# Project: Data Ingestion

### Data Source

I intended to use X(Twitter) posts and commands as source data. However, getting X data is expensive. Therefore, I switch to Quora questions and answers as my data.

图形用户界面, 应用程序

描述已自动生成

Here is the link to my data source: <https://huggingface.co/datasets/LxYxvv/quora_qa_raw>.

### A Snippet of the Dataset

An example of raw data

The following is an example of one data, I make the first-level **key** bold.

{"**qid**":9000001,"**url**":"/unanswered/How-has-Starbucks-Order-and-Pay-changed-people’s-habits-lives","**title**":"{\"sections\": [{\"spans\": [{\"text\": \"How has Starbucks Order and Pay changed people\\u2019s habits/lives?\", \"modifiers\": {}}], \"indent\": 0, \"quoted\": false, \"type\": \"plain\", \"is\_rtl\": false}]}","**creationTime**":1467388039955600,"**followerCount**":0,"**viewCount**":70,"**numAnswers**":0,"**numMachineAnswers**":0,"**isLocked**":false,"**isTrendyQuestion**":false,"**asker**":{"uid":143638334,"givenName":"Adriana","familyName":"Amaro","isMachineAnswerBot":false},"**answers**":"[]"}

Structure of raw data

If we parse the data properly, we can see its structure more clearly:

* **qid**: Int
* **url**: String
* **title**: Dictionary as String
  + **sections**: List
    - **spans**: List
      * **text**: String
      * **modifiers**: Dictionary as String
    - **indent**: Int
    - **quoted**: Boolean
    - **type**: String
    - **is\_rtl**: Boolean
* **creationTime**: Long
* **followerCount**: Int
* **viewCount**: Int
* **numAnswers**: Int
* **numMachineAnswers**: Int
* **isLocked**: Boolean
* **isTrendyQuestion**: Boolean
* **asker**: Dictionary as String
  + **uid**: Int
  + **givenName**: String
  + **familyName**: String
  + **isMachineAnswerBot**: Boolean
* **answers**: List
  + **aid**: Int
  + **url**: String
  + **content**: Same as title
  + **author**: Dictionary as String
    - **uid**: Int
    - **givenName**: String
    - **familyName**: String
    - **profileUrl**: String
  + **isSensitive**: Boolean
  + **isShortContent**: Boolean
  + **creationTime**: Int
  + **numViews**: Int
  + **numUpvotes**: Int
  + **numShares**: Int
  + **numComments**: Int
  + **comments**: List
    - **commentId**: Int
    - **creationTime**: Int
    - **url**: String
    - **author**: Same as author under answers
    - **content**: Same as title
    - **comments**: (Optional) List

### Data Profiling

I count the following invalid cases during the data profiling:

* For question json (top level):
  + Unmatched keys
  + Empty url
  + Null title
  + Negative integer (followerCount, viewCount, numAnswers, numMachineAnswers) and long (creationTime)
  + Unmatched answer number
* For answer json:
  + Unmatched keys
  + Negative integer (numViews, numUpvotes, numShares, numComments)
  + Unmatched comment number (I need to check the correct recursively)

I also count the number of different type of questions or answers:

* For question:
  + Locked questions
  + Trendy questions

For answer json:

* + Sensitive content
  + Short content

Result

Answer Comment Num Not Match 8413

Answer Key Not Match 0

Answer Negative Comment Num 0

Answer Negative Share Num 0

Answer Negative Upvote Num 0

Answer Negative View Num 0

Answer Number 2129401

Answer Number Not Match 19

Answer Sensitive 100331

Answer ShortContect 557467

Question Key Not Match 0

Question Locked 9533

Question Negative Answer Num 0

Question Negative Creation Time 0

Question Negative Machine Answer Num 0

Question Negative View Count 0

Question Negative follower Count 0

Question Null Title 75896

Question Null URL 0

Question Trendy 0

Question number 997955

### Data Cleaning

For each data, I extract data that has valid creationTime and url:

* For creationTime, I convert it to the format of "yyyy-MM-dd HH:mm:ss"
* For url, I process it to get the content of question. There are two types of url:
  + "/What-will-terrorism-look-like-in-100-years"
  + "/unanswered/How-has-Starbucks-Order-and-Pay-changed-people’s-habits-lives"

I delete '/'s and '/unanswered', replace '-' and '\t' with ' ', and convert all characters to lower case.

A Snippet of result

2016-07-01 15:47:14 what will terrorism look like in 100 years

2016-07-01 15:47:19 how has starbucks order and pay changed people’s habits lives

2016-07-01 15:47:21 what is the highest iit rank for a student from dubai in 2016

2016-07-01 15:47:22 sorry dont know u

2016-07-01 15:47:30 why does rendering videos produces so much heat in my laptop so i need to use a cooler every single time

2016-07-01 15:47:31 what are the good magazines for engineering service exam

2016-07-01 15:47:32 how many womens have foot fetish

2016-07-01 15:47:32 how is iiit sonepat campus

2016-07-01 15:47:33 my mouse cursor is out of sync with the original in screen sharing webrtc api

2016-07-01 15:47:35 need a baas chat service provider with webhook support to migrate from sendbird we built our app on sendbird but they removed the 49 per month plan