article_classifier

December 31, 2020

[1]:
```
import seaborn as sns
from pyspark import SparkContext
from pyspark.sql import SparkSession
```

0.1 Thiết lập môi trường

1. Spark master gồm 2 spark work
2. HDFS 1 namenode và 2 datanode

[2]:
```
ss = SparkSession.Builder() \
     .appName("articles") \
     .master("spark://spark-master:7077") \
     .getOrCreate()
```

[4]:
```
df = ss.read.parquet("hdfs://namenode:9000/data/articles.parquet")
```

[5]:
```
df.show()
```

```
+--------------------+--------------------+--------------------+----------------
----+--------------+------------+--------------------+--------------------+-----
--------------+--------------------+--------+
|                  id|               title|                sapo|
url|        source|pega_cate_id|         title_token|          sapo_token|
content_token|           all_token|   label|
+--------------------+--------------------+--------------------+----------------
----+--------------+------------+--------------------+--------------------+-----
--------------+--------------------+--------+
|6.366651292038185E17|Anh phát hiện 39 …|Ngày 23/10, cảnh
…|http://vnmedia.vn…|    vnmedia.vn|         102|Anh phát_hiện 39 …|Ngày
23/10 , cảnh…|Theo cảnh_sát địa…|anh phát_hiện    …|Thế giới|
|6.368640877043630…|Phát hiện két sắt…|Theo TASS ngày
23…|http://congan.com…| congan.com.vn|         102|Phát_hiện két sắt…|(
CAO ) Theo TASS…|Theo điều_tra ban…|phát_hiện két sắt…|Thế giới|
|6.370351636924579…|Máy bay rơi ở Mex…|Theo Sputnik
ngày…|http://congan.com…| congan.com.vn|         102|Máy_bay rơi ở Mex…|(
CAO ) Theo Sput…|Chiếc máy_bay đan…|máy_bay rơi ở mex…|Thế giới|
| 6.37236419836928E17|Hình ảnh Đệ nhất …|- Giữa lúc tin
```

đồ…|http://vnmedia.vn…| vnmedia.vn| 102|Hình_ảnh Đệ nhất …|-
Giữa lúc tin_đồ…|HÌnh_ảnh Đệ nhất …|hình_ảnh đệ nhất …|Thế giới|
|6.3804225427199 59E17|Thủ lĩnh cao nhất…|(CAO) Hôm
27-10,…|http://congan.com…| congan.com.vn| 102|Thủ_lĩnh cao
nhất…| (CAO) Hôm 27-…|Các nguồn_tin ở S…|thủ_lĩnh cao nhất…|Thế giới|
|6.383146015204556…|Nhóm người di cư …|Theo Daily Mail
n…|http://congan.com…| congan.com.vn| 102|Nhóm người di_cư …|(CAO
) Theo Dail…|Cảnh_sát Pháp sau…|nhóm người di_cư …|Thế giới|
|6.384845724814131…|Quân đội Iraq ban…|Ngày 28/10, quân
…|http://baotintuc…| baotintuc.vn| 102|Quân_đội Iraq ban…|Ngày
28/10 , quân…|Người biểu_tình t…|quân_đội iraq ban…|Thế giới|
|6.385290400436305…|Chân dung "Chị Bì…|Được gọi với cái
…|http://afamily.vn…| AFamily| 102|Chân_dung " Chị_B…|Được
gọi với cái …|39 thi_thể được p…|chân_dung chị_b…|Thế giới|
|6.385365374207508…|Cuộc sống khổ cực…|Khi Li Hua nộp
14…|http://danviet.vn…| danviet.vn| 102|Cuộc_sống khổ_cực…|Khi
Li_Hua nộp 14…|Trong số các nạn_…|cuộc_sống khổ_cực…|Thế giới|
|6.385404403925893…|Tài xế container …|Người lái xe
cont…|https://vtc.vn/ta…| vtc.vn| 102|Tài_xế container …|(
VTC News) - Ng…|(VTC News) - Ng…|tài_xế container …|Thế giới|
|6.385850645113978…|Một góc nhìn về t…|Luật sư Hoàng
Duy…|https://nhandan.c…|nhandan.com.vn| 102|Một góc nhìn về
t…|Luật_sư Hoàng_Duy…|Trong bài viết gử…|một góc nhìn về t…|Thế giới|
|6.385851275895357…|Đảng cầm quyền ở …|Roi-tơ dẫn thông
…|https://nhandan.c…|nhandan.com.vn| 102|Đảng cầm_quyền ở …|Roi -
tơ dẫn thôn…|Theo CNE , Tổng_t…|đảng cầm_quyền ở …|Thế giới|
|6.385851275937300…| Cơ hội để thay đổi|Đợt biểu tình
kéo…|https://nhandan.c…|nhandan.com.vn| 102| Cơ_hội để thay_đổi|Đợt
biểu_tình kéo…|Khởi_phát từ thủ_…|cơ_hội để thay_đổ…|Thế giới|
|6.385851487288279E17|Cô-lôm-bi-a: Rơi …|Không quân Cô-
lôm…|https://nhandan.c…|nhandan.com.vn| 102|Cô - lôm - bi -
a…|Không_quân Cô - l…|Trong thông_cáo ,…|cô lôm bi a…|Thế giới|
|6.385851487288279E17|Vì một châu lục k…|Những vụ tiến
côn…|https://nhandan.c…|nhandan.com.vn| 102|Vì một châu_lục
k…|Những vụ tiến_côn…|Trong bối_cảnh xu…|vì một châu_lục k…|Thế giới|
|6.386141457978736…|[Video] Cháy rừng…|Cháy rừng tại
ban…|https://www.vietn…|vietnamplus.vn| 102|[Video] Cháy
rừ…|Cháy rừng tại ban…|Ngày 27/10 , cháy…| video cháy rừ…|Thế giới|
|6.386148180357529…|Khung cảnh tan ho…|Cuộc săn lùng kẻ
…|https://www.vietn…|vietnamplus.vn| 102|Khung_cảnh tan_ho…|Cuộc
săn_lùng kẻ …|Cuộc săn_lùng kẻ …|khung_cảnh tan_ho…|Thế giới|
|6.386167924372521E17|[Video] Biểu tình…|Các cuộc biểu
tìn…|https://www.vietn…|vietnamplus.vn| 102|[Video] Biểu_tì…|Các
cuộc biểu_tìn…|Các cuộc biểu_tìn…| video biểu_tì…|Thế giới|
|6.386198699574804…|Tổng thống Chile …|Tổng thống Chile
…|https://www.vietn…|vietnamplus.vn| 102|Tổng_thống
Chile_…|Tổng_thống Chile_…|Tổng_thống Chile_…|tổng_thống chile_…|Thế
giới|

```
|6.386205023830261…|Mượn tay người Ku…|Giới phân tích
ch…|https://vtc.vn/mu…|        vtc.vn|         102|Mượn tay người Ku…|(
VTC News ) - Gi…|( VTC News ) - Gi…|mượn tay người ku…|Thế giới|
+-------------------+-------------------+-------------------+---------------
----+-------------+-----------+-------------------+-------------------+-----
--------------+--------------------+--------+
only showing top 20 rows
```

[6]:
```
df_new = df.drop('id',
                 'title',
                 'sapo',
                 'url',
                 'source',
                 'title_token',
                 'sapo_token',
                 'content_token')
```

[7]:
```
df_new = df_new.dropna()
```

[8]:
```
df_new.show(5)
```

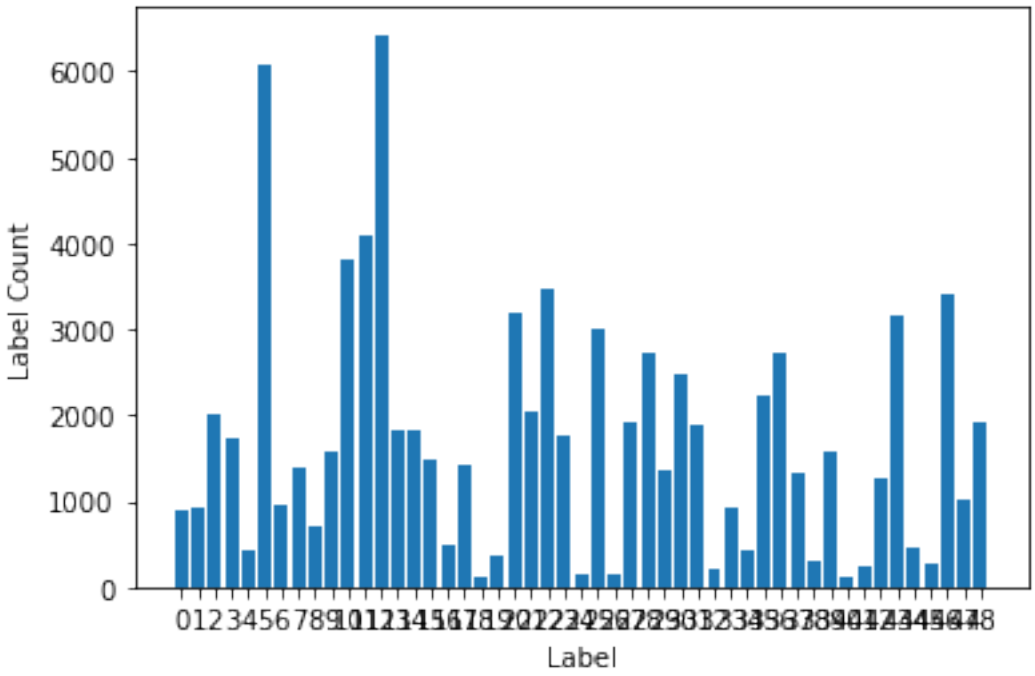
```
+-----------+--------------------+--------+
|pega_cate_id|           all_token|   label|
+-----------+--------------------+--------+
|         102|anh phát_hiện    …|Thế giới|
|         102|phát_hiện két sắt…|Thế giới|
|         102|máy_bay rơi ở mex…|Thế giới|
|         102|hình_ảnh đệ nhất …|Thế giới|
|         102|thủ_lĩnh cao nhất…|Thế giới|
+-----------+--------------------+--------+
only showing top 5 rows
```

Visualize data

[10]:
```
import numpy as np
import matplotlib.pyplot as plt
print(df.groupBy('label').count())
l = df.groupBy('label').count().collect()
x = list(zip(*l))[0]
y = list(zip(*l))[1]
x_pos = np.arange(len(x))
# slope, intercept = np.polyfit(x_pos, y, 1)
plt.bar(x_pos, y,align='center')
plt.xticks(x_pos)
plt.ylabel('Label Count')
plt.xlabel('Label')
```

```
plt.show()
```

```
DataFrame[label: string, count: bigint]
```



Tokenize text in all_token columns

[8]:
```
from pyspark.ml.feature import Tokenizer, CountVectorizer
tkn = Tokenizer().setInputCol("all_token").setOutputCol("content_tokenized")
train_df = tkn.transform(df_new)
# train_df = tokenized.drop('title_token', 'sapo_token', 'content_token')
```

[9]:
```
train_df.show(5)
```

```
+-----------+--------------------+--------+--------------------+
|pega_cate_id|           all_token|   label|   content_tokenized|
+-----------+--------------------+--------+--------------------+
|         102|anh phát_hiện    …|Thế giới|[anh, phát_hiện, …|
|         102|phát_hiện két sắt…|Thế giới|[phát_hiện, két, …|
|         102|máy_bay rơi ở mex…|Thế giới|[máy_bay, rơi, ở,…|
|         102|hình_ảnh đệ nhất …|Thế giới|[hình_ảnh, đệ, nh…|
|         102|thủ_lĩnh cao nhất…|Thế giới|[thủ_lĩnh, cao, n…|
+-----------+--------------------+--------+--------------------+
only showing top 5 rows
```

TF-IDF

[10]:
```
from pyspark.ml.feature import HashingTF, IDF
from pyspark.ml.feature import RegexTokenizer, StopWordsRemover, StringIndexer
from pyspark.ml import Pipeline
```

[11]:
```
label_stringIdx = StringIndexer(inputCol = "label", outputCol = "label_id")
hashingTF = HashingTF(inputCol="content_tokenized", outputCol="rawFeatures",␣
 ↪numFeatures=10000)
idf = IDF(inputCol="rawFeatures", outputCol="features", minDocFreq=5)␣
 ↪#minDocFreq: remove sparse terms
pipeline = Pipeline(stages=[hashingTF, idf, label_stringIdx])
pipelineFit = pipeline.fit(train_df)
dataset = pipelineFit.transform(train_df)
```

[12]:
```
dataset = dataset.withColumnRenamed("label", "label_name")
dataset = dataset.withColumnRenamed("label_id", "label")
```

[13]:
```
dataset.show(5)
```

```
+-----------+--------------------+----------+--------------------+-------------
-------+--------------------+-----+
|pega_cate_id|           all_token|label_name|   content_tokenized|
rawFeatures|            features|label|
+-----------+--------------------+----------+--------------------+-------------
-------+--------------------+-----+
|         102|anh phát_hiện    …|  Thế giới|[anh, phát_hiện,
…|(10000,[44,277,57…|(10000,[44,277,57…|  1.0|
|         102|phát_hiện két sắt…|  Thế giới|[phát_hiện, két,
…|(10000,[54,63,250…|(10000,[54,63,250…|  1.0|
|         102|máy_bay rơi ở mex…|  Thế giới|[máy_bay, rơi,
ở,…|(10000,[63,378,49…|(10000,[63,378,49…|  1.0|
|         102|hình_ảnh đệ nhất …|  Thế giới|[hình_ảnh, đệ,
nh…|(10000,[37,43,52,…|(10000,[37,43,52,…|  1.0|
|         102|thủ_lĩnh cao nhất…|  Thế giới|[thủ_lĩnh, cao,
n…|(10000,[63,70,133…|(10000,[63,70,133…|  1.0|
+-----------+--------------------+----------+--------------------+-------------
-------+--------------------+-----+
only showing top 5 rows
```

Chia dữ liệu thành 2 tập train và test với tỷ lệ 80-20

[14]:
```
df_train, df_test = dataset.randomSplit([0.8, 0.2])
```

[15]:
```
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
```

[16]:
```
evaluator = MulticlassClassificationEvaluator(labelCol='label',
                                              metricName='accuracy')
```

Huấn luyện mô hình

[17]:
```
lr = LogisticRegression(regParam=0.3, elasticNetParam=0)
lrModel = lr.fit(df_train)
pred = lrModel.transform(df_test)
```

Đánh giá mô hình

[18]:
```
evaluator.evaluate(pred)
```

[18]:
```
0.821733459805549
```

0.2 Thử nghiệm Spark Streaming

[4]:
```
static = ss.read.parquet("hdfs://namenode:9000/data/Dstream/f1.parquet")
dataSchema = static.schema
```

[5]:
```
static.printSchema()
```

```
root
 |-- id: long (nullable = true)
 |-- title_token: string (nullable = true)
 |-- sapo_token: string (nullable = true)
 |-- content_token: string (nullable = true)
 |-- tag_token: double (nullable = true)
 |-- title_postag: string (nullable = true)
 |-- sapo_postag: string (nullable = true)
 |-- content_postag: string (nullable = true)
 |-- tag_postag: double (nullable = true)
 |-- title_ner: string (nullable = true)
 |-- sapo_ner: string (nullable = true)
 |-- content_ner: string (nullable = true)
 |-- tag_ner: double (nullable = true)
 |-- update_time: string (nullable = true)
 |-- source_tracking: string (nullable = true)
 |-- __index_level_0__: long (nullable = true)
```

[6]:
```
static.show(4)
```

```
+-----------------+-------------------+-------------------+-----------------
--+---------+--------------------+--------------------+--------------------+----
------+-------------------+-------------------+-------------------+-------+--
----------------+---------------+-----------------+
```

```
|                id|         title_token|          sapo_token|
content_token|tag_token|        title_postag|         sapo_postag|
content_postag|tag_postag|           title_ner|            sapo_ner|
content_ner|tag_ner|        update_time|source_tracking|__index_level_0__|
+------------------+--------------------+--------------------+------------------
--+---------+--------------------+--------------------+--------------------+----
------+--------------------+--------------------+--------------------+-------+--
----------------+---------------+-----------------+
|783112665224601602|Huyền My hội_ngộ …|Show thời_trang m…|Tối_ngày 29/11 ở
…|     null|Huyền/Np My/Np hộ…|Show/Nb thời_tran…|Tối_ngày/N 29/11/…|
null|Huyền_My/PER Hồng…|Huyền_My/PER Hồng…|Hà_Nội/LOC Fashio…|
null|2020-12-01 00:00:01|    newsdb.news|                 0|
|783116112820854786|Chủ_tịch Quốc_hội…|Tối 30-11 , tại Q…|Uỷ_viên
Bộ_Chính_…|     null|Chủ_tịch/N Quốc_h…|Tối/N 30-11/M ,/C…|Uỷ_viên/N
Bộ_Chín…|      null|Quốc_hội/ORG Danh…|Quảng_trường_Hồ_C…|Bộ_Chính_trị/ORG
…|   null|2020-12-01 00:00:01|    newsdb.news|                 1|
|783118176766550016|Nhìn lại thành_tự…|T1 có những người…|Với việc có cho
m…|     null|Nhìn/V lại/R thàn…|T1/Ny có/V những/…|Với/E việc/N có/V…|
null|  Canna/PER Zeus/PER|              T1/PRO|Zeus/PER Canna/PE…|
null|2020-12-01 00:00:01|    newsdb.news|                 2|
|783118586843652109|Triệt_phá ổ nhóm …|Ngày 30/11 , Phòn…|Ngày 30/11 ,
Phòn…|     null|Triệt_phá/V ổ/N n…|Ngày/N 30/11/M ,/…|Ngày/N 30/11/M
,/…|      null|                null|Phòng_Cảnh_sát_Hì…|Phòng_Cảnh_sát_Hì…|
null|2020-12-01 00:00:01|    newsdb.news|                 3|
+------------------+--------------------+--------------------+------------------
--+---------+--------------------+--------------------+--------------------+----
------+--------------------+--------------------+--------------------+-------+--
----------------+---------------+-----------------+
only showing top 4 rows
```

[7]:
```
import requests
from pyvi import ViTokenizer
```

[8]:
```
streaming = ss.readStream.schema(dataSchema).option("maxFilesPerTrigger", 1)\
.parquet("hdfs://namenode:9000/data/Dstream/")
```

[9]:
```
activityCounts = streaming.groupBy("content_token").count()
```

[10]:
```
ss.conf.set("spark.sql.shuffle.partitions", 5)
```

[11]:
```
activityQuery = activityCounts.writeStream.queryName("test")\
.format("memory").outputMode("complete")\
.start()
```

[12]:
```
activityQuery.awaitTermination(timeout=1)
```

[12]: False

[13]:
```
from time import sleep
last_count = 0
for x in range(5):
    """
    TODO: xử lý API
    """
    print("Time step: ", x+1)
    df_q = ss.sql("SELECT content_token FROM test ")
    current_count = df_q.count()
    list_contents = df_q.select('content_token').rdd.flatMap(lambda x: x).
 ↪collect()
    list_contents_update = [list_contents[i] for i in range(last_count,␣
 ↪current_count)]
    print("current records: ", current_count)
    print("new records: ", current_count - last_count)
    text = '#$'.join(list_contents_update)
    ids = '#$'.join([str(i) for i in range(last_count, current_count)])
    last_count = current_count
    re = requests.post("http://localhost:8000/TextClassification", data={"text":
 ↪text, 'id':ids})
    print("processed records: ", re.json()['size'])
    print("======================================")
#     sleep(1)
```

```
Time step:  1
current records:  0
new records:  0
processed records:  1
======================================
Time step:  2
current records:  674
new records:  674
processed records:  675
======================================
Time step:  3
current records:  1250
new records:  576
processed records:  1251
======================================
Time step:  4
current records:  1714
new records:  464
processed records:  1715
======================================
Time step:  5
```

```
current records:  1984
new records:  270
processed records:  1985
=====================================
```

[14]:
```
activityQuery.stop()
```

[ ]:
```

```