CSCI 5408

Data Management, Warehousing and Analytics

Assignment 4

Name : Sagarkumar Pankajbhai Vaghasia

CSID : vaghasia

Banner ID : #B00878629

GitLab Link :
https://git.cs.dal.ca/vaghasia/csci5408_s22_sagarkumar_vaghasia_b00878629/-/tree/main/A4

Problem #1

Business Intelligence Reporting using Cognos.

Step 1: Firstly, I have downloaded the weather dataset from Kaggle website[1]. Weather stations are located in the southeast region of Brazil.

Step 2: From the first look of the downloaded dataset, it can be seen that all the tables (central_west, north, northeast, south, southeast) have same columns except two tables (columns_description and stations).

Step 3: Measurable fields and dimensions

Measurable fields and dimensions selected for the given dataset with the reasons are described below:

Field -1: max temperature in previous hour OR min temperature in previous hour.
Dimensions: “region”, “state”, “station”, “date”, “hour”.
This illustrates dicing when we want the max temperature on hourly basis of specific region for specific date. Also, slicing can be applied when we want it for any one dimension only.

Field -2: atmospheric pressure at station height.
Dimensions: “station”, “height”.
This illustrates dicing where we can measure the pressure by station and height.

Field -3: wind speed.
Dimensions: “wind direction”.
This illustrates slicing where we can measure average wind speed for specific direction.

Field -4: radiation.
Dimensions: “region”, “state”, “station”, “date”, “hour”.
We can measure radiation on hourly basis OR for specific date OR region wise OR state wise OR station wise. The slicing can be applied if we measure radiation based on hours only and dicing is applied if we measure radiation based on multiple dimensions like hour, date and state.

Field -5: total precipitation.
Dimensions: “region”, “state”, “station”, “date”, “hour”.
The reason for selecting total precipitation is same as of radiation.

Filed -6: air temperature.
Dimensions: “region”, “state”, “station”, “date”, “hour”.

Step 4: The column names in all the csv files are in different languages. So, I translated it to English by using google translate and columns_description table.
There is one cell in a column named as “abbreviation” which is blank for “region” entry which I assigned as reg.

Step 5: I created IBM Cognos account by following instructions of laboratory tutorial[2][5].

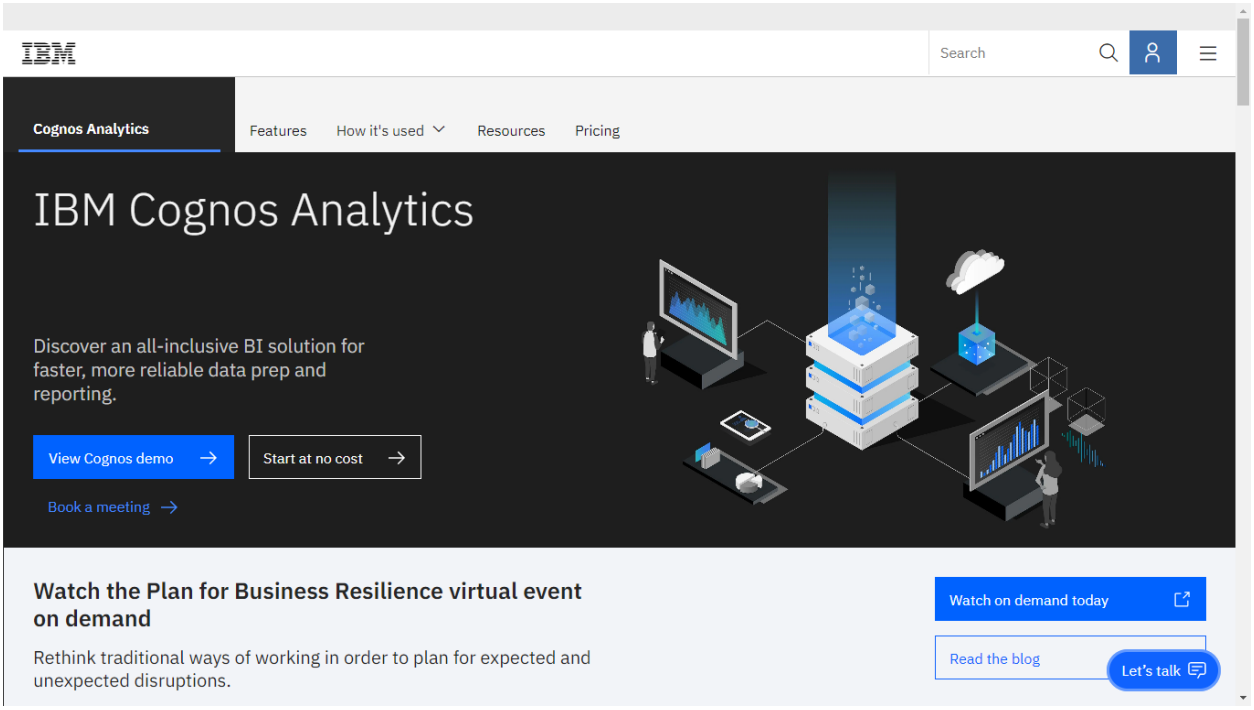


Figure 1: IBM Cognos

I decided to implement star schema using the following four dimensions :

- Date
- Region
- State
- Station

Cognos do not accept files larger than 100 MB . Thus, I manually picked 4000 rows from given CSV files.
Then, I uploaded those files to the Cognos.
I have developed those dimension tables using Microsoft Excel.

Step 6:

- I choose to go with star schema.
- The first dimension I took was date. Based on date we can get answers for the questions like “Total precipitation for specific date” and also we can visualize and analyze it.
- I imported weather table as fact table which I generated from central_west by picking random 4000 rows. I have also imported state, region, station, date tables as dimension tables.

After that, I developed star schema as shown in the below image.

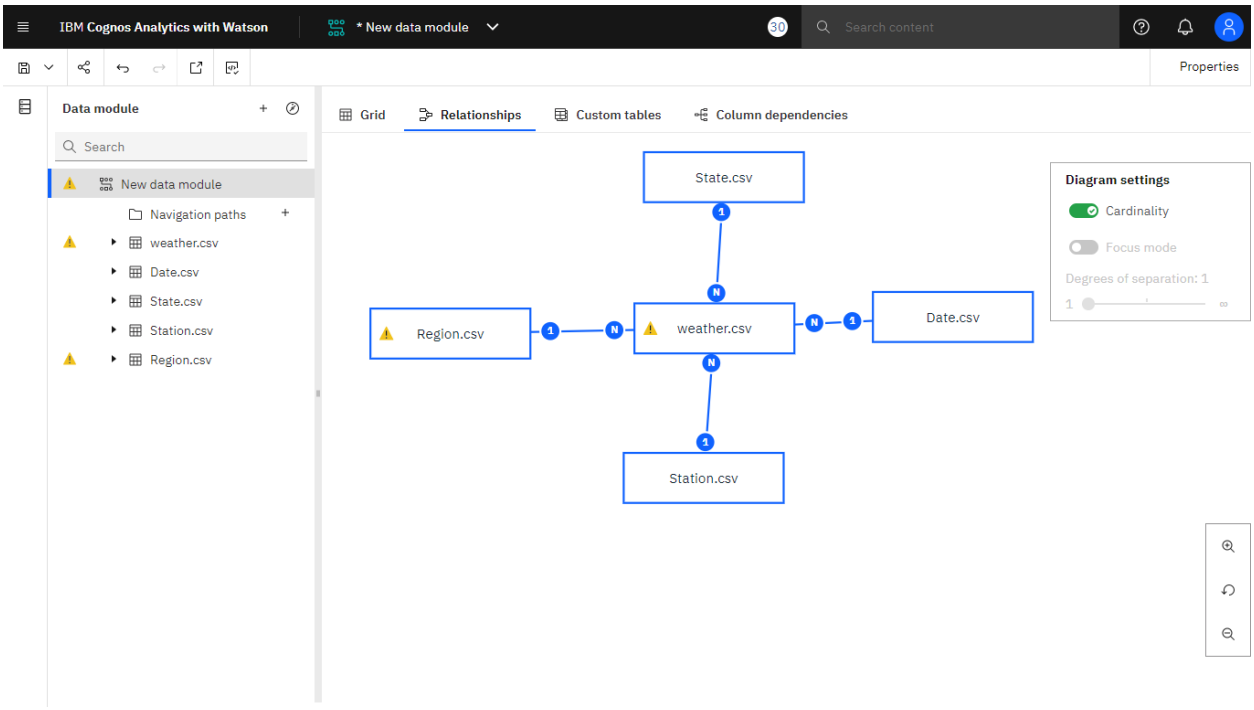


Figure 2: Star Schema

- It can be seen that there are one to many relationships between dimension table and fact table as one value of dimension table is connected with many value of fact table.

Step 7: I have performed visual analysis using IBM Cognos exploration.

Visual Analysis 1

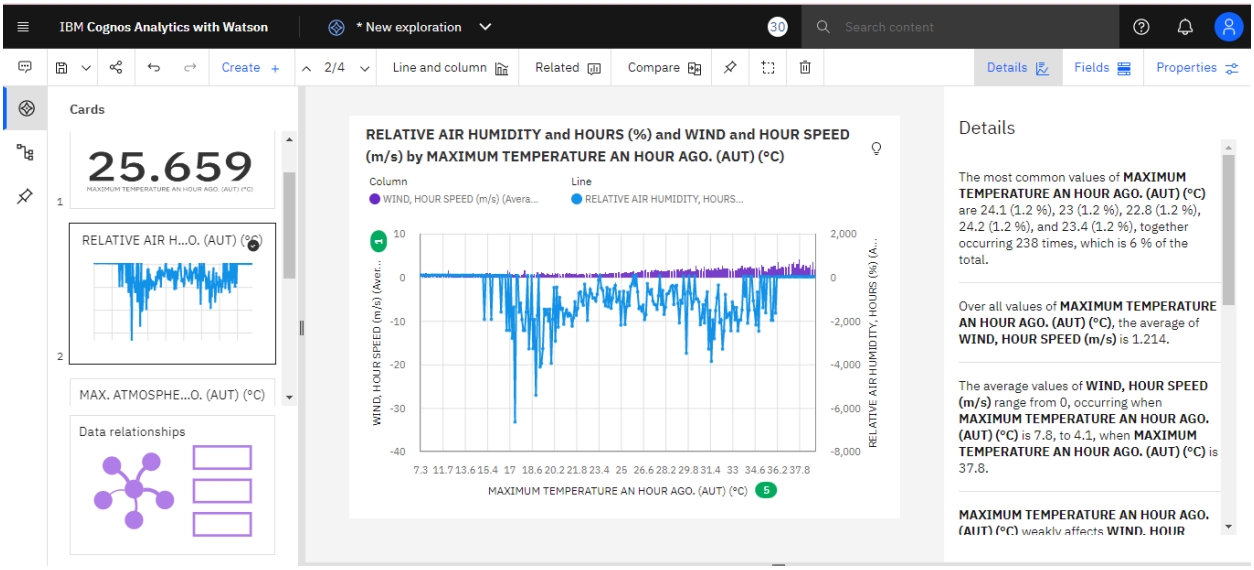


Figure 3: Visual Analysis 1

Visual Analysis 2

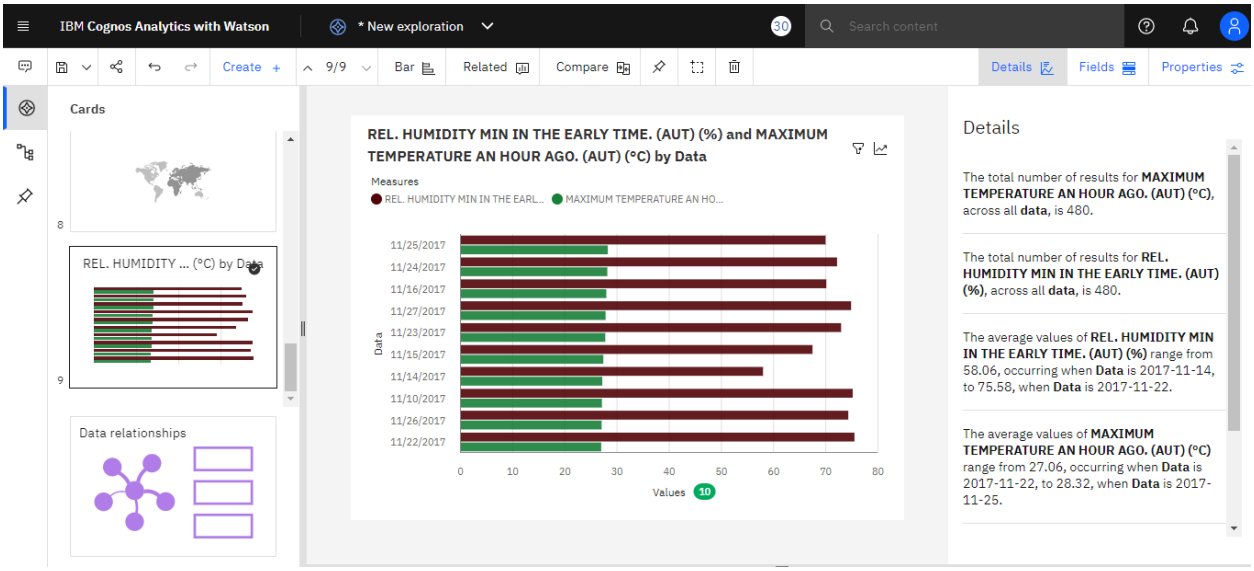


Figure 4: Visual Analysis 2 [Bar-chart]

Visual Analysis 3

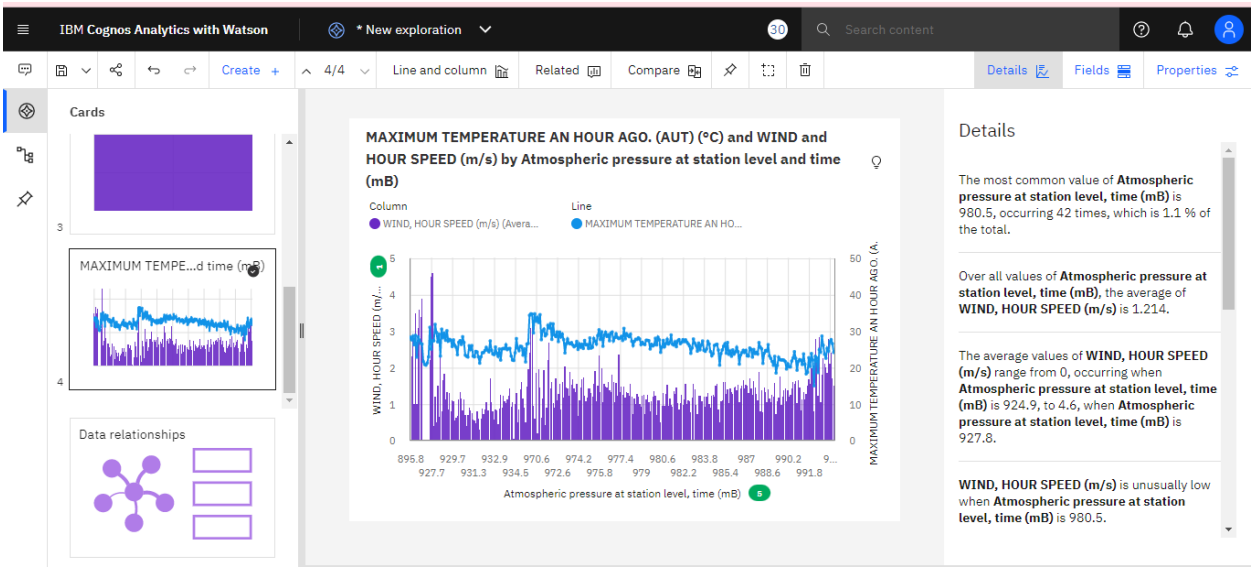


Figure 5: Visual Analysis 3

Visual Analysis 4

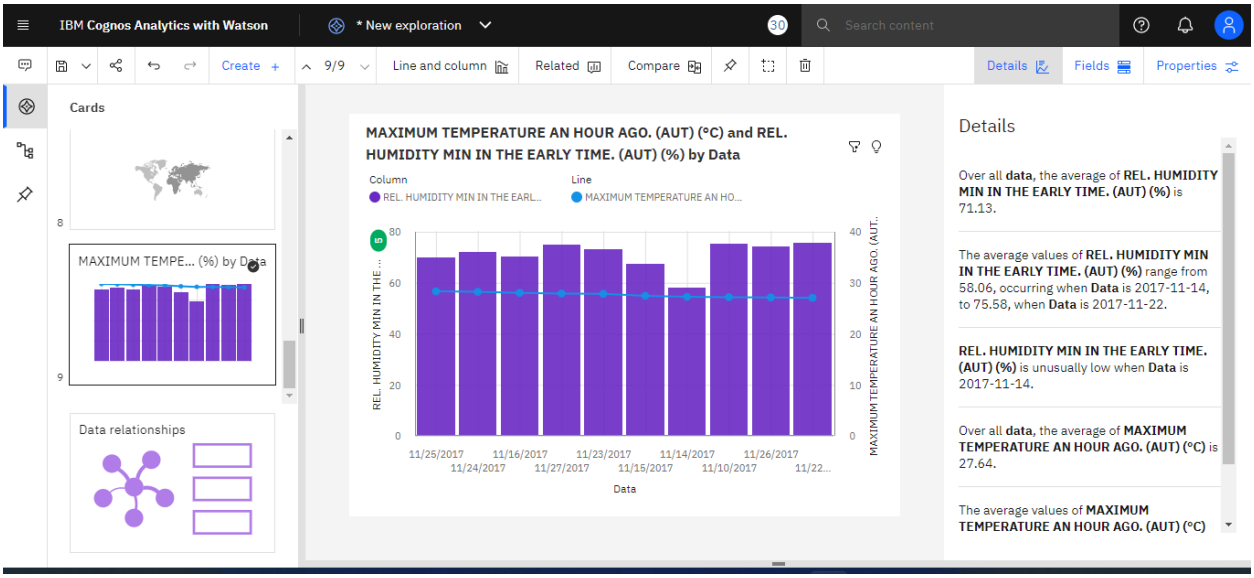


