CS777 – Assignment 2

Gaurav Tungare

------------------------------------------------------------------------------------------------------------------------------------

Task 1 & Task 2

A picture containing chart

Description automatically generated

Graphical user interface, table

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

21/09/17 17:17:43 INFO org.sparkproject.jetty.util.log: Logging initialized @3380ms to org.sparkproject.jetty.util.log.Slf4jLog

21/09/17 17:17:43 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0\_292-b10

21/09/17 17:17:43 INFO org.sparkproject.jetty.server.Server: Started @3484ms

21/09/17 17:17:43 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@799cbc07{HTTP/1.1, (http/1.1)}{0.0.0.0:35843}

21/09/17 17:17:43 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-d706-m/10.128.15.193:8032

21/09/17 17:17:44 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at cluster-d706-m/10.128.15.193:10200

21/09/17 17:17:45 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found

21/09/17 17:17:45 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.

21/09/17 17:17:46 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application\_1631898679671\_0002

21/09/17 17:17:47 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-d706-m/10.128.15.193:8030

21/09/17 17:17:49 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl:

Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.

21/09/17 17:17:51 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

**1000000**

**Get the top 20,000 words**

save top10WordsToASingleFile

21/09/17 17:20:29 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://gtungare\_bucket\_1/output\_mod2\_task1\_task2\_s50/' directory.

**Top Words in Corpus: [('the', 26451056), ('of', 12507151), ('in', 10807932), ('and', 10758007), ('a', 7988396), ('to', 7681799), ('was', 4588136), ('is', 3746524), ('for', 3145951), ('as', 3066049)]**

**Word Postions in our Feature Matrix. Last 20 words in 20k positions: [('ub', 19999), ('yarmouth', 19998), ('ami', 19997), ('divert', 19996), ('csa', 19995), ('unlocked', 19994), ('lockout', 19993), ('dent', 19992), ('misuse', 19991), ('sizeable', 19990), ('overt', 19989), ('aqueous', 19988), ('minerva', 19987), ('glamour', 19986), ('madurai', 19985), ('ethnographic', 19984), ('estadio', 19983), ('corbin', 19982), ('rida', 19981), ('melancholy', 19980)]**

Task 1 Done

Starting Task 2

**[('23541998', array([0.09648127, 0.04994325, 0.02951192, ..., 0. , 0. ,**

**0. ])), ('23542268', array([0.10638298, 0.06534954, 0.04559271, ..., 0. , 0. ,**

**0. ])), ('23167799', array([0.081103 , 0.02595296, 0.03163017, ..., 0. , 0. ,**

**0. ]))]**

21/09/17 17:34:39 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

21/09/17 17:34:40 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@799cbc07{HTTP/1.1, (http/1.1)}{0.0.0.0:0}

Graphical user interface, text, application, email

Description automatically generated

Task 2

Graphical user interface, timeline

Description automatically generated

Graphical user interface, table

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

21/09/18 20:45:34 INFO org.sparkproject.jetty.util.log: Logging initialized @3267ms to org.sparkproject.jetty.util.log.Slf4jLog

21/09/18 20:45:35 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0\_292-b10

21/09/18 20:45:35 INFO org.sparkproject.jetty.server.Server: Started @3373ms

21/09/18 20:45:35 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@6659a72f{HTTP/1.1, (http/1.1)}{0.0.0.0:37127}

21/09/18 20:45:35 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-647f-m/10.128.15.204:8032

21/09/18 20:45:36 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at cluster-647f-m/10.128.15.204:10200

21/09/18 20:45:36 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found

21/09/18 20:45:36 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.

21/09/18 20:45:37 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application\_1631996903641\_0002

21/09/18 20:45:38 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-647f-m/10.128.15.204:8030

21/09/18 20:45:40 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.

21/09/18 20:45:41 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

1000000

Get the top 20,000 words

save top10WordsToASingleFile

Top Words in Corpus: [('the', 26451056), ('of', 12507151), ('in', 10807932), ('and', 10758007), ('a', 7988396), ('to', 7681799), ('was', 4588136), ('is', 3746524), ('for', 3145951), ('as', 3066049)]

Word Postions in our Feature Matrix. Last 20 words in 20k positions: [('lane', 1999), ('retirement', 1998), ('consisted', 1997), ('legislative', 1996), ('pilot', 1995), ('aid', 1994), ('glass', 1993), ('determined', 1992), ('federation', 1991), ('hits', 1990), ('speech', 1989), ('shared', 1988), ('wins', 1987), ('poetry', 1986), ('faced', 1985), ('soccer', 1984), ('onto', 1983), ('guns', 1982), ('contest', 1981), ('kill', 1980)]

Task 1 Done

Starting Task 2

[('432001', array([0.11651132, 0.04527297, 0.03395473, ..., 0. , 0. ,

0. ])), ('432253', array([0.11693548, 0.03387097, 0.025 , ..., 0. , 0. ,

0. ])), ('456101', array([0.14671815, 0.09266409, 0.05405405, ..., 0. , 0. ,

0. ]))]

21/09/18 21:00:32 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

**getPrediction\_1**

1..myDoc

2..wordsInThatDoc

3...allDictionaryWordsInThatDoc

4...myArray

5...tf \* idf

6...distances

7...topK

8...docIDRepresented

9...numTimes

10...return

[('Articles\_containing\_Japanese-language\_text', 6), ('Articles\_with\_short\_description', 6), ('All\_set\_index\_articles', 5), ('Surnames', 5), ('All\_stub\_articles', 5), ('CS1\_Japanese-language\_sources\_(ja)', 4), ('Given\_names', 3), ('2007\_singles', 2), ('Japanese-language\_songs', 2), ('Japanese\_film\_songs', 2), ('Oricon\_Weekly\_number-one\_singles', 2), ('2007\_songs', 2), ('Articles\_needing\_additional\_references\_from\_May\_2019', 1), ('2005\_films', 1), ('Wikidata\_redirects', 1), ('2013\_songs', 1), ('2003\_singles', 1), ('Films\_about\_summer\_camps', 1), ('Articles\_with\_hAudio\_microformats', 1), ('Redirects\_connected\_to\_a\_Wikidata\_item', 1), ('Monitored\_short\_pages', 1), ('2013\_singles', 1), ('2000s\_Japanese\_single\_stubs', 1), ('Songs\_written\_by\_Sin', 1), ('Kanjani\_Eight\_songs', 1), ('Japanese\_television\_dramas\_based\_on\_manga', 1), ('All\_articles\_needing\_additional\_references', 1), ('All\_articles\_lacking\_sources', 1), ('Mika\_Nakashima\_songs', 1), ('Japanese-language\_films', 1), ('Redirects\_from\_songs', 1), ('Manga\_stubs', 1), ('Wikipedia\_requested\_photographs\_of\_anime\_and\_manga', 1), ('2010s\_Japanese\_single\_stubs', 1), ('Unprintworthy\_redirects', 1), ('All\_articles\_with\_topics\_of\_unclear\_notability', 1), ('Films\_about\_dysfunctional\_families', 1), ('2013\_Japanese\_television\_series\_debuts', 1), ('CS1\_uses\_Japanese-language\_script\_(ja)', 1), ('Japanese\_television\_stubs', 1), ('Songs\_written\_by\_Kazufumi\_Miyazawa', 1), ('Use\_mdy\_dates\_from\_July\_2014', 1), ('Romance\_anime\_and\_manga', 1), ('2000s\_Japanese\_film\_stubs', 1), ('Songs\_written\_by\_Kohmi\_Hirose', 1), ('Masafumi\_Akikawa\_songs', 1), ('2005\_manga', 1), ('2003\_songs', 1), ('Films\_about\_organized\_crime\_in\_Japan', 1), ('Avex\_Trax\_singles', 1), ('Hitomi\_Shimatani\_songs', 1), ('Seinen\_manga', 1), ('Shogakukan\_manga', 1), ('Yomiuri\_Telecasting\_Corporation\_shows', 1), ('Japan\_album\_stubs', 1), ('1992\_albums', 1), ('Billboard\_Japan\_Hot\_100\_number-one\_singles', 1), ('All\_articles\_with\_dead\_external\_links', 1), ('Articles\_with\_Japanese-language\_sources\_(ja)', 1), ('Nippon\_Television\_shows', 1), ('Manga\_series', 1), ('Album\_infoboxes\_lacking\_a\_cover', 1), ('Articles\_lacking\_sources\_from\_December\_2009', 1), ('Articles\_with\_topics\_of\_unclear\_notability\_from\_April\_2016', 1), ('Articles\_with\_dead\_external\_links\_from\_May\_2019', 1), ('Redirects\_to\_Wiktionary', 1), ('Music\_articles\_with\_topics\_of\_unclear\_notability', 1), ('The\_Stalin\_albums', 1)]

**getPrediction\_2**

1..myDoc

2..wordsInThatDoc

3...allDictionaryWordsInThatDoc

4...myArray

5...tf \* idf

6...distances

7...topK

8...docIDRepresented

9...numTimes

10...return

[('Articles\_containing\_Japanese-language\_text', 6), ('Articles\_with\_short\_description', 6), ('All\_set\_index\_articles', 5), ('All\_stub\_articles', 5), ('Surnames', 5), ('CS1\_Japanese-language\_sources\_(ja)', 4), ('Given\_names', 3), ('2007\_singles', 2), ('Japanese-language\_songs', 2), ('Japanese\_film\_songs', 2), ('Oricon\_Weekly\_number-one\_singles', 2), ('2007\_songs', 2), ('2005\_films', 1), ('Articles\_needing\_additional\_references\_from\_May\_2019', 1), ('2003\_singles', 1), ('Wikidata\_redirects', 1), ('2013\_songs', 1), ('2000s\_Japanese\_single\_stubs', 1), ('Kanjani\_Eight\_songs', 1), ('Japanese\_television\_dramas\_based\_on\_manga', 1), ('Monitored\_short\_pages', 1), ('2013\_singles', 1), ('Articles\_with\_hAudio\_microformats', 1), ('Songs\_written\_by\_Sin', 1), ('All\_articles\_needing\_additional\_references', 1), ('Films\_about\_summer\_camps', 1), ('Redirects\_connected\_to\_a\_Wikidata\_item', 1), ('Japanese-language\_films', 1), ('Redirects\_from\_songs', 1), ('All\_articles\_lacking\_sources', 1), ('Mika\_Nakashima\_songs', 1), ('2010s\_Japanese\_single\_stubs', 1), ('2013\_Japanese\_television\_series\_debuts', 1), ('CS1\_uses\_Japanese-language\_script\_(ja)', 1), ('Japanese\_television\_stubs', 1), ('Manga\_stubs', 1), ('Wikipedia\_requested\_photographs\_of\_anime\_and\_manga', 1), ('All\_articles\_with\_topics\_of\_unclear\_notability', 1), ('Unprintworthy\_redirects', 1), ('Films\_about\_dysfunctional\_families', 1), ('Songs\_written\_by\_Kohmi\_Hirose', 1), ('Songs\_written\_by\_Kazufumi\_Miyazawa', 1), ('2000s\_Japanese\_film\_stubs', 1), ('Films\_about\_organized\_crime\_in\_Japan', 1), ('Masafumi\_Akikawa\_songs', 1), ('2003\_songs', 1), ('Avex\_Trax\_singles', 1), ('Hitomi\_Shimatani\_songs', 1), ('Romance\_anime\_and\_manga', 1), ('Use\_mdy\_dates\_from\_July\_2014', 1), ('2005\_manga', 1), ('Japan\_album\_stubs', 1), ('1992\_albums', 1), ('Seinen\_manga', 1), ('Shogakukan\_manga', 1), ('Yomiuri\_Telecasting\_Corporation\_shows', 1), ('Nippon\_Television\_shows', 1), ('All\_articles\_with\_dead\_external\_links', 1), ('Articles\_with\_Japanese-language\_sources\_(ja)', 1), ('Billboard\_Japan\_Hot\_100\_number-one\_singles', 1), ('Manga\_series', 1), ('Redirects\_to\_Wiktionary', 1), ('Music\_articles\_with\_topics\_of\_unclear\_notability', 1), ('The\_Stalin\_albums', 1), ('Articles\_lacking\_sources\_from\_December\_2009', 1), ('Articles\_with\_topics\_of\_unclear\_notability\_from\_April\_2016', 1), ('Album\_infoboxes\_lacking\_a\_cover', 1), ('Articles\_with\_dead\_external\_links\_from\_May\_2019', 1)]

