# Lendo os dados e gravando em múltiplos formatos

Arquivo de dados:

https://drive.google.com/file/d/1U0e-ssOerAZwEL9-6wBduj (https://drive.google.com/file/d/1U0e-ssOerAZwEL9-6wBduj)

O arquivo fora inicialmente baixado para '../data/bkp/AppleStore.csv', mas não foi adicionado ao repositório (figura no .gitignore). Para rodar novamente este notebook, deve-se repetir o processo de gravação do arquivo de dados (.csv) de origem no diretório indicado.

In [1]:

```
import csv, io, json, requests, sqlite3
import pandas as pd

# O arquivo foi baixado para ../data/bkp/AppleStore.csv
df = pd.read_csv('../data/bkp/AppleStore.csv', index_col=0)
```

# Execução 1: tag "news" com maior quantidade de avaliações

Identifique a Aplicação da categoria News, que tiver a maior quantidade de avaliações rating_count_tot.

In [2]:

```
df_news = df[df['prime_genre'] == 'News'].filter(['track_name', 'rating_count_tot'])

# Verificando se pode haver mais de 1 ocorrência de track_name (versões diferentes)
df_news.groupby('track_name').count().sort_values(by='rating_count_tot', ascending=Fals
e)
```

Out[2]:

| track_name | rating_count_tot |
|---|---|
| 20 Minutes.fr - l'actualité en continu | 1 |
| franceinfo - l'actualité & les élections en direct | 1 |
| WSVN Hurricane Tracker | 1 |
| WIRED Magazine | 1 |
| ViRATES[バイレーツ]-面白ネタまとめの決定版！ | 1 |
| Verbrechen - echte Polizeifälle aus Deiner Umgebung | 1 |
| USA TODAY | 1 |
| US Presidential Election 2016 - Polls | 1 |
| Twitter | 1 |
| TopBuzz: Best Viral Videos, GIFs, TV & News | 1 |
| Ticket Scanner for Powerball & MegaMillions Pool | 1 |
| The Washington Post Classic | 1 |
| The Guardian | 1 |
| Tagesschau | 1 |
| State Council - Official Chinese government app | 1 |
| SmartNews - Trending News & Stories | 1 |
| Smart Channel -New Style of News Reader- | 1 |
| ZAKER 专业版 | 1 |
| n-tv Nachrichten | 1 |
| Scanner911 Pro | 1 |
| theSkimm | 1 |
| 腾讯新闻HD-最资深的阅读软件 | 1 |
| 腾讯新闻-头条新闻热点资讯掌上阅读软件 | 1 |
| 网易新闻 - 精选好内容，算出你的兴趣 | 1 |
| 新浪新闻-阅读最新时事热门头条资讯视频 | 1 |
| 搜狐新闻—新闻热点资讯掌上阅读软件 | 1 |
| 天天快报 - 最热门的新闻资讯软件 | 1 |
| 华尔街见闻（专业版）-全球财经新闻精选 | 1 |
| 凤凰新闻(专业版)-有料的军事新闻、娱乐短视频 | 1 |
| 今日头条(专业版) - 推荐热点新闻资讯、娱乐视频 | 1 |
| ... | ... |
| Daily Classifieds for iPhone | 1 |
| Conservative Talk Radio | 1 |
| Clown Spotter - Find Clowns Around You | 1 |
| CNN: Breaking US & World News, Live Video | 1 |
| CNN Politics | 1 |

| track_name | rating_count_tot |
|---|---|
| CBS News - Watch Free Live Breaking News | 1 |
| Boycott Trump Biz | 1 |
| BFMTV : l'info en continu | 1 |
| AOL: News, Email, Weather & Video | 1 |
| ABC News - US & World News + Live Video | 1 |
| Fresco — Be a part of the news | 1 |
| HuffPost - News, Politics & Entertainment | 1 |
| ST channel ［エスティーチャンネル］- 雑誌『セブンティーン』公式アプリ | 1 |
| JCnews - Anime & Game Culture | 1 |
| SPIEGEL ONLINE - Nachrichten | 1 |
| Reddit Official App: All That's Trending and Viral | 1 |
| Quartz • News in a whole new way | 1 |
| Presidential Election & Electoral College Maps | 1 |
| PodCruncher podcast app - Player and manager for podcasts | 1 |
| Pocket Casts | 1 |
| PS Deals+ - Games Price Alerts for PS4, PS3, Vita | 1 |
| OPM Alert | 1 |
| News Pro - Breitbart Edition | 1 |
| News Break - Local & World Breaking News & Radio | 1 |
| NBC News | 1 |
| MSNBC | 1 |
| Lotto Results - Mega Millions Powerball Lottery | 1 |
| LotteryHUB | 1 |
| KATWARN | 1 |
| 2 ちゃんねる for iPhone | 1 |

