

Data Mining 101

Now we have the data in a datastore, how are we going to get useful information out?

SQL Query => Simple Aggregation (mean) => Simple Statistics (standard deviation) => Hypothesis Testing => Data Mining => Artificial Intelligence

- Make sure you master SQL well
- Low hanging fruits;
- Best bang for the bucks!



SQL

Learning by doing

<https://www.w3resource.com/sql-exercises/>

SQL CheatSheet

SQL CHEAT SHEET <http://www.sqltutorial.org>



QUERYING DATA FROM A TABLE

SELECT c1, c2 FROM t;

Query data in columns c1, c2 from a table

SELECT * FROM t;

Query all rows and columns from a table

SELECT c1, c2 FROM t

WHERE condition;

Query data and filter rows with a condition

SELECT DISTINCT c1 FROM t

WHERE condition;

Query distinct rows from a table

SELECT c1, c2 FROM t

ORDER BY c1 ASC [DESC];

Sort the result set in ascending or descending order

SELECT c1, c2 FROM t

ORDER BY c1

LIMIT n OFFSET offset;

Skip *offset* of rows and return the next *n* rows

SELECT c1, aggregate(c2)

FROM t

GROUP BY c1;

Group rows using an aggregate function

SELECT c1, aggregate(c2)

FROM t

GROUP BY c1

HAVING condition;

Filter groups using HAVING clause

QUERYING FROM MULTIPLE TABLES

SELECT c1, c2

FROM t1

INNER JOIN t2 ON condition;

Inner join t1 and t2

SELECT c1, c2

FROM t1

LEFT JOIN t2 ON condition;

Left join t1 and t2

SELECT c1, c2

FROM t1

RIGHT JOIN t2 ON condition;

Right join t1 and t2

SELECT c1, c2

FROM t1

FULL OUTER JOIN t2 ON condition;

Perform full outer join

SELECT c1, c2

FROM t1

CROSS JOIN t2;

Produce a Cartesian product of rows in tables

SELECT c1, c2

FROM t1, t2;

Another way to perform cross join

SELECT c1, c2

FROM t1 A

INNER JOIN t2 B ON condition;

Join t1 to itself using INNER JOIN clause

USING SQL OPERATORS

SELECT c1, c2 FROM t1

UNION [ALL]

SELECT c1, c2 FROM t2;

Combine rows from two queries

SELECT c1, c2 FROM t1

INTERSECT

SELECT c1, c2 FROM t2;

Return the intersection of two queries

SELECT c1, c2 FROM t1

MINUS

SELECT c1, c2 FROM t2;

Subtract a result set from another result set

SELECT c1, c2 FROM t1

WHERE c1 [NOT] LIKE pattern;

Query rows using pattern matching %, _

SELECT c1, c2 FROM t

WHERE c1 [NOT] IN value_list;

Query rows in a list

SELECT c1, c2 FROM t

WHERE c1 BETWEEN low AND high;

Query rows between two values

SELECT c1, c2 FROM t

WHERE c1 IS [NOT] NULL;

Check if values in a table is NULL or not

SQL Check-list

- What are primary keys
- How to select subset data
- How to join two tables
- How to group by data

Note: We will NOT spend too much time on SQL, because many of the functions can be done in Python and you should learn it in a formal database class. However, you should make sure you know SQL well !
And I will have at least one (basic) question on SQL in Mid-term

Some useful resources

MS SQL Server 2017 Free Edition

- <https://www.microsoft.com/en-us/sql-server/sql-server-editions-express>

MS SQL Sample Database

- <https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-2017>

MS Azure Data Studio Download

- <https://docs.microsoft.com/en-us/sql/azure-data-studio/download?view=sql-server-2017>

Installer

1. MS SQL Server for Windows

<https://www.microsoft.com/en-us/sql-server/sql-server-editions-express>

1. Postgres for Mac

<https://www.postgresql.org/download/macosx/>

3. MySQL for Mac:

<https://dev.mysql.com/downloads/mysql/>