```
[ ] import pandas as pd
    metadf = pd.read_csv("drive/Shareddrives/810/metadata.csv")
```

```
/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2882: DtypeWarning: Columns (1,4,5,6,13,14,15,16) have mixed types.Specify dtype option on import or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

```
[ ] !pip3 install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 33 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 49.2 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=281853642 sha256=3f466d65faac0d64ba586b69668650367acff3cb5c218cf3e616e5fe741f31fe
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e92efd1e1d5334db05f2e83bcef74f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1
```

```
[ ] from pyspark.sql import SparkSession
```

```
MAX_MEMORY = "20g"
spark = SparkSession \
    .builder \
    .appName("EcommerceAnalysis") \
    .config("spark.executor.memory", MAX_MEMORY) \
    .config("spark.driver.memory", MAX_MEMORY) \
    .config("spark.ui.port", "4040") \
    .master("local[*]") \
    .getOrCreate()
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.2.1
Master
local[*]
AppName
EcommerceAnalysis

```
[ ] df = spark \
      .read \
      .option("header", "true") \
      .option("inferSchema", "true") \
      .csv(f"/content/drive/Shareddrives/810/metadata.csv")
```

```
[ ] df.write \
      .mode("overwrite") \
      .option("compression", "gzip")\
      .parquet("kag_pa")
```

```
[ ] df = spark.read.parquet("kag_pa").persist()
```

```
df.show()
```

```
+--------+----+--------+--------------------+--------------------+------+----------+-------+--------------------+------------+--------------------+--------------------+----------+
|cord_uid| sha|source_x|               title|                 doi|pmcid|pubmed_id|license|            abstract|publish_time|             authors|             journal|    mag_id|
+--------+----+--------+--------------------+--------------------+------+----------+-------+--------------------+------------+--------------------+--------------------+----------+
|zjsgzqgz|null| Medline|Practical home-ba...|10.23736/s0022-47...|  null|  33092335|    unk|                null|  2020-10-22|Zhang, Shiyan; Yu...|The Journal of sp...|      null|
|pjjxf5rp|null| Medline|Patients as partn...|   10.1002/anr3.12086|  null|  33392510|    unk|Insitu simulation...|        2020|Cegielski, D; Dar...| Anaesthesia reports|      null|
|fltja2n3|null| Medline|Optimising specia...|10.1177/1357633x2...|  null|  34726995|    unk|Telehealth can ef...|  2021-12-01|Haydon, Helen M; ...|Journal of teleme...|      null|
|ru7erkbm|null| Medline|Predictors of Mor...|    10.1002/jmv.26750|  null|  33350488|    unk|INTRODUCTION The ...|  2020-12-22|Merugu, Ganesh Pr...|Journal of medica...|      null|
|3yi5wx07|null| Medline|Invited Review: A...|   10.1111/nan.12529|  null|  30394574| cc-by|Despite more than...|        2019|Tzioras, M; Davie...|Neuropathology an...|      null|
|jtkct565|null| Medline|Eye health and qu...|10.1136/bmjopen-2...|  null|  32868362| cc-by|INTRODUCTION Visi...|  2020-08-30|Assi, Lama; Rosma...|           BMJ open|      null|
|49cbl5b5|null| Medline|Feasibility of a ...|10.1109/embc.2016...|  null|  28227501|    unk|Development of a ...|        2016|Opie, Nicholas L;...|Conference procee...|      null|
|akxaru5a|null| Medline|Inpatient Transit...|10.1089/dia.2020....|  null|  32396395|    unk|"Introduction: Du...| the diabetes ser...| especially in th...|          2020-05-12|Jones, Morgan S; ...|
|fg0st4lg|null| Medline|The Power of Peer...|    10.5688/ajpe8471|  null|  34283744|    unk|Regardless of a f...|  2021-02-01|Biehle, Lauren; C...|American journal ...|      null|
|xq3umt6s|null| Medline|Physiotherapy stu...|10.1186/s12909-02...|  null|  33243213| cc-by|BACKGROUND Simula...|  2020-11-26|Pritchard, Shane ...|BMC medical educa...|      null|
|1f8yx3ef|null| Medline|What makes one fe...|  10.1111/bjhp.12501|  null|  33400353|    unk|OBJECTIVE Health ...|  2021-01-05|Merino, M Dolores...|British journal o...|      null|
|wn97pah8|null| Medline|TMPRSS2 activates...|10.1128/jvi.03372-12|  null|  23536651|    unk|Infection with hu...|        2013|Bertram, Stephani...| Journal of virology|      null|
|fir831nd|null| Medline|Is saliva a relia...| 10.17219/dmp/132515|  null|  34043887|    unk|"This review aime...|""SARS-CoV-2""|""2019-nCoV""|""oral fluid""|""saliva""|
|36c5468c|null| Medline|Posttraumatic Str...|10.1016/j.biopsyc...|  null|  32709416|    unk|Posttraumatic str...|  2020-06-10|Cisler, Josh M; H...|Biological psychi...|      null|
|emn93fql|null| Medline|Upregulation of A...|10.1177/0145561132...|  null|  34130518|    unk|                null|  2021-06-15|Jain, Amit; Lampe...|Ear, nose, & thro...|      null|
|22v6aiyq|null| Medline|Willingness of in...|10.1017/s104902 3x...|  null|  25007172|    unk|BACKGROUND An ear...|        2014|Shenhar, Gilead; ...|Prehospital and d...|      null|
|3egab2ld|null| Medline|Preclinical and c...|10.23736/s0021-95...|  null|  31089086|    unk|BACKGROUND Synthe...|        2019|Pellenc, Quentin;...|The Journal of ca...|      null|
|7mkli8s3|null| Medline|Predicting asthma...|10.3399/bjgp.2020...|  null|  34133316|    unk|"BACKGROUND There...|  2021-06-15|Noble, Michael; B...|The British journ...|      null|
|q2yaqlc1|null| Medline|   Radical Practice.|10.1177/152483992...|  null|  33295210|    unk|                null|  2021-01-01|      Roe, Kathleen M|Health promotion ...|      null|
|8kaqqlfc|null| Medline|Traditional versu...|10.1007/s00268-01...|  null|  24984879|    unk|BACKGROUND Roboti...|        2014|Tranchart, Hadrie...|World journal of ...|      null|
+--------+----+--------+--------------------+--------------------+------+----------+-------+--------------------+------------+--------------------+--------------------+----------+
only showing top 20 rows
```

```
[ ] df.count()
```

```
902589
```

```
[ ] _df = df.select("pmcid").where("pmcid like '%PMC%' and pmc_json_files != 'null'")
    _df.toPandas().to_csv('pmcidlist.csv')
```

```
[ ] df.registerTempTable("kag")
```

```
/usr/local/lib/python3.7/dist-packages/pyspark/sql/dataframe.py:140: FutureWarning: Deprecated in 2.0, use createOrReplaceTempView instead.
  FutureWarning
```

```
spark.sql("select * from kag limit 5").show()
```

```
+--------+---+--------+--------------------+--------------------+-----+--------+-------+--------------------+------------+--------------------+--------------------+------+---------------+----+
|cord_uid|sha|source_x|               title|                 doi|pmcid|pubmed_id|license|            abstract|publish_time|             authors|             journal|mag_id|who_covidence_id|arxi|
+--------+---+--------+--------------------+--------------------+-----+--------+-------+--------------------+------------+--------------------+--------------------+------+---------------+----+
|zjsgzqgz|null| Medline|Practical home-ba...|10.23736/s0022-47...| null| 33092335|    unk|                null|  2020-10-22|Zhang, Shiyan; Yu...|The Journal of sp...|  null|           null|
|pjjxf5rg|null| Medline|Patients as partn...|  10.1002/anr3.12086| null| 33392510|    unk|Insitu simulation...|        2020|Cegielski, D; Dar...| Anaesthesia reports|  null|           null|
|fltja2n3|null| Medline|Optimising specia...|10.1177/1357633x2...| null| 34726995|    unk|Telehealth can ef...|  2021-12-01|Haydon, Helen M; ...|Journal of teleme...|  null|           null|
|ru7erkbm|null| Medline|Predictors of Mor...|   10.1002/jmv.26750| null| 33350488|    unk|INTRODUCTION The ...|  2020-12-22|Merugu, Ganesh Pr...|Journal of medica...|  null|           null|
|3yi5wx07|null| Medline|Invited Review: A...|   10.1111/nan.12529| null| 30394574|  cc-by|Despite more than...|        2019|Tzioras, M; Davie...|Neuropathology an...|  null|           null|
+--------+---+--------+--------------------+--------------------+-----+--------+-------+--------------------+------------+--------------------+--------------------+------+---------------+----+
```

```
spark.sql("""

select count(pmcid) from kag where pmcid like 'PMC%'

""").show()
```

```
+------------+
|count(pmcid)|
+------------+
|      325252|
+------------+
```

```
spark.sql("""

select count(pmcid) from kag where pmcid like 'PMC%' and publish_time>'2019-12-01'

""").show()
```

```
+------------+
|count(pmcid)|
+------------+
|      273983|
+------------+
```

```
spark.sql("""

select pmcid, publish_time, title, authors, abstract from kag where pmcid like 'PMC%' and publish_time>='2019-12-01'
order by publish_time

""").show()
```

```
+----------+------------+--------------------+--------------------+--------------------+
|     pmcid|publish_time|               title|             authors|            abstract|
+----------+------------+--------------------+--------------------+--------------------+
|PMC7327228|  2019-12-01|Implementing the ...|Seal, Hayley E.; ...|Children with con...|
|PMC8422197|  2019-12-01|Message from Depu...|    Leung, Gabriel M|                null|
|PMC8422199|  2019-12-01|One Hundred Years...|Zhang, Ran; Dong,...|Almost 100 years ...|
|PMC7122531|  2019-12-01|Complication from...|      Han, Duck Jong|Sensitization to ...|
|PMC7247748|  2019-12-01|Immunogenicity of...|Thibau, Arno; Dic...|The current probl...|
|PMC6874263|  2019-12-01|MERS-CoV in Camel...|Farag, Elmoubashe...|We tested samples...|
|PMC7358814|  2019-12-01|2019 NOVEL CORONA...|     Fagbule, O.F.|                null|
|PMC6863388|  2019-12-01|Characterisation ...|Lawson, J.S.; Sym...|The Crandell-Rees...|
|PMC6962603|  2019-12-01|Antiviral immunit...|Singanayagam, Ara...|Patients with fre...|
|PMC7133260|  2019-12-01|Graphical Abstrac...|                null|                null|
|PMC6926486|  2019-12-01|Enzootic patterns...|El-Kafrawy, Sheri...|BACKGROUND: The M...|
|PMC7097340|  2019-12-01|Antiviral effect ...|Ben-Shabat, Shimo...|Viral infections ...|
|PMC6899506|  2019-12-01|A systematic revi...|Dighe, Amy; Jomba...|Human infection w...|
|PMC8422198|  2019-12-01|Morbidity Analysi...|Dong, Shuaibing; ...|What is already k...|
|PMC6874236|  2019-12-01|Half-Life of Afri...|Stoian, Ana M.M.;...|African swine fev...|
|PMC6964808|  2019-12-01|Implementation Sy...|Kim, Amanda J.; T...|Laws are fundamen...|
|PMC7195332|  2019-12-01|Discovery of nove...|Grädler, Ulrich; ...|Abstract Fragment...|
|PMC6874235|  2019-12-01|Middle East Respi...|Zheng, Jian; Hass...|A high percentage...|
|PMC8048528|  2019-12-01|Spatiotemporal re...|Harvey, William T...|Effective control...|
|PMC7126094|  2019-12-01|Synthesis and bio...|Yoon, Ji Hye; Lee...|3-Acyl-2-phenylam...|
+----------+------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Import the title SIA result