Figure 6: Visual Analysis 4 [Bar-chart]

Visual Analysis 5

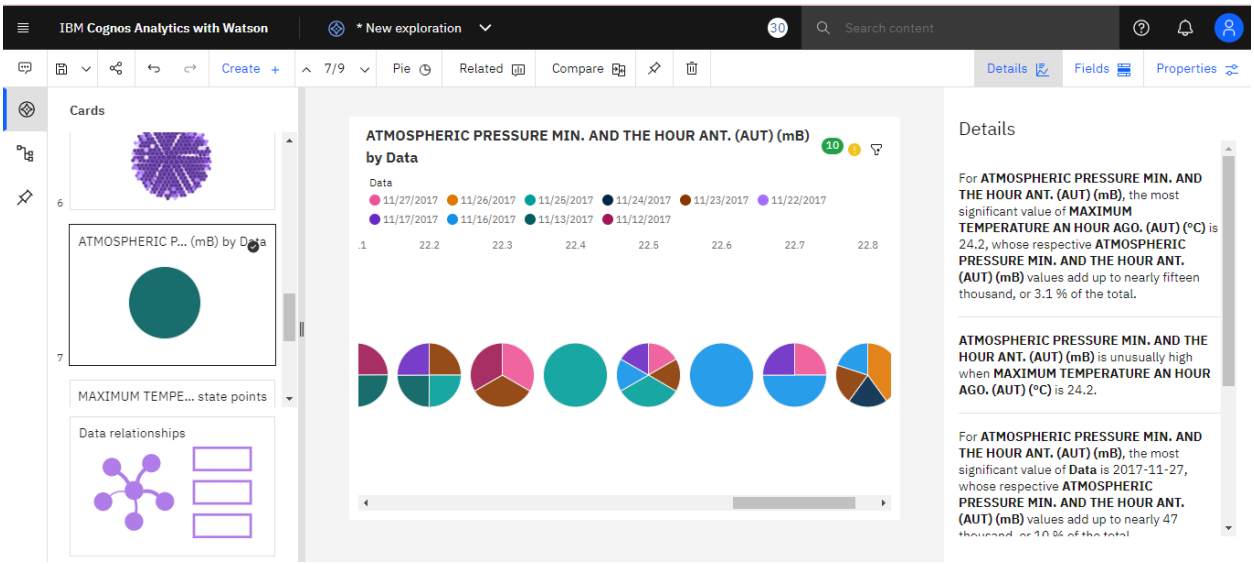


Figure 7: Visual Analysis 5 [Pie-chart]

Problem #2

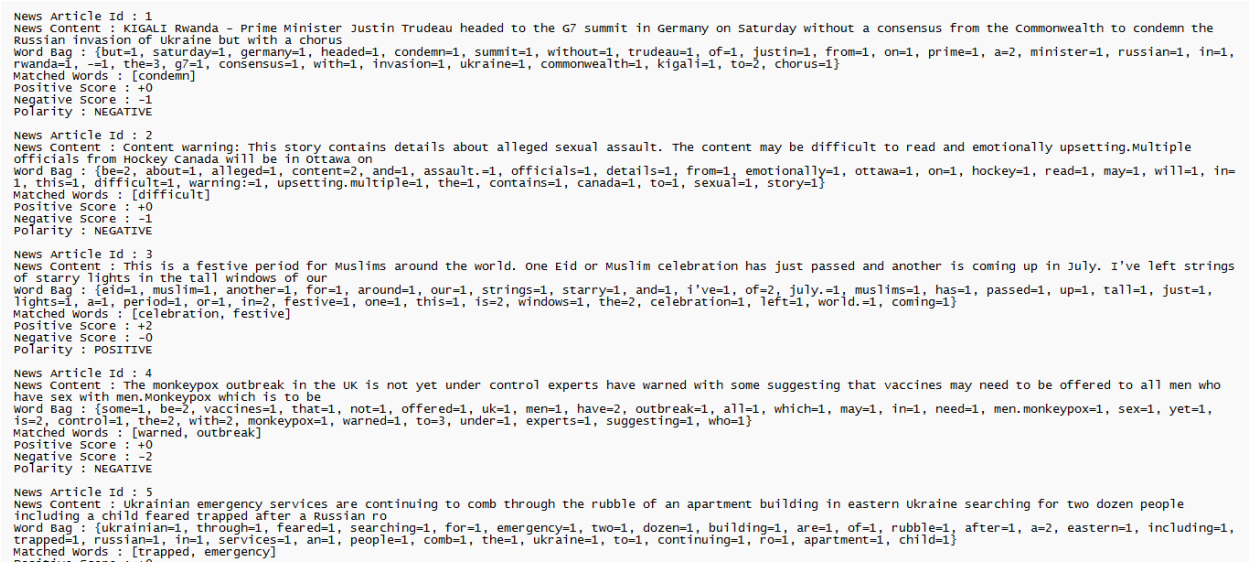
Sentiment Analysis

For sentiment analysis, I have developed a script without using any additional java external library.

My code will generate txt file as an output.

Steps:

- I have taken the reference of the parser which I used in the last assignment. Now using that parser, the code will parse news articles into array list.
- Now, I will generate list of bags of words where each bag of words contains news description.
- After that, I developed a Hashset for positive and negative words[3][4].
- I have also developed one model class having following variables: id, content, contentFrequencyCount, matchedWords, polarity, positiveScore, negativeScore.
- In the logic of Sentiment Analysis, code will go through list of bag words and it will compare with the previously developed hashset of positive and negative words which will ultimately store the result.
- At the end, I simply generated a txt file for the output.
- The image of the output txt file is attached below.



```
News Article Id : 1
News Content : KIGALI Rwanda - Prime Minister Justin Trudeau headed to the G7 summit in Germany on Saturday without a consensus from the Commonwealth to condemn the
Russian invasion of Ukraine but with a chorus
Word Bag : {but=1, saturday=1, germany=1, headed=1, condemn=1, summit=1, without=1, trudeau=1, of=1, justin=1, from=1, on=1, prime=1, a=2, minister=1, russian=1, in=1,
rwanda=1, -=1, the=3, g7=1, consensus=1, with=1, invasion=1, ukraine=1, commonwealth=1, kigali=1, to=2, chorus=1}
Matched Words : [condemn]
Positive Score : +0
Negative Score : -1
Polarity : NEGATIVE

News Article Id : 2
News Content : Content warning: This story contains details about alleged sexual assault. The content may be difficult to read and emotionally upsetting.Multiple
officials from Hockey Canada will be in Ottawa on
Word Bag : {be=2, about=1, alleged=1, content=2, and=1, assault.=1, officials=1, details=1, from=1, emotionally=1, ottawa=1, on=1, hockey=1, read=1, may=1, will=1, in=
1, this=1, difficult=1, warning:=1, upsetting.multiple=1, the=1, contains=1, canada=1, to=1, sexual=1, story=1}
Matched Words : [difficult]
Positive Score : +0
Negative Score : -1
Polarity : NEGATIVE

News Article Id : 3
News Content : This is a festive period for Muslims around the world. One Eid or Muslim celebration has just passed and another is coming up in July. I've left strings
of starry lights in the tall windows of our
Word Bag : {eid=1, muslim=1, another=1, for=1, around=1, our=1, strings=1, starry=1, and=1, i've=1, of=2, july.=1, muslims=1, has=1, passed=1, up=1, tall=1, just=1,
lights=1, a=1, period=1, or=1, in=2, festive=1, one=1, this=1, is=2, windows=1, the=2, celebration=1, left=1, world.=1, coming=1}
Matched Words : [celebration, festive]
Positive Score : +2
Negative Score : -0
Polarity : POSITIVE

News Article Id : 4
News Content : The monkeypox outbreak in the UK is not yet under control experts have warned with some suggesting that vaccines may need to be offered to all men who
have sex with men.Monkeypox which is to be
Word Bag : {some=1, be=2, vaccines=1, that=1, not=1, offered=1, uk=1, men=1, have=2, outbreak=1, all=1, which=1, may=1, in=1, need=1, men.monkeypox=1, sex=1, yet=1,
is=2, control=1, the=2, with=2, monkeypox=1, warned=1, to=3, under=1, experts=1, suggesting=1, who=1}
Matched Words : [warned, outbreak]
Positive Score : +0
Negative Score : -2
Polarity : NEGATIVE

News Article Id : 5
News Content : Ukrainian emergency services are continuing to comb through the rubble of an apartment building in eastern Ukraine searching for two dozen people
including a child feared trapped after a Russian ro
Word Bag : {ukrainian=1, through=1, feared=1, searching=1, for=1, emergency=1, two=1, dozen=1, building=1, are=1, of=1, rubble=1, after=1, a=2, eastern=1, including=1,
trapped=1, russian=1, in=1, services=1, an=1, people=1, comb=1, the=1, ukraine=1, to=1, continuing=1, ro=1, apartment=1, child=1}
Matched Words : [trapped, emergency]
```

Problem #2

Semantic Analysis

For sentiment analysis I have developed a script without using any additional java external library.

My code will generate txt file as an output.

I have developed TF-IDF and TermFrequency table in the form of txt file using core JAVA.

Steps:

- I have taken the reference of the parser which I used in the last assignment. Now using that parser, the code will parse news articles into array list.
- Then, I developed two classes: SemanticAnalysis and SemanticAnalysisModel.
- We will travel through all the documents and calculate how many documents contains the search word.
- After that, based on the list I deveoped a tabluar data in txt file.
- You can better understand the working of the code by looking at it.
- At the end, I simply generated a txt file for the output.
- The image of the output txt file is attached below.

```
TF-IDF (Term Frequency - Inverse Document Frequency
=====================================================
Total Document: 652
Search Query    df      N/df                        Log10(N/df)
condition       5       652/5=130.4                 2.1152775913959014
weather         4       652/4=163.0                 2.2121876044039577
people          24      652/24=27.166666666666668   1.4340363540203143

Calculation of term people
==========================
Term: people
people appeared in 24 documents         Total Words(m)      Frequency(f)
Article: #30                            31                  1
Article: #42                            30                  1
Article: #122                           34                  1
Article: #158                           36                  1
Article: #244                           32                  1
Article: #296                           30                  1
Article: #352                           30                  1
Article: #364                           28                  1
Article: #365                           35                  1
Article: #378                           31                  1
Article: #380                           35                  1
Article: #413                           31                  1
Article: #421                           34                  1
Article: #473                           32                  1
Article: #481                           32                  1
Article: #495                           32                  1
Article: #513                           28                  1
Article: #525                           37                  1
Article: #532                           32                  1
Article: #533                           32                  1
Article: #568                           32                  1
Article: #575                           37                  1
Article: #642                           34                  1
Article: #648                           30                  1

News article with highest relative frequency
============================================
Content: A second plane carrying Ukrainian refugees fleeing war has arrived in Newfoundland and Labrador.The flight chartered by the provincial government andcarrying
177 people and their pets landed Tu
f/m: 0.035714287
```

References :

[1] P. em I.-D. Mestrado, “Climate weather surface of Brazil – hourly,” *Kaggle.* [Online]. Available: https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region?resource=download [Accessed: 20-Oct-2021]

[2] “Lab9_IBM_Cognos_data_module,” *Brightspace.com*. [Online]. Available: https://dal.brightspace.com/d2l/le/content/221749/viewContent/3063667/View. [Accessed: 20-Jul-2022].

[3] Marcin, *Negative-words.Txt*. .

[4] Marcin, *Positive-words.Txt*. .

[5] “IBM Cognos Analytics,” *IBM Cognos Analytics*. [Online]. Available: https://www.ibm.com/products/cognos-analytics. [Accessed: 23-Jul-2022].