Bid Data Mining Homework 2

Team Member:

SID: 109598033 SID: 109598001

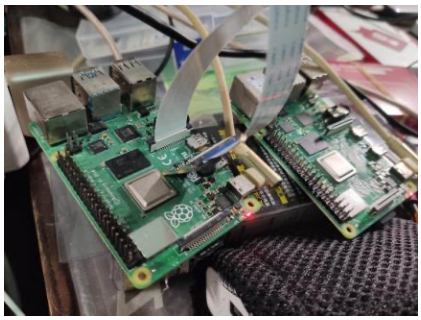
Spark Platform:

The platform consists of:
1. Raspberry Pi 4 Model B x2

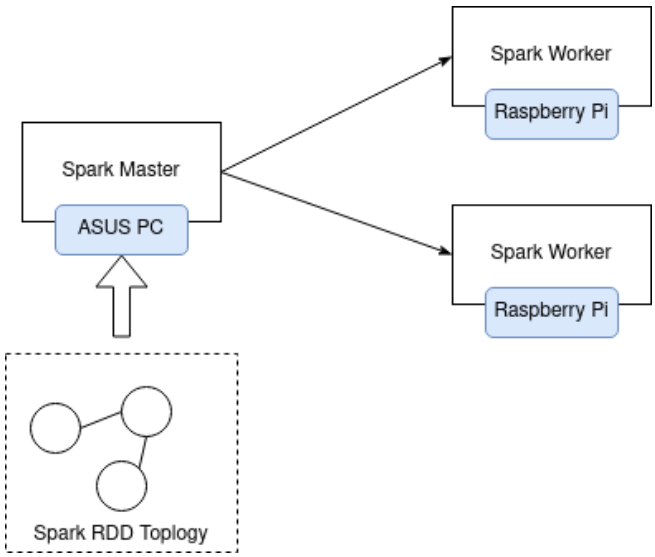
- OS: Linux Ubuntu 20.04 Server
- CPU architecture: aarch64
- RAM: 8GB
- CPU: Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
- Number of CPU: 4C (CPU) 1T (Thread Per CPU)

2. Asus-vivobook notebook

- OS: Linux Ubuntu 20.04 LTS
- CPU architecture: x86_64
- RAM: 8GB
- CPU: Intel(R) Core(TM) i3-8130U CPU @ 2.20GHz
- Number of CPU: 4C (CPU) 2T (Thread Per CPU)



The simple architecture of spark cluster:



Task arrangement for Team:

SID: 109598033

1. Q1

2. Q4

SID: 109598001

1. Q2

2. Q3

The description of tree directory:

Explain where the files put into and how it does.

- **README file explains in detail for what the steps of spark implementation in this homework.**
- **'outputs' directory shows the result of the homework.**
- **park project is 'hw2.ipynb' which codes by Python.**

The generated output:

Q1

q1_news_title_total

word	total
economy	26198
obama	22576
microsoft	17570
obamas	5203
us	4713
palestine	3812
new	3740
says	3518
president	3039
economic	2951
microsofts	2939
windows	2734
2016	2428
global	2197
10	2150
growth	1936
trump	1721

q1_news_title_total_cate

Topic	Total_Title_cate	count
economy	economy	24685
economy	us	2773
economy	economic	2623
economy	global	1864
economy	growth	1747
economy	says	1626
economy	china	1204
economy	2016	1112
economy	chinas	1074
economy	new	999
economy	brexit	965
economy	uk	852
economy	could	772
economy	boost	767
economy	world	723
economy	quarter	697
economy	oil	689

q1_news_title_total_date

PublishDate	Total_Title_day	count
2002-04-02	wreath	1
2002-04-02	arlington	1
2002-04-02	obama	1
2002-04-02	cemetery	1
2002-04-02	lays	1
2002-04-02	national	1
2008-09-20	health	1
2008-09-20	look	1
2008-09-20	chinese	1
2008-09-20	economy	1
2012-01-28	nouriel	1
2012-01-28	roubini	1
2012-01-28	economy	1
2012-01-28	global	1
2012-01-28	back	1
2012-01-28	2008	1
2015-03-01	microsoft	65

q1_news_title_total q1_news_title_total_cate q1_news_title_total_date

q1_news_headline_t

word	total
obama	26004
economy	25211
president	22494
microsoft	20551
barack	13091
us	9940
said	9342
new	9267
economic	8942
year	5687
first	5073
one	4902
windows	4736

q1_news_headline_total

q1_news_headline_total_cate

Topic	Total_Headline_cate	count
economy	economy	23931
economy	economic	8232
economy	growth	4351
economy	said	4131
economy	us	3605
economy	year	3381
economy	global	3075
economy	new	2754
economy	percent	2573
economy	quarter	2183
economy	minister	1801
economy	government	1731

q1_news_headline_total_cate

q1_news_headline_total_date

PublishDate	Total_Headline_day	count
2002-04-02	wreath	2
2002-04-02	obama	2
2002-04-02	national	1
2002-04-02	president	1
2002-04-02	unknowns	1
2002-04-02	honor	1
2002-04-02	arlington	1
2002-04-02	laid	1
2002-04-02	tomb	1
2002-04-02	barack	1
2002-04-02	cemetery	1
2002-04-02	lays	1

q1_news_headline_total_date

Q2

df_popular_d

ID	average
ID1	2.0
ID2	0.0
ID3	0.0
ID4	10.0
ID5	0.0
ID6	99.5
ID7	0.0
ID8	0.0
ID9	11.0
ID10	27.5
ID11	9.0
ID12	25.0

q2_popular_day

df_popular_hour

ID	average
ID1	1.4791666666666700
ID2	-0.041666666666666700
ID3	-0.3125
ID4	6.729166666666670
ID5	-0.3125
ID6	70.125
ID7	-0.020833333333333300
ID8	-0.16666666666666700
ID9	7.604166666666670
ID10	13.166666666666700
ID11	6.375
ID12	15.583333333333300

q2_popular_hour

Q3

q3_Title_sentiment_score

topic	sum	average
economy	-336.9370044373350	-0.010475919672833200
microsoft	49.43849052234930	0.00231042576513456
obama	-15.743686315455400	-0.000581484259119260
palestine	-164.48440896913700	-0.019865266783712200

q3_Title_sentiment_score

q3_Headline_sentiment_score

topic	sum	average
economy	-1271.3909442082200	-0.039529613040083900
microsoft	-318.81900083682000	-0.014899476625704300
obama	-481.8388358106560	-0.01779644822938710
palestine	-363.16995277670900	-0.043861105407815200

q3_Headline_sentiment_score

Q4

col_col	2015	2016	ahead	amid	back	bad	bank	better	big	boost	brexit
rate	9	15	20	13	0	0	29	5	2	15	2
warns	0	15	2	1	0	0	22	0	1	1	76
economy	423	863	159	197	202	208	419	212	190	682	737
despite	14	10	3	1	0	3	11	6	2	1	5
trump	0	2	0	0	1	8	1	4	3	6	10
business	9	12	0	5	6	2	5	2	6	3	9
years	23	6	0	2	3	0	11	0	0	6	2
crisis	0	5	0	11	4	3	2	0	1	4	8
president	6	4	0	5	1	0	3	7	2	9	4
economys	5	5	0	0	0	8	3	2	2	0	3
imf	1	18	1	2	2	2	8	0	12	1	52
india	7	23	2	6	2	3	8	5	8	5	7
prices	5	5	0	7	0	6	5	3	2	5	5

q4_Title_economy_co-occurrence_matrix

col_col	10	2	2016	3	365	4	5	950	adds	ai	amazon	android
cortana	34	9	0	0	0	0	0	0	5	2	1	45
stock	3	1	9	1	0	1	3	1	0	0	3	0
business	11	0	2	1	15	2	1	0	3	0	1	9
support	63	5	10	0	8	3	1	0	44	0	5	18
services	1	0	3	1	6	0	0	0	5	0	3	2
ios	18	1	1	2	6	2	1	1	11	0	1	67
review	13	37	3	4	0	44	1	33	0	0	1	0
hololens	10	2	8	2	0	1	3	0	2	1	0	0
amazon	0	6	4	2	0	0	0	8	1	0	277	0
server	4	0	89	0	0	1	0	0	4	0	1	0
4	14	10	10	13	0	408	12	3	1	0	0	0
surface	29	26	32	114	1	379	88	5	0	0	2	0
uk	3	12	8	9	2	6	2	22	0	1	13	0

q4_Title_microsoft_co-occurrence_matrix

col_col	2016	action	address	administration	america	americans	attack	attacks	back	barack	bill	brexit
legacy	15	0	0	1	2	0	1	0	0	11	5	3
trump	12	3	7	4	13	7	9	17	15	65	6	17
republicans	3	1	1	4	1	1	2	4	3	3	2	0
trumps	0	0	0	1	2	0	3	6	0	4	1	0
leaders	1	12	0	1	2	0	0	0	2	12	1	1
court	1	11	0	27	0	3	0	1	8	2	0	0
tells	0	0	0	5	11	9	2	2	10	7	0	3
donald	5	0	1	0	2	1	6	2	5	60	3	3
barack	12	3	11	4	11	3	10	7	15	1187	8	15
zika	2	6	0	7	0	0	0	0	2	2	18	0
shooting	0	5	3	1	0	3	5	0	1	5	0	0
president	36	10	42	13	21	10	11	11	10	166	21	9
makes	1	1	2	2	4	0	2	3	1	7	1	7