```
result = spark \
      .read \
      .option("header", "true") \
      .option("inferSchema", "true") \
      .csv(f"/content/drive/Shareddrives/810/title_result.csv")


result.write \
    .mode("overwrite") \
    .option("compression", "gzip")\
    .parquet("kag_lb")

result = spark.read.parquet("kag_lb").persist()

result.registerTempTable("result")
```

```
/usr/local/lib/python3.7/dist-packages/pyspark/sql/dataframe.py:140: FutureWarning: Deprecated in 2.0, use createOrReplaceTempView instead.
  FutureWarning
```

```
spark.sql("""

select * from result

""").show()
```

```
+---+-----+-----+-----+--------+--------------------+----------+-----+
|_c0|  neg|  neu|  pos|compound|               title|  paper_id|label|
+---+-----+-----+-----+--------+--------------------+----------+-----+
|  0|  0.0|  1.0|  0.0|     0.0|Timing of surgery...|PMC8206995|  0.0|
|  1|  0.0|  1.0|  0.0|     0.0|      Poster Sessions|PMC7111423|  0.0|
|  2|  0.0|  1.0|  0.0|     0.0|Cardiovascular Ac...|PMC7122603|  0.0|
|  3|  0.0|  1.0|  0.0|     0.0|               Posters|PMC7162159|  0.0|
|  4|  0.0|  1.0|  0.0|     0.0|               Posters|PMC7130089|  0.0|
|  5|  0.0|  1.0|  0.0|     0.0|Foot & mouth disease|PMC7111224|  0.0|
|  6|  0.0|  1.0|  0.0|     0.0|SARS-CoV-2 infect...|PMC8652887|  0.0|
|  7|  0.0|0.697|0.303|  0.7269|SARS-CoV-2 vaccin...|PMC7995008|  1.0|
|  8|  0.0|  1.0|  0.0|     0.0|Fusarium: more th...|PMC8379525|  0.0|
|  9|  0.0|  1.0|  0.0|     0.0|Infektionsschutz ...|PMC7152143|  0.0|
| 10|  0.0|  1.0|  0.0|     0.0|Fungal diversity ...|PMC8648402|  0.0|
| 11|0.157|0.843|  0.0| -0.3818|The Japanese Clin...|PMC8304927| -1.0|
| 12|  0.0|  1.0|  0.0|     0.0|Non-neoplastic di...|PMC7339753|  0.0|
| 13|  0.0|  1.0|  0.0|     0.0|Disorders of the ...|PMC7158344|  0.0|
| 14|  0.0|  1.0|  0.0|     0.0|2019 HRS/EHRA/APH...|PMC7223859|  0.0|
| 15|0.111|0.889|  0.0|  -0.128|The proteasome as...|PMC7236745| -1.0|
| 16|  0.0|  1.0|  0.0|     0.0|Enfermedades infe...|PMC7271218|  0.0|
| 17|  0.0|  1.0|  0.0|     0.0|Single-Stranded R...|PMC7158200|  0.0|
| 18|  0.0|  1.0|  0.0|     0.0|The Immunophysiol...|PMC7158304|  0.0|
| 19|  0.0|0.725|0.275|  0.2263|Exploration of th...|PMC7112061|  1.0|
+---+-----+-----+-----+--------+--------------------+----------+-----+
only showing top 20 rows
```

```
spark.sql("""
with merged_df as (
    select pmcid, publish_time, title as md_title, authors, abstract
    from kag
    where pmcid like 'PMC%' and publish_time>='2019-12-01'
    order by kag.publish_time
  )

select pmcid, publish_time, md_title, authors, abstract, neg, neu, pos, compound, label
from merged_df
INNER JOIN result ON merged_df.pmcid=result.paper_id;

""").show()
```

```
+----------+------------+--------------------+--------------------+--------------------+-----+-----+-----+--------+-----+
|     pmcid|publish_time|            md_title|             authors|            abstract|  neg|  neu|  pos|compound|label|
+----------+------------+--------------------+--------------------+--------------------+-----+-----+-----+--------+-----+
|PMC6874235|  2019-12-01|Middle East Respi...|Zheng, Jian; Hass...|A high percentage...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6890879|  2019-12-03|Evaluation of ant...|Peng, Ju-Yi; Horn...|Bacillus lichenif...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6898929|  2019-12-06|Risk factors for ...|Gupta, Ena; Hosse...|BACKGROUND: Clini...|0.244|0.756|  0.0| -0.6705| -1.0|
|PMC6899506|  2019-12-01|A systematic revi...|Dighe, Amy; Jomba...|Human infection w...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6901795|  2019-12-03|Zebrafish TRIM25 ...|Jin, Yilin; Jia, ...|RIG-I-like recept...|  0.0|0.796|0.204|  0.5574|  1.0|
|PMC6902615|  2019-12-09|Spontaneous breat...|Xia, Jingen; Gu, ...|BACKGROUND: The u...|0.316|0.684|  0.0| -0.7184| -1.0|
|PMC6905523|  2019-12-11|IL-4/IL-13 polari...|Rogers, Kai J.; B...|BACKGROUND: Ebola...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6909074|  2019-12-05|Effect of carboni...|Esfandiari, Neda;...|BACKGROUND: Prist...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6910952|  2019-12-13|Structural insigh...|Sato, Yusuke; Tsu...|Npl4 is likely to...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6912106|  2019-12-12|Recombinant Rotav...|Papa, Guido; Vend...|Rotavirus (RV) re...|  0.0|0.859|0.141|  0.4215|  1.0|
|PMC6917592|  2019-12-11|Interferon-Indepe...|Ashley, Caroline ...|The critical role...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6924143|  2019-12-20|Prediction of nov...|Khanna, Varun; Li...|BACKGROUND: Toll-...|  0.0|0.796|0.204|  0.3182|  1.0|
|PMC6925858|  2019-12-21|A review on the e...|Matei, Ioana A.; ...|Anaplasma phagocy...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6928167|  2019-12-23|Herpes simplex vi...|Hraiech, Sami; Bo...|BACKGROUND: Herpe...|0.178|0.822|  0.0| -0.3818| -1.0|
|PMC6934550|  2019-12-27|In Vivo Activity ...|DeWald, Lisa Evan...|During the Ebola ...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6934710|  2019-12-27|A natural polymor...|Ávila-Pérez, Gine...|Zika virus (ZIKV)...|  0.0|0.676|0.324|  0.5859|  1.0|
|PMC6935106|  2019-12-27|Prevalence and ph...|Zhang, Fanfan; Lu...|BACKGROUND: In Ch...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6936066|  2019-12-30|Roles of transfor...|Song, Dongli; Tan...|BACKGROUND: Teloc...|  0.0| 0.86| 0.14|  0.3818|  1.0|
|PMC6939335|  2020-01-02|Dexmedetomidine i...|Nakashima, Tsuyos...|BACKGROUND: Dexme...|0.132|0.711|0.157|   0.128|  1.0|
|PMC6941262|  2020-01-02|Whole genome sequ...|Kamau, Everlyn; O...|BACKGROUND: Human...|0.145|0.855|  0.0| -0.296| -1.0|
+----------+------------+--------------------+--------------------+--------------------+-----+-----+-----+--------+-----+
only showing top 20 rows
```

```
spark.sql("""
with merged_df as (
    select pmcid, publish_time, title as md_title, authors, abstract
    from kag
    where pmcid like 'PMC%' and publish_time>='2019-12-01'
    order by kag.publish_time),

 df as (
  select pmcid, publish_time, md_title, authors, abstract, neg, neu, pos, compound, label
from merged_df
INNER JOIN result ON merged_df.pmcid=result.paper_id)

select  pmcid, EXTRACT(YEAR FROM publish_time) as publish_year, md_title, authors, abstract, neg, neu, pos, compound, label
from df

