Reddit DataFrame Schema

This is the schema for the Reddit comments when reading a Spark DataFrame:

```python
from pyspark.sql import types

comments_schema = types.StructType([
    types.StructField('archived', types.BooleanType()),
    types.StructField('author', types.StringType()),
    types.StructField('author_flair_css_class', types.StringType()),
    types.StructField('author_flair_text', types.StringType()),
    types.StructField('body', types.StringType()),
    types.StructField('controversiality', types.LongType()),
    types.StructField('created_utc', types.StringType()),
    types.StructField('distinguished', types.StringType()),
    types.StructField('downs', types.LongType()),
    types.StructField('edited', types.StringType()),
    types.StructField('gilded', types.LongType()),
    types.StructField('id', types.StringType()),
    types.StructField('link_id', types.StringType()),
    types.StructField('name', types.StringType()),
    types.StructField('parent_id', types.StringType()),
    types.StructField('retrieved_on', types.LongType()),
    types.StructField('score', types.LongType()),
    types.StructField('score_hidden', types.BooleanType()),
    types.StructField('subreddit', types.StringType()),
    types.StructField('subreddit_id', types.StringType()),
    types.StructField('ups', types.LongType()),
    #types.StructField('year', types.IntegerType()),
    #types.StructField('month', types.IntegerType()),
])
```

You can then read the data with:

```python
comments = spark.read.json(inputs, schema=comments_schema)
```

Updated Fri Aug. 25 2023, 15:55 by ggbaker.

CourSys is hosted by SFU Computing Science.

Admission
Programs
Learning
Research
Community
About

Maps + directions
Library
Academic Calendar
Road Report
Give to SFU

CONNECT WITH US
Facebook
Twitter
YouTube

CONTACT US
Simon Fraser University
8888 University Drive
Burnaby, B.C.
Canada V5A 1S6

Terms and conditions
© Simon Fraser University