Create datasets for the Content-based Filter

This notebook builds the data we will use for creating our content based model. We'll collect the data via a collection of SQL queries from the publicly available Kurier.at dataset in BigQuery. Kurier.at is an Austrian newsite. The goal of these labs is to recommend an article for a visitor to the site. In this lab we collect the data for training, in the subsequent notebook we train the recommender model.

This notebook illustrates

- how to pull data from BigQuery table and write to local files
- how to make reproducible train and test splits

In [1]:
```
import os
import tensorflow as tf
import numpy as np
from google.cloud import bigquery

PROJECT = 'qwiklabs-gcp-00-573e51f4471d' # REPLACE WITH YOUR PROJECT ID
BUCKET = 'qwiklabs-gcp-00-573e51f4471d' # REPLACE WITH YOUR BUCKET NAME
REGION = 'us-central1' # REPLACE WITH YOUR BUCKET REGION e.g. us-central1

# do not change these
os.environ['PROJECT'] = PROJECT
os.environ['BUCKET'] = BUCKET
os.environ['REGION'] = REGION
os.environ['TFVERSION'] = '2.1'
```

In [2]:
```
%%bash
gcloud  config  set project $PROJECT
gcloud config set compute/region $REGION
```

```
Updated property [core/project].
Updated property [compute/region].
```

We will use this helper function to write lists containing article ids, categories, and authors for each article in our database to local file.

In [3]:
```
def write_list_to_disk(my_list, filename):
  with open(filename, 'w') as f:
    for item in my_list:
        line = "%s\n" % item
        f.write(line)
```

Pull data from BigQuery

The cell below creates a local text file containing all the article ids (i.e. 'content ids') in the dataset.

Have a look at the original dataset in BigQuery. Then read through the query below and make sure you understand what it is doing.

In [4]:
```
sql="""
#standardSQL

SELECT
  (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(hits.customDimensions)) AS
FROM `cloud-training-demos.GA360_test.ga_sessions_sample`,
  UNNEST(hits) AS hits
WHERE
  # only include hits on pages
  hits.type = "PAGE"
  AND (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(hits.customDimensions))
GROUP BY
  content_id

"""

content_ids_list = bigquery.Client().query(sql).to_dataframe()['content_id'].tol
write_list_to_disk(content_ids_list, "content_ids.txt")
print("Some sample content IDs {}".format(content_ids_list[:3]))
print("The total number of articles is {}".format(len(content_ids_list)))
```

```
Some sample content IDs ['299824032', '299865757', '299918857']
The total number of articles is 15634
```

There should be 15,634 articles in the database.
Next, we'll create a local file which contains a list of article categories and a list of article authors.

Note the change in the index when pulling the article category or author information. Also, we are using the first author of the article to create our author list.
Refer back to the original dataset, use the `hits.customDimensions.index` field to verify the correct index.

In [5]:
```
sql="""
#standardSQL
SELECT
  (SELECT MAX(IF(index=7, value, NULL)) FROM UNNEST(hits.customDimensions)) AS c
FROM `cloud-training-demos.GA360_test.ga_sessions_sample`,
  UNNEST(hits) AS hits
WHERE
  # only include hits on pages
  hits.type = "PAGE"
  AND (SELECT MAX(IF(index=7, value, NULL)) FROM UNNEST(hits.customDimensions))
GROUP BY
  category
"""
categories_list = bigquery.Client().query(sql).to_dataframe()['category'].tolist
write_list_to_disk(categories_list, "categories.txt")
print(categories_list)
```

```
['News', 'Stars & Kultur', 'Lifestyle']
```

The categories are 'News', 'Stars & Kultur', and 'Lifestyle'.
When creating the author list, we'll only use the first author information for each article.

In [6]:
```
sql="""
#standardSQL
SELECT
  REGEXP_EXTRACT((SELECT MAX(IF(index=2, value, NULL)) FROM UNNEST(hits.customDi
FROM `cloud-training-demos.GA360_test.ga_sessions_sample`,
  UNNEST(hits) AS hits
WHERE
  # only include hits on pages
  hits.type = "PAGE"
  AND (SELECT MAX(IF(index=2, value, NULL)) FROM UNNEST(hits.customDimensions))
GROUP BY
  first_author
"""
authors_list = bigquery.Client().query(sql).to_dataframe()['first_author'].tolis
write_list_to_disk(authors_list, "authors.txt")
print("Some sample authors {}".format(authors_list[:10]))
print("The total number of authors is {}".format(len(authors_list)))
```

```
Some sample authors ['Alexander Huber', 'Christine Klafl', 'Philipp Albrechtsber
ger', 'Wolfgang Atzenhofer', 'Marlene Patsalidis', 'Peter Temel', 'Stefan Hofe
r', 'Michael Pammesberger', 'Helmut Brandstätter', 'Martina Salomon']
The total number of authors is 385
```

There should be 385 authors in the database.

Create train and test sets.

In this section, we will create the train/test split of our data for training our model. We use the concatenated values for visitor id and content id to create a farm fingerprint, taking approximately 90% of the data for the training set and 10% for the test set.

In [7]:
```
sql="""
WITH site_history as (
  SELECT
      fullVisitorId as visitor_id,
      (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(hits.customDimensions))
      (SELECT MAX(IF(index=7, value, NULL)) FROM UNNEST(hits.customDimensions))
      (SELECT MAX(IF(index=6, value, NULL)) FROM UNNEST(hits.customDimensions))
      (SELECT MAX(IF(index=2, value, NULL)) FROM UNNEST(hits.customDimensions))
      SPLIT(RPAD((SELECT MAX(IF(index=4, value, NULL)) FROM UNNEST(hits.customDi
      LEAD(hits.customDimensions, 1) OVER (PARTITION BY fullVisitorId ORDER BY h
  FROM
    `cloud-training-demos.GA360_test.ga_sessions_sample`,
    UNNEST(hits) AS hits
   WHERE
     # only include hits on pages
      hits.type = "PAGE"
      AND
      fullVisitorId IS NOT NULL
      AND
      hits.time != 0
      AND
      hits.time IS NOT NULL
      AND
      (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(hits.customDimensions))
)
SELECT
```

```
  visitor_id,
  content_id,
  category,
  REGEXP_REPLACE(title, r",", "") as title,
  REGEXP_EXTRACT(author_list, r"^[^,]+") as author,
  DATE_DIFF(DATE(CAST(year_month_array[OFFSET(0)] AS INT64), CAST(year_month_arr
  (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(nextCustomDimensions)) as n
FROM
  site_history
WHERE (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(nextCustomDimensions))
      AND ABS(MOD(FARM_FINGERPRINT(CONCAT(visitor_id, content_id)), 10)) < 9
"""
training_set_df = bigquery.Client().query(sql).to_dataframe()
training_set_df.to_csv('training_set.csv', header=False, index=False, encoding='
training_set_df.head()
```

Out[7]:

	visitor_id	content_id	category	title	author	months_since_epo
0	1038643850985118087	299828023	News	Glyphosat geht in die Verlängerung	Andreas Anzenberger	5
1	1083265653486482344	299828023	News	Glyphosat geht in die Verlängerung	Andreas Anzenberger	5
2	1090462532616257705	299696307	News	Patient mit Taxi heimgeschickt	Johannes Weichhart	5
3	1090462532616257705	299793337	News	"Sittenwächter" an See in Niederösterreich: Ha...	None	5
4	1122220186524713655	299915364	News	Glyphosat-Verlängerung: SPD wirft CDU Vertraue...	None	5

In [8]:

```
sql="""
WITH site_history as (
  SELECT
      fullVisitorId as visitor_id,
      (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(hits.customDimensions))
      (SELECT MAX(IF(index=7, value, NULL)) FROM UNNEST(hits.customDimensions))
      (SELECT MAX(IF(index=6, value, NULL)) FROM UNNEST(hits.customDimensions))
      (SELECT MAX(IF(index=2, value, NULL)) FROM UNNEST(hits.customDimensions))
      SPLIT(RPAD((SELECT MAX(IF(index=4, value, NULL)) FROM UNNEST(hits.customDi
      LEAD(hits.customDimensions, 1) OVER (PARTITION BY fullVisitorId ORDER BY h
  FROM
    `cloud-training-demos.GA360_test.ga_sessions_sample`,
    UNNEST(hits) AS hits
   WHERE
     # only include hits on pages
      hits.type = "PAGE"
      AND
      fullVisitorId IS NOT NULL
      AND
      hits.time != 0
      AND
      hits.time IS NOT NULL
```

```
      AND
      (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(hits.customDimensions))
)
SELECT
  visitor_id,
  content_id,
  category,
  REGEXP_REPLACE(title, r",", "") as title,
  REGEXP_EXTRACT(author_list, r"^[^,]+") as author,
  DATE_DIFF(DATE(CAST(year_month_array[OFFSET(0)] AS INT64), CAST(year_month_arr
  (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(nextCustomDimensions)) as n
FROM
  site_history
WHERE (SELECT MAX(IF(index=10, value, NULL)) FROM UNNEST(nextCustomDimensions))
      AND ABS(MOD(FARM_FINGERPRINT(CONCAT(visitor_id, content_id)), 10)) >= 9
"""
test_set_df = bigquery.Client().query(sql).to_dataframe()
test_set_df.to_csv('test_set.csv', header=False, index=False, encoding='utf-8')
test_set_df.head()
```

Out[8]:

	visitor_id	content_id	category	title	author	months_since_e
0	1162477138335769203	299928807	News	Missbrauchsvorwürfe: Tiroler Skiverband durchs...	Mirad Odobasic	
1	1162477138335769203	299982579	News	VIDEO: Basejumper springen von Berg in Flugzeug	Mathias Kainz	
2	1237559046481705676	299798977	News	Postler als Geisterfahrer in der Rettungsgasse...	Thomas Sendlhofer	
3	1237559046481705676	299831571	News	Chinesen investieren 3 Mrd. Euro in Osteuropa	Peter Temel	
4	1237559046481705676	299825001	News	Auslieferung: Mutmaßlicher Sechsfach-Mörder ha...	Michaela Reibenwein	

Let's have a look at the two csv files we just created containing the training and test set. We'll also do a line count of both files to confirm that we have achieved an approximate 90/10 train/test split.

In the next notebook, **Content Based Filtering** we will build a model to recommend an article given information about the current article being read, such as the category, title, author, and publish date.

In [9]:

```
%%bash
wc -l *_set.csv
```

```
  25599 test_set.csv
 232308 training_set.csv
 257907 total
```

In [10]:

```
!head *_set.csv
```

```
==> test_set.csv <==
1162477138335769203,299928807,News,Missbrauchsvorwürfe: Tiroler Skiverband durch
suchte Heim-Protokolle,Mirad Odobasic,574,299918253
1162477138335769203,299982579,News,VIDEO: Basejumper springen von Berg in Flugze
ug,Mathias Kainz,574,299921761
1237559046481705676,299798977,News,Postler als Geisterfahrer in der Rettungsgass
e unterwegs,Thomas Sendlhofer,574,299865757
1237559046481705676,299831571,News,Chinesen investieren 3 Mrd. Euro in Osteurop
a,Peter Temel,574,299828023
1237559046481705676,299825001,News,Auslieferung: Mutmaßlicher Sechsfach-Mörder h
at Flugangst,Michaela Reibenwein,574,299866366
1237559046481705676,299831571,News,Chinesen investieren 3 Mrd. Euro in Osteurop
a,Peter Temel,574,299802565
1905837605298170342,299830996,News,Wie die Schule in der Neuzeit ankommen könnt
e,Martina Salomon,574,299805494
1905837605298170342,299692362,Stars & Kultur,"""Unsere Zivilisation zerbrich
t""",Peter Jarolin,574,299817990
1905837605298170342,299817990,News,Kolumne Anstoß: Das Allerheiligste,Philipp Al
brechtsberger,574,299821998
2051487466804948450,299912151,News,NÖ: Beißender Geruch in der Klasse,Jürgen Zah
rl,574,299824032

==> training_set.csv <==
1038643850985118087,299828023,News,Glyphosat geht in die Verlängerung,Andreas An
zenberger,574,299425707
1083265653486482344,299828023,News,Glyphosat geht in die Verlängerung,Andreas An
zenberger,574,299957318
1090462532616257705,299696307,News,Patient mit Taxi heimgeschickt,Johannes Weich
hart,574,299793337
1090462532616257705,299793337,News,""Sittenwächter"" an See in Niederösterreic
h: Hauptverdächtiger in Haft",,574,299836255
1122220186524713655,299915364,News,Glyphosat-Verlängerung: SPD wirft CDU Vertrau
ensbruch vor,,574,299949290
1122220186524713655,299949290,News,Nationalbank: 35 Tonnen Gold wieder zurück in
Wien,Stefan Berndl,574,299949290
1137730095009895236,299925086,News,Marihuana-Adventkalender findet in Kanada rei
ßenden Absatz,,574,299826775
1137730095009895236,299950903,News,Vienna: OGH bestätigt Zwangsabstieg in 2. Lan
desliga,Stefan Berndl,574,299935266
1225476921202045176,299912085,News,Erster ÖBB-Containerzug nach China unterwegs,
Stefan Hofer,574,299939900
1225476921202045176,299930032,News,Google steigt bei Wiener IT-Firma ein,Stefan
Hofer,574,299939900
```

In [ ]: