Spark Assignment - Articles

I. Extract: Load the data

- Read data as json via spark dataframe

II.Transform: Exploratory data analysis using spark df

- Unique Id count
- Remove the html tags column "Article_Description" and "Full_Article"
- Merge the columns "Heading", "Article_Description" and "Full_Article" separated by space and place the merged text in a new column name "Preprocessed_Text"
- select columns Id Preprocessed_Text; Article_Type; Tonality; Outlets
- new column outlet_tags based on outlet text with .com as website and rest as App
- show df
- GroupBy Article_Type, Tonality and count

III. Load: Save analysis report

- show df, save as files(TXT)

In [18]:
```
# To install Spark
# !pip install spark
```

In [19]:
```
# Importing necessary libraries
import os
import spark
from pyspark import SparkConf,SparkContext
from pyspark.sql import SparkSession, types
from pyspark.sql.types import StringType
from pyspark.sql.functions import regexp_replace, split, explode, struct, col, concat, lit, when
import pandas as pd
```

In [20]:
```
# Creating a spark session
spark = SparkSession.builder.master('local').appName('Json File').getOrCreate()
```

Extract: Load the data

In [21]:
```
# Reading JSON file into a dataframe
df = spark.read.json("articles.json", multiLine=True)
df.printSchema()
```

```
root
 |-- Article.Banner.Image: string (nullable = true)
 |-- Article.Description: string (nullable = true)
 |-- Article_Type: string (nullable = true)
 |-- Full_Article: string (nullable = true)
 |-- Heading: string (nullable = true)
 |-- Id: string (nullable = true)
 |-- Outlets: string (nullable = true)
 |-- Tonality: string (nullable = true)
```

In [22]:
```
# To see the dataframe
df.show()
```

```
+--------------------+--------------------+------------+--------------+--------------------+--------------------+--------+
|Article.Banner.Image| Article.Description|Article_Type|  Full_Article|             Heading|                  Id|Outlets|Tonality|
+--------------------+--------------------+------------+--------------+--------------------+--------------------+--------+
|                    |<p>The helicopter...|  Commercial|<p>The helicop...|A Puzzling Maneuv...|d6995462-5e87-453...|  Essex Caller|Negative|
|                    |<p>A year after t...|  Commercial|<p>A year after t...|Bell Nexus Air ...|8b05e939-a89e-454...|Aviation Week Net...|Positive|
|http://images.tmt...|<p>Bell released ...|  Commercial|<p>Bell released ...|Bell Helicopter S...|697c0490-bceb-425...|      TMTPost|Positive|
|http://www.fredzo...|<p>Bell est une s...|  Commercial|<p>Bell est une s...|BELL DOUILE LA C...|13793578-c11b-414...|    Fredzone|Positive|
|                    |<p>It was still a...|  Commercial|<p>It was still a...|Les premiers reto...|f33c7b11-5f77-4a9...|  FrenchWeb|Positive|
|                    |<p>The LG Signatu...|  Commercial|<p>The LG Signatu...|Highlights of CES...|1422d70c-cf18-42d...|The Daily Star|Positive|
|                    |<p>Le concept Vis...|  Commercial|<p>Le concept Vis...|Le Concept Vision...|f096edd3-13db-4ae...|  Eric Hougue|Positive|
|http://upload.can...|<p>Bell recently ...|  Commercial|<p>Bell recently ...|Bell Company Anno...|f8f917ec-0cb0-4a4...|    CanNews|Positive|
|                    |<p>Bell Helicopte...|  Commercial|<p>Bell Helicopte...|Bell Helicopter M...|1702dec7-7424-469...|  AviationPros|Positive|
|                    |<p>Bell Helicopte...|  Commercial|<p>Bell Helicopte...|Bell Helicopter U...|1d110da1-05c7-467...|  Opulent Club|Positive|
|                    |<p>While the rest...|  Commercial|<p>While the rest...|Bell Nexus The ...|36380a7e-4c5f-4f9...|  OmniGeekEmpire|Positive|
|https://www.sae.o...|<p>Bell, Safran, ...|  Commercial|<p>Bell, Safran, ...|Bell Nexus full-s...|e1902fb9-ba3e-426...|  SAE International|Positive|
|                    |<p>The Vertical F...|  Commercial|<p>The Vertical F...|Bell Reveals Nexu...|b15a8436-45a2-48d...| Electric VTOL News|Positive|
|http://imagesvc.t...|<p>The 2019 CES r...|  Commercial|<p>The 2019 CES r...|Bell Reveals the ...|a13694b7-8fee-49a...|    The Drive|Positive|
|                    |<p>Although not a...|  Commercial|<p>Although not a...|Bell says its sup...|c84d4436-243f-483...|     Driving|Positive|
|https://www.ainon...|<p>Bell returned ...|  Commercial|<p>Bell returned ...|Bell Unveils Nexu...|965334c3-cccc-430...|Aviation Internat...|Positive|
|                    |<p>Bell Helicopte...|  Commercial|<p>Bell Helicopte...|Bell unveils Nexu...|0178e3bb-b033-44f...|      AWuAV|Positive|
|https://www.flyin...|<p>Traffic-hoppin...|  Commercial|<p>Traffic-hoppin...|Bell Unveils Nexu...|e6cabe53-fcb5-437...|  Flipboard|Positive|
|                    |<p>Bell Nexus cou...|  Commercial|<p>Bell Nexus cou...|Bell Flying Car...|6f6396b1-bf06-4cf...|1 News Day|Positive|
|                    |<p>Across the way...|  Commercial|<p>Across the way...|Best of CES: Harl...|a463beea-7ee0-498...|    Electrek|Positive|
+--------------------+--------------------+------------+--------------+--------------------+--------------------+--------+
only showing top 20 rows
```

Transform: Exploratory data analysis using spark df

In [23]:
```
## Q1. Unique ID count
print('Unique number of ids present: ',df.select('Id').distinct().count())
```

```
Unique number of ids present:  4305
```

In [24]:
```
# Changing the name of the column to avoid errors
df = df.withColumnRenamed('Article.Description','Article_Description')
df.printSchema()
```

```
root
 |-- Article.Banner.Image: string (nullable = true)
 |-- Article_Description: string (nullable = true)
 |-- Article_Type: string (nullable = true)
 |-- Full_Article: string (nullable = true)
 |-- Heading: string (nullable = true)
 |-- Id: string (nullable = true)
 |-- Outlets: string (nullable = true)
 |-- Tonality: string (nullable = true)
```

In [25]:
```
## Q2. Function to remove the html tags from column “Article_Description” and “Full_Article”

def remove_html(df,colname):
    df = df.withColumn(colname, regexp_replace(colname,"[<p></p>&nbs]",""))
    return df

df = remove_html(df,'Article_Description')
df = remove_html(df,'Full_Article')

# To see if the tags are removed
df.show()
```

```
+--------------------+--------------------+------------+--------------+--------------------+--------------------+--------+
|Article.Banner.Image| Article_Description|Article_Type|  Full_Article|             Heading|                  Id|Outlets|Tonality|
+--------------------+--------------------+------------+--------------+--------------------+--------------------+--------+
|                    |The helicoter tha...|  Commercial|The helicoter tha...|A Puzzling Maneuv...|d6995462-5e87-453...|  Essex Caller|Negative|
|                    |A year after teai...|  Commercial|A year after teai...|Bell Nexus Air ...|8b05e939-a89e-454...|Aviation Week Net...|Positive|
|http://images.tmt...|Bell released the ...|  Commercial|Bell released the ...|Bell Helicopter S...|697c0490-bceb-425...|      TMTPost|Positive|
|http://www.fredzo...|Bell et ue oceiac...|  Commercial|Bell et ue oceiac...|BELL DOUILE LA C...|13793578-c11b-414...|    Fredzone|Positive|
|                    |It wa till aecdot...|  Commercial|It wa till aecdot...|Les premiers reto...|f33c7b11-5f77-4a9...|  FrenchWeb|Positive|
|                    |The LG Sigature O...|  Commercial|The LG Sigature O...|Highlights of CES...|1422d70c-cf18-42d...|The Daily Star|Positive|
|                    |Le cocet Viio Ura...|  Commercial|Le cocet Viio Ura...|Le Concept Vision...|f096edd3-13db-4ae...|  Eric Hougue|Positive|
|http://upload.can...|Bell recetly aouc...|  Commercial|Bell recetly aouc...|Bell Company Anno...|f8f917ec-0cb0-4a4...|    CanNews|Positive|
|                    |Bell Helicoter, a...|  Commercial|Bell Helicoter, a...|Bell Helicopter M...|1702dec7-7424-469...|  AviationPros|Positive|
|                    |Bell Helicoter ha...|  Commercial|Bell Helicoter ha...|Bell Helicopter U...|1d110da1-05c7-467...|  Opulent Club|Positive|
|                    |While the ret of ...|  Commercial|While the ret of ...|Bell Nexus The ...|36380a7e-4c5f-4f9...|  OmniGeekEmpire|Positive|
|https://www.sae.o...|Bell, Safra, EPS,...|  Commercial|Bell, Safra, EPS,...|Bell Nexus full-s...|e1902fb9-ba3e-426...|  SAE International|Positive|
|                    |The Vertical Flig...|  Commercial|The Vertical Flig...|Bell Reveals Nexu...|b15a8436-45a2-48d...| Electric VTOL News|Positive|
|http://imagesvc.t...|The 2019 CES reve...|  Commercial|The 2019 CES reve...|Bell Reveals the ...|a13694b7-8fee-49a...|    The Drive|Positive|
|                    |Although ot a car...|  Commercial|Although ot a car...|Bell says its sup...|c84d4436-243f-483...|     Driving|Positive|
|https://www.ainon...|Bell retured thi ...|  Commercial|Bell retured thi ...|Bell Unveils Nexu...|965334c3-cccc-430...|Aviation Internat...|Positive|
|                    |Bell Helicoter, a...|  Commercial|Bell Helicoter, a...|Bell unveils Nexu...|0178e3bb-b033-44f...|      AWuAV|Positive|
|https://www.flyin...|Traffic-hoig eVTO...|  Commercial|Traffic-hoig eVTO...|Bell Unveils Nexu...|e6cabe53-fcb5-437...|  Flipboard|Positive|
|                    |Bell Nexu could e...|  Commercial|Bell Nexu could e...|Bell Flying Car...|6f6396b1-bf06-4cf...|1 News Day|Positive|
|                    |Acro the way howe...|  Commercial|Acro the way howe...|Best of CES: Harl...|a463beea-7ee0-498...|    Electrek|Positive|
+--------------------+--------------------+------------+--------------+--------------------+--------------------+--------+
only showing top 20 rows
```

In [26]:
```
## Q3. Merge the columns “Heading”, “Article_Description” and “Full_Article” separated by space and place the merged text in a new column name “Preprocessed_Text”

def col_combine(df, lst, colname):
    return df.withColumn(colname, concat(col(lst[0]), lit(" "), col(lst[1]), lit(" "), col(lst[2])))
lst_cols = ['Heading', 'Article_Description', 'Full_Article']
df = col_combine(df, lst_cols, 'Preprocessed_Text')
df.printSchema()
```

```
root
 |-- Article.Banner.Image: string (nullable = true)
 |-- Article_Description: string (nullable = true)
 |-- Article_Type: string (nullable = true)
 |-- Full_Article: string (nullable = true)
 |-- Heading: string (nullable = true)
 |-- Id: string (nullable = true)
 |-- Outlets: string (nullable = true)
 |-- Tonality: string (nullable = true)
 |-- Preprocessed_Text: string (nullable = true)
```

In [27]:
```
# Checking the first record to see the transformation
df.select('Preprocessed_Text').collect()[0]
```

Out[27]:
```
Row(Preprocessed_Text='A Puzzling Maneuver, Then Freefall: NTSB Report Provides New Details in Southeast Alaska Helicopter Crash That Killed 3 The helicoter that cra hed i Southeat Alaka i late Setember, killig three eole, etered a 500-foot freefall efore droig to a Glacier Bay Natioal Park each, accordig to y the Natioal Traortat io Safety Board. The reliiary NTSB reort released Friday offer o official roale caue. That determiatio wolquo;t e made util ext year at the earliet. The helicoter th at crahed i Southeat Alaka i late Setember, killig three eole, etered a 500-foot freefall efore droig to a Glacier Bay Natioal Park each, accordig to y the Natioal Tr aortatio Safety Board.;The reliiary NTSB reort released Friday offer o official roale caue. That determiatio wolquo;t e made util ext year at the earliet.')
```

In [28]:
```
## Q4. Select columns :  Id, Preprocessed_Text, Article_Type, Tonality outlet
clean_df = df.select('Id', 'Preprocessed_Text', 'Article_Type', 'Tonality', 'Outlets')
clean_df.printSchema()
```

```
root
 |-- Id: string (nullable = true)
 |-- Preprocessed_Text: string (nullable = true)
 |-- Article_Type: string (nullable = true)
 |-- Tonality: string (nullable = true)
 |-- Outlets: string (nullable = true)
```

In [29]:
```
## Q5. Creating new column outlet_tags based on outlet text with .com as website and rest as App
clean_df = clean_df.withColumn('outlet_tags', when(clean_df.Outlets.like('%.%'), \
                                                   lit('website')).otherwise(lit('app')))

## Q6. Show Dataframe
clean_df.show()
```

```
+--------------------+--------------------+------------+--------+--------------------+-----------+
|                  Id|  Preprocessed_Text|Article_Type|Tonality|             Outlets|outlet_tags|
+--------------------+--------------------+------------+--------+--------------------+-----------+
|d6995462-5e87-453...|A Puzzling Maneuv...|  Commercial|Negative|        Essex Caller|        app|
|8b05e939-a89e-454...|Bell Nexus Air ...|  Commercial|Positive|Aviation Week Net...|        app|
|697c0490-bceb-425...|Bell Helicopter S...|  Commercial|Positive|      TMTPost|        app|
|13793578-c11b-414...|BELL DOUILE LA C...|  Commercial|Positive|    Fredzone|        app|
|f33c7b11-5f77-4a9...|Les premiers reto...|  Commercial|Positive|  FrenchWeb|        app|
|1422d70c-cf18-42d...|Highlights of CES...|  Commercial|Positive|The Daily Star|        app|
|f096edd3-13db-4ae...|Le Concept Vision...|  Commercial|Positive|  Eric Hougue|        app|
|f8f917ec-0cb0-4a4...|Bell Company Anno...|  Commercial|Positive|    CanNews|        app|
|1702dec7-7424-469...|Bell Helicopter M...|  Commercial|Positive|  AviationPros|        app|
|1d110da1-05c7-467...|Bell Helicopter U...|  Commercial|Positive|  Opulent Club|        app|
|36380a7e-4c5f-4f9...|Bell Nexus The ...|  Commercial|Positive|  OmniGeekEmpire|        app|
|e1902fb9-ba3e-426...|Bell Nexus full-s...|  Commercial|Positive|  SAE International|        app|
|b15a8436-45a2-48d...|Bell Reveals Nexu...|  Commercial|Positive| Electric VTOL News|        app|
|a13694b7-8fee-49a...|Bell Reveals the ...|  Commercial|Positive|    The Drive|        app|
|c84d4436-243f-483...|Bell says its sup...|  Commercial|Positive|     Driving|        app|
|965334c3-cccc-430...|Bell Unveils Nexu...|  Commercial|Positive|Aviation Internat...|        app|
|0178e3bb-b033-44f...|Bell unveils Nexu...|  Commercial|Positive|      AWuAV|        app|
|e6cabe53-fcb5-437...|Bell Unveils Nexu...|  Commercial|Positive|  Flipboard|        app|
|6f6396b1-bf06-4cf...|Bell Flying Car...|  Commercial|Positive|1 News Day|        app|
|a463beea-7ee0-498...|Best of CES: Harl...|  Commercial|Positive|    Electrek|        app|
+--------------------+--------------------+------------+--------+--------------------+-----------+
only showing top 20 rows
```

In [30]:
```
## Q7. GroupBy Article_Type, Tonality and count
file = clean_df.groupBy('Article_Type', 'Tonality').count()
file.show()
```

```
+------------------+--------+-----+
|      Article_Type|Tonality|count|
+------------------+--------+-----+
|          Training|Positive|    2|
|         Financing| Neutral|    1|
|          Military| Neutral|  178|
|          Military|        |  185|
|          Military|Negative|   53|
|Support & Services|Positive|   26|
|            Others|Positive|   45|
|            Others| Neutral|    7|
|          Training|        |    4|
|          Military|Positive| 1261|
|         Executives|        |    6|
|         Financing|        |    2|
|        Commercial| Neutral|  668|
|        Commercial|        |  235|
|         Financing|Positive|    6|
|        Commercial|Negative|  278|
|        Commercial|Positive| 1889|
|         Executives|Positive|   57|
|         Executives| Neutral|    2|
+------------------+--------+-----+
```

Load: Save analysis report

In [31]:
```
file = file.toPandas()
file.to_csv('Output.txt', index=False)
print('Successfully saved at ',os.getcwd())
```

```
Successfully saved at  D:\M.Sc BDA\Coding\BD3P#\BD3P2_PySpark
```