**getPrediction\_3**

1..myDoc

2..wordsInThatDoc

3...allDictionaryWordsInThatDoc

4...myArray

5...tf \* idf

6...distances

7...topK

8...docIDRepresented

9...numTimes

10...return

[('Articles\_containing\_Japanese-language\_text', 6), ('Articles\_with\_short\_description', 6), ('All\_set\_index\_articles', 5), ('Surnames', 5), ('All\_stub\_articles', 5), ('CS1\_Japanese-language\_sources\_(ja)', 4), ('Given\_names', 3), ('2007\_singles', 2), ('Japanese-language\_songs', 2), ('Japanese\_film\_songs', 2), ('Oricon\_Weekly\_number-one\_singles', 2), ('2007\_songs', 2), ('2003\_singles', 1), ('2005\_films', 1), ('2013\_songs', 1), ('Wikidata\_redirects', 1), ('Articles\_needing\_additional\_references\_from\_May\_2019', 1), ('2000s\_Japanese\_single\_stubs', 1), ('Kanjani\_Eight\_songs', 1), ('Japanese\_television\_dramas\_based\_on\_manga', 1), ('Articles\_with\_hAudio\_microformats', 1), ('Songs\_written\_by\_Sin', 1), ('All\_articles\_needing\_additional\_references', 1), ('Films\_about\_summer\_camps', 1), ('Redirects\_connected\_to\_a\_Wikidata\_item', 1), ('Monitored\_short\_pages', 1), ('2013\_singles', 1), ('Redirects\_from\_songs', 1), ('All\_articles\_lacking\_sources', 1), ('Mika\_Nakashima\_songs', 1), ('Japanese-language\_films', 1), ('Unprintworthy\_redirects', 1), ('All\_articles\_with\_topics\_of\_unclear\_notability', 1), ('Manga\_stubs', 1), ('Wikipedia\_requested\_photographs\_of\_anime\_and\_manga', 1), ('2010s\_Japanese\_single\_stubs', 1), ('Films\_about\_dysfunctional\_families', 1), ('2013\_Japanese\_television\_series\_debuts', 1), ('CS1\_uses\_Japanese-language\_script\_(ja)', 1), ('Japanese\_television\_stubs', 1), ('Masafumi\_Akikawa\_songs', 1), ('2003\_songs', 1), ('Avex\_Trax\_singles', 1), ('Hitomi\_Shimatani\_songs', 1), ('Romance\_anime\_and\_manga', 1), ('Use\_mdy\_dates\_from\_July\_2014', 1), ('2005\_manga', 1), ('Songs\_written\_by\_Kohmi\_Hirose', 1), ('Songs\_written\_by\_Kazufumi\_Miyazawa', 1), ('2000s\_Japanese\_film\_stubs', 1), ('Films\_about\_organized\_crime\_in\_Japan', 1), ('1992\_albums', 1), ('Japan\_album\_stubs', 1), ('Seinen\_manga', 1), ('Shogakukan\_manga', 1), ('Yomiuri\_Telecasting\_Corporation\_shows', 1), ('All\_articles\_with\_dead\_external\_links', 1), ('Nippon\_Television\_shows', 1), ('Billboard\_Japan\_Hot\_100\_number-one\_singles', 1), ('Articles\_with\_Japanese-language\_sources\_(ja)', 1), ('Album\_infoboxes\_lacking\_a\_cover', 1), ('Articles\_lacking\_sources\_from\_December\_2009', 1), ('Articles\_with\_topics\_of\_unclear\_notability\_from\_April\_2016', 1), ('Articles\_with\_dead\_external\_links\_from\_May\_2019', 1), ('Manga\_series', 1), ('Redirects\_to\_Wiktionary', 1), ('Music\_articles\_with\_topics\_of\_unclear\_notability', 1), ('The\_Stalin\_albums', 1)]

21/09/18 21:12:53 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@6659a72f{HTTP/1.1, (http/1.1)}{0.0.0.0:0}

A screenshot of a computer

Description automatically generated

**Task 3**

Graphical user interface, text, application

Description automatically generated

Timeline

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Output of the console

21/09/19 22:42:15 INFO org.sparkproject.jetty.util.log: Logging initialized @3806ms to org.sparkproject.jetty.util.log.Slf4jLog

21/09/19 22:42:15 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0\_292-b10

21/09/19 22:42:15 INFO org.sparkproject.jetty.server.Server: Started @3917ms

21/09/19 22:42:15 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@52a869dc{HTTP/1.1, (http/1.1)}{0.0.0.0:37925}

21/09/19 22:42:16 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-51a3-m/10.128.15.205:8032

21/09/19 22:42:16 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at cluster-51a3-m/10.128.15.205:10200

21/09/19 22:42:17 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found

21/09/19 22:42:17 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.

21/09/19 22:42:19 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application\_1632091245495\_0001

21/09/19 22:42:20 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at cluster-51a3-m/10.128.15.205:8030

21/09/19 22:42:22 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.

21/09/19 22:42:24 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

+---+--------------------+

| \_1| \_2|

+---+--------------------+

| 10|Redirects\_from\_moves|

| 10|Redirects\_with\_ol...|

| 10|Unprintworthy\_red...|

| 12| Anarchism|

| 12| Anti-capitalism|

| 12| Anti-fascism|

| 12|Articles\_containi...|

| 12|Articles\_containi...|

| 12|Articles\_prone\_to...|

| 12|Articles\_with\_sho...|

| 12|CS1:\_long\_volume\_...|

| 12| Economic\_ideologies|

| 12| Far-left\_politics|

| 12| Good\_articles|

| 12|Libertarian\_socia...|

| 12| Libertarianism|

| 12| Political\_culture|

| 12|Political\_ideologies|

| 12| Political\_movements|

| 12| Social\_theories|

+---+--------------------+

only showing top 20 rows

**max\_count\_pages**

+---------------+

|max\_count\_pages|

+---------------+

| 2586023|

+---------------+

**avg\_count\_pages**

+---------------+

|avg\_count\_pages|

+---------------+

|79.943160886687|

+---------------+

+----------+

**|med\_count)|**

+----------+

| 6|

+----------+

**stddev\_pages**

+------------------+

| stddev\_pages|

+------------------+

|4441.5084186380545|

+------------------+

**top 10 mostly used wikipedia categories**

+--------------------+-------+

| \_2| count|

+--------------------+-------+

| Noindexed\_pages|2586023|

| All\_stub\_articles|2243344|

|WikiProject\_Biogr...|1650535|

|Articles\_with\_sho...|1516225|

|Redirects\_from\_moves|1489008|

|Unprintworthy\_red...|1412217|

|Coordinates\_on\_Wi...|1048235|

|Biography\_article...| 970853|

|Stub-Class\_biogra...| 939454|

| Living\_people| 938708|

+--------------------+-------+

only showing top 10 rows

21/09/19 23:09:46 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@52a869dc{HTTP/1.1, (http/1.1)}{0.0.0.0:0}

**Task 4** - Removing Stop Words, do Stemming and redo the task 2 (2 points)

**Task 4.1 - Remove Stop Words (1 point)**

So removing the English Stop words (most common words like ”a, the, is, are, i, you, ...”) would not change the final kNN results since these words are stop words and do not carry any value since they appear almost consistently is each document.

Per the standard convention , removal of stop words are necessary for efficient process

**Task 4.2 - Do English word stemming (1 point)**

So stemming will change the result heavily since we will keep the root words removing the other similar words resulting in changing the TF for each document resulting in distances being changed for each