데이터 수집

연구를 하는 과정에서 데이터 수집은 정말 중요합니다.

데이터를 수집하는 과정은 정말 다양하지만, 제가 생각하는 수집 방법은 크게 2가지 입니다.

1. 제공된 데이터 다운 ex) AI-Hub, bigkinds (뉴스 데이터 제공)
2. 데이터 직접 수집

   2.1 API가 존재하는가? ex) reddit, bigkinds (유료인가..?)

   2.2 다른 사람이 먼저 구현한 코드 (github)

   2.3 * 크롤링 (최후의 보루) 제공된 데이터를 다운 받을 때도 크롤링을 사용하는 경우가 존재 (너무 많을 때 Selenium을 이용해서 사람이 하는 것처럼 다운)

google-play-scraper (API)

설치

https://play.google.com/store/games?hl=en

In [1]:

```
!pip install google-play-scraper
```

```
Requirement already satisfied: google-play-scraper in /usr/local/lib/python3.10/dist-packages (1.2.4)
```

데이터 수집

In [2]:

```
from google_play_scraper import Sort, reviews_all
# API에서 제공되는 데이터는 application의 모든 리뷰를 가져오는 것은 아닙니다.
# 또한 시도마다 서버에서 제공되는 데이터 개수가 달라질 수 있습니다.
# reviews_all안의 값들은 영어 데이터를 수집하기 위해 설정되어 있습니다.
raw_data = reviews_all(
    'com.centr.app', # app
    # 'com.mojang.minecraftpe',
    lang='en', # defaults to 'en'
    country='us', # defaults to 'us'
    sort=Sort.MOST_RELEVANT # defaults to Sort.MOST_RELEVANT
)

# 댓글, 별점, 공감수, 시간
```

수집된 데이터 출력

In [3]:

```
from pprint import pprint
pprint(raw_data[:5])
```

```
[{'appVersion': '6.4.1.20230612.1',
  'at': datetime.datetime(2023, 8, 26, 13, 34, 7),
  'content': 'The app has a lot of content, from meals to meditation to '
             'exercise. However, I would be very useful if the app and website '
             'worked. There are constant crashes stating, "My session has '
             'expired." Apparently, this has been an issue before. Usually it '
             'fixes itself within a short amount time, few minutes, but today '
             "has been different mirroring peoples' issues for the past few "
             'years.',
  'repliedAt': None,
  'replyContent': None,
  'reviewCreatedVersion': '6.4.1.20230612.1',
  'reviewId': 'ff90e664-0623-45be-bf39-c4f1f3fa339e',
  'score': 2,
  'thumbsUpCount': 12,
  'userImage': 'https://play-lh.googleusercontent.com/a/ACg8ocIBqHLERcTnil73za4_YLEuuizcbCfGydJPZ6-SOWTr=mo',
  'userName': 'Jacob Savage'},
 {'appVersion': '6.4.1.20230612.1-WearApp',
  'at': datetime.datetime(2023, 8, 21, 18, 58, 42),
  'content': 'I have a subscription, but as I navigate through the app I keep '
             'receiving a message saying my session has ended and I keep '
             'needing to log back in. I need to do this MULTIPLE times within '
             "a few minutes. It's making the app unusable. Extremely "
             'disappointed by this. Also, please add the ability to keep the '
             "timer going on workouts when I lock my phone. I'm not trying to "
             'keep my phone screen on for an entire 40 minute period.',
  'repliedAt': None,
  'replyContent': None,
  'reviewCreatedVersion': '6.4.1.20230612.1-WearApp',
  'reviewId': 'cc0d9c7c-3474-4508-8dc3-444e738df969',
  'score': 1,
  'thumbsUpCount': 21,
  'userImage': 'https://play-lh.googleusercontent.com/a-/ALV-UjXHILvMv3FKVoNOvEzEr0pjyIxDECvnlh7KuBit28GH7868',
  'userName': 'Joshua Pauselius'},
 {'appVersion': '6.4.1.20230612.1',
  'at': datetime.datetime(2023, 8, 25, 11, 54, 49),
  'content': 'The program and workouts are great. The main issue is that the '
             'app is borderline unusable. I have a Galaxy Ultra and it kicks '
             'me out to log in almost every time I select a workout. Some '
             "mornings I can't even use it all together (about half). Don't "
             "buy it until this is fixed. Customer service's response was "
             '"Use our free YouTube channel".',
  'repliedAt': None,
  'replyContent': None,
  'reviewCreatedVersion': '6.4.1.20230612.1',
  'reviewId': '1a7c043d-54a8-4204-9e59-54b9865dfb30',
  'score': 1,
  'thumbsUpCount': 45,
  'userImage': 'https://play-lh.googleusercontent.com/a-/ALV-UjVQhYDjOL74hairQGmwONfv2pMvMBRued0sF5Phm5XNGTE',
  'userName': 'Michael Koska'},
 {'appVersion': '6.4.1.20230612.1',
  'at': datetime.datetime(2023, 8, 21, 15, 18, 24),
  'content': "Doesn't work half the time. You can't open most of the workouts "
             'and it randomly signs me out when I try to access the app. '
             "Clearing cache and uninstalling it doesn't work. It sometimes "
             "takes days before it'll work again. I've emailed support "
             'numerous times and was told they are aware of the issue and a '
             'fix will be out eventually. What BS is that?! For the cost of '
             'this app, I should be able to expect that it works every time. '
             "Even when it does work, it's good but not great.",
  'repliedAt': None,
  'replyContent': None,
  'reviewCreatedVersion': '6.4.1.20230612.1',
  'reviewId': 'db54488c-92d8-4299-a9c0-6f1517b1039f',
  'score': 1,
  'thumbsUpCount': 48,
  'userImage': 'https://play-lh.googleusercontent.com/a-/ALV-UjUcqqfmaG3MxCtMwn6dV2PWDLI6yExQeCbFxfZQP8bm0AU',
  'userName': 'Felicia Lopes'},
 {'appVersion': '6.5.1.20230626.2',
  'at': datetime.datetime(2023, 7, 24, 12, 50, 26),
  'content': 'Could use some more personalized features. Would be nice if it '
             'had better descriptions available prior to the coached workouts. '
             'I\'m traveling alot and my internet connections are "unreliable" '
             'so would be nice to stage/download workouts on my phone. They do '
             "seem to refine and update often. All in all, I've found it's "
             'generally been a great app for my needs.',
  'repliedAt': datetime.datetime(2023, 7, 26, 0, 40, 55),
  'replyContent': 'Thanks for your great review! We are continuing development '
                  'on the app with lots of new features to come. Offline video '
                  'support is very high on our priority list! Be sure to check '
                  'back regularly for updates! For anything else feel free to '
                  'get in touch at hello@centr.com.',
  'reviewCreatedVersion': '6.5.1.20230626.2',
  'reviewId': '2458ba79-38a3-46e9-844d-98f742601613',
  'score': 4,
  'thumbsUpCount': 19,
  'userImage': 'https://play-lh.googleusercontent.com/a-/ALV-UjWB1y3BOVJFaMZCfIvY6JOJk3tgmMN7lIgEJzeiSHf2l34',
  'userName': 'Mark'}]
```

데이터 저장 (csv)

In [4]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [5]:

```
import csv

data_li = [(info['content'], info['score'], info['thumbsUpCount'], info['at']) for info in raw_data]

# mount 과정을 통해서 자신의 구글 드라이브에 데이터를 저장합니다.
# 혹은 경로 설정없이 "application_data.csv"로 저장해도 되지만, 런타임이 끝나면 데이터가 사라질 위험이 있습니다.
with open("/content/drive/MyDrive/기계학습특론/3주차/application_data.csv", 'wt', encoding='utf-8', newline='') as file:
        csv_writer = csv.writer(file)
        csv_writer.writerow(["contents", "star", "agree", "date"])

        for data in data_li:
            csv_writer.writerow(data)
```

In [6]:

```
import pandas as pd

review_df = pd.read_csv("/content/drive/MyDrive/기계학습특론/3주차/application_data.csv")
review_df
```

Out[6]:

	contents	star	agree	date
0	The app has a lot of content, from meals to me...	2	12	2023-08-26 13:34:07
1	I have a subscription, but as I navigate throu...	1	21	2023-08-21 18:58:42
2	The program and workouts are great. The main i...	1	45	2023-08-25 11:54:49
3	Doesn't work half the time. You can't open mos...	1	48	2023-08-21 15:18:24
4	Could use some more personalized features. Wou...	4	19	2023-07-24 12:50:26
...	...	...	...	...
5412	Chrisssss!	5	0	2023-04-03 06:09:27
5413	✖️	1	0	2023-01-07 11:57:17
5414	❤❤❤❤😍😍😍😍😍	5	0	2023-04-12 23:01:10
5415	Ok	5	0	2022-12-29 03:47:29
5416	مزخرفه	1	0	2020-11-11 07:57:42

5417 rows × 4 columns

In [ ]: