CS/DATA 322 - Big Data Analytics

Final Exam - December 18th, 2021

Each question is worth 5 points.

Sentiment analysis with the Natural Language API.

We will perform sentiment analysis on "The Awful German Language", an 1880 essay by Mark Twain.

Create a BigQuery table - `final` - that will store the `german_response.json` file and answer the questions below.

The file is available here: **gs://cs-322-elena-manilich/german_response.json**

Note: There is no need to process the file with Sentiment API. The `german_response.json` is the result json file returned by Google Sentiment API.

The `german_response.json` file includes sentiment values for the document broken down by **sentence**. The sentiment method returns two values:

- *score* - is a number from -1.0 to 1.0 indicating how positive or negative the statement is.
- *magnitude* - is a number ranging from 0 to infinity that represents the weight of sentiment expressed in the statement, regardless of being positive or negative.

Longer blocks of text with heavily weighted statements have higher magnitude values.

If terms were mentioned more than once, the API would return a different sentiment score and magnitude for each mention, along with an aggregate sentiment for the entity.

Question 1

How many sentences were identified by the Sentiment API.

In [1]:
```
%%bigquery
select count(*)
from nokia.final
```

Out[1]:

	f0_
0	228

Question 2

A. How many sentences contain a positive statement? Use 0 as a threshold - all sentences with a score above 0 will be considered positive.

In [3]:
```
%%bigquery
select count(*)
from nokia.final
where sentiment.score > 0
```

Out[3]:

	f0_
0	61

B. Display the most positivie sentences in the essay, text with the highest score.

In [7]:
```
%%bigquery
select text.content
from nokia.final
where sentiment.score > 0
order by sentiment.score desc
limit 5
```

Out[7]:

	content
0	They impart a martial thrill to the meekest su...
1	In this way I have made quite a valuable colle...
2	Whenever I come across a good one, I stuff it ...
3	Joy, joy, with flying Feet the she-Englishwoma...
4	Yes!

Question 3

A. How many sentences contain a negative statement? Use 0 as a threshold - all sentences with a score below 0 will be considered negative.

In [8]:
```
%%bigquery
select count(*)
from nokia.final
where sentiment.score < 0
```

Out[8]:

	f0_
0	137

B. Display the most negative sentences in the essay, text with the lowest score.

In [9]:
```
%%bigquery
select text.content
from nokia.final
where sentiment.score < 0
order by sentiment.score asc
limit 5
```

Out[9]:

	content
0	"Freundshaftsbezeigungen" seems to be "Friends...
1	That is manifestly absurd.
2	It is as bad as Latin.
3	So, as an added E often signifies the plural, ...
4	Now there are more adjectives in this language...

Question 4

A. Display all sentences with the highest magnitude (weight).

In [13]:
```
%%bigquery
select text.content, sentiment.magnitude -- sqrt(sentiment.score*sentiment.score) as magnitude
from nokia.final
order by magnitude desc
limit 5
```

Out[13]:

	content	magnitude
0	They impart a martial thrill to the meekest su...	0.9
1	In this way I have made quite a valuable colle...	0.9
2	Whenever I come across a good one, I stuff it ...	0.9
3	Joy, joy, with flying Feet the she-Englishwoma...	0.9
4	Yes!	0.9

Entity Sentiment Analysis

Entity sentiment analysis inspects the given text for known entities (proper nouns and common nouns), returns information about those entities, and identifies the prevailing emotional opinion of the entity within the text, especially to determine a writer's attitude toward the entity as positive, negative, or neutral.

- salience indicates the importance or relevance of this entity to the entire document text.

This file is the result of the Entity Sentiment analysis for the essay by Mark Twain.

gs://cs-322-elena-manilich/german_response_entity.json

Create a table in BigQuery that stores the result of the Entity Sentiment Analysis and answer the following questions:

Question 5

A. Select the top ten distinct entity names with the highest importance.

In [18]:
```
%%bigquery
select distinct name, salience
from nokia.finalentity
order by salience desc
limit 10
```

Out[18]:

	name	salience
0	North German	0.700033
1	master	0.108559
2	inventor	0.021102
3	curiosity	0.010563
4	rain	0.010221
5	Awful German	0.005633
6	case	0.004747
7	person	0.002634
8	Fishwife	0.002246
9	language	0.001492

B. What is the lowest and the highest sentiment score for each of the most important entity?

In [19]:
```
%%bigquery
select distinct name, salience, m.sentiment.score as sent -- HIGHEST
from nokia.finalentity, unnest (mentions) as m
order by salience desc, sent desc
limit 10
```

Out[19]:

	name	salience	sent
0	North German	0.700033	0.0
1	master	0.108559	-0.1
2	inventor	0.021102	-0.1
3	inventor	0.021102	-0.2
4	curiosity	0.010563	0.7
5	curiosity	0.010563	0.0
6	rain	0.010221	0.0
7	rain	0.010221	-0.1
8	rain	0.010221	-0.3
9	Awful German	0.005633	0.0

In [20]:
```
%%bigquery
select distinct name, salience, m.sentiment.score as sent -- LOWEST
from nokia.finalentity, unnest (mentions) as m
order by salience desc, sent asc
limit 10
```

Out[20]:

	name	salience	sent
0	North German	0.700033	0.0
1	master	0.108559	-0.1
2	inventor	0.021102	-0.2
3	inventor	0.021102	-0.1
4	curiosity	0.010563	0.0
5	curiosity	0.010563	0.7
6	rain	0.010221	-0.3
7	rain	0.010221	-0.1
8	rain	0.010221	0.0
9	Awful German	0.005633	0.0

Question 6

Which entities have the highest variation in terms of the lowest and the highest sentiment score? List the top 5 entity names with the highest absolute difference in sentiment scores.

For example, the lowest sentiment score for the name (entity) 'language' is -0.7 and the highest sentiment score is 0.4. The absolute difference is equal to 1.1.

In [43]:
```
0.4 - (-0.7)
```

Out[43]:
```
1.1
```

In [21]:
```
%%bigquery
select distinct name1, abs(high_sent - low_sent) as variation
from ( # Lowest sentiment score.
select distinct name as name1, salience, m.sentiment.score as low_sent
from nokia.finalentity, unnest (mentions) as m
order by salience desc, low_sent asc
), ( # Highest sentiment score.
select distinct name as name2, salience, m.sentiment.score as high_sent
from nokia.finalentity, unnest (mentions) as m
order by salience desc, high_sent desc
)
order by variation desc
limit 10
```

Out[21]:

	name1	variation
0	folly	1.8
1	friends.	1.8
2	aspects	1.8
3	friends	1.8
4	idea	1.8
5	performance	1.8
6	collection	1.8
7	she-Englishwoman	1.8
8	one	1.8
9	way	1.8

Video Intelligence Analysis

Next, we will analyze the results of the Video Intelligence API that processed a popular Youtube video. Processed json result file is available here:

gs://cs-322-elena-manilich/winnie.json

Question 7

How many unique categories were identified by The Video Intelligence APIs?
Hint: look into the categories only, there are only few categories identified in this video.

In [27]:
```
%%bigquery
select count(distinct ce.description)
from nokia.finalvideo, unnest (categoryentities) as ce
```

Out[27]:

	f0_
0	19

Question 8

How many segments are in the video?

In [28]:
```
%%bigquery
select count(s.segment)
from nokia.finalvideo, unnest (segments) as s
```

Out[28]:

	f0_
0	329

Question 9

A. Which entities appeared in most segments? List the top ten entities (descriptions) that appear frequently.

In [29]:
```
%%bigquery
select entity.description, count(s.segment) as qtysegments
from nokia.finalvideo, unnest(segments) as s
group by entity.description
order by qtysegments desc
limit 10
```

Out[29]:

	description	qtysegments
0	animation	38
1	animal	35
2	illustration	29
3	grassland	15
4	grass	15
5	meadow	14
6	text	13
7	song	12
8	font	12
9	graphic design	11

B. Can you spot `hot air ballooning` in the video? What is the offset of the segments where the `hot air ballooning1 appears?

In [31]:
```
%%bigquery
select entity.description, s.segment.endTimeOffset, s.segment.startTimeOffse
t
from nokia.finalvideo, unnest(segments) as s
where entity.description = "hot air ballooning"
```

Out[31]:

	description	endTimeOffset	startTimeOffset
0	hot air ballooning	439s	433.280s
1	hot air ballooning	447.360s	440.400s
2	hot air ballooning	500.920s	493.480s
3	hot air ballooning	514.720s	512s
4	hot air ballooning	523.880s	520.880s

Question 10

What is the length of the video in minutes?

In [32]:
```
%%bigquery
select endv - startv as durv
from ( # Get latest ending time for any segment.
select
  cast(substr(s.segment.endTimeOffset, 0, length(s.segment.endTimeOffset) -
1) as numeric) as endv,
  #cast(substr(s.segment.startTimeOffset, 0, length(s.segment.startTimeOffse
t) - 1) as numeric) as startv,
from nokia.finalvideo, unnest (segments) as s
order by endv desc
), ( # Get earliest starting time for any segment.
select
  #cast(substr(s.segment.endTimeOffset, 0, length(s.segment.endTimeOffset) -
1) as numeric) as endv,
  cast(substr(s.segment.startTimeOffset, 0, length(s.segment.startTimeOffse
t) - 1) as numeric) as startv,
from nokia.finalvideo, unnest (segments) as s
order by startv asc
)
order by durv desc
limit 1
```

Out[32]:

	durv
0	616.840000000

Can you guess what kind of video is it?

Enjoy at : https://youtu.be/bEwE4wyz00o (https://youtu.be/bEwE4wyz00o)

Submit your work in a pdf and an html format to Canvas.

In [ ]: