as it classified zero instance correctly on 98.

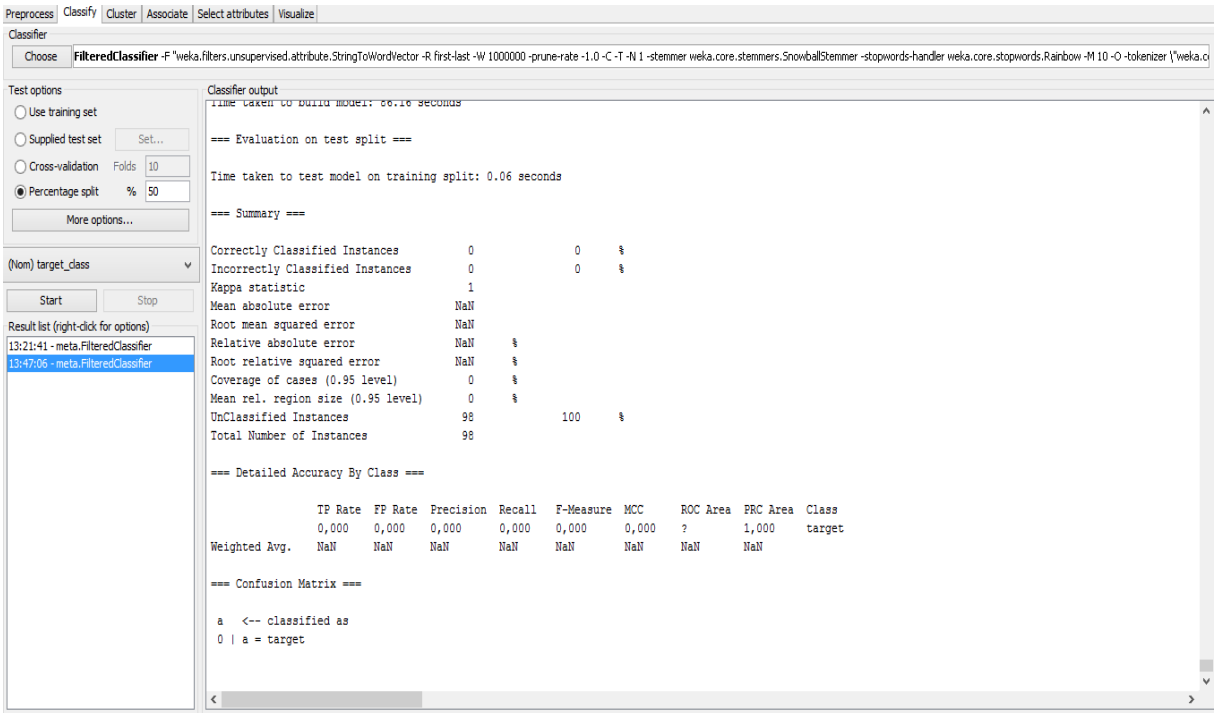


Figure 25 Disambiguation classifier holdout evaluation

2.1.2 Gender classifier evaluation

The Gender classifier, based on a naïve Bayes classifier (see Figure 26), on the other side was clearly working as it classified 93 instances of 98 correctly (94% of exactitude).