75 rows × 1 columns

In [3]:

```
# Obtendo a aplicação com maior número de ratings em sua história
df_news.sort_values(by='rating_count_tot', ascending=False).head(1)
```

Out[3]:

| | track_name | rating_count_tot |
|---|---|---|
| **251** | Twitter | 354058 |

## Resultado: Twitter, com 354058 avaliações

# Execução 2: 10 mais avaliados entre as tags "music" e "book"

Identificar quais são as 10 Aplicações do gênero Music e Book que possuem a maior quantidade de avaliações no arquivo csv apple_store.

In [10]:

```
df_music_book = df[df['prime_genre'].isin(['Music', 'Book'])]

df_topten = df_music_book.sort_values(by='rating_count_tot', ascending=False).head(10)
df_topten
```

Out[10]:

| | id | track_name | size_bytes | currency | price | rating_count_tot | rating_count_ver |
|---|---|---|---|---|---|---|---|
| **8** | 284035177 | Pandora - Music & Radio | 130242560 | USD | 0.0 | 1126879 | 3594 |
| **202** | 324684580 | Spotify Music | 132510720 | USD | 0.0 | 878563 | 8253 |
| **20** | 284993459 | Shazam - Discover music, artists, videos & lyrics | 147093504 | USD | 0.0 | 402925 | 136 |
| **40** | 290638154 | iHeartRadio – Free Music & Radio Stations | 116443136 | USD | 0.0 | 293228 | 110 |
| **93** | 302584613 | Kindle – Read eBooks, Magazines & Textbooks | 169747456 | USD | 0.0 | 252076 | 80 |
| **270** | 336353151 | SoundCloud - Music & Audio | 105009152 | USD | 0.0 | 135744 | 594 |
| **816** | 421254504 | Magic Piano by Smule | 55030784 | USD | 0.0 | 131695 | 1102 |
| **1431** | 509993510 | Smule Sing! | 109940736 | USD | 0.0 | 119316 | 33 |
| **803** | 418987775 | TuneIn Radio - MLB NBA Audiobooks Podcasts Music | 101735424 | USD | 0.0 | 110420 | 370 |
| **1438** | 510855668 | Amazon Music | 77778944 | USD | 0.0 | 106235 | 4605 |

# Execução 3: 10 mais citados entre as tags "music" e "book"

Após encontrar a aplicação do tipo News utilize a sua API, para identificar quais das 10 aplicações do tipo Music e Book, possuem o maior número de citações nessa API.

Consultar o número de citações de cada track do item anterior na API aplicação mais avaliada (twitter).

## Heurística das citações

Não há indicativo de usuário no dataframe original. Uma alternativa seria buscar por parte do nome em # ou @ em um conjunto de tweets. No entanto, isso provocaria um efeito colateral, com aumento no número real em casos como Shazam (que adicionaria as menções e citações ao herói).

Em função disso, busquei manualmente os usuários no contexto dos aplicativos e então fiz uma busca pelas citações (@usuario), adicionando os resultados ao dataframe.

## Scraping limitado

Para fins do desafio, economizando tempo e recursos, além de evitar o bloqueio das chamadas pelo twitter, a busca será limitada aos tweets do ano de 2019, limitando a 10000 ocorrências de qualquer dos usuários que são objetos da análise (citados). Em uma situação real, a ingestão inicial de tweets seria feita em um job (bash) usando a memsa biblioteca, gravando em um arquivo para consumo posterior. Esse job poderia ser agendado para