q4_Title_obama_co-occurrence_matrix

col_col	2015	2016	4	abbas	arab	arrested	bank	boys	call	calls	children	city
killed	0	0	40	0	0	0	3	0	0	0	11	0
israels	0	0	0	1	1	0	2	0	0	1	3	0
trump	0	0	0	0	0	0	3	0	2	2	0	1
support	0	0	0	7	1	0	3	0	2	2	1	0
school	0	0	0	0	1	6	1	0	0	0	2	0
gaza	0	0	0	1	0	0	12	0	1	3	6	3
children	0	0	6	0	1	1	0	0	2	0	79	0
palestinian	0	4	0	24	17	6	34	0	3	18	29	10
hamas	0	0	0	1	0	0	1	0	0	1	1	1
city	0	0	0	0	0	0	1	0	0	0	0	57
shooting	0	0	0	1	0	17	1	0	0	0	0	0
president	0	3	0	22	0	0	0	0	1	12	0	0
talks	0	2	0	0	0	0	1	0	1	5	0	0

q4_Title_palestine_co-occurrence_matrix

col_col	2015	2016	according	also	bank	business	cent	central	china	chinas	chinese
rate	146	101	81	42	228	25	185	144	43	105	19
economy	1434	1264	1220	668	1392	866	886	778	1031	1422	678
despite	46	51	23	10	56	33	23	23	58	42	35
business	60	46	81	28	45	1280	9	12	33	23	17
people	13	17	24	18	20	30	10	14	9	14	7
years	140	105	99	19	84	43	76	49	65	116	36
president	24	46	25	22	137	21	4	74	22	37	56
reserve	17	22	19	16	212	11	16	62	8	12	13
prices	64	73	33	52	100	20	25	46	49	51	34
domestic	142	61	74	32	48	18	81	24	48	54	13
said	287	227	88	154	567	145	150	281	196	302	104
market	57	72	52	50	55	31	26	43	158	158	59
union	13	9	40	18	31	33	7	15	50	15	3

q4_Headline_economy_co-occurrence_matrix

col_col	10	2015	2016	365	4	according	also	android	announced	app	apple
business	72	37	58	67	12	31	32	17	147	44	33
support	161	7	38	10	8	21	26	26	118	61	39
services	22	14	53	65	2	11	24	20	156	22	8
people	112	11	8	8	10	15	16	8	31	27	14
years	55	20	26	5	8	18	14	14	50	20	37
hololens	38	23	48	0	1	10	30	1	55	50	2
giant	59	11	15	3	13	20	15	15	59	33	18
operating	332	26	13	1	6	23	12	31	45	29	12
corp	33	21	75	12	6	20	26	6	131	4	26
companies	26	16	22	16	2	22	12	7	54	10	77
said	175	12	42	23	19	28	51	26	69	32	45
market	42	31	41	9	21	35	12	19	41	7	57
4	34	16	48	0	617	12	29	1	42	2	43

q4_Headline_microsoft_co-occurrence_matrix

col_col	2016	address	administration	american	americans	announced	ap	barack	called	campaign
sunday	14	136	14	34	53	15	14	398	26	9
trump	45	29	25	34	14	2	18	438	38	115
republicans	16	15	31	17	14	5	44	317	24	11
people	18	52	87	120	64	30	82	554	68	14
leaders	7	15	25	43	12	6	64	490	40	20
years	19	42	83	53	29	27	44	477	15	36
court	21	4	118	10	8	36	72	652	18	16
donald	44	24	24	28	10	3	32	492	31	123
barack	413	468	346	372	339	256	914	12782	319	416
president	566	785	521	641	540	458	1001	12143	480	591
said	63	126	226	110	123	38	162	2012	92	120
union	28	383	23	19	36	9	19	421	12	22
meeting	26	13	10	17	7	12	30	420	17	19

q4_Headline_obama_co-occurrence_matrix

col_col	2015	2016	abbas	according	affairs	agency	arab	authority	bank	called	center	city
sunday	16	4	8	5	27	5	3	3	7	2	2	8
killed	6	6	2	8	1	1	2	2	20	2	0	17
organization	8	26	4	4	3	6	5	13	9	5	2	3
israels	5	5	6	6	1	8	0	3	19	5	0	2
support	9	16	7	3	6	17	1	7	5	3	1	8
school	5	3	0	3	0	1	4	0	7	0	5	6
gaza	17	27	3	6	3	8	3	9	38	4	3	48
people	10	9	3	5	4	4	8	2	12	11	1	12
palestinian	152	165	251	105	44	64	86	286	286	69	17	145
city	12	17	7	29	3	2	5	2	70	3	4	425
years	3	15	6	9	0	4	5	4	7	2	2	14
president	20	20	247	10	21	8	7	74	14	16	29	10
said	25	57	39	9	24	26	22	24	59	10	7	34

q4_Headline_palestine_co-occurrence_matrix