Project Documentation

Alon Fixler 038183836 [alonfixler@gmail.com](mailto:alonfixler@gmail.com)

Yariv Zur 039627831 [yarivtzur@gmail.com](mailto:yarivtzur@gmail.com)

Specifications

1) Ant installed and configured.

2) JDK 7.0 and up.

3) MySQL Server.

Set-up instructions

1. ant compile
2. ant jar

Running instructions

1. Type in command line : java –jar run.jar full.txt pos.txt [username] [password]

Where username and password are your Mysql username and password (type them without the brackets).

1. The program will now start processing the input files and fill the db with them. This should take a few minutes.
2. When the program’s ready you’ll see a line that says: “Please insert your query”. You can insert one of the next 7 queries. Make sure you type them in **exactly** as they are written (including gaps, commas, lower/upper case letters). Make sure to insert names with a capital letter for the first letter of each part of the name. Insert numbers with no commas:
3. when were X born
4. which people who have profession X, were born in year Y
5. when did X die
6. who is artist X
7. what are the top-k populated countries
8. who are the artists that were born in the X decade

insert X as “1950”,”1960”,”1970”…

1. who are all the artists that are from country X
2. To exit the program, type in “exit”.

5 use cases

1**) Query**: which people who have profession politician, were born in 1948

**Results**:

Levy Mwanawasa Probability: 0.7584

Gerry Adams Probability: 0.7584

Pim Fortuyn Probability: 0.7584

Paul Murphy Probability: 0.7584

David Davis Probability: 0.7584

**Query execution time**: 25 mS

**Way of computing**: this query was computed by selecting people names and the year they born probability from the persons table. We selected only the people that have profession fields equals “politician” and “born in” field equals 1948. We chose to present the results in descending order.

2) **Query**: who is artist Ray Charles

**Results**:

Type: singer

Nationality: American

Genre: r&b

Probability: 0.45119999999999993

Type: singer

Nationality: American

Genre: Unknown

Probability: 0.333312

**Query execution time**: 64

**Way of computing:** in this query we selected all the information about a certain artist from the musicians table: his type, nationality and genre. The selected artist is an artist that we have at least one piece of information about him (besides his name). The reason we have 2 tuples as results his that this artist was added to the db from 2 different heuristics with different kinds of genre extracted (in the second one we couldn’t find any genre). Probability was calculated as a multiplication of the probabilities of the values that were extracted.

3) **Query**: who are the artists that were born in the 1960 decade

**Results**:

Whitney Houston Born In: 1963 Probability: 0.99

Jon Brion Born In: 1963 Probability: 0.99

Steve Rowe Born In: 1965 Probability: 0.99

Bret Michaels Born In: 1963 Probability: 0.99

Taylor Dayne Born In: 1962 Probability: 0.99

Bret Michaels Born In: 1963 Probability: 0.99

Trent Reznor Born In: 1965 Probability: 0.99

Gillian Welch Born In: 1967 Probability: 0.99

Diana Krall Born In: 1964 Probability: 0.99

Alison Moyet Born In: 1961 Probability: 0.99

Maxi Priest Born In: 1962 Probability: 0.99

Adam Sandler Born In: 1966 Probability: 0.99

Dave Mustaine Born In: 1961 Probability: 0.99

Serj Tankian Born In: 1967 Probability: 0.99

Tim Whitnall Born In: 1961 Probability: 0.99

Johnny Marr Born In: 1963 Probability: 0.99

Anthony Kiedis Born In: 1962 Probability: 0.99

Johnny Marr Born In: 1963 Probability: 0.99

Jack Irons Born In: 1962 Probability: 0.99

Krist Novoselic Born In: 1965 Probability: 0.99

Lars Ulrich Born In: 1963 Probability: 0.99

James Marsters Born In: 1962 Probability: 0.99

Criss Angel Born In: 1967 Probability: 0.99

Whitney Houston Born In: 1963 Probability: 0.99

Courtney Love Born In: 1964 Probability: 0.96

Paula Abdul Born In: 1962 Probability: 0.96

Jason Everman Born In: 1967 Probability: 0.96

Jon Bon Jovi Born In: 1962 Probability: 0.96

Criss Angel Born In: 1967 Probability: 0.96

Bret Michaels Born In: 1963 Probability: 0.96

Mike Tramp Born In: 1961 Probability: 0.96

Lenny Kravitz Born In: 1964 Probability: 0.96

Johnny Marr Born In: 1963 Probability: 0.96

Jon Bon Jovi Born In: 1962 Probability: 0.96

Maynard James Keenan Born In: 1964 Probability: 0.96

Bret Michaels Born In: 1963 Probability: 0.96

Trey Parker Born In: 1969 Probability: 0.96

Bono Born In: 1960 Probability: 0.96

Johnny Marr Born In: 1963 Probability: 0.96

Maxi Priest Born In: 1962 Probability: 0.96

Toni Braxton Born In: 1967 Probability: 0.96

Gillian Welch Born In: 1967 Probability: 0.96

Lisa Gerrard Born In: 1961 Probability: 0.96

Adam Clayton Born In: 1960 Probability: 0.96

Dinah Washington Born In: 1963 Probability: 0.96

Janet Jackson Born In: 1966 Probability: 0.96

Whitney Houston Born In: 1963 Probability: 0.96

Alison Moyet Born In: 1961 Probability: 0.96

Lisa Gerrard Born In: 1961 Probability: 0.96

Adam Clayton Born In: 1960 Probability: 0.96

Sheryl Crow Born In: 1962 Probability: 0.96

Whitney Houston Born In: 1963 Probability: 0.96

Sertab Erener Born In: 1964 Probability: 0.96

Timo Kotipelto Born In: 1969 Probability: 0.96

The Edge Born In: 1961 Probability: 0.96

Serj Tankian Born In: 1967 Probability: 0.96

Garth Brooks Born In: 1962 Probability: 0.96

Sertab Erener Born In: 1964 Probability: 0.96

Timo Kotipelto Born In: 1969 Probability: 0.96

The Edge Born In: 1961 Probability: 0.96

Axl Rose Born In: 1962 Probability: 0.96

Ara Gevorgyan Born In: 1960 Probability: 0.96

Adam Sandler Born In: 1966 Probability: 0.96

Sarah Brightman Born In: 1960 Probability: 0.96

Larry Mullen Jr. Born In: 1961 Probability: 0.96

Anthony Kiedis Born In: 1962 Probability: 0.96

James Hetfield Born In: 1963 Probability: 0.96

Dave Mustaine Born In: 1961 Probability: 0.96

Taylor Dayne Born In: 1962 Probability: 0.96

Chad Smith Born In: 1962 Probability: 0.96

James Hetfield Born In: 1963 Probability: 0.96

Tim Whitnall Born In: 1961 Probability: 0.96

Amy Grant Born In: 1960 Probability: 0.96

Hillel Slovak Born In: 1962 Probability: 0.96

Jon Brion Born In: 1963 Probability: 0.96

TobyMac Born In: 1964 Probability: 0.96

Strawberry Alarm Clock Born In: 1966 Probability: 0.96

Jack Irons Born In: 1962 Probability: 0.96

Elliott Smith Born In: 1969 Probability: 0.96

Chris Cornell Born In: 1964 Probability: 0.96

Golden Earring Born In: 1961 Probability: 0.96

Lars Ulrich Born In: 1963 Probability: 0.96

Elliott Smith Born In: 1969 Probability: 0.96

Kim Thayil Born In: 1960 Probability: 0.96

Golden Earring Born In: 1961 Probability: 0.96

George Michael Born In: 1963 Probability: 0.96

Steve Rowe Born In: 1965 Probability: 0.96

Kylie Minogue Born In: 1968 Probability: 0.96

Matt Cameron Born In: 1962 Probability: 0.96

Fleetwood Mac Born In: 1967 Probability: 0.96

James Marsters Born In: 1962 Probability: 0.96

Billy Ray Cyrus Born In: 1961 Probability: 0.96

Kurt Cobain Born In: 1967 Probability: 0.96

Ben Shepherd Born In: 1968 Probability: 0.96

Trent Reznor Born In: 1965 Probability: 0.96

Lisa Loeb Born In: 1968 Probability: 0.96

Billy Ray Cyrus Born In: 1961 Probability: 0.96

Paula Abdul Born In: 1962 Probability: 0.96

Ben Shepherd Born In: 1968 Probability: 0.96

Diana Krall Born In: 1964 Probability: 0.96

Toby Keith Born In: 1961 Probability: 0.96

Courtney Love Born In: 1964 Probability: 0.96

Serj Tankian Born In: 1967 Probability: 0.58

Query execution time: 334

**Way of computing:** We selected all the names from the musicians table that their “Born In” field from table persons was between decade and decade+9. We used the join statement on name value from both tables to extract the information we wanted.

4) **Query**: who are all the artists that are from country France

**Results**:

Charles Aznavour Probability: 0.62

Camille Dalmais Probability: 0.62

Brigitte Bardot Probability: 0.62

Alizֳ©e Probability: 0.62

Yelle Probability: 0.62

Depeche Mode Probability: 0.5

Yelle Probability: 0.5

**Query execution time:** 6 mS

**Way of computing:** we selected all the names from the musicians table that their nationality field fitted the country inserted in the query. We did by using an external file with country names and their nationality that was loaded into a properties data structure at the beginning of the program. Upon inserting an artist to the musicians table we searched for his birth country in the properties ds according to the nationality we extracted before according to one of the heuristics and inserted that value into the musician “origin” field. We used the join statement on the origin field from table musicians and on the name field from table countries.

5) **Query**: what are the top -5 populated countries

**Results**:

Peoples Republic of China Population: 1306314812 Probability: 0.99

Union of Soviet Socialist Republics Population: 300000000 Probability: 0.99

United States Population: 300000000 Probability: 0.99

Indonesia Population: 241973879 Probability: 0.99

Brazil Population: 188078261 Probability: 0.99

**Query execution time**: 54

**Way Of Computing:** We selected the population and population probability from the countries table, ordered them by population in descending order and limited the result to 5, according to the top-k value that was inserted.

**Information about our DB**

At the beginning of our program we built 3 db tables: Persons, Musicians and Countries. The tables have the following fields:

Persons

"(Ind INT not NULL AUTO\_INCREMENT," +

"Name VARCHAR(255) not NULL," +

"BornIn INT," +

"BornIn\_probability DOUBLE," +

"DiedIn INT," +

"DiedIn\_probability DOUBLE," +

"Profession VARCHAR(255)," +

"Profession\_probability DOUBLE," +

"PRIMARY KEY (Ind))";

Musicians

"(Ind INT not NULL AUTO\_INCREMENT," +

"Name VARCHAR(255) not NULL," +

"Type VARCHAR(255)," +

"Type\_probability DOUBLE," +

"Nationality VARCHAR(255)," +

"Nationality\_probability DOUBLE," +

"Origin VARCHAR(255)," +

"Genre VARCHAR(255)," +

"Genre\_probability DOUBLE," +

"PRIMARY KEY (Ind))";

Countries

"(Ind INT not NULL AUTO\_INCREMENT," +

"Name VARCHAR(255) not NULL," +

"Population INT," +

"Population\_probability DOUBLE," +

"PRIMARY KEY (Ind))";

All the tables were built so that their primary key is an auto increment index. We did it because we wanted the option to insert multiple tuples with the same name because different heuristics can extract different information/probability about a certain person/musician/country. We made sure to make a connection between the tables according to the persons name and musicians name and by the musicians nationality field and the country name field as it was explained in the way of computing query number 6 earlier in this document.