""").show()
```

```
+----------+------------+--------------------+--------------------+--------------------+-----+-----+-----+--------+-----+
|     pmcid|publish_year|            md_title|             authors|            abstract|  neg|  neu|  pos|compound|label|
+----------+------------+--------------------+--------------------+--------------------+-----+-----+-----+--------+-----+
|PMC6874235|        2019|Middle East Respi...|Zheng, Jian; Hass...|A high percentage...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6890879|        2019|Evaluation of ant...|Peng, Ju-Yi; Horn...|Bacillus lichenif...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6898929|        2019|Risk factors for ...|Gupta, Ena; Hosse...|BACKGROUND: Clini...|0.244|0.756|  0.0| -0.6705| -1.0|
|PMC6899506|        2019|A systematic revi...|Dighe, Amy; Jomba...|Human infection w...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6901795|        2019|Zebrafish TRIM25 ...|Jin, Yilin; Jia, ...|RIG-I-like recept...|  0.0|0.796|0.204|  0.5574|  1.0|
|PMC6902615|        2019|Spontaneous breat...|Xia, Jingen; Gu, ...|BACKGROUND: The u...|0.316|0.684|  0.0| -0.7184| -1.0|
|PMC6905523|        2019|IL-4/IL-13 polari...|Rogers, Kai J.; B...|BACKGROUND: Ebola...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6909074|        2019|Effect of carboni...|Esfandiari, Neda;...|BACKGROUND: Prist...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6910952|        2019|Structural insigh...|Sato, Yusuke; Tsu...|Npl4 is likely to...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6912106|        2019|Recombinant Rotav...|Papa, Guido; Vend...|Rotavirus (RV) re...|  0.0|0.859|0.141|  0.4215|  1.0|
|PMC6917592|        2019|Interferon-Indepe...|Ashley, Caroline ...|The critical role...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6924143|        2019|Prediction of nov...|Khanna, Varun; Li...|BACKGROUND: Toll-...|  0.0|0.796|0.204|  0.3182|  1.0|
|PMC6925858|        2019|A review on the e...|Matei, Ioana A.; ...|Anaplasma phagocy...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6928167|        2019|Herpes simplex vi...|Hraiech, Sami; Bo...|BACKGROUND: Herpe...|0.178|0.822|  0.0| -0.3818| -1.0|
|PMC6934550|        2019|In Vivo Activity ...|DeWald, Lisa Evan...|During the Ebola ...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6934710|        2019|A natural polymor...|Ávila-Pérez, Gine...|Zika virus (ZIKV)...|  0.0|0.676|0.324|  0.5859|  1.0|
|PMC6935106|        2019|Prevalence and ph...|Zhang, Fanfan; Lu...|BACKGROUND: In Ch...|  0.0|  1.0|  0.0|     0.0|  0.0|
|PMC6936066|        2019|Roles of transfor...|Song, Dongli; Tan...|BACKGROUND: Teloc...|  0.0| 0.86| 0.14|  0.3818|  1.0|
|PMC6939335|        2020|Dexmedetomidine i...|Nakashima, Tsuyos...|BACKGROUND: Dexme...|0.132|0.711|0.157|   0.128|  1.0|
|PMC6941262|        2020|Whole genome sequ...|Kamau, Everlyn; O...|BACKGROUND: Human...|0.145|0.855|  0.0| -0.296| -1.0|
+----------+------------+--------------------+--------------------+--------------------+-----+-----+-----+--------+-----+
only showing top 20 rows
```

```
spark.sql("""
with merged_df as (
    select pmcid, publish_time, title as md_title, authors, abstract
    from kag
    where pmcid like 'PMC%' and publish_time>='2020-01-01' and publish_time<='2023-01-01'
    order by kag.publish_time),

 df as (
  select pmcid, publish_time, md_title, authors, abstract, neg, neu, pos, compound, label
from merged_df
INNER JOIN result ON merged_df.pmcid=result.paper_id)

select EXTRACT(YEAR FROM publish_time) as publish_year,avg(compound)
from df
group by publish_year

order by publish_year


""").show()
```

```
+------------+--------------------+
|publish_year|       avg(compound)|
+------------+--------------------+
|        2020|-0.00853388939756...|
|        2021|-0.00409004964988...|
|        2022|0.004802362984040585|
+------------+--------------------+
```

```
spark.stop()
```