جامعة السودان للعلوم والتكنولوجيا

كلية علوم الحاسوب وتقانة المعلومات

تحليل الآراء من التغريدات السودانية

Opinion Mining For The Sudanese Tweets

أكتوبر 2016 م

مشروع علمي مقدم كأحد متطلبات الحصول على بكالريوس الشرف في هندسة البرمجيات

بسم الله الرحمن الرحيم

جامعة السودان للعلوم والتكنولوجيا

كلية علوم الحاسوب وتقانة المعلومات

قسم هندسة البرمجيات

تحليل الآراء من التغريدات السودانية

Opinion Mining For The Sudanese Tweets

أكتوبر 2016 م

إعداد الطالبات:

أفراح متوكل إبراهيم طه

ملاك عادل محمد عيسى

توقيع المشرف :

د.هشام عبدالله

.....

الآية

قال تعالى :

(قلْ لَوْ كَانَ الْبَحْرُ مِدَادًا لِكَلِمَاتِ رَبِّي لَنَفِدَ الْبَحْرُ قَبْلَ أَنْ تَنْفَدَ كَلِمَاتُ رَبِّي وَلَوْ جِئْنَا بِمِثْلِهِ مَدَدًا)

" سورة الكهف الآية " ١٠٩

صدق الله العظيم

الحمد

الـحمد لله أقصى مَبلَغِ الـحَمدِ ... والشُكرُ لله مِن قَبـلٍ ومِن بَعـدِ
الـحمد لله عن سَـمعٍ وعن بَصَـرٍ ... الـحَمدُ لله عن عقـلٍ وعَن جسـدِ
الـحمد لله عن سـاقٍ وعن قـدمٍ ... الـحمد لله عن كَتِفـي وعنك يـدي
الـحمد لله عن قلبـي وعن رِئتـي ... الـحمد لله عن كِلتـي وعن كبـدي
الـحمد لله عن أُمـي وعن أبتي ... والحمد لله عن أخوات ذا العبدِ
الـحمد لله فـي سـرِّي وفـي علنـي ... والحمد لله فـي حُزنـي وفـي سَـعدي
الـحمد لله عمّـا كـنت أعلَمُـهُ ... والحمد لله عَمَّـا غـابَ عن خَلَدي
الـحمد لله من عمَّـت فضـائلُهُ ... وأنـعُمُ الله أعيـت منـطِق العددِ
فالحمد لله ثُمَّ الشُـكرُ يتبَعُـهُ ... والحمد لله عن شـكري وعن حمدي

الإهداء

بسم الله الرحمن الرحيم

(قل إعملوا فسيرى الله عملكم ورسوله والمؤمنون)

صدق الله العظيم

إلهي لايطيب الليل إلا بشكرك ولايطيب النهار إلى بطاعتك .. ولاتطيب اللحظات إلا بذكرك .. ولا تطيب الآخرة إلا بعفوك .. ولا تطيب الجنة إلا برؤيتك الله جل جلاله

****

إلى من بلغ الرسالة وأدى الأمانة .. ونصح الأمة .. إلى نبي الرحمة ونور العالمين..

سيدنا محمد صلى الله عليه وسلم

****

يا من أحمل اسمك بكل فخر

يا من أفتقدك منذ الصغر

يا من يرتعش قلبي لذكرك

يا من أودعتني لله أهديك هذا البحث **أبي**

****

إلى حكمتي .....وعلمي

إلى أدبي .....وحلمي

إلى طريقي .... المستقيم

إلى طريق..... الهداية

إلى ينبوع الصبر والتفاؤل والأمل

إلى كل من في الوجود بعد الله ورسوله **أمي الغالية**

****

إلى القلوب الطاهرة الرقيقة والنفوس البريئة إلى رياحين حياتي **إخوتي**

****

إلى توأم روحي ورفيقة دربي .. إلى صاحبة القلب الطيب والنوايا الصادقة إلى من رافقتني منذ أن حملنا حقائب صغيرة ومعها سرت الدرب خطوة بخطوة وما تزال ترافقني حتى الآن صديقتي مودة محمد بابكر المقلي

****

الى من ساهم في وصولنا لطريق النهايه
الى كل من علمني شيئا جديدا
وغذى فكري بالعلم والمعرفه
الى كل من وقف بجانبنا وساعدنا
في كل المصاعب
الى أساتذتنا الأجلاء

****

إلى من كانوا ملاذي وملجئي
إلى من تذوقت معهم أجمل اللحظات
إلى من سأفتقدهم ..... وأتمنى أن يفتقدوني
إلى من جعلهم الله أخوتي بالله .....و من أحببتهم بالله اصدقاءنا في الجامعه

****

الشكر و التقدير

الحمد لله ذي المن والفضل والإحسان، حمداً يليق بجلال عظمته، وصَلِّ اللهم على خاتم الرسل، من لا نبي بعده.

الشكر أولاً وأخيراً، على حسن توفيقه، وكريم عونه، وعلى ما منَّ وفتح به علَيَّنا من إنجاز هذا البحث.

****

ويسرنا أن نخص بالشكر والعرفان الدكتور هشام عبدالله .

****

وأخيراً، نتقدم بجزيل الشكر إلى كل من مدوا لنا يد العون والمساعدة في إخراج هذه البحث على أكمل وجه.

****

المستخلص

شهدت مواقع التواصل الإجتماعي تطوراً كبيراً في السنين الأخيرة و أصبحت مصدراً مهما للمعلومات . وكذلك أتاحت فرصة كبيرة للتواصل و مشاركة المعلومات بين الأشخاص حول العالم . ويمكن الإستفادة من الكمية الهائلة من البيانات التي يتم نشرها يومياً لمعرفة آراء الناس حول منتج معين أو خدمة أو شخصية .

يهدف هذا المشروع الى إسترجاع التغريدات السودانية من موقع التواصل الإجتماعي (تويتر) وتخزينها في قاعدة البيانات لإستخراج أكثر الكلمات تداولاً في التغريدات السودانية وأكثر المواضيع التي يتم طرحها ومن ثم تطبيق خوارزميات التجميع في الذكاء الإصطناعي لصنيف التغريدات السودانية الى فئات محددة (دينية،سياسية،ترفيه،رياضية،أخرى) وإستخراج عدد التغريدات لكل فئة.

Abstract

The massive developments of social media sites in the in recent years and have become an important source of information. And also it allowed a great opportunity to communicate and share information between people around the world. And you can take advantage of the large quantity of data that are published daily to find out people's views about a particular product, service or personality.

This project is aimed at retrieving Sudanese tweets from the social networking site (Twitter) and stored in the database to extract the most words traded in the Sudanese tweets and more topics that are put forward and then apply assembly algorithms in artificial intelligence to classify the Sudanese tweets to specific groups (religious, political, entertainment, Sports, Other) and extract the number of tweets for each category

الإختصارات

الاختصار	المصطلح	شرح المصطلح
API	Application Program Interface	فكرة (API) هي أن تَجلِب البيانات والخدمات بتنسيقات تسمح لنا بإستخدامها مرة أخرى من أي مكان ، وأشهر هذه التنسيقات هي JSON و XML .
OAuth	Open Authorization	بروتوكول مفتوح يُمكّن من عمل تحقق مُؤمَن بطريقة بسيطة وقياسية
JSON	JavaScript Object Notation	عبارة عن صيغة متسلسلة لنقل البيانات
PHP	Hypertext Preprocessor	لغة PHP هي واحدة من أشهر لغات البرمجة التي يتم إستخدامها في إنشاء مواقع الويب و هي من اللغات التي يقوم خادم الويب بتفسير و تنفيذ الكود الخاص بها ثم يرسل النتيجة ليتم عرضها في متصفح المستخدم
AT	Access Token	رمز يتم استخدامه من قبل تطبيق تويتر للوصول الى الموارد المحمية نيابة عن المستخدم

فهرس الأشكال

الشكل(2.8.2.2) الطريقة الهرمية لتجميع البيانات:.....13

الشكل(3.8.2.2) يمثل الاسلوب القائم على الكثافة لتجميع البيانات : 14

الشكل (9.2.2) العنقدة أو التجميع:.....15

الشكل (1.3.2) نتيجة تحليل الأراء حول منتج أيفون6 :.....23

الشكل (4.3.2) أداة 25Trends :.....26

الشكل (6.3.2) أداة As Tweeted :.....28

الشكل (7.3.2) نتيجة تحليل الرأي العام الإلكتروني عقب إستهداف الأماكن المقدسة :.....28

الشكل (3.8.3) بيئة تطوير لغة الR :.....32

الشكل (4.4) يوضح إنشاء تطبيق تويتر 45

الشكل (1.4.4) يوضح الصلاحيات التي تعطى للمستخدم :.....47

الشكل (6.4) الخطوات الأساسية للحصول على البيانات وتخزينها :.....53

الشكل (3.5) يمثل اكثر الكلمات تكرارا في شاشة التنفيذ R :..... 56

الشكل(4.5) رسم ال WordCloud للتغريدات السودانية :.....57

الشكل(5.5) نتيجة تنفيذ خوارزمية Hierarchical clustering على التغريدات:..... 58

الشكل (1.6.5) تحديد مراكز التجمعات عشوائيا لكل فئة باستخدام R :.....60

الشكل(2.6.5) نتيجة تطبيق خوارزمية Kmeans على التغريدات :.....68

الشكل (7.5) أكثر الموضوعات تداولاً في التغريدات السودانية :.....69

فهرس الجداول

الجدول(2.3.2) يوضح نتائج تحليل الآراء حول الطاقة النووية بين عامي 2012 و2013 :.....24

الجدول (3.4) الإختيار بأسلوب صحيح :.....44

الجدول(2.5.4) أنواع دوال JSON :.....50

الجدول (1.6.5) يمثل جزء من البيانات و Indexالمقابل :..... 59

الجدول (2.6.5) يمثل مركز التجمع العشوائي لكل مجموعة :.....60

الجدول (2.6.5) يمثل تجميع البيانات في المجموعات اعتمادا على اقرب مركز لها:.....63

الجدول (3.6.5) يمثل مركز التجمع الجديد لكل مجموعة :.....64

الجدول (4.6.5) يمثل تجميع البيانات في المجموعات اعتمادا على اقرب مركز لها(بالنسبة لمراكز التجمع الجديدة) : 67

الجدول (5.6.5) عدد التغريدات لكل مجموعة :.....69

فهرس الرسم البياني

رسم بياني (1.9.2.2) لعمل خوارزمية ال K-Medoids:.....17

رسم بياني (2.9.2.2) لعمل خوارزمية الk-mesns :.....19

رسم بياني (1.8.8.3) المخطط المبعثر في لغه ال R :.....39

رسم بياني (2.8.8.3) المدرج التكراري في لغه ال R :.....40

رسم بياني (2.8.8.3) المخطط الصندوقي في لغه ال R :.....41

رسم البياني (3.5) لاكثر الكلمات تكرارا في التغريدات السودانية :.....60

فهرس الموضوعات

المقدمة 1

1.1 المقدمة :.....2

2.1 مشكلة البحث :.....2

3.1 أهداف البحث :.....2

4.1 حدود البحث :.....3

5.1 منهجية البحث :.....3

6.1 هيكله البحث :.....3

1.6.1 الباب الأول :.....3

2.6.1 الباب الثاني :.....3

3.6.1 الباب الثالث :.....4

4.6.1 الباب الرابع : 4

5.6.1 الباب الخامس :.....4

6.6.1 الباب السادس :.....4

الباب الثاني :.....5

الخلفية النظرية والدراسات السابقة :.....5

الفصل الأول :.....6

1.2 المقدمة :.....7

2.2 الخلفية النظرية :.....7

1.2.2 تعدين الآراء : 7

2.2.2 الشبكات الإجتماعية :.....7

3.2.2 تويتر 8

4.2.2 واجهات برمجة التطبيقات :.....9
5.2.2 واجهات برمجة التطبيقات لتويتر :.....9
1.5.2.2 Search API :.....9
2.5.2.2 REST API :.....9
3.5.2 Streaming API..... 10
الفصل الثاني :.....19
تجميع البيانات (Data Clustering):.....19
6.2.2 المقدمة :.....21
7.2.2 تجميع البيانات :.....21
8.2.2 طرق تجميع البيانات 21
1.8.2.2 طريقة التقسيم (Partitioning Method) :.....22
2.8.2.2 الطريقة الهرمية (Hierarchical Method) :.....22
3.8.2.2 الإسلوب القائم على الكثافة :.....23
4.8.2.2 الأسلوب القائم على الشبكة :.....24
5.8.2.2 الأسلوب القائم على نموذج :.....24
6.8.2.2 الأسلوب القائم على القيد :.....24
9.2.2 خوارزميات العنقدة إو تجميع:.....24
1.9.2.2 خوارزمية K-Medoids clustering:.....25
2.9.2.2 خوارزمية K- means clustering:.....28
3.9.2.2 خوارزمية CLARA clustering :.....29
11.2.2 تطبيقات خوارزمية K-means:.....29
3.2 الفصل الثالث :.....11
3.2 الدراسات السابقة :.....11
1.3.2 الدراسة الأولى : تحليل الآراء عن أيفون 6 :.....12
2.3.2 الدراسة الثانية : تحليل الآراء حول الطاقة النووية :.....13
3.3.2 الدراسة الثالثة : تحليل الآراء السلبية :.....14
4.3.2 الدراسة الرابعة :25Trends: 14

5.3.2 الدراسة الخامسة : موقع Twitalyzer:.....15
6.3.2 الدراسة السادسة :AsTweeted :.....16
7.3.2 الدراسة السابعة : تحليل آراء إستهداف الأراضي المقدسة :..... 17
الباب الثالث :.....29
الأدوات و التقنيات المستخدمة:.....29
1.3 المقدمة :.....33
2.3 لغة R والتحليل الإحصائي :.....33
1.2.3 مقدمة الي لغة ال R :.....33
2.2.3 ماهي لغة ال R :.....33
3.2.3 كيفة تثبيت لغة ال R :.....34
4.2.3 ما هي مميزات R :.....35
5.2.3 ماهي حدود ال R :.....36
6.2.3 إضافة حزم جديدة في ال R :.....36
7.2.3 أنواع المتغيرات في لغة ال R :.....37
8.2.3 المخططات البيانية الإحصائية في R :.....38
1.8.2.3 أنواع المخططات البيانيه في لغة R :.....38
9.2.3 الفرق بين ال R و Studio :.....42
3.3 MYSQL :.....42
1.3.3 مميزات MySQL :.....42
4.3 PHP :.....43
1.4.3 مميزات PHP :.....43
5.3 WampServer :.....43
1.5.3 مميزات WampServer :.....44
الباب الرابع :.....45
جمع البيانات Data Collection :.....45
1.4 المقدمة :.....46
2.4 إنشاء حساب في من موقع التواصل تويتر :.....46

3.4 المصادقة والتفويض في تويتر :.....46
1.3.4 ميزات برتكول اوث في تويتر :.....47
4.4 إنشاء تطبيق في تويتر :.....48
1.4.4 الصلاحيات بعد إنشاء تطبيق تويتر:.....48
5.4 الحصول على البيانات :.....50
1.5.4 عملية الطلب :.....51
2.5.4 نتيجة الطلب :.....51
3.5.4 نبذة عن ال JSPN :.....54
4.5.4 مميزات ال JSPN :.....55
6.4 تخزين البيانات في قاعدة بيانات :.....55
1.6.4 إنشاء قاعدة بيانات باستخدام MSQL :.....55
2.6.4 ربط PHP مع قاعدة البيانات MSQL :.....55
3.6.4 التحويل من تنسيق JSON الى مصفوفة PHP :.....55
4.6.4 إستخراج قيم المصفوفة :.....55
5.6.4 إدخال القيم الى قاعدة البيانات بكود PHP:.....56
الباب الخامس :.....57
التطبيق Implementation :.....57
1.5 المقدمة :.....58
2.5 ربط قاعدة البيانات بال R :.....58
3.5 إيجاد أكثر الكلمات تداولاً في التغريدات السوانيه :.....58
4.5 إنشاء الWordCloud للتغريدات السودانية :.....60
5.5 تطبيق خوارزمية Hierarchical clustering :.....61
6.5 تطبيق خوارزمية الk-means :.....62
7.5 تطبيق Topic modeling على التغريدات :.....70
الباب السادس :.....72
النتائج والتوصيات Results and Recommendations.....72
1.6 المقدمة:.....73

2.6 النتائج :.....73
3.6 التوصيات :.....73
الخاتمة :.....74
الملاحق :.....75
المراجع :.....100

المقدمة

الباب الأول

الإطار العام للبحث

1.1 المقدمة

التطور الذي شهدته تقنيات الشبكات الإجتماعية في القرن الواحد والعشرين أحدث نقله نوعيه في الاتصال البشري في عالمنا اليوم. فمع كثرت مرتاديها ومستخديمها، تجاوزت الشبكات الإجتماعية مرحلة الترفيه في أهميتها لتدخل حقبة جديدة في تأثيرها في مختلف الميادين كالسياسي والإجتماعي والإقتصادي والتسويقي .

تقنية تحليل الآراء هي وسيلة فعالة لإكتشاف الرأي العام ويمكن أن تسفر عن نتائج مثيرة للإهتمام فيمكن معرفة الرأي العام حول أي منتج أو خدمة أو شخصية.

يقوم المستخدمون بالتعبير عن كل مايجول في خواطرهم عن الطريق الكتابة والتصريح تجاه القضايا التي يتناولونها في مواقع الشبكات الاجتماعية. ومن هنا جاء تحليل الآراء كأحد الإجراءات التي يستخدمها الباحثون في مجال الإتصال الرقمي لقراءة مواقف وإتجاهات الرأي العام على تلك المنصه أي أنه إستراتيجية إتصالية للوصول الى مايتداوله المستخدمون على شبكات التواصل.
ومن أشهر شبكات التواصل الأجتماعي التويتر لذلك يعتبر تويتر مصدر غني بالمعلومات لإتخاذ القرارات وتحليل الآراء.

2.1 مشكلة البحث

الطرق التقليدية لمعرفة الرأي العام ليست فعالة لأنها تضم عدد أو فئة محدودة من الأشخاص بالإضافة الى أنه ليس هنالك ما يضمن مصداقيتهم في الإجابة كما أنها تحتاج الى إعداد وصياغة .

3.1 أهداف البحث

الهدف من هذا المشروع هو الإستفادة من البيانات المتاحة من مواقع التواصل الإجتماعي (تويتر) ومعالجة التغريدات السودانية وتصنيفها الى (سياسية , دينية , رياضية , ترفيهيه , أخرى) وإيجاد أكثر الكلمات تكراراً في التغريدات السودانية أكثر المواضيع تداولاً.

4.1 حدود البحث

هذا البحث يعتمد على وسائل التواصل الإجتماعي، وسيتم جمع البيانات من موقع التواصل الإجتماعي تويتر(twitter) ، وسيتم تحليل التغريدات باستخدام خوارزمية (K means) وخوارزمية (Hierarchical).

5.1 منهجية البحث .

تجميع التغريدات السودانية من موقع التواصل الإجتماعي(تويتر). تخزين التغريدات في قاعدة البيانات وذلك بعد معالجته. تحليل التغريدات وذلك من خلال إيجاد أكثر الكلمات تكراراَ، أكثر المواضيع تداولاً. وتصنيف التغريدات إلى (سياسية , دينية , رياضية , ترفيهيه ,اخرى) بإستخدام خوارزميات التجميع في الذكاء الإصطناعي.

6.1 هيكل البحث:

1.6.1 الباب الأول

يتناول هذا الباب مقدمة عن مشكلة و أهداف ومنهجية المشروع الذي سنتطرق له فى البحث ونبذة بسيطة عما سيتم تناوله لاحقا في البحث.

2.6.1 الباب الثاني

يتكون هذا الباب من ثلاثة فصول :الفصل الاول الأطار النظري ،الفصل التاني تجميع البيانات والفصل الثالث الدراسات السابقة.

3.6.1 الباب الثالث

يتناول هذا الباب التقنيات والأدوات التي تم إستخدامها لعمل المشروع من لغات البرمجة و الخوادم و صيغ البيانات المساعدة.

4.6.1 الباب الرابع

يتناول هذا الباب كيفية الحصول على البيانات من التويتر وتخزينها في قاعدة البيانات .

5.6.1 الباب الخامس

يتناول هذا الباب خوارزميات التجميع والتصنيف في الذكاء الإصطناعي.

6.6.1 الباب السادس

يتناول هذا الباب ما توصلنا إليه من نتائج و ما نقدمه من توصيات.

الملاحق و المراجع

الباب الثاني

الخلفية النظرية و الدراسات السابقه

الفصل الأول

الخلفية النظرية

1.2 المقدمه

البحوث والدراسات السابقه لها قدر كبير من الأهميه في إقناع القارئ بأن الباحث مطلع على مسبق من البحوث، لذلك سوف يحتوي هذا الباب على دراسات سابقه لها علاقه بهذا البحث، كما يحتوي على خلفيه نظريه عن البحث .

2.2 الخلفيه النظريه :

1.2.2 تعدين الآراء (Opinion mining)

هو إستخدام معالجة اللغات الطبيعية ، وعلم اللغة الحاسوبي و التحليل النصي من أجل الكشف عما يحمله النص من اراء سواء إيجابية أو سلبية أو محايدة تجاه موضوع النص. يستخدم تحليل الآراء على نطاق واسع على سبيل المثال في مجالات التسويق و خدمة العملاء و غيرها من المجالات الأُخرى[1].

2.2.2 الشبكات الإجتماعية (Social network)

الشبكات الإجتماعية هي مصطلح يطلق على مجموعة من المواقع على شبكة الإنترنت ظهرت مع الجيل الثاني للويب أو ما يعرف بإسم ويب 2.0 تتيح التواصل بين الأفراد في بيئة مجتمع إفتراضي يجمعهم حسب مجموعات إهتمام أو شبكات انتماء (بلد ، جامعة ، مدرسة ، شركة ... إلخ) كل هذا يتم عن طريق خدمات التواصل المباشر مثل إرسال الرسائل أو الإطلاع على الملفات الشخصية للآخرين ومعرفة أخبارهم ومعلوماتهم التي يتيحونها للعرض .

تصنّف مواقع الشبكات الإجتماعية ضمن مواقع الويب 2.0 لأنها بالدرجة الأولى تعتمد على مستخدميها في تشغيلها وتغذية محتوياتها. كما تتنوّع أشكال وأهداف تلك الشبكات الإجتماعية ، فبعضها عام يهدف إلى التواصل العام وتكوين الصداقات حول العالم وبعضها الآخر يتمحور حول تكوين شبكات إجتماعية في نطاق محدد ومنحصر في مجال معين مثل شبكات المحترفين وشبكات المصورين ومصممي الجرافكس.

كما أن الشبكة الإجتماعية هي موقع على شبكة الإنترنت يجمع الناس معاً للحديث وتبادل الأفكار والمصالح أو تكوين صداقات جديدة ، هذا النوع من التعاون وتبادل البيانات وكثيراً ما يشار إليه وسائل الإعلام كما الإجتماعي ، على عكس وسائل الإعلام التقليدية التي غالبا ما يتم إنشاؤه من قبل ما لا يزيد عن 10 أشخاص ، ومواقع وسائل الإعلام الإجتماعية تحتوي على المحتوى الذي تم إنشاؤه من قبل مئات أو حتى الملايين من الناس مختلفة.

وفيما يلي قائمة صغيرة من بعض أكبر الشبكات الإجتماعية المستخدمة اليوم :

- الفيسبوك : (facebook)(http://www.facebook.com) موقع الشبكات الإجتماعية الأكثر شعبية على شبكة الإنترنت . الفيسبوك هو مقصد للمستخدمين لإعداد صفحات الويب الخاصة بهم، والتواصل مع الأصدقاء، وتبادل الصور، والحديث عن ما يفعلونه ، الخ..
- ماي سبيس : (Myspace)(http://www.myspace.com) الموقع الذي كان الأكثر رواجاً قبل أن يدخل في منافسة شديدة مع الفيس بوك مؤخراً، وهو تطبيق يقدم شبكة تفاعلية بين الأصدقاء المسجّلين في التطبيق، ويمكّن المستخدمين من نشر الصور، وكتابة المدونات، ونشر الموسيقى ومقاطع الفيديو، وإرسال الرسائل.
- تويتر :(Twitter) (http://www.Twitter.com) أحد الشبكات الأسرع نمواً . ويقوم تويتر بتقديم خدمة التدوين المصغر برسالة واحدة لا تتجاوز المائة و الأربعون حرف، المعروفة بإسم التغريدات، كما يمكنك متابعة الأشخاص الذين تعرفهم أو الذين كنت مهتما بهم، وتبادل الرسائل النصية.
- لنكد إن :LinkedIn (http://www.linkedln.com) شبكة إجتماعية للمحترفين. يضم الموقع قرابة مليونين محترف ومحترفة في مجالات متنوعة ومختلفة يتشاركون في مجموعات إهتمام . الخاصية المتميزة في الموقع هي خاصية التزكيات ، فبإمكان مديرك أو زملائك السابقين في وظيفة معينة عملت بها تزكيتك عن عملك في الشركة[2] .

3.2.2 تويتر

تويتر (Twitte) أحد أشهر شبكات التواصل الإجتماعية و وسائل التواصل الإجتماعي ، ويقدم خدمة التدوين المصغر والتي تسمح لمستخدميهِ بإرسال (تغريدات)عن حالتهم أو عن أحداث حياتهم أو إبداء أراءهم بحد أقصى 140 حرف للرسالة الواحدة . وذلك مباشرة عن طريق موقع تويتر أو عن طريق إرسال رسالة نصية قصيرة SMS أوعن طريق برامج المحادثة الفورية أو التطبيقات التي يوفرها الموقع .

وتظهر تلك التحديثات في صفحة المستخدم، ويمكن للأصدقاء قراءتها مباشرة من صفحتهم الرئيسية أو زيارة ملف المستخدم الشخصي وأصبح موقع تويتر متوفر باللغة العربية منذ مارس 2012.

يتميز تويتر بأنه يسمح للمستخدم بأن ينشر آراءه و أفكاره، فموقع تويتر ليس مجرد موقع للتعارف فقط بل هو موقع للتدوين أيضاً، فانت في تويتر تعتبر مُدوّن تكتب آراءك و أفكارك لتناقشها مع الجميع[3] .

4.2.2 واجهة برمجة التطبيقات (API)

واجهة برمجة التطبيقات (Application Programming Interface) وتختصر ب (API) هي واجهة برمجيه تُقدم للمطورين إمكانية فهم التطبيقات وما تُقدمه هذه التطبيقات، والوصول لتلك التطبيقات وتبادل البيانات معها. فمثلاً المبرمج الذي يُريد أن يطور تطبيقاً يستخدم تويتر فعليه أن يَفهم الAPI الخاصة بتويتر ثم برمجة تطبيقه ليتفاعل مع الواجهة البرمجية لتويتر ويتبادل البيانات معها .

الهدف الرئيسي منها توفير قائمة من الوظائف المستقلة تماماً عن الآلية التي نُفذت منها، لتتيح للآخرين التواصل معها من خلال أي آلية أخرى.

فكرة (API) هي أن تَجلب البيانات والخدمات بتنسيقات تسمح لنا بإستخدامها مرة أخرى من أي مكان، وأشهر هذه التنسيقات هي JSON و XML .

5.2.2 واجهة برمجة التطبيقات لتويتر (Twitter API)

هي طريقة للوصول الى واجهات برمجية للتطبيقات لتَستطيع قراءة وكتابة البيانات من حسابك في تويتر، والتعرف على صاحب الحساب ومعلومات عن المتابعين . تويتر لديه ثلاثه واجهات برمجيه للتطبيقات هي :

- Search API
- REST API
- Streaming API

Search API 1.5.2.2

يَسمح للمستخدمين البحث في محتوى تويتر وإسترجاع التغريدات وفقاً لِشروط معينة مثل إسترجاع التغريدات بلغة معينة أو من بلد معين أو تحديد عدد معين للتغريدات أوالبحث عن كلمات محددة ضمن التغريدات، وإيجاد التغريدات التى تعود الى مستخدم معين أو العثور على تغريدة من مستخدم معين وهناك حالات يفضل فيها إستخدام (Streaming API) .

REST API 1.5.2.2

REST هي إختصار (Representational State Transfe) وهو نمط من بنية البرمجيات المُنسقة ، يتكون من مبادئ وتوجيهات قابلة للتطوير والمستخدمة لإنشاء خدمات الويب ، وهو مَبني بطريقة تُوافق مبدأ خدمات الويب (WebServices) .

سميت بهذا الاسم نظراً لأنها وسيلة لنقل المعلومات بين ال (client) و (server) عن طريق بروتوكلات ال(HTTP)، فجميع العمليات تتم من خلال هذه البروتوكولات بإختلاف أنواعها ومبادئها .

يتكون ال(Rest API) من مجموعة من ال(Services) وكل واحد منها يعمل بشكل منفصل ويُمكنها الإرتباط لتعمل كمجموعة واحدة.

يُوفر للمبرمجين الوصول إلى بيانات موقع التواصل الإجتماعي تويتر(Twitter) لقراءتها مع إمكانية كتابة البيانات. أيضاً يُوفر قراءة الملف الشخصي لمنشئ التغريده المعينه و متابعين البيانات وأكثر من ذلك.

(Rest API) يحدد تطبيقات تويتر (Twitter) والمستخدمين بإستخدام (OAuth)، والإستجابه للطلبات تتوفر في شكل (JSON) .

Streaming API 3.5.2.2

هو على عكس (Search API) حيث يتم فيها الحصول علي البيانات من التغريدات التي حدثت بالفعل، يتم فيها الحصول علي البيانات من التغريدات التي حدثت في الوقت الحقيقي القريب.
مع (Streaming API) يُمكن للمستخدمين تسجيل مجموعه من المعايير (كلمات مفتاحيه ، أسماء المستخدمين، المواقع، أسماء الأماكن وألخ ..)، إذا طابقت معايير التغريده المطلوبه سيتم دفعها مباشره للمستخدم[4] .

الفصل الثانـي

تجميع البيانـات

(Data Clustering)

6.2.2 مقدمة

أيضاً وكما هو معروف إن عملية تجميع البيانات في سجلات بأحجام كبيرة مثل (Data Warehouse) قد تحتوي علي معلومات كبيرة وقد تكون مفيدة لأصحابها من خلال معرفة العديد من الإحصاءات المطموبة مستقبلاً. لذلك نحتاج تقنيات مختلفة للحصول على هذة المعلومات. فالتقنيات العادية قد لا تكون مفيدة وغير عملية في مثل هذة الأحجام الكبيرة من البيانات، لذلك وجدت تقنيات التنقيب في البيانات التي تستخدم الذكاء الإصطناعي لتسهيل وتحسين عملية البحث الإستنباط. ومن الطرق الهامة المستخدمة في التنقيب طريقة العنقدة Clustering التي تفيد في تجميع البيانات المتماثلة في عناقيد يسهل التعامل معها وإجراء البحث ضمنياً للوصول إلى الهدف المنشود[12].

7.2.2 تجميع البيانات (Data Clustering)

تجميع البيانات هي عملية وضع البيانات في مجموعات متشابهه، وهي أحد فروع التنقيب عن البيانات. تقسم خوارزمية التجميع مجموعة بيانات إلى عدة تجمعات، حيث أن التشابه بين النقاط ضمن تجمع معين أكبر من التشابه بين نقطتين ضمن تجمعين مختلفين. فكره تجميع البيانات بسيطة في طبيعتها وقريبة جدَاً من الإنسان في طريقة تفكيره حيث أننا كلما تعاملنا مع كمية كبيرة من البيانات نميل إلى تلخيص الكم الهائل من البيانات إلى عدد قليل من المجموعات أو الفئات، وذلك من أجل تسهيل عملية التحليل[13].

8.2.2 طرق تجميع البيانات

يمكن تصنيف طرق التجميع الى الفئات التالية :

- طريقة التقسيم (Partitioning Method)
- الطريقة الهرمية (Hierarchical Method)
- الأسلوب القائم على الكثافة(Density-based Method)
- الأسلوب القائم على الشبكة(Grid-Based Method)
- الأسلوب القائم على نموذج (Model-Based Method)
- الأسلوب القائم على القيد(Constraint-based Method)

1.8.2.2 طريقة التقسيم(Partitioning Method)

لنفترض انه لدينا مجموعة من البيانات "N" وعدد مجموعات التقسيم "K " وكل قسم يمثل العنقودية بحيث $k \leq n$ وهذا يعني أنه سوف يتم تصنيف البيانات في مجموعات K، والتي تلبي المتطلبات التالية:

- تحتوي كل مجموعة كائن واحد على الأقل.
- كل كائن يجب أن ينتمي إلى مجموعة واحدة بالضبط.

2.8.2.2 الطريقة الهرمية(Hierarchical Method)

هذا الأسلوب يقوم بإنشاء التحليل الهرمي لمجموعة معينة من البيانات. ويمكن تصنيف الأساليب الهرمية على أساس كيفية تشكيل التحليل الهرمي .هناك طريقتان لتقسيم البيانات بالطريقة الهرمية:

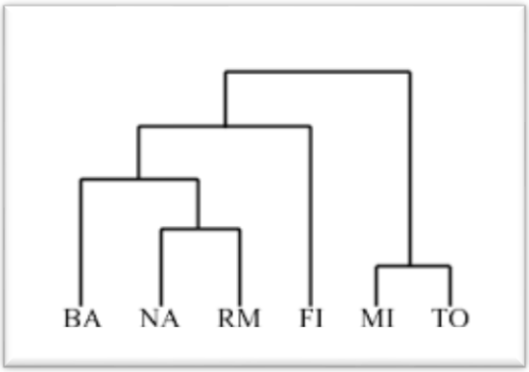
- Agglomerative Approach

ويعرف هذا النهج أيضا بإسم نهج من أسفل إلى أعلى. يتم البدأ مع نقطة من المجموعات الفردية، وفي كل خطوة يتم دمج أقرب زوج من المجموعات.

- Divisive Approach

ويعرف هذا النهج أيضا بإسم نهج من أعلى إلى أسفل . يتم البدء مع كافة الكائنات في نفس المجموعة .

يستمر في التكرار، يتم تقسيم المجموعة إلا أن تصل إلى مجموعات أصغر.

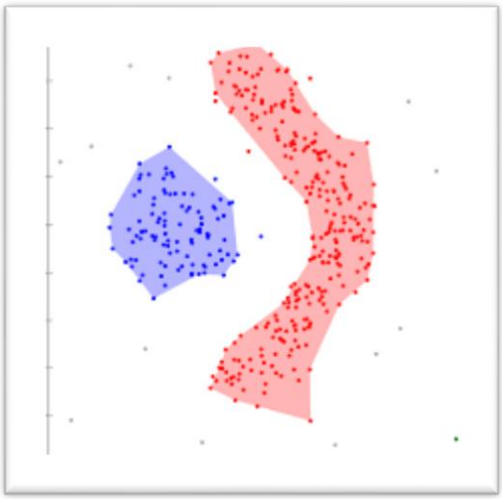


الشكل(2.8.2.2) الطريقة الهرمية لتجميع البيانات

3.8.2.2 الأسلوب القائم على الكثافة(Density-base Method)

ويستند هذا الأسلوب على فكرة الكثافة. الفكرة الأساسية هي أن يستمر النمو في مجموعة بعينها طالما أن

الكثافة لم تتجاوز حد معين، أي لكل نقطة من نقاط البيانات داخل مجموعة معينة، في دائرة نصف قطرها من المجموعة المعينة يجب أن تحتوي على ما لا يقل عن الحد الأدنى من النقاط.



الشكل(3.8.2.2) يمثل الاسلوب القائم على الكثافة لتجميع البيانات

4.8.2.2 الأسلوب القائم على الشبكة(Grid-Based Method)

يتم تحديد مجموعة من شبكة الخلايا وتعيين البيانات (الكائنات) إلى خلية الشبكة المناسبة وحساب الكثافة من كل خلية. القضاء على الخلايا، ذات الكثافة الأقل في المجموعة معينة. تشكيل المجموعات من جماعات متجاورة (المجاورة) من خلايا كثيفة .

5.8.2.2 الأسلوب القائم على نموذج(Model-Based Method)

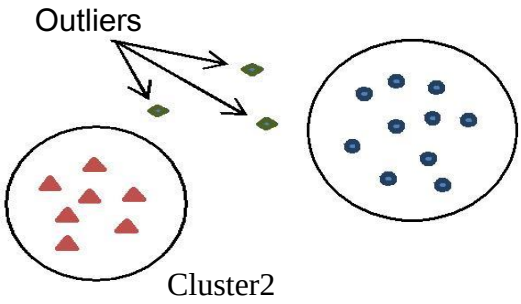
في هذه الطريقة يتم إفتراض نموذجاً لكل مجموعة للعثور على أفضل تناسب للبيانات لنموذج معين. هذه الطريقة تضع المجموعات عن طريق تجميع دالة الكثافة. وهو يعكس التوزيع المكاني للنقاط للبيانات. يوفر هذا الأسلوب أيضاً وسيلة للتحديد التلقائي لعدد من المجموعات استناداً إلى الإحصاءات القياسية، بأخذ الضوضاء في الاعتبار. وبالتالي فإنه ينتج أساليب تجميع قوية.

6.8.2.2 الأسلوب القائم على القيد(Constraint-based Method)

في هذه الطريقة يتم تجميع البيانات عن طريق دمج المستخدم أو قيود تطبيق المنحى. ويشير القيد إلى توقعات المستخدم أو خصائص نتائج التكتل المنشود .القيود توفر لنا طريقة تفاعلية للتواصل مع عملية التجميع .القيود يمكن أن تكون محددة من قبل المستخدم أو شرط التطبيق.

9.2.2 خوارزميات العنقدة أو التجميع

العنقدة أو التجميع هي عملية تجميع للعناصر المتشابهة على شكل عناقيد Clusters حيث يتم تجميع جميع العناصر المتشابهة ضمن عنقود واحد له خصائص معينة يختلف من خلالها عن باقي العناقيد الأخرى وقد يكون هناك عناصر شاذة لا يمكن أن تنتمي إلى اي تجمع. يوضح الشكل التالي عملية العنقدة أو التجميع[12]:



الشكل (9.2.2) العنقدة أو التجميع

طرق التقسيم او التجميع للبيانات :

تقوم بتقسيم البيانات المؤلفة من (n) كائن (غرض) الى (k) قسم حيث يمثل كل قسم عنقود فهي تجمع البيانات في k عنقود التي تحقق المتطلبات التالية :

- كل عنقود يجب ان يحتوي على الاقل يحتوي كائناً واحداً .
- كل كائن (غرض) يجب ان ينتمي الي عنقود واحد فقط.

هناك عدد من الخوارزميات المستخدمة في عملية تجميع البيانات أو تقسيم البيانات، ومن هذه الخوارزميات التي سوف يتم الحديث عنها بشكل مفصل:

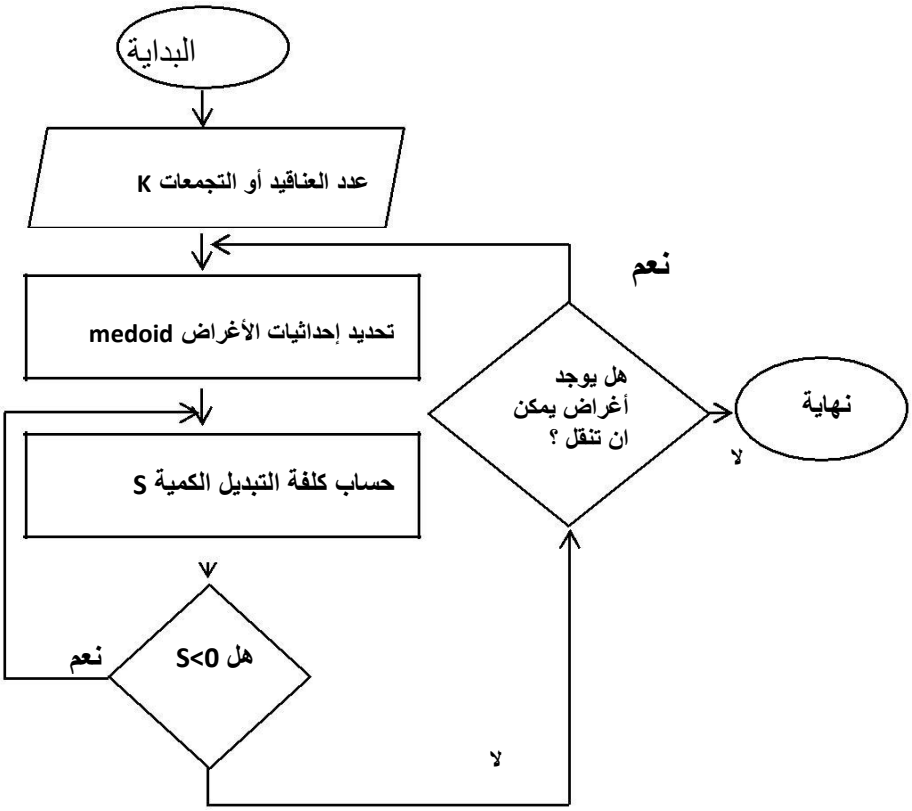
- خوارزمية (K-Medoids Clustering)
- خوارزمية (K-means Clustering)
- خوارزمية CLARA

1.9.2.2 خوارزمية (K-Medoids Clustering)

تقوم هذه الخوارزمية على مبدأ تقسيم قاعدة البيانات D التي تحتوي N سجل إلى K عنقود (مجموعة) بحيث يتم تقليل المسافات بين البيانات ومراكز العناقيد Cluster Centroid [2] .

Medoids : وهي عبارة عن أغراض ممثّلة بمجموعة من البيانات أو عنقود بمجموعة من البيانات والتي يكون فيها متوسط الإختلاف لكل الأغراض في العنقود أقل ما يمكن.

تستخدم هذة الخوارزمية لعنقدة أو تجميع بعض البيانات اعتماداً على خصائصها إلى K عنقود، حيث تتم عملية العنقدة من خلال تقليل المسافات بين الأغراض ومركز العنقود وكل تجمع يمثل بأحد الأغراض الموجودة فيه، ويوضح المخطط (1.9.2.2) عمل الخوارزمية :



رسم بياني (1.9.2.2) لعمل خوارزمية ال K-Medoids

أما خطوات هذه الخوارزمية فهي على النحو الآتي :

1. تحديد عدد التجمعات K. وهي تعتبر خطوة تهيئة أولية.
2. تحديد إحداثيات الأغراض medoids عشوائياً.
3. حساب كلفة تبديل الغرض medoid القديم O بالغرض الجديد'O. بإستخدام معيار الخطأ التربيعي الذي يعطى بالعلاقة :

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left\| p - m_i \right\|^2$$

يعبر m عن المسافة بين medoid والغرض المختار P. ويتم حساب كلفة التبديل S بالعلاقة:

$$S = E_{o'} - E_o$$

4. إذا كان S<0 تقوم الخوارزمية بعملية التبديل لتشكيل التجمع الجديد.
5. تكرار الخطوات من 2 إلى 4 حتى يتم الوصول إلى وضع ثابت لا يتغير.

❖ مزايا خوارزمية K- Medoids :

تمتاز خوارزمية K- Medoids بالمزايا الآتية :

- فعالة جدًا في حجم البيانات الصغيرة.
- تتعامل مع النقاط الشاذة تعاملاً جيداً.
- التغلب على العشوائية في أختيار النقاط الأولية.

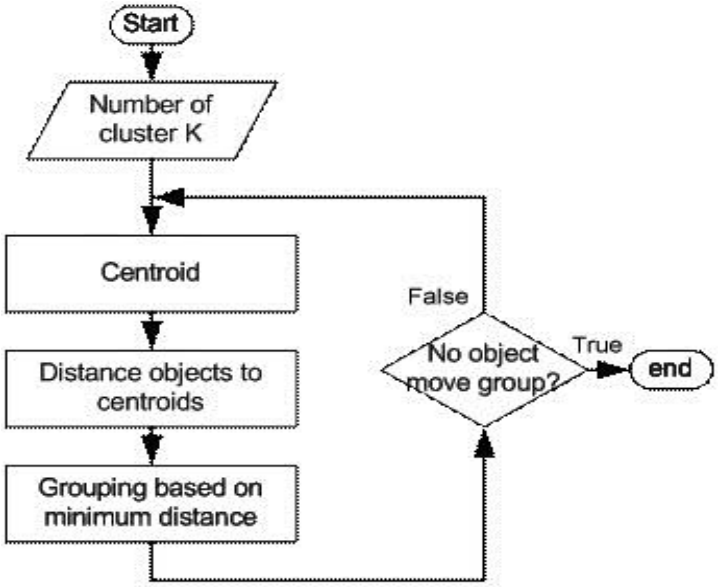
❖ مساوئ خوارزمية K- Medoids :

- بطيئة في حجم البيانات الكبيرة.
- لا تناسب الأحجام الكبيرة للبيانات.
- درجة التعقيد في كل تكرار كبيرة نسبياً وتقدر بـ($O(k(n-k)^2$,تمثل n حجم البيانات و k عدد العناقيد.
- درجة تعقيد إيجاد مصفوفة المسافات في كل مرة تقدر بـ($O(n^2$, حيث تمثل n حجم البيانات. وهي درجة تعقيد عالية نسبياً وتؤدي لإشغال حجم كبير في الذاكرة[12] .

2.9.2.2 خوارزمية K-Means Clustering

تستخدم هذه الخوارزمية لتجميع عدة بيانات(أمثلة) إعتماداً على خصائصها إلى K تجمع، وتتم عمليه التجميع من خلال تقليل المسافات بين البيانات ومركز التجمع (centroid cluster).

يبين الشكل التالي خوارزمية K-Means Cluster



رسم بياني (2.9.2.2) لعمل خوارزمية الk-mesns

أما خطوات هذه الخوارزمية فهي :

1. تحديد عدد التجمعات K، وهي تعتبر خطوة تهيئة أولية.
2. تحديد إحداثيات مراكز التجمعات Centroid عشوائياً لأول مرة ويتم حسابه عن طريق إيجاد (متوسط النقاط التي تنتمي للمركز) لباقي المرات.
3. حساب المسافة بين كل مثال وبين جميع المراكز، ويتم استخدام البعد الإقليدي. يعطى البعد الإقليدي d_{ij} بين مثالين i,j بالعلاقة التالية :

$$(1)\ d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

حيث إن :

n: عدد خصائص المثال.

x_{ik} : إحداثيات الخاصية k للمثال i.

x_{jk} : إحداثيات الخاصية k للمثال j (يكون في العادة إحداثيات المركز).

4. تجميع البيانات (الأمثلة) مع أقرب مركز لها.

5. تكرار الخطوات من 2 إلى 4 حتى حصول الاستقرار (عدم وجود كائنات تنتقل ضمن التجمعات) أو حتى التكرار عدد معين من المرات.

يعتمد أداء هذه الخوارزمية على المواقع الأولية لمراكز التجمعات (Centroid)، ومن المستحسن تنفيذ هذه الخوارزمية عدة مرات مع إختلاف المراكز في كل مرة عن المرات السابقة.

مزايا خوارزمية K-Means Clustering:

1. ذات فعالية عالية.
2. سهلة التنفيذ .
3. تتعامل مع القيم المستمرة والقيم المتقطعة (الأسمية) .

مساوئ خوارزمية K-Means Clustering:

1. تحديد عدد التجمعات K يتم عشوائياً وقبل معالجة الأمثلة.
2. حساسة للحالة الأولية. يؤدي اختيار حالات أولية متعددة إلى إعطاء نتائج مختلفة للتجمعات. ونتيجة لذلك قد تقع الخوارزمية في مشكلة النهاية المحلية.
3. شكل التجمع دائري لأنه يعتمد على حساب المسافة[13].

3.9.2.2 خوارزمية CLARA

ألية عمل هذا الخوارزمية تقوم على التالي : بدلاً من تطبيق عملية التجميع على كامل مجموعة البيانات يتم أخذ عينة تعبر عن البيانات و من ثم تطبيق خوارزميات (K-means, K-mediods) و بعد الإنتهاء يتنج لدنيا مراكز العناقيد الملطوبة و من ثم تتم عملية ضم باقي البينانات الموجودة في قاعدة البيانات إلى العناقيد المناسبة و إن خوارزمية (CLARA) تستطيع التعامل مع مجموعات من البيانات أكبر بكثير من خوارزميات (K-means,K-medoids) و إن فعالية و كفاءة خوارزمية (CLARA) تعتمد على حجم العينة .

تعقيد هذه الخوارزمية هو :

$$O\left(ks^2 + k\,(n-k)\right)$$

حيث s هو حجم العينة و k هو عدد العناقيد و n هو العدد الكلي للكائنات.

11.2.2 تطبيقات خوارزمية K-means Clustering

تعتبر خوارزمية K-Means Clustering من أبسط خوارزميات التجميع وأكثرها فعالية، وهناك الكثير من تطبيقاتها، أهمها :

- التعليم بدون إشراف Unsupervised learning للشبكات العصبية.
- تمييز النماذج Pattern recognitions.
- التصنيف Classification.
- معالجة الصورة Image processing.
- الرؤية بواسطة الآلة Machine vision.
- الذكاء الصنعي Artificial intelligent.

وبشكل أساسي، إذا كان هناك عدة أمثلة، وكل مثال يملك عدة خصائص، ونرغب بتقسيم الأمثلة وفق خصائصها، يمكن تطبيق هذه الخوارزمية.

نبين فيما يلي مبدأ تطبيق خوارزمية K-Means Clustering على تعلم الآلة machine learning أو التنقيب عن البيانات:

يتم إسناد كل مثال اعتماداً على خاصية التجميع تلقائياً إلى أحد التجمعات، وهذا يدعى التعليم بدون إشراف unsupervised learning، لأن الخوارزمية تصنف المثال تلقائياً إعتماداً على شرط معطى (مثلاً أقرب مسافة إلى مركز التجمع). لا نحتاج إلى الإشراف على الخوارزمية بأن نقول لها أن عملية التصنيف صحيحة أو خاطئة. تعتمد عملية التعلم على أمثلة التدريب المقدمة للخوارزمية، ولدينا خياران في عملية التدريب هذه :

- تدريب غير منتهي: تعتبر جميع البيانات المقدمة إلى الخوارزمية تلقائياً أمثلة تدريب.
- التدريب المنتهي: بعد أن يتم إعتبار عملية التدريب منتهية (إعطاء معلومات صحيحة عن المتوسط). نبدأ عند ذلك بجعل الخوارزمية تصنف تجمع الأمثلة الجديدة. وهذا يتم ببساطة من خلال إسناد المثال الجديد إلى أقرب مركز تجمع دون إعادة حساب المراكز الجديدة. وفي هذه الحالة تكون مراكز التجمعات الناتجة بعد إنتهاء عملية التدريب مراكز ثابتة[13].

الفصل الثالث

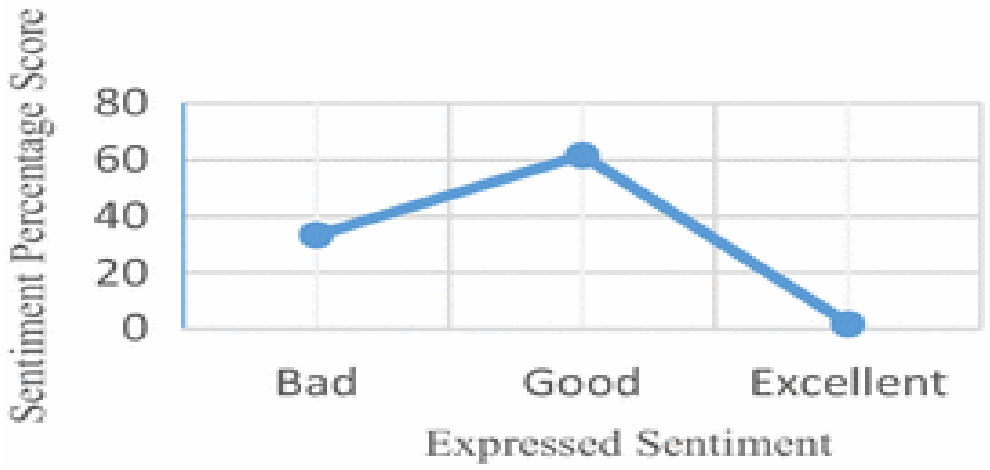
الدراسات السابقة

Previous Studies

3.2 الدراسات السابقه :

1.3.2 تحليل الآراء باستخدام تويترعن هاتف ايفون6

تم إجراء هذه الدراسه على سبعه مدن موجوده في الولايات المتحدة الأمريكية حيث كان عدد التغريدات التي تم جمعها من خلال واجهة برمجه التطبيقات لتويتر(Twitter API) 940 تغريدة، من تلك التغريدات 530 كانوا من الذكور و 410 كانوا إناث، وبعد تصفية التغريدات أصبحت 442 تغريدة التي لها علاقة بهذه الدراسة ، وبعد التحليل التام للآراء كانت النتيجه 60% من الاشخاص كانت آرائهم بأن هاتف ايفون6 هو منتج جيد بينما 40% كانت آرائهم أنه منتج غير جيد.



الشكل (1.3.2) نتيجة تحليل الأراء حول منتج أيفون6

اما نتيجه تحليل الآراء حول خصائص ايفون6 كانت آرائهم أن كاميرا ايفون6 جودة صورتها مماثله لكاميرات الDESL وإيضاً كانت آرائهم حول خاصية نظام الموسيقى أنه سهل في الاستعمال وإدارة المحتوى. الكاميرا ونظام الموسيقى كانوا من الخصائص المميزه في ايفون6 حسب نتيجه التحليل[5] .

2.3.2 تحليل الآراء على تويتر حول الطاقة النووية

الطاقة النووية هي قضية وطنية كبيره لأنها سلاح ذُو حدين , الكفاءه الإقتصادية لتوليد الطاقة (من أهم فوائد الطاقة النووية) و المخاطر المحتمله لتسرب الإشعاع (من أضرار الطاقة النووية)، لذلك بناء محطة نووية يحتاج الى موافقه من السكان في مساحه البناء، إستخدام الدراسة التقليدية لمعرفة الرأي العام حول الطاقه النووية تحتاج الى وقت، تحليل الآراء توفر وسيلة بديلة لمراقبة الرأي العام حول الطاقة النووية حيث تم إستخدام البيانات من تويتر المتعلقه بالطاقة النووية باللغة الكورية (أي التغريدات التي تحتوي على مصطلح

"نوويه " أو " طاقة نووية") بين عامي 2009 و2013 , وبعدها تم تصفية البيانات (اي إستبعاد البيانات التي ليس لها صلة بالطاقة النووية) و من تم إستخراج العبارات الإيجابية والسلبية من النص .

من 2009 الى 2011 كان عدد العبارات الإيجابية 1,012 وعدد العبارات السلبية 3,219 وهذا يعكس أن التغريدات السلبية هي الأكثر في تويتر حول الطاقة النوويه وذلك بسبب كارثة فوكوشيما النووية التي كانت في مارس 2010 والتي اظهرت تغير جزري في الرأي حول الطاقة النووية، ولتعقب التغيرات على الرأي العام في الطاقة النووية تم إقتراح قياس مؤشر الرأي شهرياً، والجدوال التالي(2.6.2) يوضح دقة النتائج بين عامي 2012 و2013 ,حيث كانت دقة التنبؤ على تويتر الإيجابية والسلبية أكثر من 50% [6] .

Sentiment	2012		2013	
	Accuracy Rate	No .of Tweets	Accuracy Rate	No .of Tweets
Positive	51.58%	948	50.55%	991
Negative	61.19%	2067	64.08%	2289
Neutral	38.96%	2066	21.67%	1375

الجدول (2.3.2) يوضح نتائج تحليل الآراء حول الطاقة النووية بين عامي 2012 و2013

3.3.2 تحليل التغريدات السلبية على تويتر مؤشر للإصابة بأمراض القلب

هذة الدراسة أعدها فريق من الباحثين في "جامعة بنسلفانيا" ، وحلل فيها عالم النفس ، يوهانس أيكستت ، وزملاؤه ، تحتوي على 826 مليون تغريدة في «تويتر» ، كتبها أشخاص ينتمون إلى 1400 مقاطعة أميركية ، بين عامي 2009 و 2010 تضم نحو 90% من تعداد الولايات المتحدة وتم ربطها بمعلومات حول الوفاة بسبب أمراض القلب .و من أجل تحليل هذه المعلومات وربطها بِالإصابة بأمراض القلب ، قام الفريق بِتحليل لغة التغريدات للمساعدة في التنبؤ بالوفيات الناجمة عن الإصابة بأمراض القلب .

حيث شهدت المقاطعات ـ التي تضمنت تغريدات سكانها كلمات تتصل بالخصومة والعدائية والكراهية والإرهاق، مثل "أحمق" و"غيور" و"ضَجِر" ـ معدلات أعلى بشكل ملحوظ من الوفيات الناجمة عن تصلب

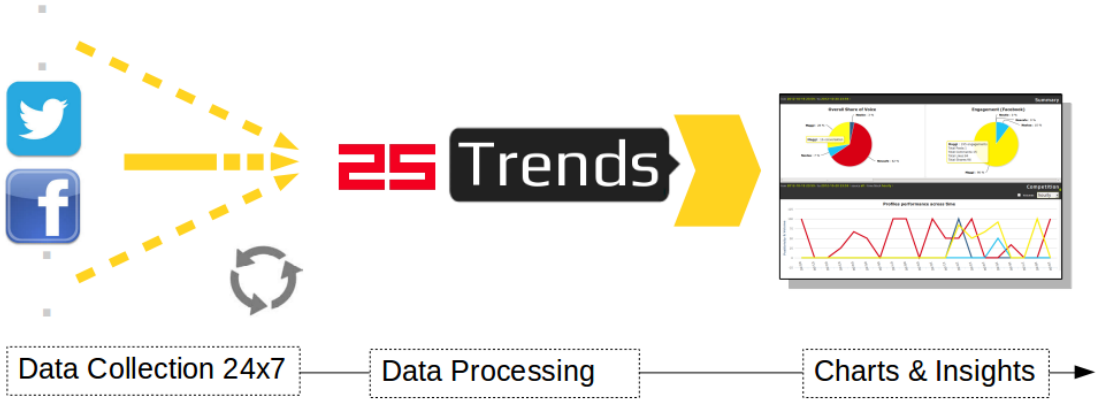
الشرايين، ومن ذلك النوبات القلبية والسكتات .وعلى النقيض من ذلك , كانت أمراض القلب أقل شيوعاً في المناطق التي عكست تغريدات سكانها مشاعر أكثر إيجابية ومشاركة . وأثبت نموذج التنبؤ المعتمد على «تويتر» تفوقه على النماذج الأخرى المستندة إلى العوامل الديموغرافية والاجتماعية والاقتصادية والمخاطر الصحية.

وعلى الرغم مما هو معروف منذ فترة طويلة عن إسهام الإجهاد والغضب والوحدة في زيادة مخاطر الإصابة بالنوبات القلبية وغيرها من أمراض القلب المسببة للوفاة ، فإن ذلك لا يُقلل من غرابة نتائج الدراسة، مع تحذير الباحثين من أن « الأشخاص الذين يكتبون التغريدات ليس هم من يموتون " وذلك حسب ما وضحه العلماء أن المشاعر السلبية تؤدي الى سلوكيات إجتماعية وسلوكية مختلفة ، حيث يصبح الشخص أقل قدرة على تناول الطعام وينعزل عن الآخرين، مما يؤدي بدوره الى رفع خطر الاصابة بامراض القلب" »، الأمر الذي يعني أن تغريدات الشخص لا تتنبأ بخطر تعرضه لأمراض القلب، بل إن التغريدات السلبية الجماعية في أجزاء معينة من البلاد تتوافق مع معدلات مرتفعة من الوفيات فيها[7] .

4.3.2 25Trends أول أداة مصرية لتحليل المشاركات الإجتماعية

تعد Trends25 من أولى الأدوات التي تم تطويرها في مصر والتي تعمل على تحليل ومراقبة المشاركات الإجتماعية على مواقع التواصل الإجتماعي مثل فيسبوك وتويتر ويوتيوب. وهي بهذا حصلت على الريادة في سوق صاعد وواعد جداً في الدول الغربية وينتظر له النجاح والصعود في الوطن العربي بشكل أكبر خلال الفترة القادمة. تأتي أهمية مراقبة وتحليل مشاركات المستخدمين سواء للشركات أو في المجال السياسي من كونها تساعد على تحديد الموضوعات الأكثر رواجاً بين مستخدمي الشبكات الإجتماعية ، والتعرف أيضاً على آرائهم وتحليلها ما بين سلبي وإيجابي ومحايد وهو ما يسمى بتحليل المشاعر أو sentiment analysis.

وتتيح الخدمة المجانية بTrends25 للمستخدم أن يبحث عن هاشتاج أو كلمة مفتاحية معينة على تويتر ليرى نتائج البحث مصنفة حسب الطبيعة (سلبي – إيجابي – محايد) وحسب اللغة وحسب الروابط الأكثر انتشاراً والتغريدات الأكثر حصولاً على إعادة تغريد، بالإضافة إلى أكثر المستخدين استخداماَ لهذا الهاشتاج أو الكلمة التي بحثت عنها . أما الخدمة المجانية الآخرى فهي Twitter Trends أو إتجاهات تويتر، والتي تمكن المستخدم من معرفة الموضوعات الرائجة بحسب البلد، والمُمَيز هنا أنك تستطيع الإختيار ما بين البلدان العربية كمصر والسعودية والإمارات والأردن والكويت وغبرها , وهو ما لا تجده في موقع تويتر نفسه[8].



الشكل (4.3.2) أداة 25Trends

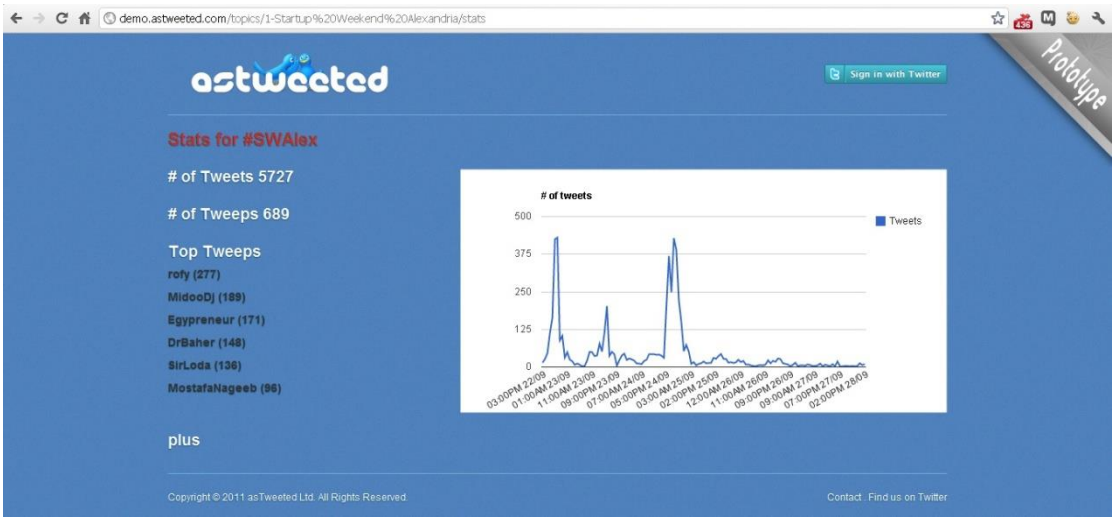
5.3.2 موقع Twitalyzer خدمة لتحليل محتوى الشبكات الإجتماعية

Twitalyzer وهو موقع إلكتروني يقوم بتحليل نتائج البيانات الموجودة على صفحة تويتر ، ويستند هذا الموقع إلى عدة مميزات، تجعل تحليل البيانات تحليلاً متقدماً ودقيقاً . فهذا الموقع يقوم بعرض كافة التحليلات بصورة سهلة القراءة ، إذ أنه يقدم نسباً مئوية للعوامل مثل : الأثر، والمشاركة، والتأثير، ووفرة المعلومات، والنفوذ . كما أن الموقع يقوم بعرض النتائج على شكل رسوم بيانية تفاعلية ، تستند إلى الجاذبية في المنظر، كما أن أدوات التحليل لا تشتمل على تحليل بيانات تويتر فحسب ، بل يمكن أيضا إدماج أدوات أخرى لتحليل مضمون الشبكات الإجتماعية بشكل عام .والسبب الأخير الذي يجعل هذا الموقع مفضلاً لدى الكثيرين هو أن كافة المعلومات والأدوات الموجودة فيه هي بالمجان[9] .

6.3.2 AsTweeted أول أداة مصرية لأرشفة التغريدات

تُعد AsTweeted من الأدوات المجانية الأفضل في مجال أرشفة أو حفظ التغريدات حيث تتيح لك تتبع موضوع معين عن طريق هاشتاج أو كلمة مفتاحية ، مثل #عرب_نت، وهي تتميز بأنها "طُورت في مصر".

ومن مميزات هذه الأداة أنها تتيح لك حفظ التغريدات وفتحها لاحقاً لقراءتها أو إستخدامها في أي وقت، وهو ما لا يتيحه تويتر نفسه حيث يسمح لك بالرجوع في البحث لمدة أسبوع فقط . وقد أضافت AsTweeted مؤخراً ميزة التحليل ، بالإضافة للتتبع والأرشفة، حيث يمكن للمستخدم معرفة عدد التغريدات، وعدد المغردين، وأكثر المغردين نشراً لهذه الكلمة أو الهاشتاج الذي تتبعه. بالإضافة لتحليل الآراء ما بين سلبي وإيجابي[10] .



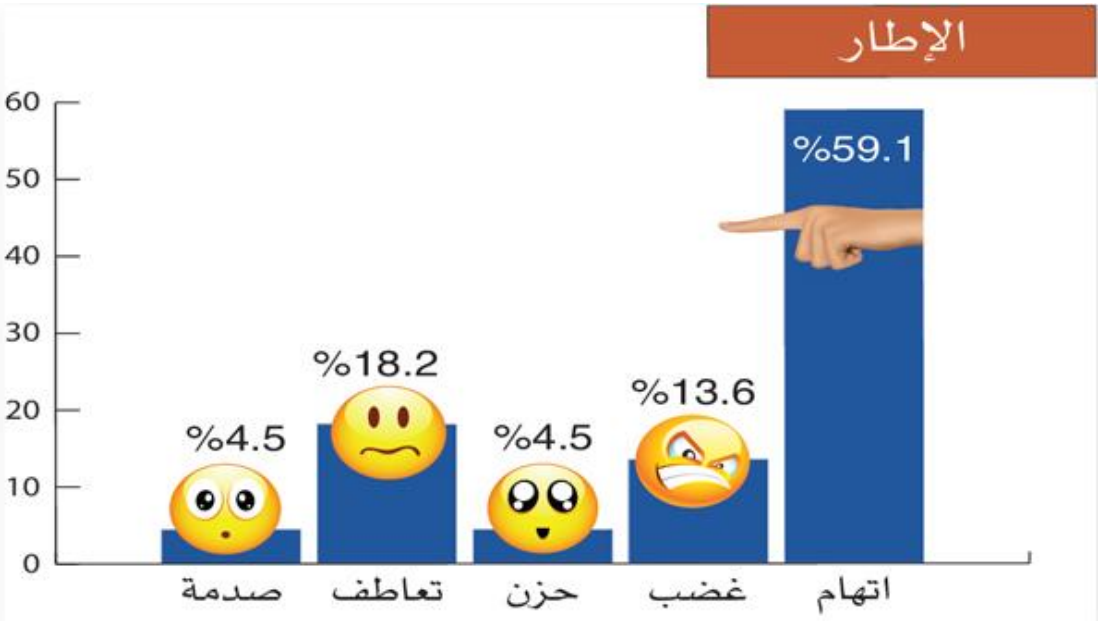
الشكل (6.3.2) أداة AsTweeted

7.3.2 تحليل للرأي العام الإلكتروني عقب إستهداف الأماكن المقدسة

هي دراسة لتوجهات الرأي العام العربي في مواقع التواصل الإجتماعي لحادثة التفجير التي وقعت بقرب المسجد النبوي الشريف بالمدينة المنورة والتي تم وصفها بأنها إنتهاك للأراضي المقدسة . تم إعداد هذه الدراسة لتحليل ورصد تعليقات مستخدمي موقع التواصل الاجتماعي (تويتر) حول حادثة تفجير المدينة المنورة لتحليل تعليق الرأي العام العربي على هذا الحادث وكيف تم وصف القائم به.

وبينت النتائج أن أكثر من ثلث المستخدمين (36.4%) وصفوا ما حدث أنه تفجير، بينما ربط ربع المستخدمين 25% الحادث بداعش و20.5% وصفوا الحادث بالإرهاب . وأظهرت الدراسة أن 84.1% من المغردين تحدثوا عن نتائج التفجير فيما غرد 4.5% فقط عن الحادث نفسه .كما بينت أن غالبية المغردين (51.1%) إستخدموا إطار الإتهام في وصف الحادث و18.2% إطار التعاطف مع المملكة.

من جانب آخر، أشارت النتائج أن 29.3% من المغردين يشيرون الى أن سبب التفجير بعض الحكومات العربية بينما أشار 22% أن هناك مؤامرة ما وراء التفجير .غَطت الدراسة نشاط المستخدمين خلال عشر أيام من التفجير وإستَخلصت الآف التغريدات المتعلقة بالحادثة بطريقة تحليل البيانات الكبيرة (big data) وذلك لأخذ عينات بأسلوب أكاديمي لتحليل إتجاهات الرأي العام بمجمله حول آراء المغردين العرب التي تم التعبير عنها الكترونياً . ومن جانب آخر عبر 4.5% عن حزنهم بما حدث و4.5% كانوا تحت تأثير الصدمة خصوصاً أن إستهداف الارهاب للمدينة المنورة والمسجد النبوي وفي الشهر الكريم كان صادماً للمغردين وعملية دموية يصعب تفسيرها[11] .



الشكل (7.3.2) نتيجة تحليل الرأي العام الإلكتروني عقب إستهداف الأماكن المقدسة

الباب الثالث

الأدوات والتقنيات المستخدمة

(Tools and Techniques)

1.3 المقدمه

في هذا الفصل سنتناول وصف للغات البرمجة والأدوات والتقنيات المستخدمة لتنفيذ المشروع مع بيان خصائصها ومميزاتها .

2.3 لغة R والتحليل الإحصائي

1.2.3 مقدمة الى لغة R

تعد لغة R من اللغات التي صعد نجمها حديثا وبشكل سريع بمجال البرمجة العلمية في قطاعي الإحصاء والمعلوماتية الحيوية (bioinformatics) حيث باتت معتمدة على نطاق واسع في كثير من الجامعات ومراكز البحث العلمية، وأصبحنا نرى استخدامها والإشارة إليها في المقالات المنشورة بالمجلات العلمية المحكّمة يزداد بشكل طردي ومتسارع، هذا عدى عن حقيقة كونها لغة حرة مفتوحة المصدر يخضع توزيعها لترخيص GPL الشهير. كل ذلك أدى إلى تزايد ما هو متوفر ومتاح على الشبكة (الإنترنت) من مصادر لها على توزع طيف تلك المصادر، فهناك الكتب الإلكترونية والدروس التعليمية وحتى المناهج الأكاديمية والدورات التدريبية إضافة إلى البرامج الجاهزة والمكتوبة بلغة R لتنفيذ هذه المهمة أو تلك، حتى أنها باتت تحظى ببعض الإمتياز مقارنة بالعديد من العمالقة في قطاع البرمجة الرياضياتية العلمية والإحصائية مثل SAS و SPSS خصوصاً في مجال توافر الجديد من الطرق والخوارزميات الحديثة، حيث يقاد هذا التوجه في معظمه من طرف الجامعات ممثلة بطلاب الدراسات العليا يحفّزهم على ذلك سهولة بناء الإضافات لهذه اللغة، ويعتبر هذا الأسلوب رغم ما قد يشوبه من نقاط ضعف تتعلق بموثوقية وجودة وغزارة تلك الإضافات الجديدة، والتي تتبع خبرة ومهارة مطوريها وناشريها، لكنها تبقى في القطاع العلمي والأكاديمي أفضل كثيرا من البدائل التجارية التي يعيبها إرتفاع ثمنها من جهة، ومن جهة أخرى بطئ إضافة التحديثات التي تعكس تطور القطاعات العلمية المختلفة، حيث أنها عادة ما تتبع دورة تجارية تتحكم بها الشركات المنتجة.

2.2.3 ماهي لغة R

الـ R كما يعرفها موقعها الرسمي " : عبارة عن مجموعة متكاملة من البرمجيات التي تسمح بمعالجة البيانات, القيام بعمليات حسابية و إظهار البيانات الرسومية."

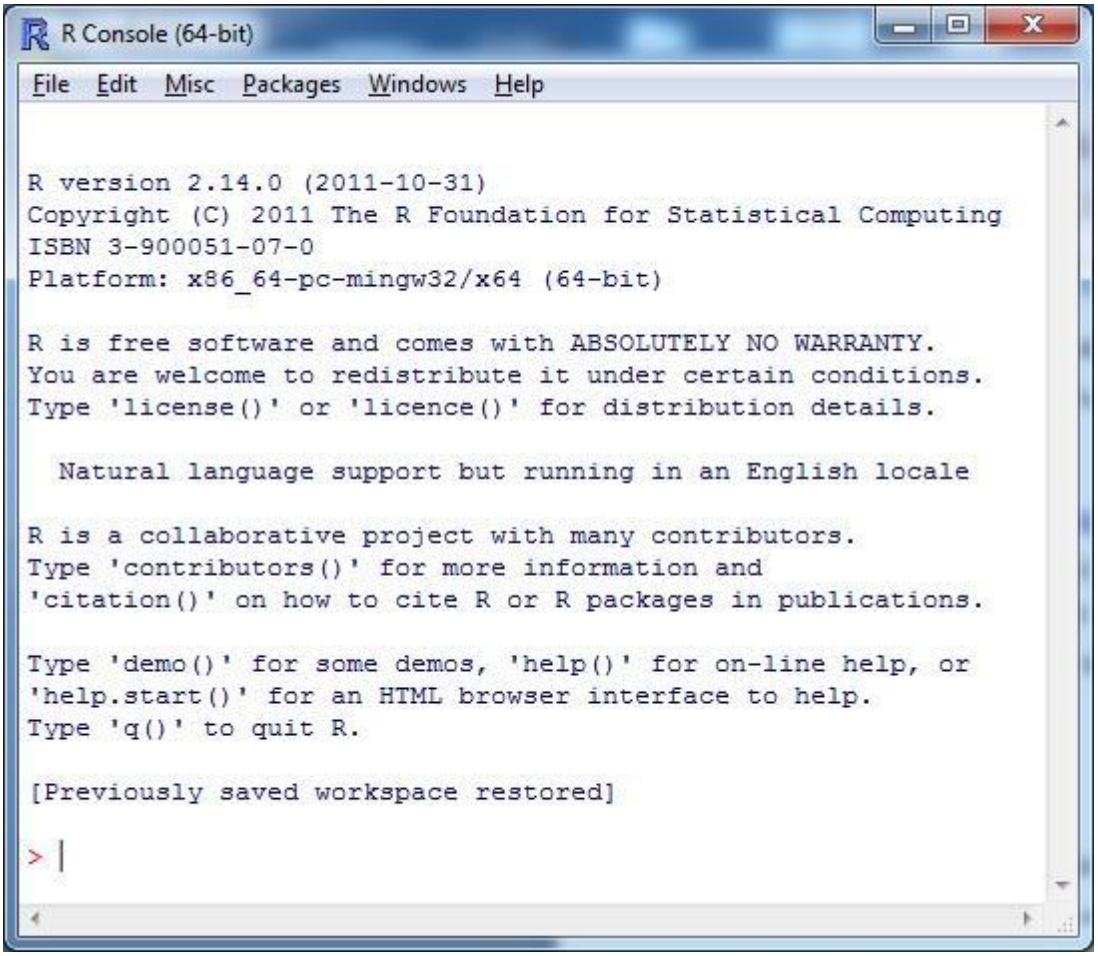
ينظر للـ R على أنها إمتداد للغة الـ S و الـ S-PLUS المطورة في بداية التسعينيات من القرن الماضي. تتميز لغة الـ R بكثرة إستعمالها من طرف الاحصائيين حتى أن الكثير من الناس يظن أنها حكر لهم مثل برنامج SPSS لكن يتميزالـ R بكونه يتمتع بصفات لغات البرمجة الاخرى حيث يمكن القيام بالبرمجة الكائنية وحتى إنشاء صفحات ويب. لكن مايميزه هو أنه يحتوي على العديد من الحزم الاحصائية مما جذب إليه العديد من الإحصائيين.

"GNU R"أو المعروفة اختصارًا ب R هي لغة وبيئة تطوير (Language and development environment) متخصصة في تحليل وتمثيل البيانات والحوسبة الإحصائية. وتتكون R من حزمة رئيسية "core"يمكن توسيعها بواسطة حزم أخرى موجودة بمستودعات المشروع الرئيسي"CRAN". يبلغ عدد هذه الحزم حتى الآن 2456 . وهي تقدم مكتبات للطرق الإحصائية الأساسية والمقدمة كالإحصاء الوصفي، والأختبارات الإحصائية، وتخطيط التجارب، وتحليل الإرتباطات الخطية، واللاخطية، وتحليل المتتاليات الزمنية، والتحليل متعدد المتغيرات، وتحليل الإشارات.

في السنوات العشر الأخيرة، تجاوزت R الدوائر الأكاديمية إلى قطاعات تكنولوجية متعددة حيث تقوم بإستعمالها شركات كبرى بما في ذلك قوقل ومايكروسوفت وموقع فايسبوك (لتحليل البيانات الشخصية). ثم إن شركات أدوية عالمية (AstraZeneca،Pfizer، Johnson & Johnson، …) تعتمده لتحليل البيانات السريرية . .

3.2.3 كيفية تثبيت لغة R

يمكن تحميل لغة R من الموقع الرسمي لها على الشبكة والموجود على العنوان http://www.r-project.org حيث توجد إصدارات منها لمعظم أنظمة التشغيل الشائعة ومنها Windows و Linux وحتى Apple. إن عملية التنصيب سهلة وتخلو من التعقيدات، وعند الإنتهاء منها يمكن تشغيل بيئة عمل لغة R بالنقر على الأيقونة الخاصة بالبرنامج سواء تلك الموجودة على سطح المكتب أو من خلال قائمة البرامج، وحينها ستظهر شاشة سطر الأوامر الخاصة بلغة Rوهو المكان المعتاد لكتابة الأوامر الخاصة بهذه اللغة كما هو ملاحظ في الشكل التالي:



الشكل (3.8.3) بيئة تطوير لغة الR

4.2.3 ماهي مميزات R ؟

- مجانية، مفتوحة المصدر، ومتاحة للجميع.
- متعددة المنصات يعمل على أنظمة لينوكس ويونكس وماك ووينذوز .
- مختصة في التحليل الإحصائي وبنائها " syntax" سهل ملائم جدا لهذه الغاية.
- ذات أداء عال وقابلة للموازاة (Parallel computing) وهو أمر هام لعمليات حوسبة معقدة مثل نمذجة ومحاكاة المناخ والنظم الأحيائية.
- لغة مفسرة ولغة لكتابة السكربتات مثل بايثون.
- التمثيل بياني ذو جودة عالية مع إمكانية إنتاج مخططات ثلاثية الأبعاد بإستعمال مكتبة"OpenGL"
- نظام وثائق ودعم محكم ومتنوع .

5.2.3 ماهي حدود ال R ؟

- لكل لغة برمجة حدود، فمن الواضح أن ليس من الممكن مثلاً عمل لعبة ثلاثية الأبعاد بواسطة Rكما أنه لن نستطيع حل كل المشاكل الرياضية وتحليل جميع أنواع البيانات.
- بما أنها لغة مفسرة، ستكون R أبطأ في بعض البرامج الضخمة وذات البنية معقدة.
- قد تكون ذاكرة R غير مهيأة لتعامل مع بيانات ذات سعة فائقة لأنها تقوم بتحميل كافة البيانات والمكتبات في الذاكرة الحيّة.

6.2.3 إضافة حزم جديدة في ال R

تحتوي النسخة المحملة للـ R على مجموعة من الحزم التي تمكننا من قراءة البيانات وتحليلها والقيام بمجموعة من الإختبارات الاحصائية، لكن في بعض الأحيان نحتاج إلى القيام ببعض العمليات التي لاتوجد بصفة تلقائية في الـ R وليست سهلة البرمجة أو ربما ليست هي لب الموضوع، لهذا تم إنشاء مواقع أو مخازن (Repository) تحتوي على مجموعة حزم مفتوحة المصدر مطورة من قبل أشخاص ومجموعات أبحاث.

1. **مخزن الـCRAN**

المخزن الرسمي للـ R في موقع الـ CRAN ويحتوي على العديد من الحزم الخاصة بعدة مجالات بينها المعلوماتية الحيوية.

2. **مخزن الـBioconductor**

تهدف منصة اللـ Bioconductor إلى إنشاء موقع يحتوي على حزم متعلقة بمجال معالجة البيانات الحيوية وتسهيل الوصول إليها واستعمالها. بالإضافة إلى الحزم يحتوي الموقع على مجموعة من البيانات المستعملة بكثرة، مثلا أسماء الجينات مع أرقام تسلسلها في عدة قواعد بيانات. على الرغم من أن منصة الـ Bioconductor مفتوحة المصدر إلا أنها لا تقبل الحزم إلى إذا خضعت لشروط الموقع، وهي إضافة ملفات لشرح كيفية إستعمال الحزمة مع أمثلة وشرح لمختلف الدوال المتاحة ويكون الملف مرتبة بطريقة تتناسب مع شروط الموقع .

7.2.3 أنواع المتغيرات في لغة R :

- المتغيرات البسيطة

تتقبل لغة الـ R أنواع المتغيرات البسيطة مثل الاعداد الطبيعية، الأعداد الحقيقية و الجمل النصية.

- المصفوفات

بالإضافة إلى هذه الأنواع يمكن إنشاء متغيرات ذات بنية أكثر تعقيدا كالأشعة (جدول بصف واحد) أو المصفوفات .

- العوامل(Factors)

في بعض الأحيان يكون لدينا مصفوفة تحتوي على بيانات متكررة أو بيانات مصنفة (مثلا ذكر،انثى)

في هذه الحالة يمكن إستعمال نوع اخر من البيانات وهي العوامل والتي تسمح لك يتصنيف البيانات

- **أطر البيانات(Data Frames) :**

يمكن أيضا تسميتها بالجداول، هي شبيهة بالمصفوفات لكن كل عمود عبارة عن عامل (Factor) أو مصفوفة.

- **قراءة البيانات من الملفات :**

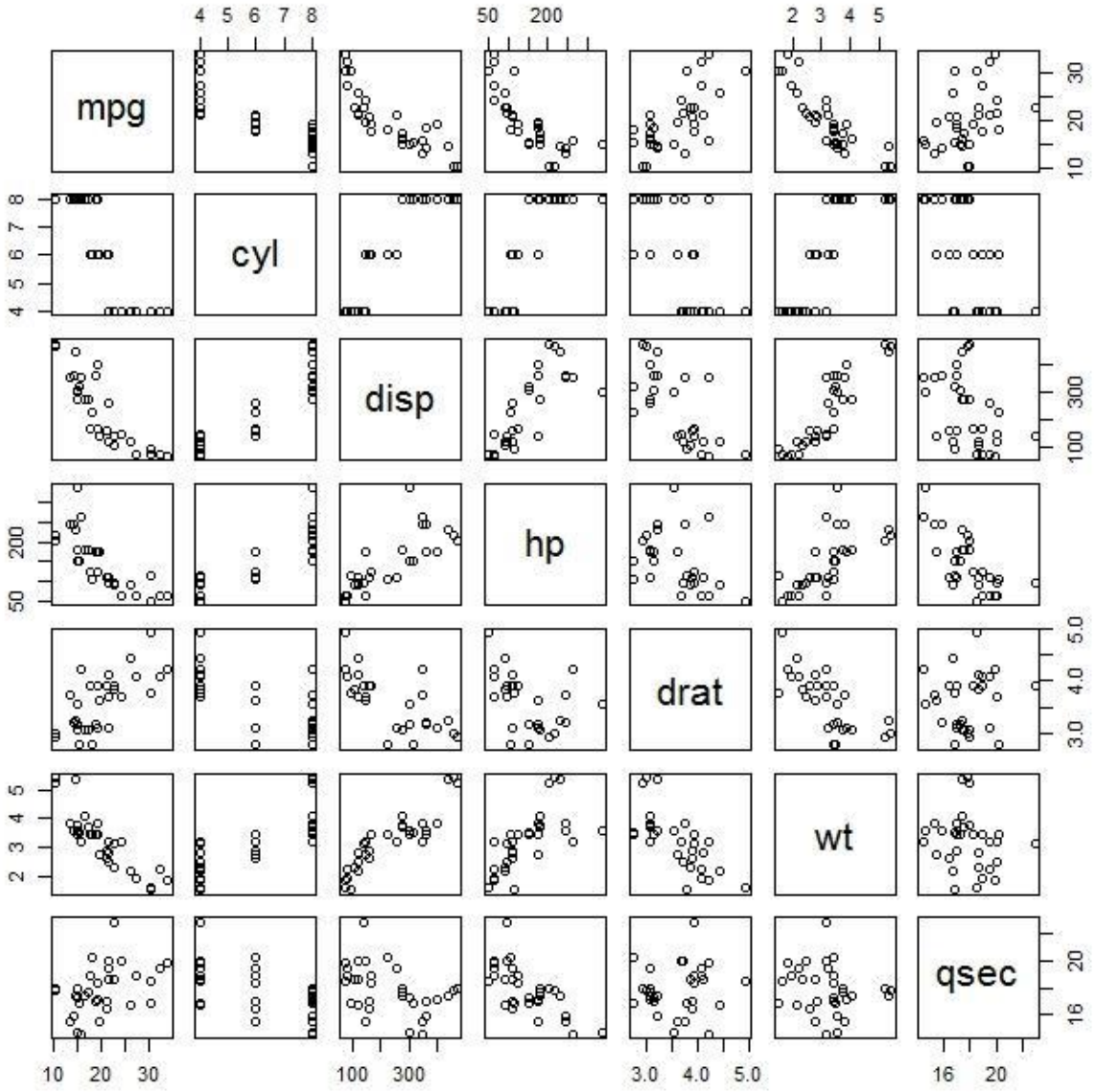
يحتوي الـ R على أوامر تسمح بقراءة البيانات من الملفات وتحميلها على شكل جداول.

8.2.3 المخططات البيانية الإحصائية في لغة R

تبدأ أول خطوة من أي تحليل إحصائي بإستكشاف ما لدينا من بيانات وذلك من خلال إلقاء نظرة سريعة على بعض المخططات البيانية والرسوم التوضيحية ذات الصيغة الإحصائية والتي عليها أن تقوم بتنفيذ تلك المهمة بشكل كامل .

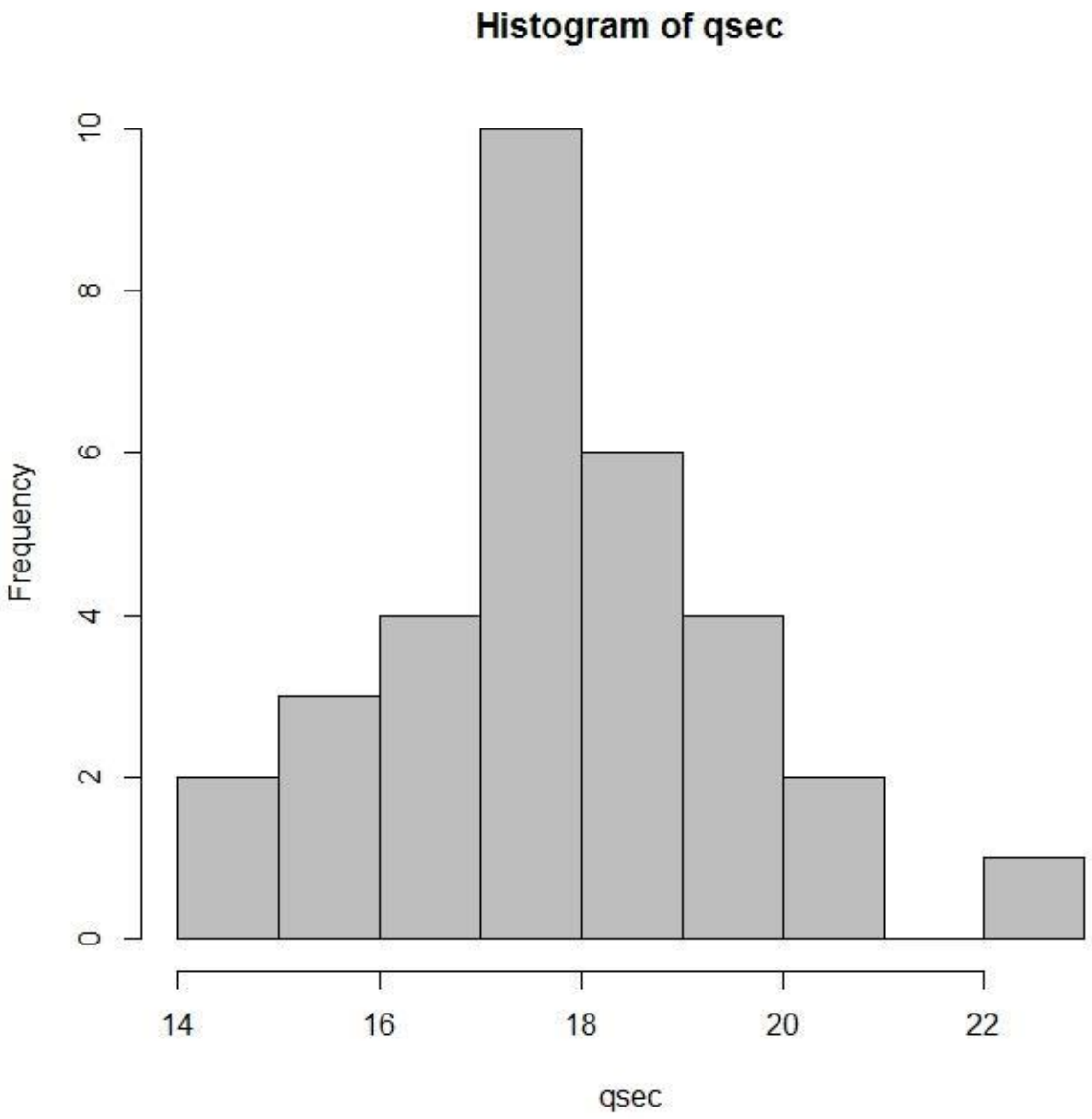
1.8.2.3 أنواع المخططات البيانية في لغة R

لدى لغة R تعليمة بسيطة بالصيغة لكنها في ذات الوقت تقدم خدمة عظيمة في إطار عرض البيانات وتوضيح ما فيها من علاقات محتملة، تدعى هذه التعليمة pairs وتقبل كدخل لها إسم إطار البيانات الذي لدينا كاملا، لتقوم بعدها برسم مصفوفة من المخططات البيانية لكل زوج ممكن من هذه البيانات على شكل مخطط مبعثر scatter plot) بحيث يظهر كل زوج في مخططين بيانيين يتبادلان فيه مكان التمثيل على المحورين x و y .



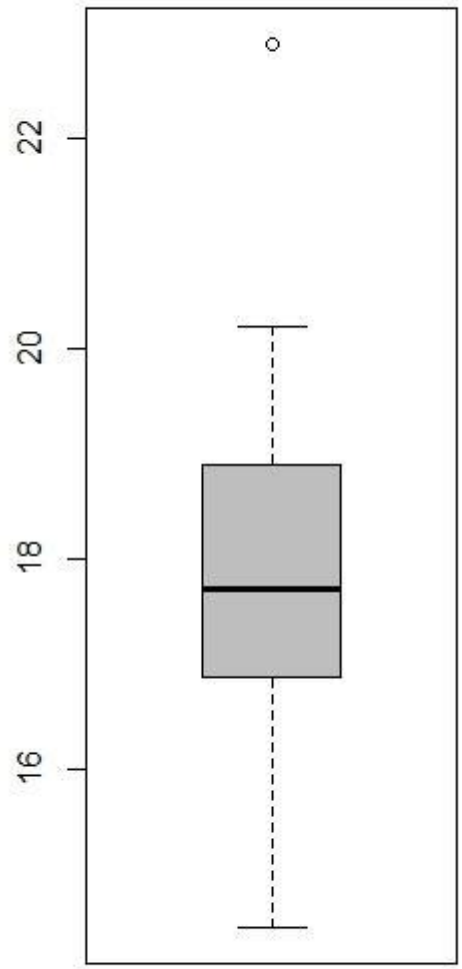
الرسم البياني (1.8.2.3) المخطط المبعثر في لغة R

هناك مخطط بياني آخر ذي طبيعة استخدام إحصائية وهو المدرج التكراري(Histogram) ، وهو يبين طبيعة توزع ما لدينا من قيم على المجال المحصور ما بين الحد الأدنى والأقصى، ففي بعض الأحيان لا يكون تلخيص البيانات بمساعدة الحد الأدنى والأقصى والمتوسط كافياً، حينها يتم اللجو إلى هذا المخطط البياني والذي يقسم فيه المجال الكلي ما بين الحد الأدنى والحد الأقصى إلى فئات أو مجموعات، ومن ثم نرسم أعمدة بيانية توضح عدد مرات تكرار ظهور القيم ضمن كل واحدة من هذه الفئات أو المجموعات.



الرسم البياني (2.8.2.3) المدرج التكراري في لغة R

وهنالك نوع آخر من المخططات البيانية ذات الصبغة الإحصائية وهو المخطط الصندوقي ، يمثل بالشكل التالي :



الرسم البياني (3.8.2.3) المخطط الصندوقي في لغة R

حيث يوضح الخطين الأفقيين على طرفي الرسم في الأعلى والأسفل كل من القيمة الصغرى (في الأسفل) والعظمى (في الأعلى)، أما الصندوق الموجود بينهما فتوضح بدايته من الأسفل ما يسمى بحد الربع الأول[14] .

9.2.3 الفرق بين ال R و RStudio

R هو عبارة عن تطبيق يتم تثبيته على جهاز الكمبيوتر ويستخدم موارد الكمبيوتر الشخصية لمعالجة لغة البرمجة R . أما Studio فهي تتكامل مع الR وتعتبر بيئة تطوير متكاملة (IDE) لتوفير مزيد من الوظائف فهي تجمع بين محرر شفرة المصدر، وبناء أدوات التشغيل الآلي والمصحح.

MySQL 3.3

هي قاعده بيانات مفتوحة المصدر (Open Source) ، والأكثر شهرة ضمن قواعد البيانات مفتوحة المصدر، تم إنشاءها بواسطة شركة (MySQL AB) السودية . تعرف MySQL بأنها قاعدة بيانات، وهي طريقة من طرق الاحتفاظ بالبيانات، وقاعدة البيانات تتكون من جداول والجداول بدورها تتكون من صفوف وأعمدة وخلايا. تتكون أي قاعدة بيانات غالباً من عدة جداول، ويحمل كل جدول إسماً مختلفاً يميزه مثلاً "customers"أو "products" أو"orders" ، وكل جدول يتكون من صفوف تحتوي البيانات.

أحد الحقائق المذهلة عن MySQL أنها تقوم بتخزين البيانات الضمنية الموجودة في التطبيقات و هذه الحقيقة مذهلة بسبب أن معظم الناس تعتقد أن MySQL تستخدم فقط مع التطبيقات التي تخزن بيانات صغيرة أو متوسطة.

يعتقد البعض أن قاعدة البيانات MySQL يستخدم في المواقع الصغيرة والمتوسطة التي تستخدم قاعدة بيانات صغيرة. ولكن الصحيح أن نظام قاعدة البيانات MySQL يستخدم في المواقع الإلكترونية التي تحتاج قاعدة بيانات كبيرة جداً وعدد كبير من المستخدمين (مثل جوجل، ياهو، تويتر...).

1.3.3 مميزات MySQL :

1. كفاءة في الأداء .
2. الأمن و الإعتمادية.
3. موثوقية عالية.
4. سهولة في الستخدام.
5. التوافق مع المعايير الموجودة.
6. سياسة ترخيص سهل.
7. القابلية للتوسع و النقل[15].

PHP 4.3

PHPهي إختصار **P**ersonal **H**ome **P**ages، لكنها الآن تعني .PHP: **H**ypertext **P**reprocessorال PHP تم إختراعها و إنشاؤها وتطويرها في عام 1994 بواسطة راسموس ليدروف ، لغة الـ PHP هي لغة مفتوحة المصدر و يتم تطويرها بإستمرار عن طريق مجموعة من المتطوعين تحت رخصة PHP و هذا من أحد الأسباب الرئيسية لشهرتها.

هي لغة برمجة تم صناعتها أساساً من أجل استخدامها لتطوير و برمجة مواقع الويب، لغة PHP هي واحدة من أشهر لغات البرمجة التي يتم استخدامها في إنشاء مواقع الويب و هي من اللغات التي يقوم خادم الويب بتفسير و تنفيذ الكود الخاص بها ثم يرسل النتيجة ليتم عرضها في متصفح المستخدم، و تستخدم لغة PHP لإنشاء صفحات ويب ديناميكية أي صفحات متغيرة المحتوي و هذا المحتوي يتغير نتيجة التفاعل مع المستخدم ،أن PHP كلغة تعمل في الخادم (server) فقط.

1.4.3 مميزات PHP :

1. يمكن تشغيلها على مختلف أنظمة التشغيل .
2. يمكن تشغيلها على بيئات عمل مختلفة .
3. تدعم مجموعة واسعة من قواعد البيانات مثل (MySQL 3.x/4.x/5.x , Oracle , ODBC and SQL).
4. تطبيقها على العديد من خوادم الويب (Web Servers).
5. مجانية ولا تحتاج لرخصة إستخدام.
6. بالإضافة الى السرعة والسهولة والقوة والسلاسة التوافقية والحماية و قابلية التوسع[16] .

WampServer 5.3

هو بيئة لتطوير تطبيقات الويب على نظام التشغيل ويندوز (Windows)، تم تصميمه بواسطة (Romain Bourdon). هذه البيئة متاحة مجاناً بموجب ترخيص البرنامج العالمي (GPML) في نسختين متميزتين : 32 و 64 بت . تسمح البيئة للمطورين بإنشاء تطبيقات الويب بإستخدام (Apache2) ، لغة البرمجة (PHP) وقاعدة البيانات (MySQL) . كما تحتوي البيئة على جزء يُسمى (PhpMyAdmin) الذي يسمح للمطور بسهولة إدارة قواعد البيانات .

1.5.3 مميزات WampServer :

1. يوفر طريقة سهلة للحصول على حلول لتطبيقات ويندوز التي تستخدم (MySQL and Apache PHP,).
2. يوفر العديد من الإضافات في عدد قليل من الحزم.
3. السماح بالإنتقال عبر الإنترنت غير المتصل.
4. إدارة الإعدادات الخاصة بالشبكة.

5. كما يتميز هذا البرنامج بأكثر من إصدار يتوافق مع جميع الأجهزة والسيرفرات بالإضافة الى أنة يتعامل مع إصدارات متعدده من الـApache ، PHP، .MySQL [17]

الباب الرابع

جمع البيانات

Data Collection

1.4 المقدمة

يمكن إستخدام مواقع التواصل الاجتماعي كمصدر غني بالبيانات ولذلك يتناول هذا الفصل كيفية جمع البيانات من التويتر وتخزينها في قاعدة البيانات.

2 .4 إنشاء حساب في موقع تويتر

قبل إسترجاع البيانات من تويتر لابد أن يتوفر لدى الشخص حساب في موقع تويتر ويتم توثيقه من قبل شركة تويتر.

3.4 المصادقة و التفويض في تويتر

(Authentication and authorization in twitter)

المصادقة : هي عملية التحقق من هوية المستخدم.

التفويض أو الإذن : هو عملية التحقق من أن المستخدم لديه الحق في تنفيذ بعض الجراءات، مثل قراءة وثيقة أو الوصول إلى حساب البريد الإكتروني.

تويتر يدعم أساليب قليلة من المصادقة (Authentication) ، و مع مجموعة أنماط و أساليب المصادقة أوث (OAuth authentication) قد يتساءل المط ور عن الطريقة أو الْسلوب الذي يجب أن يستخدمه . وعندما يريد إختيار أسلوب المصادقة الذي سيستخدمه يجب أن يفهم المط ور كيفية تأثير ذلك الْسلوب على المستخدمين و كذلك تأثيره على طريقة كتابته للتطبيق الخاص به.

ماهو بروتوكول أوث (OAuth) ؟

هو إختصار ل (Open Authorization) و هو بروتوكول مفتوح يُمكن من عمل تفويض مؤمّن بطريقة بسيطة وقياسية من قِبل تطبيقات الويب والهاتف الجوال و تطبيقات سطح المكتب.

و إن فهم الطريقة التي يعمل بها البروتوكول أوث (OAuth) يساعد على إنشاء و تصحيح التطبيقات التي تستخدم واجهة برمجة تطبيقات تويتر (Twitter's API) .و إذا كان المطور يعلم ما هو أسلوب المصادقة الذي سيستخدمه ، يقوم موقع تويتر بتقديم المساعدة لتمكين المط ور من معرفة أنه قام قال بإختيار الْسلوب الصحيح ؛ وذلك من خلال هذا الجدول المبسط :

If you use the…	Send…
REST API	OAuth signed orapplication-only auth requests
Search API	OAuth signed orapplication-only auth requests
Streaming API	OAuth signed

الجدول (3.4) الإختيار بأسلوب صحيح

و من أجل إستخدام بروتوكول أوث (OAuth) ، يجب أن يُصمم المط ور التطبيق بحيث:

- يحصل على رموز الوصول للحساب من أجل أن يتصرف التطبيق في حساب المستخدم نيابة عن المستخدم.
- يعطي الذن لجميع طلبات ال HTTP التي يرسلها إلى واجهات برمجة تطبيقات تويتر.

1.3.4 مميزات بروتوكول أوث (OAuth) في تويتر:

1. إعتماد بروتوكول أوث (OAuth) في موقع تويتر و هو موقع تواصل إجتماعي شهير يدل على أهمية هذا البروتوكول و الذي يلبي إحتياجات المستخدمين في المحافظة على سرية كلمة السر الخاصة بهم و بالتالي يزيد من الموثوقية.

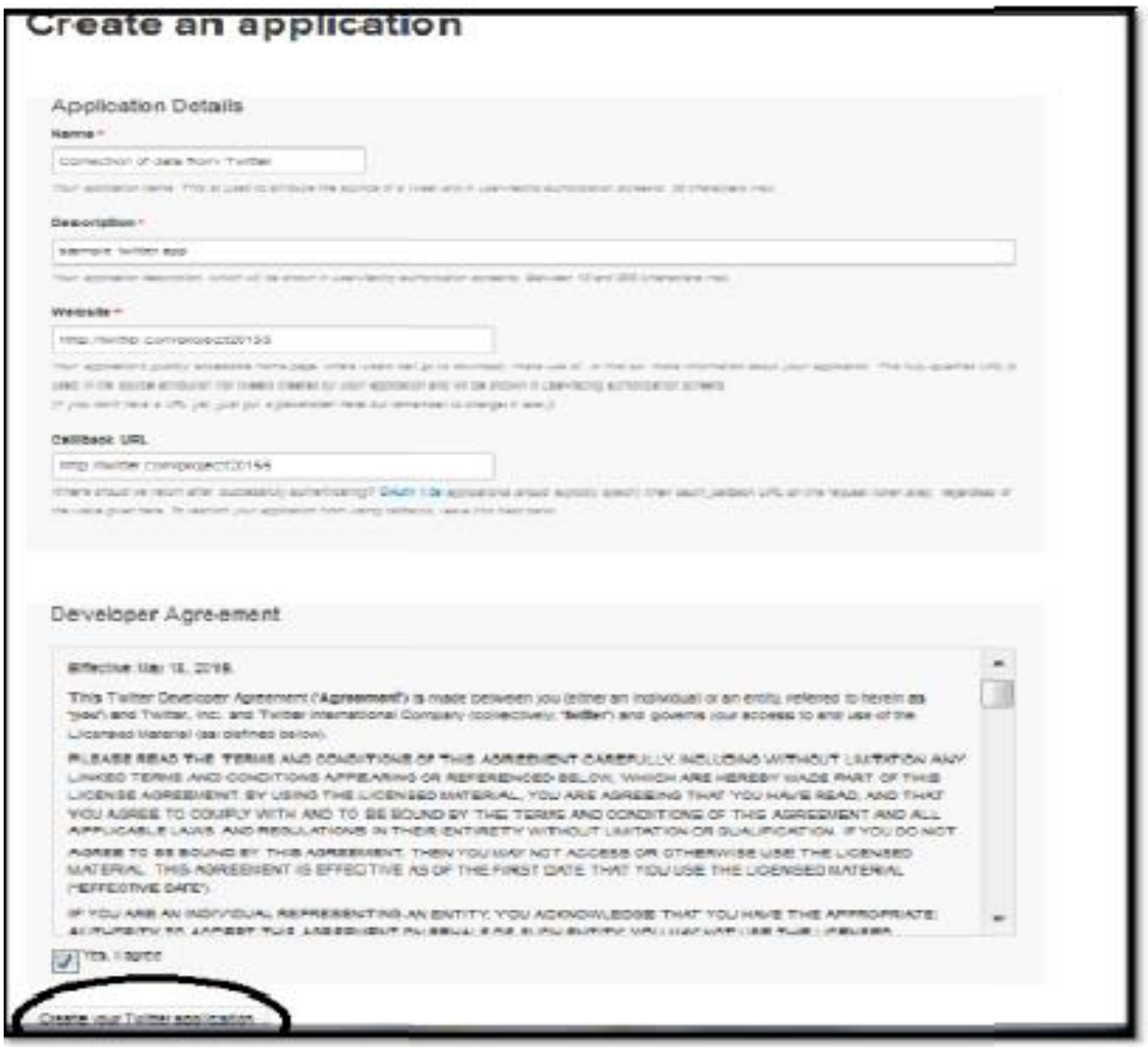
2. تقليل المسؤولية الواقعة على المط ور أو المبرمج ؛ و ذلك لَنه لَ يحصل على كلمة السر الخاصة بحساب المستخدم ، كما أنه يتمكن من أخد المعلومات التي يحتاجها فقط من الحساب ؛ و بالتالي لا يقوم بأي معالجات إضافية لكمية كبيرة من المعلومات الزائدة.

4.4 إنشاء تطبيق تويتر

يجب تسجيل أي تطبيق يحتاج إلى تفاعل مع واجهة برمجة تطبيقات تويتر بإستخدام وحدة إدارة تويتر للتطبيقات وذلك بعد عملية تسجيل الدخول , يمكنك إنشاء التطبيق من خلال هذا الرابط :

https://apps.twitter.com

يتطلب التطبيق إدخال إسم التطبيق ووصف عنه ورابط موقعك .يجب الموافقة على الشروط المكتوبة ثم الضغط على إنشاء تطبيق (Create your twitter application).



الشكل (4.4) يوضح إنشاء تطبيق تويتر

1.4.4 الصلاحيات بعد إنشاء تطبيق تويتر

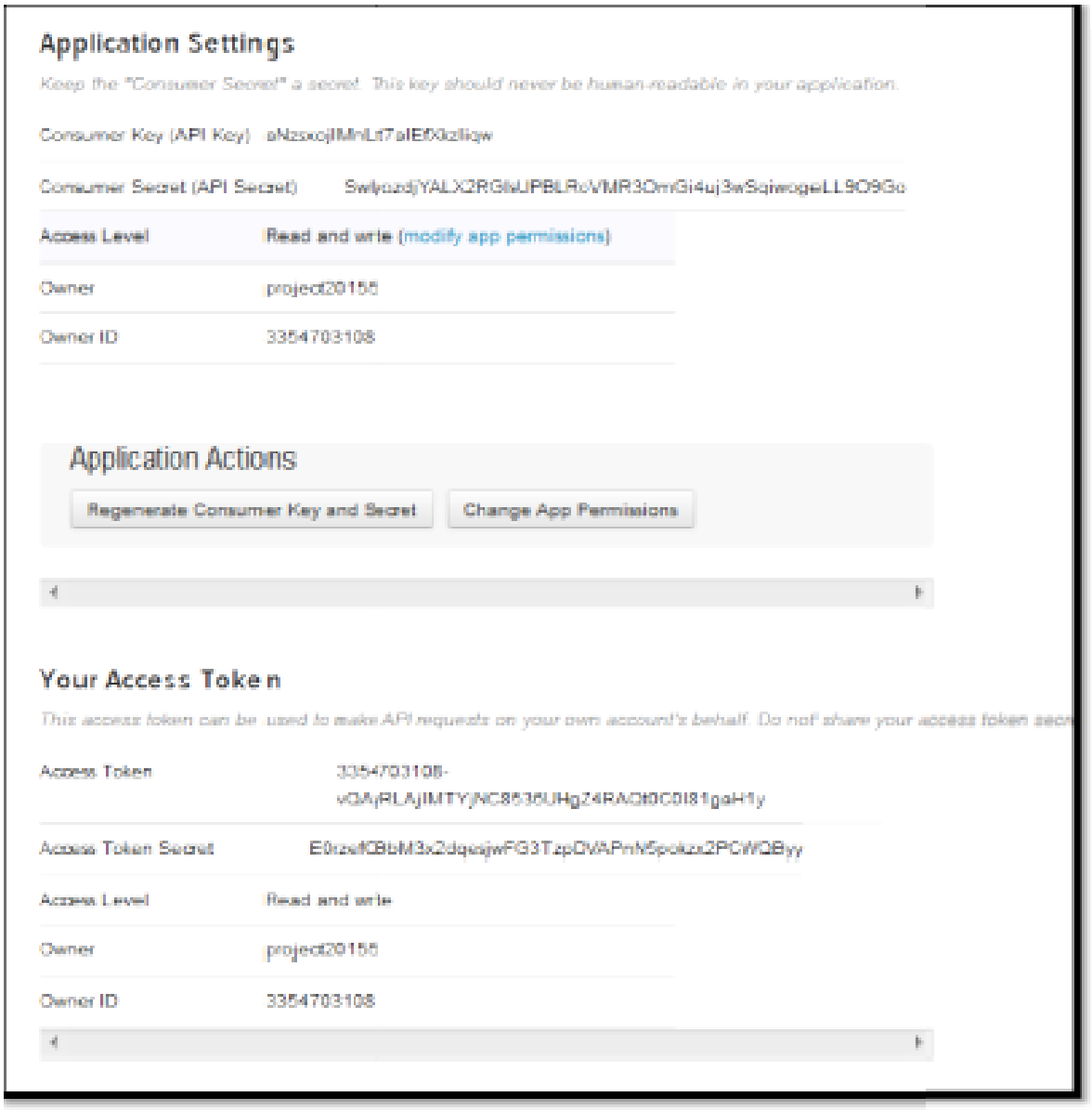
بعد إنشاء التطبيق يعطي المستخدم الصلاحيات التي تمكنه من الوصول إلى جميع المعلومات وهي :

- Consumer key مفتاح المستهلك.
- Consumer secret الرقم السري للمستهلك.
- Access token رمز الوصول.
- Access secret الرقم السري للوصول.