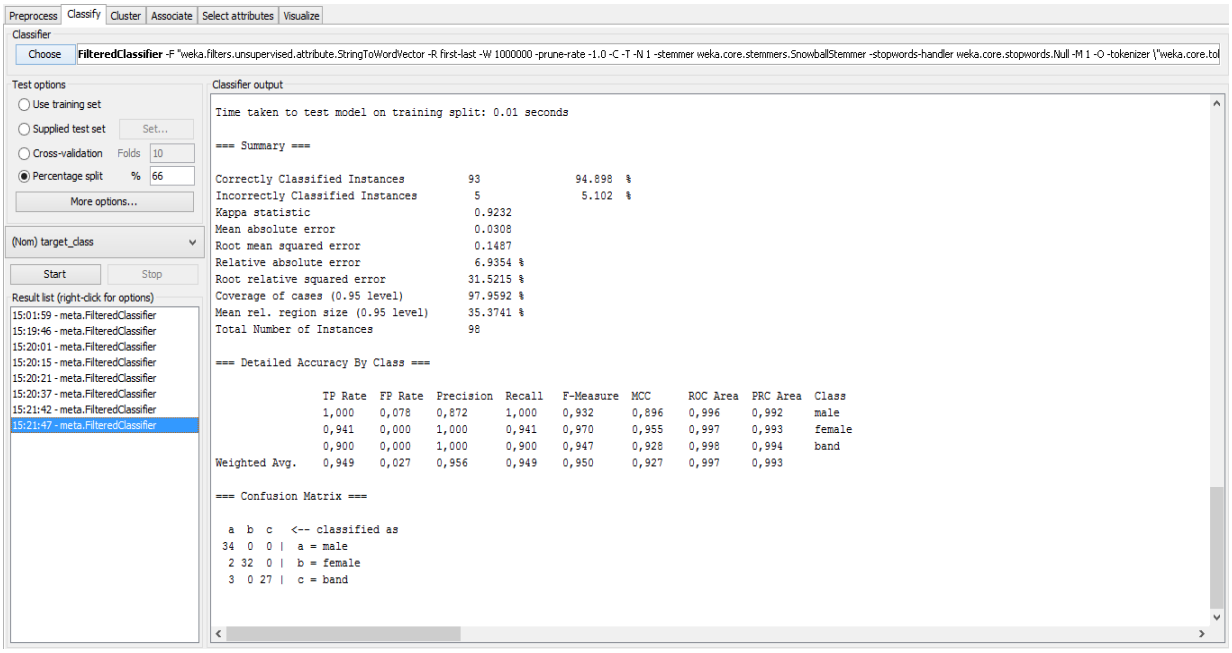


Figure 26 Gender classifier holdout evaluation

2.2 Cross-validation evaluation method

To bear on the issue of the holdout method, we use the cross-validation technique, and in our case more particularly the k-folds cross-validation: In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, which we will use in our case [9].

2.2.1 Disambiguation classifier evaluation

With the cross-validation technique, it improved the evaluation results (see Figure 27) of our one class classifier to 28 of 289 instances (9, 6889 % of exactitude).

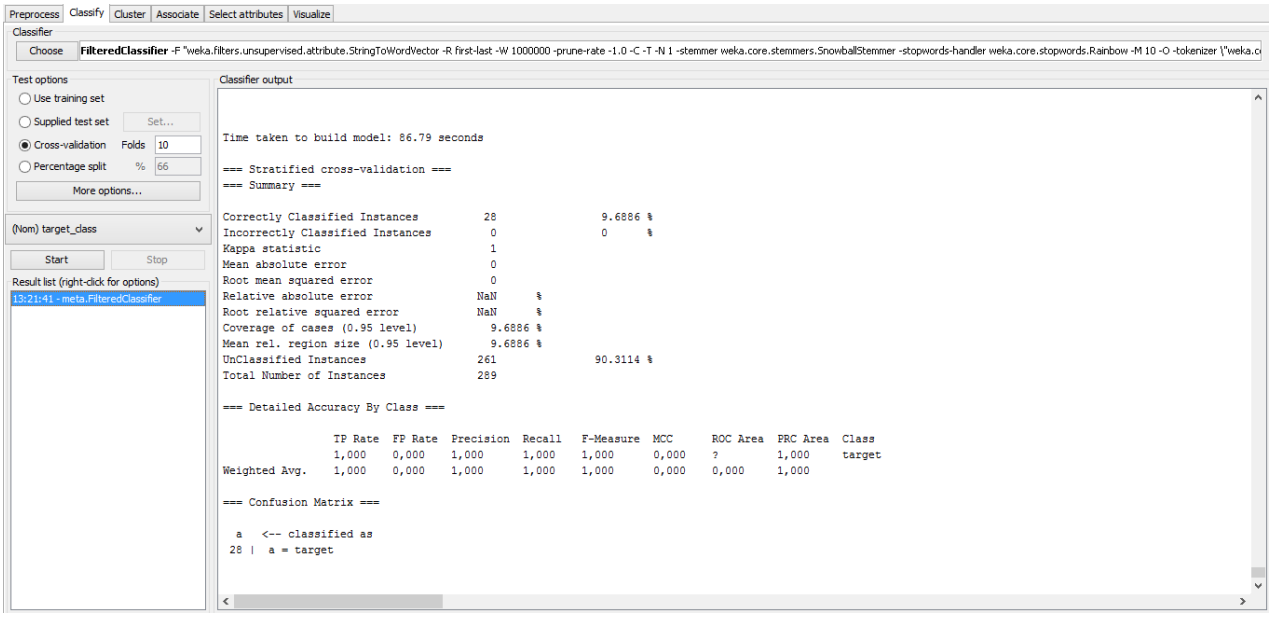


Figure 27 Disambiguation classifier cross-validation evaluation

2.2.2 Gender classifier evaluation

The results of the Naïve Bayes Classifier with the cross validation technique becomes 278 correctly classified instances of 289 (see Figure 28).

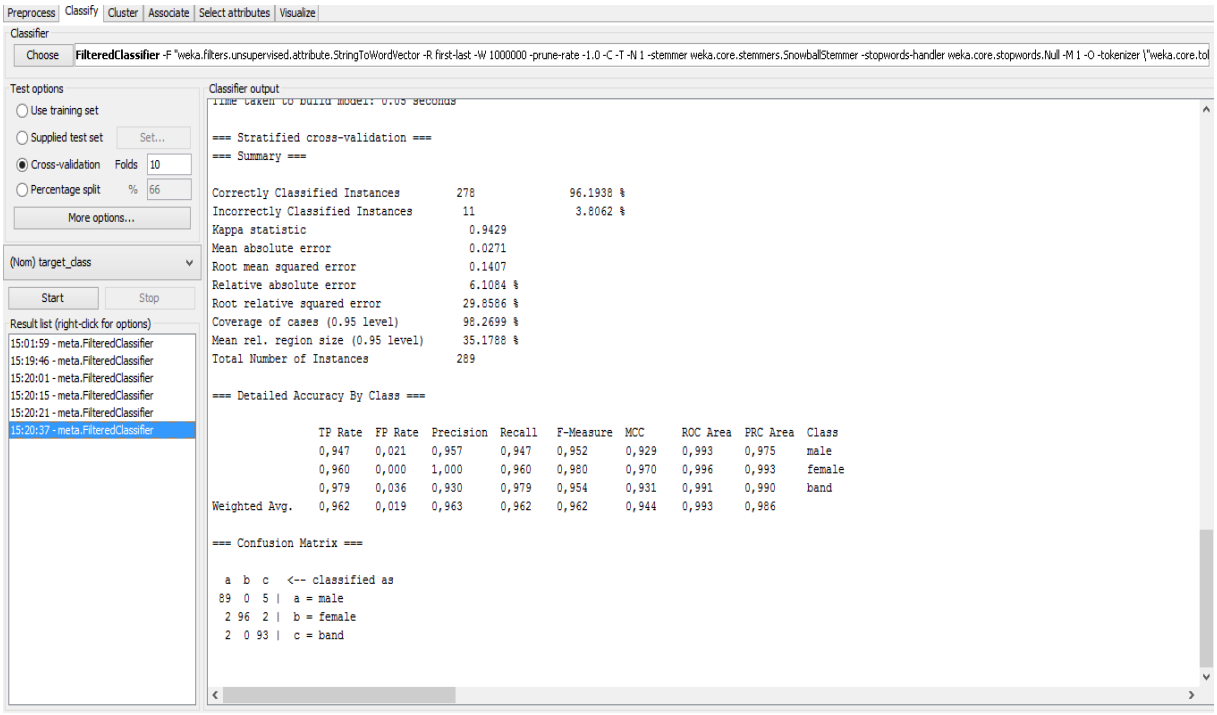


Figure 28 Gender classifier cross-validation evaluation

2.3 Evaluation analysis

As we said earlier that the one class classification part related to the disambiguation process was not successful, and even if the gender process was, the disambiguation step is a vital one, because we need to know if the article is really about an artist or not before we process in other classifiers. That is why we did not explore further this solution.