In [11]:

```
## Run only if twitterscraper is not installed
# import sys
# !{sys.executable} -m pip install twitterscraper

dict_users = {
    "id": [284035177, 324684580, 284993459, 290638154, 302584613, 336353151, 421254504,
509993510, 418987775, 510855668],
    "twitter_user": ["@pandoramusic", "@Spotify", "@Shazam", "@iHeartRadio", "@AmazonKi
ndle", "@SoundCloud", "@MagicPianoApp", "@smule", "@tunein", "@amazonmusic"],
    'n_citacoes': [0 for i in range(10)]
}

from twitterscraper import query_tweets
qry = " OR ".join(dict_users['twitter_user'])

# Getting all mentions in 2019, up to 10000.
import datetime as dt
for tweet in query_tweets(qry, 10000, begindate=dt.date(2019, 1, 1)):
    for i in range(10):
        if dict_users['twitter_user'][i] in tweet.text:
            dict_users['n_citacoes'][i] = dict_users['n_citacoes'][i] + 1

df_mention = pd.DataFrame.from_dict(dict_users)

# Faz o merge com o top10
df_topten = df_topten.merge(df_mention)

# # Ordenando pelo n_citacoes DESC
df_topten = df_topten.sort_values(by='n_citacoes', ascending=False)

df_topten
```

```
INFO: queries: ['@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @
AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunein OR @ama
zonmusic since:2019-01-01 until:2019-01-18', '@pandoramusic OR @Spotify OR
@Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp
OR @smule OR @tunein OR @amazonmusic since:2019-01-18 until:2019-02-04',
'@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR
@SoundCloud OR @MagicPianoApp OR @smule OR @tunein OR @amazonmusic since:2
019-02-04 until:2019-02-21', '@pandoramusic OR @Spotify OR @Shazam OR @iHe
artRadio OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @smule OR @t
unein OR @amazonmusic since:2019-02-21 until:2019-03-11', '@pandoramusic O
R @Spotify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @
MagicPianoApp OR @smule OR @tunein OR @amazonmusic since:2019-03-11 until:
2019-03-28', '@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @Ama
zonKindle OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunein OR @amazon
music since:2019-03-28 until:2019-04-14', '@pandoramusic OR @Spotify OR @S
hazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp OR
@smule OR @tunein OR @amazonmusic since:2019-04-14 until:2019-05-01', '@pa
ndoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR @Sou
ndCloud OR @MagicPianoApp OR @smule OR @tunein OR @amazonmusic since:2019-
05-01 until:2019-05-19', '@pandoramusic OR @Spotify OR @Shazam OR @iHeartR
adio OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunei
n OR @amazonmusic since:2019-05-19 until:2019-06-05', '@pandoramusic OR @S
potify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @Magi
cPianoApp OR @smule OR @tunein OR @amazonmusic since:2019-06-05 until:2019
-06-22', '@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @AmazonK
indle OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunein OR @amazonmusi
c since:2019-06-22 until:2019-07-09', '@pandoramusic OR @Spotify OR @Shaza
m OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @sm
ule OR @tunein OR @amazonmusic since:2019-07-09 until:2019-07-27', '@pando
ramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundC
loud OR @MagicPianoApp OR @smule OR @tunein OR @amazonmusic since:2019-07-
27 until:2019-08-13', '@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadi
o OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunein O
R @amazonmusic since:2019-08-13 until:2019-08-30', '@pandoramusic OR @Spot
ify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @MagicPi
anoApp OR @smule OR @tunein OR @amazonmusic since:2019-08-30 until:2019-09
-16', '@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR @AmazonKind
le OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunein OR @amazonmusic s
ince:2019-09-16 until:2019-10-04', '@pandoramusic OR @Spotify OR @Shazam O
R @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @smule
OR @tunein OR @amazonmusic since:2019-10-04 until:2019-10-21', '@pandoramu
sic OR @Spotify OR @Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud
OR @MagicPianoApp OR @smule OR @tunein OR @amazonmusic since:2019-10-21 un
til:2019-11-07', '@pandoramusic OR @Spotify OR @Shazam OR @iHeartRadio OR
@AmazonKindle OR @SoundCloud OR @MagicPianoApp OR @smule OR @tunein OR @am
azonmusic since:2019-11-07 until:2019-11-24', '@pandoramusic OR @Spotify O
R @Shazam OR @iHeartRadio OR @AmazonKindle OR @SoundCloud OR @MagicPianoAp
p OR @smule OR @tunein OR @amazonmusic since:2019-11-24 until:2019-12-12']
INFO: Got 504 tweets (504 new).
INFO: Got 1009 tweets (505 new).
INFO: Got 1510 tweets (501 new).
INFO: Got 2021 tweets (511 new).
INFO: Got 2535 tweets (514 new).
INFO: Got 3036 tweets (501 new).
INFO: Got 3550 tweets (514 new).
INFO: Got 4059 tweets (509 new).
INFO: Got 4574 tweets (515 new).
INFO: Got 5088 tweets (514 new).
INFO: Got 5601 tweets (513 new).
INFO: Got 6119 tweets (518 new).
INFO: Got 6625 tweets (506 new).
```

```
INFO: Got 7136 tweets (511 new).
INFO: Got 7654 tweets (518 new).
INFO: Got 8171 tweets (517 new).
INFO: Got 8685 tweets (514 new).
INFO: Got 9199 tweets (514 new).
INFO: Got 9706 tweets (507 new).
INFO: Got 10216 tweets (510 new).
```

Out[11]:

| | id | track_name | size_bytes | currency | price | rating_count_tot | rating_count_ver | us |
|---|---|---|---|---|---|---|---|---|
| **1** | 324684580 | Spotify Music | 132510720 | USD | 0.0 | 878563 | 8253 | |
| **3** | 290638154 | iHeartRadio – Free Music & Radio Stations | 116443136 | USD | 0.0 | 293228 | 110 | |
| **2** | 284993459 | Shazam - Discover music, artists, videos & lyrics | 147093504 | USD | 0.0 | 402925 | 136 | |
| **8** | 418987775 | TuneIn Radio - MLB NBA Audiobooks Podcasts Music | 101735424 | USD | 0.0 | 110420 | 370 | |
| **5** | 336353151 | SoundCloud - Music & Audio | 105009152 | USD | 0.0 | 135744 | 594 | |
| **0** | 284035177 | Pandora - Music & Radio | 130242560 | USD | 0.0 | 1126879 | 3594 | |
| **9** | 510855668 | Amazon Music | 77778944 | USD | 0.0 | 106235 | 4605 | |
| **4** | 302584613 | Kindle – Read eBooks, Magazines & Textbooks | 169747456 | USD | 0.0 | 252076 | 80 | |
| **7** | 509993510 | Smule Sing! | 109940736 | USD | 0.0 | 119316 | 33 | |
| **6** | 421254504 | Magic Piano by Smule | 55030784 | USD | 0.0 | 131695 | 1102 | |

# Entregável: Gravando em múltiplos formatos

O output esperado é a criação de um CSV, um JSON e uma base de dados local com as respectivas colunas: id, track_name, n_citacoes, size_bytes, price, prime_genre. Os dados relativos às Aplicações estão disponíveis no arquivo abaixo.

In [12]:

```
# Filtro do dataframe para gravar em diversos formatos com um número limitado de colunas
df_table = df_topten.filter(['id', 'track_name', 'n_citacoes', 'size_bytes', 'price', 'prime_genre'])
df_table.set_index('id')

df_table
```

Out[12]:

| | id | track_name | n_citacoes | size_bytes | price | prime_genre |
|---|---|---|---|---|---|---|
| **1** | 324684580 | Spotify Music | 1800 | 132510720 | 0.0 | Music |
| **3** | 290638154 | iHeartRadio – Free Music & Radio Stations | 775 | 116443136 | 0.0 | Music |
| **2** | 284993459 | Shazam - Discover music, artists, videos & lyrics | 616 | 147093504 | 0.0 | Music |
| **8** | 418987775 | TuneIn Radio - MLB NBA Audiobooks Podcasts Music | 412 | 101735424 | 0.0 | Music |
| **5** | 336353151 | SoundCloud - Music & Audio | 326 | 105009152 | 0.0 | Music |
| **0** | 284035177 | Pandora - Music & Radio | 106 | 130242560 | 0.0 | Music |
| **9** | 510855668 | Amazon Music | 99 | 77778944 | 0.0 | Music |
| **4** | 302584613 | Kindle – Read eBooks, Magazines & Textbooks | 55 | 169747456 | 0.0 | Book |
| **7** | 509993510 | Smule Sing! | 5 | 109940736 | 0.0 | Music |
| **6** | 421254504 | Magic Piano by Smule | 0 | 55030784 | 0.0 | Music |

In [13]:

```
# [ENTREGÁVEL] Criação de uma tabela no SQLite em memória
conn = sqlite3.connect(":memory:")
df_table.to_sql('summary_store', conn, if_exists='replace', index=False)
# pd.read_sql('select * from summary_store', conn)
conn.commit()
conn.close()

# [ENTREGÁVEL] Criação de uma tabela no SQLite em arquivo
conn = sqlite3.connect("../data/db.sqlite")
df_table.to_sql('summary_store', conn, if_exists='replace', index=False)
# pd.read_sql('select * from summary_store', conn)
conn.commit()
conn.close()

# [ENTREGÁVEL] Gravando o dataframe em CSV
df_table.to_csv("../data/summary_store.csv", quoting=csv.QUOTE_NONNUMERIC)

# [ENTREGÁVEL] Gravando o dataframe em JSON
df_table.to_json('../data/summary_store.json', orient="records")
```

In [ ]:

```

```