Task 1

1.structured data

```
In [ ]: ID      class   name    course  score
        1372557 A       John    math    72
        1425474 C       Ben     math    85.0
        1372558 a       John    math    72
        1325390 A       Anna    mathematics     76
        1492872 B       Mark    math    633
```

Here is an example of poor quality structured data. Though it has only 5 samples, there are a lot of problems.
1.Two samples are duplicated, where the data both indicate John's math scores.
2.The second sample of John, his class is wrongly written 'a'.
3.Anna's course is wrongly written 'mathematics', while others are 'math'.
4.Ben's math score is kept to one decimal place, which is inconsistent with other data.
5.Mark's math score is far more than 100, which seems to be recorded incorrectly.

For a good structured data, it should be processed as some tables. These tables should meet with the three normal form of database (1NF, 2NF, 3NF). For example, the primary key in a table needs to be unique, while the foreign key needs to point to the primary key of another table. Each feature in the table needs to be specified the data type (such as: int、string）.

2.unstructured data

┻┳|
┳┻| _
┻┳| •.•) Rockstar, Are The Hackers Gone?
┳┻|⊂ﾉ
┻┳|
/﹋\
(҂`_´) -AH HELL NO! GO BACK INSIDE!
<,︻╦╤─ ҉
/﹋\

https://steamcommunity.com/id/BrownShuggah/recommended/271590/

This is an example of poor quality unstructured data from Steam's game commons. From this character painting, We can easily see that a man is poking his head out of the wall and another man is shooting fiercely. They are talking about *Rockstar* which is the company made game *GTA5* (Grand Theft Auto V). If you are not familiar with this game, you have no idea what they are talking about. When we are prepare to do NLP, this will be even worse. Because computer is hard to read these characters and know its meaning.

A good non-structure data need to include the information that we want to know. For example, A document for sentiment analysis should contain sentiment tendency words (such as: good、nice、bad、worse). These data should also be easily analyzed by automated programs. If a comment using ‘gooood’ ‘niccce’ instead of normal words, which is difficult to be processed.