3 API solution evaluation

After the disappointing results of the data mining process we decided to deploy our API solution on the system, so the results that follow are on a database level.

3.1 Wikidata import results

As we see on the Figure 29, with the wikidata API importer, recovered 9 345 information about artists that it found on wikidata, over 25 528 artists available on the database.

```
1  SELECT count(*) FROM soundytics_sandbox.artist WHERE `provider`='VLGroup';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

count(*)
25528

```
1  SELECT count(*) FROM soundytics_sandbox.artist WHERE `provider`='VLGroup' and type is not null;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

count(*)
9345

Figure 29 Artists count request

The results were inserted into the database as shown in the Figures 31 to 33 below. Since we inserted the birth date, the country, the gender, the type and the real name of an artist if available in the artist table. The biography, the websites and the artist aliases table were also loaded.

```
1  SELECT * FROM soundytics_sandbox.artist WHERE `provider`='VLGroup' and type is not null;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

	name	provider	real_name	firstname	lastname	country	death_year	birth_year	birth_month	birth_day	start_year	end_year	gender	type	ipi
9:33:36	Jami Sieber	VLGroup	NULL	NULL	NULL	GB	NULL	1957	4	17	NULL	NULL	Female	Person	NULL
9:39:26	Martinho Da Vila	VLGroup	NULL	NULL	NULL	BR	NULL	1938	2	12	NULL	NULL	Male	Person	NULL
9:40:20	George Michael	VLGroup	NULL	NULL	NULL	GB	NULL	1963	6	25	NULL	NULL	Male	Person	NULL
9:40:22	Fausto Papetti	VLGroup	NULL	NULL	NULL	IT	NULL	1923	1	28	NULL	NULL	Male	Person	NULL
9:40:24	Pussy Galore	VLGroup	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Band	NULL
9:40:26	Spike Jones	VLGroup	NULL	NULL	NULL	US	NULL	1911	12	14	NULL	NULL	Male	Person	NULL
9:40:30	Amanda Marshall	VLGroup	NULL	NULL	NULL	CA	NULL	1972	8	29	NULL	NULL	Female	Person	NULL
9:40:34	Belinda Carlisle	VLGroup	NULL	NULL	NULL	US	NULL	1958	8	17	NULL	NULL	Female	Person	NULL
9:40:36	Sidney Bechet	VLGroup	NULL	NULL	NULL	US	NULL	1897	5	14	NULL	NULL	Male	Person	NULL
9:40:38	Clair Marlo	VLGroup	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Female	Person	NULL
9:40:40	Amanda Lear	VLGroup	NULL	NULL	NULL	FR	NULL	1946	11	18	NULL	NULL	Female	Person	NULL
9:40:42	Ronny Jordan	VLGroup	NULL	NULL	NULL	GB	NULL	1962	11	29	NULL	NULL	Male	Person	NULL
9:40:47	Jimmy Withersp...	VLGroup	NULL	NULL	NULL	US	NULL	1923	8	8	NULL	NULL	Male	Person	NULL
9:40:50	Louise Attaque	VLGroup	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	Band	NULL

Figure 30 Artist table

```
1 • SELECT * FROM soundytics_sandbox.artist_alias;
```

Result Grid | Filter Rows: | Edit: | Export/Import:

id	update	artist_id	alias	language
4205	2015-06-10 09:40:17	3702880	Georgios Kyriacos Panayiotou	NULL
4225	2015-06-10 09:42:35	3702926	H.I.M.	NULL
4227	2015-06-10 09:42:35	3702926	HER	NULL
4229	2015-06-10 09:42:35	3702926	His Infernal Majesty	NULL
4285	2015-06-10 09:51:04	3703000	Catharina Hagen	NULL
4286	2015-06-10 09:51:06	3703001	Antonia Christina Basilotta	NULL
4287	2015-06-10 09:51:25	3703013	Quincy Delight Jones, Jr.	NULL
4288	2015-06-10 09:51:30	3703016	+eRa+	NULL
4289	2015-06-10 09:51:49	3703030	Red Headed Stranger	NULL
4290	2015-06-10 09:51:49	3703030	Shotgun Willie	NULL
4291	2015-06-10 09:51:49	3703030	Willie Hugh Nelson	NULL
4292	2015-06-10 09:52:09	3703035	William Martin Joel	NULL

Figure 31 Artist alias table

```
1 • SELECT * FROM soundytics_sandbox.artist_biography;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

id	update	artist_id	source	language	biography
2019535	2015-06-10 09:33:36	3702878	Wikidata	en	Jami Sieber is an American cellist, vocalist and composer. She has received several positive reviews for her work. She plays acou...
2019537	2015-06-10 09:39:25	3702879	Wikidata	en	Martinho da Vila (born Martinho José Ferreira on February 12, 1938 in Duas Barras, Rio de Janeiro) is a Brazilian samba musician ...
2019558	2015-06-10 09:40:18	3702880	Wikidata	en	Georgios Kyriacos Panayiotou (born 25 June 1963), better known by his stage name George Michael, is an English singer, songw...
2019578	2015-06-10 09:40:21	3702881	Wikidata	en	Fausto Papetti (28 January 1923 – 15 June 1999) was an Italian alto saxophone player. His recordings, sometimes under the ps...
2019598	2015-06-10 09:40:24	3702884	Wikidata	en	Pussy Galore was an American garage rock band that formed in Washington, D.C. in 1985. They had a constantly fluid line-up un...
2019618	2015-06-10 09:40:26	3702886	Wikidata	en	Lindley Armstrong "Spike" Jones (December 14, 1911 – May 1, 1965) was an American musician and bandleader specializing in pe...
2019638	2015-06-10 09:40:30	3702888	Wikidata	en	Amanda Meta Marshall (born August 29, 1972) is a Canadian pop-rock singer. She has released three studio albums, the first wa...
2019658	2015-06-10 09:40:32	3702890	Wikidata	en	Belinda Jo Carlisle (born August 17, 1958) is an American singer who gained worldwide fame as the lead vocalist of The Go-Go's, ...
2019678	2015-06-10 09:40:36	3702891	Wikidata	en	Sidney Bechet (May 14, 1897 – May 14, 1959) was an American jazz saxophonist, clarinetist, and composer.
2019698	2015-06-10 09:40:38	3702893	Wikidata	en	Clara Veseliza Baker, known professionally as Clair Marlo, is a Croatian-American composer, performer, arranger and record prod...
2019718	2015-06-10 09:40:40	3702894	Wikidata	en	Amanda Lear (née Tapp; born 18 November in 1939,) is a French singer, lyricist, painter, television presenter, actress and forme...
2019738	2015-06-10 09:40:42	3702895	Wikidata	en	Ronald Laurence Albert Simpson, known as Ronny Jordan (29 November 1962 – 13 January 2014) was a British guitarist at the f...

Figure 32 Artist biography table

```
1 • SELECT * FROM soundytics_sandbox.artist_website;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: | Fetch rows:

id	update	artist_id	source	hostname	title	url
9494763	2015-06-10 09:40:20	3702880	Wikidata	georgemichael.com	GeorgeMichael.com	http://georgemichael.com
9494783	2015-06-10 09:40:34	3702890	Wikidata	belindacarlisle.tv	Belinda Carlisle	http://belindacarlisle.tv
9494803	2015-06-10 09:41:18	3702914	Wikidata	www.sistercarol.com	Sister Carol Web Site	http://www.sistercarol.com/
9494823	2015-06-10 09:41:25	3702917	Wikidata	www.aliceinchains.com	Alice in Chains	http://www.aliceinchains.com
9494843	2015-06-10 09:42:45	3702929	Wikidata	georgebenson.com	George Benson	http://georgebenson.com
9494863	2015-06-10 09:46:04	3702939	Wikidata	www.wendymatthews.com	Wendy Matthews Wendy Matthews	http://www.wendymatthews.com/
9494864	2015-06-10 09:50:14	3702966	Wikidata	www.bloodhoundgang.com	www.bloodhoundgang.com	http://www.bloodhoundgang.com/
9494865	2015-06-10 09:51:16	3703005	Wikidata	www.rasrecords.com	Account Suspended	http://www.rasrecords.com/blackuhuru/
9494866	2015-06-10 09:51:34	3703019	Wikidata	www.crashtestdummies.com	CrashTestDummies.com: the official site of the ...	http://www.crashtestdummies.com
9494867	2015-06-10 09:51:39	3703021	Wikidata	www.robyn.com	Robyn	http://www.robyn.com
9494868	2015-06-10 09:51:52	3703030	Wikidata	www.willienelson.com	Home - Willie Nelson	http://www.willienelson.com/
9494869	2015-06-10 09:52:12	3703035	Wikidata	www.billyjoel.com	Billy Joel	http://www.billyjoel.com/
9494870	2015-06-10 09:52:53	3703056	Wikidata	lisaekdahl.com	Lisa Ekdahl - Official website	http://lisaekdahl.com

Figure 33 Artist website table

3.2 Lyric language detection results

The lyrics language detection API importer was also fine, as on 197 061 tracks we could detect the language of 48 536 of them as shown in the Figure 34. The Figure 35 shows the insertion of the results into the database.

```
1 • SELECT count(*) FROM soundytics_sandbox.track WHERE `provider`='VLGroup';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

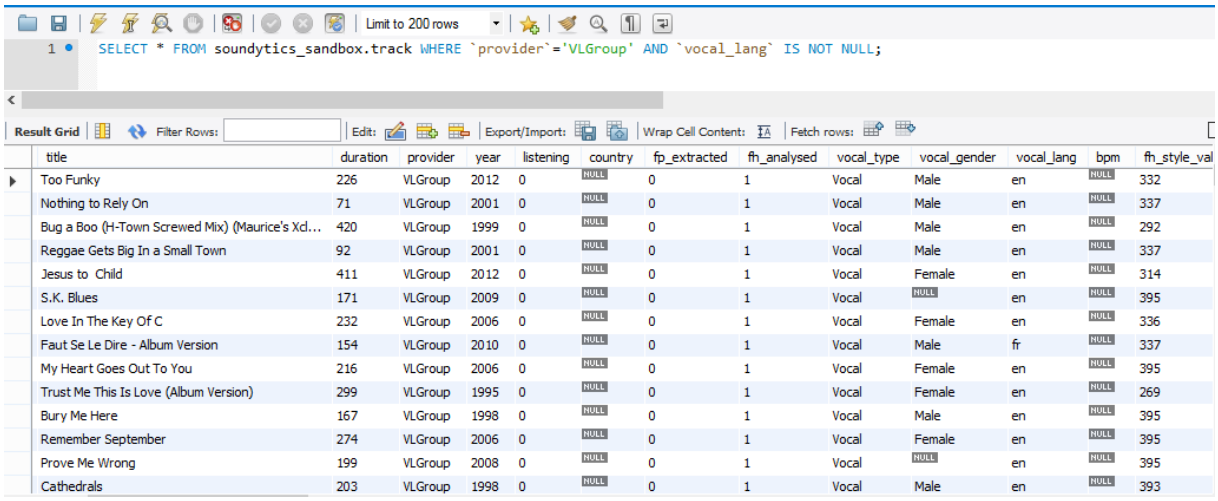
count(*)
197061

```
1 • SELECT count(*) FROM soundytics_sandbox.track WHERE `provider`='VLGroup' AND `vocal_lang` IS NOT NULL;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

count(*)
48536

Figure 34 Track count request



```
SELECT * FROM soundytics_sandbox.track WHERE `provider`='VLGroup' AND `vocal_lang` IS NOT NULL;
```

Result Grid

title	duration	provider	year	listening	country	fp_extracted	fh_analysed	vocal_type	vocal_gender	vocal_lang	bpm	fh_style_val
Too Funky	226	VLGroup	2012	0	NULL	0	1	Vocal	Male	en	NULL	332
Nothing to Rely On	71	VLGroup	2001	0	NULL	0	1	Vocal	Male	en	NULL	337
Bug a Boo (H-Town Screwed Mix) (Maurice's Xcl...	420	VLGroup	1999	0	NULL	0	1	Vocal	Male	en	NULL	292
Reggae Gets Big In a Small Town	92	VLGroup	2001	0	NULL	0	1	Vocal	Male	en	NULL	337
Jesus to Child	411	VLGroup	2012	0	NULL	0	1	Vocal	Female	en	NULL	314
S.K. Blues	171	VLGroup	2009	0	NULL	0	1	Vocal	NULL	en	NULL	395
Love In The Key Of C	232	VLGroup	2006	0	NULL	0	1	Vocal	Female	en	NULL	336
Faut Se Le Dire - Album Version	154	VLGroup	2010	0	NULL	0	1	Vocal	Male	fr	NULL	337
My Heart Goes Out To You	216	VLGroup	2006	0	NULL	0	1	Vocal	Female	en	NULL	395
Trust Me This Is Love (Album Version)	299	VLGroup	1995	0	NULL	0	1	Vocal	Female	en	NULL	269
Bury Me Here	167	VLGroup	1998	0	NULL	0	1	Vocal	Male	en	NULL	395
Remember September	274	VLGroup	2006	0	NULL	0	1	Vocal	Female	en	NULL	395
Prove Me Wrong	199	VLGroup	2008	0	NULL	0	1	Vocal	NULL	en	NULL	395
Cathedrals	203	VLGroup	1998	0	NULL	0	1	Vocal	Male	en	NULL	393

Figure 35 Track table

Conclusion

This chapter was about the evaluation of our solution and the results found, after an overview of our work environment. As we have seen, the data mining solution was unsatisfactory compared to the API solution, even if this one is not treating all the artists and tracks.

General conclusion and perspectives

This internship was very interesting and productive as it helped sharpen my R&D skills especially in the data mining area, and consuming APIs.

I did not just learn about implementing code, but also about how to evaluate my results and especially how to tidy up my code and organize it to fit in an existent system, and yet be always accessible for maintenance and readjustments.

Since the data mining solution was not very satisfying, one of our perspectives is to readjust our inputs and options on the one class classifier, and make it ready to practice for further experiments.

References

[1] I.H.Witten, E.Frank and M.A.Hall, “Data Mining Practicle Machine Learning Tools and Techniques,” 2011.

[2] D.L. Olson and D. Delen, "Advanced Data Mining Technique," Berlin, 2008.

[3] J. Leskovec, A. Rajaraman and J. D. Ullman, "Mining of Massive Datasets," 2011.

[4] Y. H. Li and A. K. Jain, "Classification of text documents," in Computer Journal, 1998, pp. 537-46.

[5] H. Zhang, "The Optimality of Naive Bayes," in American Association for Artificial Intelligence, 2004.

[6] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in 23rd International Conference on Machine Learning, Pittsburgh, 2006.

[7] Z. Hua, J. Yang, S. Coulibaly, and B. Zhang, "Integration TRIZ with problem-solving tools: a literature review from 1995 to 2006," in International Journal of Business Innovation and Research, 2006, pp. 111-128.

[8] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in Soviet Physics, Doklady, 1966.

[9] M. Geoffrey, D. Kim-Anh, and A. Christophe, "Analyzing microarray gene expression data," 2004, pp. X-Y.