Data Scraping

December 9, 2025

[5]:
```
from google_play_scraper import reviews_all, Sort
import pandas as pd
import numpy as np
```

[6]:
```
genshin_reviews = reviews_all(
    "com.miHoYo.GenshinImpact",
    sleep_milliseconds=0,
    lang='id',
    country='id',
    sort=Sort.MOST_RELEVANT)
```

[7]:
```
df = pd.DataFrame(np.array(genshin_reviews), columns=['review'])
df = df.join(pd.DataFrame(df.pop('review').tolist()))

df = df[df['content'].apply(lambda x: isinstance(x, str) and len(x.split()) >=␣
↪3)]

df = df[['content', 'score']].rename(columns={
    'content': 'review',
    'score': 'rating'
})

positif = df[df['rating'].isin([4, 5])].sample(n=750, random_state=1234)
negatif = df[df['rating'].isin([1, 2, 3])].sample(n=750, random_state=1234)
```

[8]:
```
balanced_df = pd.concat([positif, negatif]).sample(frac=1, random_state=4321).
↪reset_index(drop=True)
```

[9]:
```
balanced_df.to_excel("C:/!! Kuliah/Semester 6/Data Mining/proyek akhir/
↪reviews_balanced.xlsx", index=False)
```