يتم نسخها ووضعها في الكود .

يقوم برتوكول أوث بمطابقة مفتاح المستهلك (Consumer Key) و الرقم السري للمستهلك Consumer (secret) حيث يتم عن طريقهما معا التحقق من المستهلك(التطبيق وليس المستخدم)من قِبل مُقدم الخدمة . عملية تحديد المستهلك تسمح لمقدم الخدمة بجعل مستويات الوصول إلى المستهلكين مختلفة.

رمز الوصول Access token: يتم إستخدامه من قبل المستهلك (التطبيق) للوصول إلى الموارد المحمية نيابة عن العضو (المستخدم).



الشكل (1.4.4) يوضح الصلاحيات التي تعطى للمستخدم بعد إنشاء تطبيق تويتر

5.4 الحصول على البيانات

بعد عملية إنشاء التطبيق والحصول على الصلاحيات نقوم بتحميل (OAuth files) ومن ثم إنشاء ملف من نوع PHP يتم فيه إستخدام الصلاحيات التي تم الحصول عليها من الخطوة السابقة.

تم إستخدام search API لإسترجاع مجموعة من التغريدات وفقاً لشروط معينة.

1.5.4 عملية الطلب (Request)

get("https://api.twitter.com/1.1/search/tweets.json?lang=ar&

وذلك count=100&geocode=15.599127,32.512071,25mi)

لإسترجاع التغريدات السودانية المكتوبة باللغة العربية

بحيث يستخدم **Lang = ar** لتحديد اللغة وتم إختيار ar للغة العربية

Count=100 لتحديد عدد التغريدات في كل عملية طلب والعدد 100 يمثل أقصى عدد من التغريدات يمكن الحصول عليها في عملية الطلب الواحدة ويمكن الوصول إلى 180 عملية طلب خلال 15 دقيقة وتم جمع ما يقارب8 الف تغريدة.

geocode=15.599127, 32.512071, 25mi لتحديد المكان وهذا يمثل خطوط الطول ودوائر العرض لمدينة الخرطوم.

2.5.4 النتيجة (Result)

{"statuses":[{"created_at":"Tue Aug 23 17:44:52 +0000 2016","id":7.6814197379455e+17,"id_str":"768141973794553856","text":"من كان آخر كلامه لا إله إلا الله دخل الجنة https:\/\/t.co\/e2KCY3GHG9","truncated":false,"entities":{"hashtags":[],"symbols":[],"user_mentions":[],"urls":[{"url":"https:\/\/t.co\/e2KCY3GHG9","expanded_url":"http:\/\/du3a.org","display_url":"du3a.org","indices":[43,66]}]},"metadata":{"iso_language_code":"ar","result_type":"recent"},"source":"تطبيق دعـــــــاء <\/a>","in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":191011947,"id_str":"191011947","name":"Muhammad Amin","screen_name":"M7mdAmin","location":"khartoum","description":"just livin ;)","url":null,"entities":{"description":{"urls":[]}},"protected":false,"followers_count":242,"friends_count":705,"listed_count":1,"created_at":"Wed Sep 15 11:59:25 +0000 2010","favourites_count":297,"utc_offset":null,"time_zone":null,"geo_enabled":false,"verified":false,"statuses_count":5873,"lang":"en","contributors_enabled":false,"is_translator":false,"is_translation_enabled":false,"profile_backgro

und_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/568741340469805056\/xcV6VqAg_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/568741340469805056\/xcV6VqAg_normal.jpeg","profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"has_extended_profile":false,"default_profile":true,"default_profile_image":false,"following":false,"follow_request_sent":false,"notifications":false},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"favorited":false,"retweeted":false,"possibly_sensitive":false,"lang":"ar"}],"search_metadata":{"completed_in":0.028,"max_id":7.6814197379455e+17,"max_id_str":"768141973794553856","next_results":"?max_id=768141973794553855&q=&geocode=15.599127%2C32.512071%2C25mi&lang=ar&count=1&include_entities=1","query":"","refresh_url":"?since_id=768141973794553856&q=&geocode=15.599127%2C32.512071%2C25mi&lang=ar&include_entities=1","count":1,"since_id":0,"since_id_str":"0"}}

التغريدات المسترجعة تحتوي على الآتي:

- **created_at** : يمثل تاريخ و وقت التغريدة.
- **Favorited:** هل تم تفضيل التغريدة ام لا.
- **id_str:** رقم التغريدة.
- **Text:** يمثل نص التغريدة.
- **retweet_count:** عدد المرات التي تم فيها إعادة التغريد.

اما بالنسبة للشخص الذي قام بالتغريد :

- **Name:** يمثل إسم المستخدم.
- **profile_image_url:** رابط الصورة الشخصية للمستخدم.
- **created_at:** يمثل التاريخ عندما تم إنشاء الحساب في تويتر.

- **location:** موقع المستخدم كما هو مبين في صفحته الشخصية.
- **URL:** الروابط الموجودة في الصفحة الشخصية للمستخدم.
- **followers_count:** عدد الأشخاص الذين قام بتتبعهم.
- **protected:** هل التغريدة محمية أم لا؟
- **statuses_count:** عدد التغريدات المرسلة.
- **friends_count:** يمثل عدد الأصدقاء للمستخدم.
- **screen_name:** إسم المستخدم الموجود على الشاشة.

وما سبق هو كيفية هيكلة البيانات بإستخدام أسلوب ال (JSON) والتالي الدوال التي تتعامل مع هذا التنسيق في لغة PHP :

الدالة	الوظيفة
json_encode	تستخدم للتحويل إلي تنسيق JSON.
json_decode	تستخدم لتحويل تنسيق JSON إلي كائنات و مصفوفات يمكن التعامل معها في (PHP).
json_last_error	تستخدم لإسترجاع اخر خطأ حدث.

الجدول(2.5.4) أنواع دوال JSON

3.5.4 نبذة عن JSON

Json :هي صيغة سلسة لنقل البيانات - سهلة للأشخاص قراءتها واستيعابها، وسهلة للأجهزة والحواسيب تحليلها وتوليدها، هي صيغة بيانات مستقلة تماماً عن اللغة المتسخدمة أو الإطار المتبع فيه، يمكن لكثير من المبرمجين لعائلة لغات السي C والتي تشمل C++ ، C#، جافا، جافا سكريبت ، بيرل، بايثون، والعديد من اللغات استخدامها، تلك المميزات تجعل من جاسون الأفضل في نقل البيانات (data) من لغة لأخرى.
وأيضاً تستخدم هذه التقنية في جلب البيانات من مواقع شهيرة كموقع twitter من خلال twitter json API وحالة الطقس من موقع yahoo من خلال - yahoo json weather API .

بنيت Json باستخدام نظريتين على أن تكون:

- مجموعة من الأزواج (pairs) الأسماء (names) والقيم(values) ، كما في مختلف لغات البرمجة المعروفة كالعنصر (object) ، أو القاموس (dictionary) ، السلاسل المتعلقة (associative array).
- سلسلة مرتبة من القيم (values) ، كما هو المعروف والمتسخدم في كثير من لغات البرمجة كالسلاسل (arrays) ، فيكتور (vector) ، والقائمة (list) ، أو المصنفات(sequence) .

تلك المواصفات تجعل التركيب مكافئة لنقل البيانات ، كل لغات البرمجة الحديثة تقريبا تدعم هذه التراكيب بشكل من الأشكال ، ولذلك تجعل من جاسون الأفضل في نقل البيانات من لغة لأخرى.

في JSON التراكيب تأتي بالأشكال الأتية :

- عنصر **(**object**)** : كمجموعة غير مرتبة من أزواج الأسماء والقيم وتوضع العناصر بين الأقواس {} .
- سلاسل (Array) *:* هي مجموعة مرتبة من السلاسل، وتوضع عناصر المصفوفة بين الأقواس [] .

يمكن للقيم أن تكون مجموعة من الأحرف **(**String**)** منفصلة بعلامة ترقيم "أو رقم ، أو true أو false أو null، **أو** سلسلة **أو** عنصر **و** يتم إسناد القيم للعناصر بإستخدام الرمز ":" ويتم الفصل بين العناصر بإستخدام الفاصلة "," .
وأيضا ممكن جعل التراكيب أكثر تعقيداً بحيث تستخدم كل هذه التراكيب معاً.

4.5.4 مميزات JSON:

- سلهلة الاستخدام .
- سهلة التوليد والتحليل.
- تدعم كافة لغات البرمجة.
- أفضل خيار لخدمات الويب .[18]

6.4 تخزين البيانات في قاعدة البيانات

الخطوات التالية توضح كيفية تخزين البيانات في قاعدة البيانات :

1.6.4 إنشاء قاعدة بيانات باستخدام MySQL

في هذه الخطوة تم إنشاء قاعدة بيانات باستخدام Wampserver وتم عمل جدول ووضع فيه عمودين الاول يمثل رقم التغريدة والاخر يمثل النص.

2.6.4 ربط PHP مع قاعدة البيانات MySQL

كخطوة اولى وقبل كل شيء يتم ربط PHP مع قاعدة البيانات MySQL وذلك باستخدام الدالة () mysql_connect.

3.6.4 التحويل من تنسيق JSON الى مصفوفة PHP

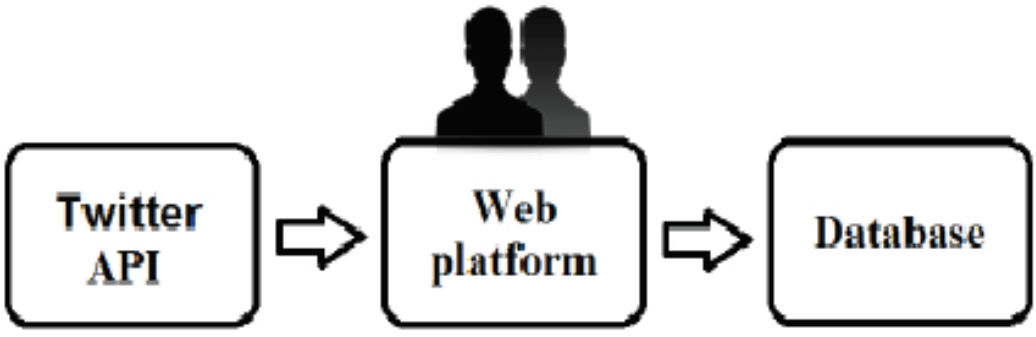
الخطوة التالية هي تحويل JSON إلى مصفوفة ليسهل التعامل معها في (PHP) وذلك بإستخدام دالة json_decode.

4.6.4 إستخراج قيم المصفوفة:

في هذه الخطوة يتم إستخراج قيم المصفوفة من المصفوفة التي تم الحصول عليها من الخطوة السابقة وتخزينها في متغيرات PHP وهنا تم إختيار النص (statuses) وتخزينها في متغيرPHP .

5.6.4 إدخال القيم إلى قاعدة بيانات بإستخدام كود PHP

الخطوة الاخيرة هي إدراج قيم النص (statuses) المتحصلة عليها من الخطوة السابقة في جدول MYSQL المنشىء مسبقاَ وقبل تخزين النص يتم التخلص من الروابط والحروف الغير مرغوب فيها مثل(#,@) وذلك بإستخدام دوال موجودة في PHP .



الشكل (6.4) الخطوات الأساسية للحصول على البيانات وتخزينها

الباب الخامس

التطبيق

Implementation

1.5 مقدمة

تجمع البيانات هي عملية وضع ال data في مجموعات متماثلة. خوارزمية التجمع تقسم مجموعة من البيانات إلى عدة مجموعات. حيث أن التشابه بين النقاط ضمن مجموعة معينة أكبر من التشابه بين نقطتين ضمن مجموعتين مختلفتين. فكره تجمع البيانات هي فكره بسيطه في طبيعتها وهي قريبه جداً من الإنسان في طريقه تفكيره حيث اننا كلما تعاملنا مع كمية كبيره من البيانات نميل إلى تلخيص الكم الهائل من البيانات إلى عدد قليل من المجموعات أوالفئات، وذلك من أجل تسهيل عمليه التحليل. خوارزميات التجمع تستخدم على نطاق واسع ليس فقط لتنظيم وتصنيف البيانات وإنما هي مفيده لضغط البيانات وبناء نموذج ترتيب البيانات. حيث أنه إذا كان بإمكاننا أن نجد مجموعات من البيانات، فإنه بالإمكان بناء نموذج للمشكلة على أساس تلك المجموعات. في هذا لباب سنتطرق الى كيفية تطبيق خوارزمية التجمع(K means) على التغريدات التي تم تخزينها في قاعدة البيانات وايجاد أكثر الكلمات تكراراً .

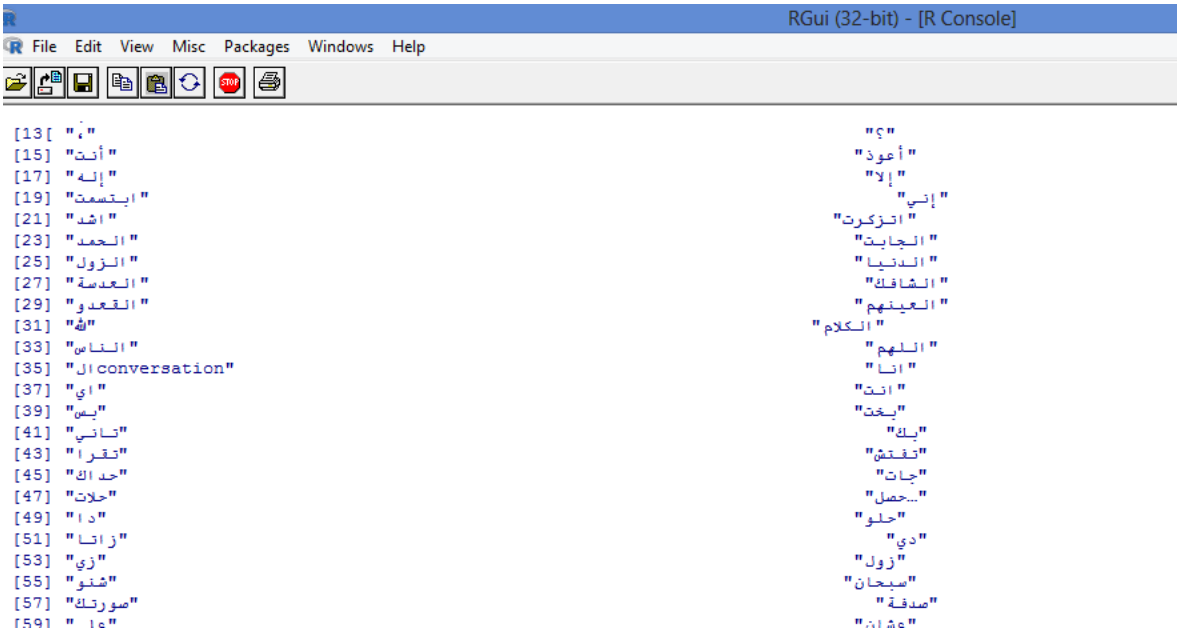
2.5 ربط قاعدة البيانات بال R

حتى يتم ربط قاعدة البيانات بال R يتم تحميل وتثبيت حزمة (RMYSQL) والمكتبات المتضمنة ومن ثم كتابة queryمناسب لاسترجاع التغريدات من الجدول المسمى Tweets الذي تم انشاؤه .

3.5 إيجاد أكثر الكلمات تداولاً في التغريدات السودانية

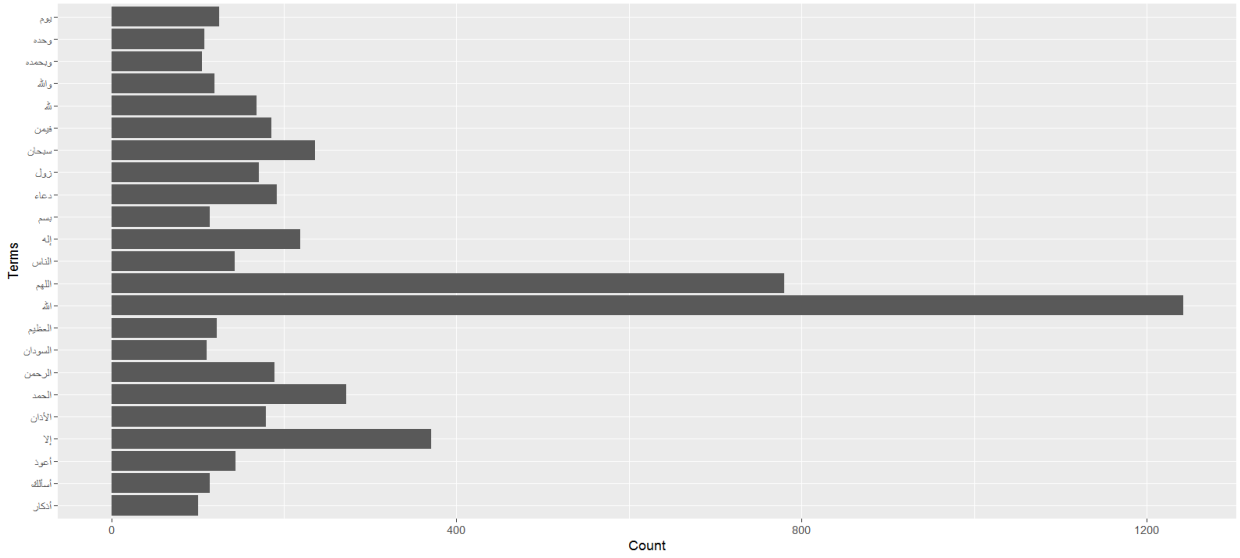
توفر لغة البرمجة R إمكانية تحليل البيانات وإجراء حسابات إحصائية وتحاليل إستكشافية غلى البيانات للحصول على نتائج معينة. قمنا بالإستفادة من هذه الخاصية لإيجاد أكثر الكلمات تداولاً في التغريدات السودانية وذالك بإتباع الخطوات التاية :

- تحميل الحزم المطلوبة وهي (tm,ggplot) وإستدعاء مكتباتها .
- تحديد رقم يمثل أقل ظهور للكلمة وقمنا بإختيار الرقم 150 بحيث يتم جلب كل الكلمات التي تكررت على الأقل 150 مرة وعدد الكلمات هي 150 كلمة بالشكل التالي .



الشكل (3.5) يمثل أكثر الكلمات تكراراً في شاشة التنفيذ R

- إنشاء المخطط البياني لأكثر الكلمات تكراراً في التغريدات السودانية.



الرسم البياني (3.5) لأكثر الكلمات تكراراً في التغريدات السودانية

من المخطط السابق نجد أن أكثر الكلمات هي "الله" وقد تكررت أكثر من 1400 مرة تليها كلمة "اللهم" بمعدل 900 مرة.

4.5 إنشاء الWordCloud للتغريدات السودانية

WordCloud يمكن أن تكون واحدة من أفضل الأدوات التي تسمح لنا تصور معظم الكلمات والمصطلحات الواردة في التغريدات على الرغم من الإستخدام الرئيسي لها هو لأغراض إستكشافية، فهي لديها ميزة أن تكون مفهومة من قبل معظم المستخدمين، وأن تكون جذابة بصرياً إلى العين البشرية. وقمنا بإنشاء الWordCloud بإتباع الخطوات التالية :

- تحميل الحزم المطلوبة وهي (WordCloud,RColorBrewer) وإستدعاء مكتباتها .
- حساب عدد الظهور لأكثر الكلمات تكراراً وترتيبها تنازلياً.
- رسم الWordCloud لأكثر الكلمات تكراراً بالشكل التالي:

شكل(4.5)رسم ال WordCloud للتغريدات السودانية

5.5 تطبيق خوارزمية Hierarchical clustering

تعتمد هذه الخوارزمية على تقسيم البيانات بشكل هرمي.

5.5.1 كيفية عمل الخوارزمية

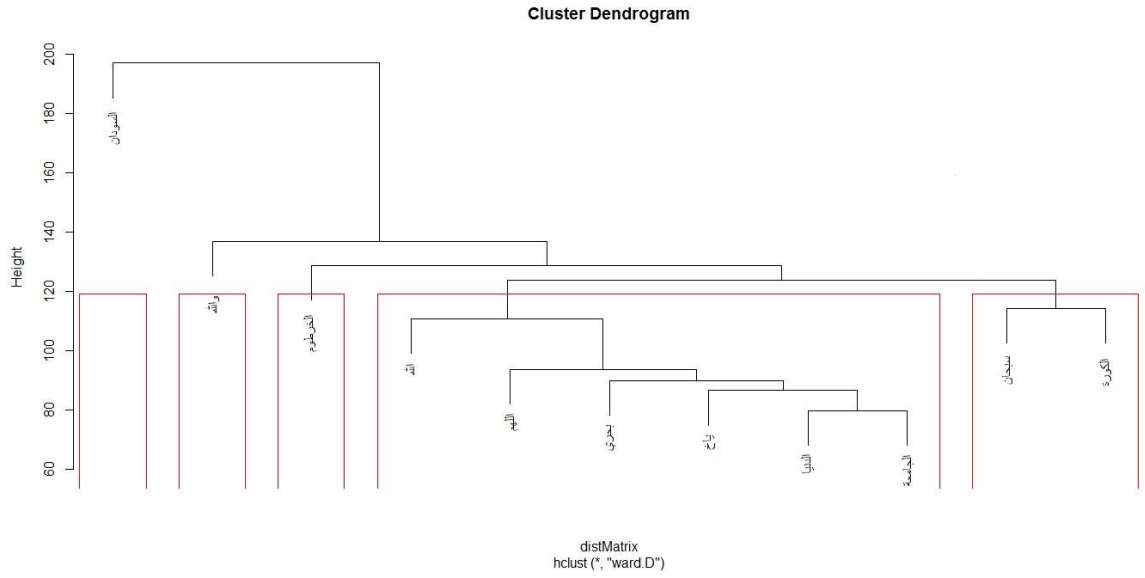
نفترض أنه لدينا مجموعة من البيانات "N"ومصفوفة مسافة N*N (المصفوفة تحتوي على الممسافة بين البيانات)

نقوم بإتباع الخطوات التاية :

1. نبدأ بوضع كل البيانات في مجموعة على حدة بالتالي إذا لدينا N من البيانات تصبح لدينا N من المجموعات.
2. كل أقرب زوج من البيانات يتم دمجها في مجموعة واحدة.
3. يتم حساب المسافة للبيانات التي تم دمجها.
4. تكرار الخطوات 2 و 3 إلى أن تصبح البيانات كلها في مجموعة واحدة.

ملاحظة:

1. لحساب المسافة تستخدم طريقة تسمى single linkage clustering والتي تعتمد على حساب المسافة من أقرب نقطة لل clustering الجديد.
2. عند تطبيق الخوارزمية على التغريدات مصفوفة المسافة تكون لأكثر الكلمات تكراراَ.



الشكل(5.5) نتيجة تنفيذ خوارزمية Hierarchical clustering على التغريدات

6.5 تطبيق خوارزمية (K Means)

1. تحديد عدد التجمعاتK، وهي تعتبر خطوة تهيئة أولية. وفي هذه الحالة تصبح عدد التجمعات تساوي 6 لاننا نريد تقسيم التغريدات الى 5 فئات (سياسية , دينية , رياضية , ترفيه ,اخرى) ، k = 5.
2. تحديد إحداثيات مراكز التجمعات Centroid عشوائياً لكل كلمة لأكثر الكلمات تكرارًا. وذلك لكل فئة من الفئات للمرة الأولى فقط.
3. -حساب المسافة بين كل البيانات ومراكز التجميع. قبل حساب المسافة يتم ترميز البيانات الى أرقام على اساس هذه الأرقام يتم حساب المسافة بين مركز التجمع والبيانات.
4. تجميع البيانات وتنظيمها في 5 مجموعات بناءً على أقل المسافات بين المركز ونقاط البيانات.

عاده يتم تنفيذ الخطوات من 2– 3 حتى الوصول إلى حاله الثبات.

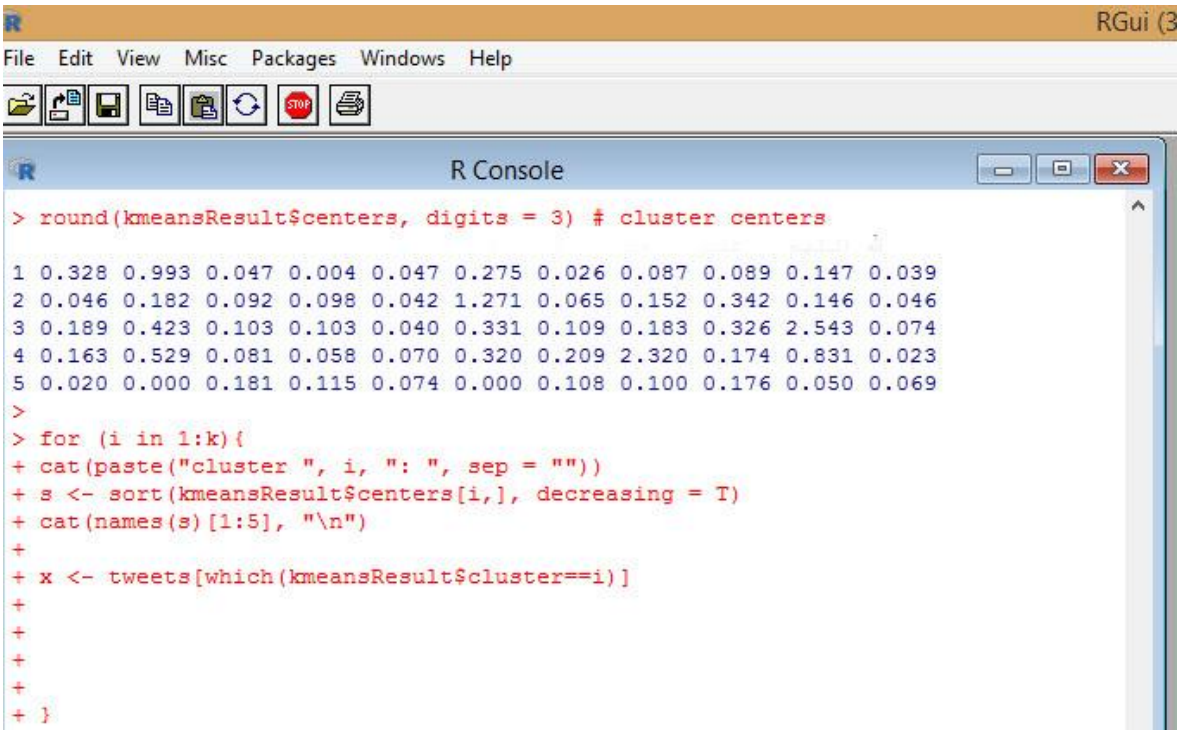
على سبيل المثال نفترض البيانات التالية :

Index	Data
0.9	(ولله ما في السماوات وما في الأرض ۚ وإلى الله ترجع الأمور)
0.3	اغلب الظن موظفي المنشآت للحكومية في السودان بيفكروا انك تكون مستفز و تافه= انت زول جادي
0.6	خلاص انا قررت أقول لكم الحقيقة واللي يحصل يحصل، انا النمر المقنع
0.1	لا إله إلا الله وحده لا شريك له له الملك وله الحمد يحيي ويميت وهو على كل شئ قدير
0.2	مشروع بئر الريان
0.7	تنزل المطر الشبكة تطش النت يقطع الكهرباء تقطع على قولة صحبتي مطر هي ولا موية نار
0.8	أستغفر الله العظيم وأتوب إليه أذكار الأذان:يدعو لنفسه بين الأذان والإقامة فإن الدعاء حينئذٍ لا يرد
0.5	الدولار فات ال17 كلها كم يوم و حيقدم للرخصة
0.4	السباح السوداني احمد جبريل المركز 47 من اصل 47 مشارك ؟! الحمد لله انو ماغرق...
0.0	كان النبي صلى الله عليه وسلم يعتكف في كل رمضان عشرة أيام فلما كان العام الذي قبض فيه اعتكف عشرين يوما - صحيح البخاري

الجدول (1.6.5) يمثل جزء من البيانات و Indexالمقابل

المجموعة	مركز التجمع
الاولى (الاراء السياسية)	C1 =0.9
التانية (الاراء الدينية)	C2 =0.7
التالتة (الترفيه)	C3 =0.5
الرابعة (الرياضية)	C4 =0.3
الخامسة(اخرى)	C5 =0.1

الجدول (2.6.5) يمثل مركز التجمع العشوائي لكل مجموعة



الشكل (1.6.5) تحديد مراكز التجمعات عشوائياً لكل فئة باستخدام ال R

نقوم بحساب المسافة بين البياتات ومراكز التجمع بالشكل التالي :

Index	الفرق بين Index و (Centroid)
0.9	0.9 - 0.9 =0
	0.9 - 0.3=0.6
	0.9 - 0.1 = 0.8
	0.9 -0.7 = 0.2 اقل مسافة
	0.9 -0.5 = 0.4

0.3	0.3 - 0.9 =-0.6 0.3 - 0.3=0 **0.3 - 0.1 = 0.2** 0.3 -0.7 = -0.4 0.3 -0.5 = -0.2
0.6	0.6 - 0.9 =-0.3 0.6 - 0.3=0.3 0.6 - 0.1 = 0.6 0.6 -0.7 = -0.1 **0.6 -0.5 = 0.1**
0.1	0.1 - 0.9 =-0.8 **0.1 - 0.3=0.2** 0.1 - 0.1 = 0 0.1 -0.7 = -0.6 0.1 -0.5 = -0.4
0.2	0.2 - 0.9 =-0.7 0.2 - 0.3= -0.1 **0.2 - 0.1 = 0.1** 0.2 -0.7 = -0.5 0.2 -0.5 = -0.3
0.7	0.7 - 0.9 =-0.2 0.7 - 0.3=0.4 0.7 - 0.1 = 0.6 0.7 -0.7 = -0 **0.7 -0.5 = 0.2**
0.8	0.8 - 0.9 =-0.1 0.8 - 0.3=0.5

	0.8 - 0.1 = 0.7 **<u>0.8 -0.7 = 0.1</u>** 0.8 -0.5 = 0.3
0.5	0.5 - 0.9 =-0.4 **<u>0.5 - 0.3=0.2</u>** 0.5 - 0.1 = 0.4 0.5 -0.7 = -0.2 0.5 -0.5 = 0
0.4	0.4 - 0.9 =-0.5 **<u>0.4 - 0.3=0.1</u>** 0.4 - 0.1 = 0.3 0.4 -0.7 = -0.3 0.4 -0.5 = -0.1
0.0	0.0 - 0.9 =-0.9 0.0 - 0.3=0.3 **<u>0.0 - 0.1 = 0.1</u>** 0.0 -0.7 = -0.7 0.0 -0.5 =0.5

نحيل كل نقطة إلى مركز تجمع بالاعتماد على أقل مسافة

Index	Data	Group
0.9	(ولله ما في السماوات وما في الأرض ۚ وإلى الله ترجع الأمور)	C2 = 0.7 (الآراء الدينية)
0.3	اغلب الظن موظفي المنشآت للحكومية في السودان بيفكروا انك تكون مستفز و تافه= انت زول جادي	C5 =0.1 (اخرى)
0.6	خلاص انا قررت أقول لكم الحقيقة واللي يحصل يحصل، انا النمر المقنع	C3 =0.5 (الترفيهيه)
0.1	لا إله إلا الله وحده لا شريك له له الملك وله الحمد يحيي ويميت وهو على كل شئ قدير	C4 =0.3 (الرياضية)

C5 =0.1 (اخرى)	مشروع بئر الريان	0.2
C3 =0.5 (الترفيهيه)	تنزل المطر الشبكة تطش النت يقطع الكهرباء تقطع على قولة صحبتي مطر هي ولا موية نار	0.7
C2 =0.7 (الدينية)	أستغفر الله العظيم وأتوب إليه أذكار الأذان:يدعو لنفسه بين الأذان والإقامة فإن الدعاء حينئذٍ لا يرد	0.8
C4 =0.3 (الرياضية)	الدولار فات ال17 كلها كم يوم و حيقدم للرخصة	0.5
C4 =0.3 (الرياضية)	السباح السوداني احمد جبريل المركز 47 من اصل 47 مشارك ؟! الحمد لله انو ماغرق...	0.4
C5 =0.1 (اخرى)	كان النبي صلى الله عليه وسلم يعتكف في كل رمضان عشرة أيام فلما كان العام الذي قبض فيه اعتكف عشرين يوما - صحيح البخاري	0.0

الجدول (2.6.5) يمثل تجميع البيانات في المجموعات اعتمادا على اقرب مركز لها

من الجدول السابق نجد ان:

C1	C2	C3	C4	C5
	0.9	0.6	0.1	0.2
	0.8	0.7	0.4	0.0
			0.5	0.3

نقوم بحساب مراكز التجمع الجديدة والتي تساوي : مجموع البيانات لكل مركز تجمع

عدد البيانات في كل مركز تجمع

C1	C2	C3	C4	C5
0.9	$1.2=\frac{0.8+0.9}{2}$	$0.6=\frac{0.7+0.6}{2}$	$=\frac{0.5+0.4+0.1}{2}$ / 0.5	$0.1=\frac{0.2+0.3+0.0}{3}$

المجموعة	مركز التجمع
الاولى (الآراء السياسية)	C1 = 0.9
التانية (الآراء الدينية)	C2 = 1.2
التالتة (الترفيهيه)	C3 = 0.6
الرابعة (الرياضية)	C4 =0.5
الخامسة(اخرى)	C5 = 0.1

الجدول (3.6.5) يمثل مركز التجمع الجديدة لكل مجموعة

نقوم مرة أخرى بحساب المسافة بين البياتات ومراكز التجمع الجديدة بالشكل التالي :

Index	الفرق بين Index و (Centroid)
0.9	0.9 - 0.9 =0 0.9 – 1.2=0.3 0.9 - 0.1 = 0.8 **اقل مسافة 0.9 -0.6= 0.2** 0.9 -0.5 = 0.4
0.3	0.3 - 0.9 =-0.6 0.3 – 1.2=0.9 **0.3 - 0.1 = 0.2** 0.3 -0.6= -0.4 0.3 -0.5 = -0.2

0.6	0.6 - 0.9 =-0.3 0.6 – 1.2=0.6 0.6 - 0.1 = 0.5 0.6 -0.6 = 0 **0.6 -0.5 = 0.1**
0.1	0.1 - 0.9 =-0.8 0.1 – 1.2=1.1 0.1 - 0.1 = 0 **0.1 -0.6= -0.5** 0.1 -0.5 = -0.4
0.2	0.2 - 0.9 =-0.7 0.2 – 1.2= 1 **0.2 - 0.1 = 0.1** 0.2 -0.6 = -0.4 0.2 -0.5 = -0.3
0.7	0.7 - 0.9 =-0.2 0.7 – 1.2=0.5 0.7 - 0.1 = 0.6 **0.7 -0.6 = 0.1** 0.7 -0.5 = 0.2
0.8	**0.8 - 0.9 =-0.1** 0.8 – 1.2=0.4 0.8 - 0.1 = 0.7 0.8 -0.6 = 0.2 0.8 -0.5 = 0.3
0.5	0.5 - 0.9 =-0.4 0.5 – 1.2=0.7

	0.5 - 0.1 = 0.4 **0.5 -0.6 = -0.1** 0.5 -0.5 = 0
0.4	0.4 - 0.9 =-0.5 0.4 – 1.2=0.8 0.4 - 0.1 = 0.3 0.4 -0.6 = -0.2 **0.4 -0.5 = -0.1**
0.0	0.0 - 0.9 =-0.9 0.0 – 1.2=1.2 **0.0 - 0.1 = 0.1** 0.0 -0.6 = -0.6 0.0 -0.5 =0.5

Index	Data	Group
0.9	(ولله ما في السماوات وما في الأرض وإلى الله ترجع الأمور)	C2 = 0.7 (الآراء الدينية)
0.3	اغلب الظن موظفي المنشآت للحكومية في السودان بيفكروا انك تكون مستفز و تافه= انت زول جادي	C5 = 0.1(اخرى)
0.6	خلاص انا قررت أقول لكم الحقيقة واللي يحصل يحصل، انا النمر المقنع	C4 =0.5 (الرياضية)
0.1	لا إله إلا الله وحده لا شريك له له الملك وله الحمد يحيي ويميت وهو على كل شئ قدير	C4 =0.3 (الرياضية)
0.2	مشروع بئر الريان	C5 =0.1 (اخرى)
0.7	تنزل المطر الشبكة تطش النت يقطع الكهرباء تقطع على قولة صحبتي مطر هي ولا موية نار	C3 = 0.6(الترفيهيه)
0.8	أستغفر الله العظيم وأتوب إليه أذكار الأذان:يدعو لنفسه بين الأذان والإقامة فإن الدعاء حينئذٍ لا يرد	C2 =0.7 (الدينية)
0.5	الدولار فات ال17 كلها كم يوم و حيقدم للرخصة	C5 = 0.1(اخرى)

C4 =0.5 (الرياضية)	السباح السوداني احمد جبريل المركز 47 من اصل 47 مشارك ؟! الحمد لله انو ماغرق...	0.4
C5 = 0.1(اخرى)	كان النبي صلى الله عليه وسلم يعتكف في كل رمضان عشرة أيام فلما كان العام الذي قبض فيه اعتكف عشرين يوما - صحيح البخاري	0.0

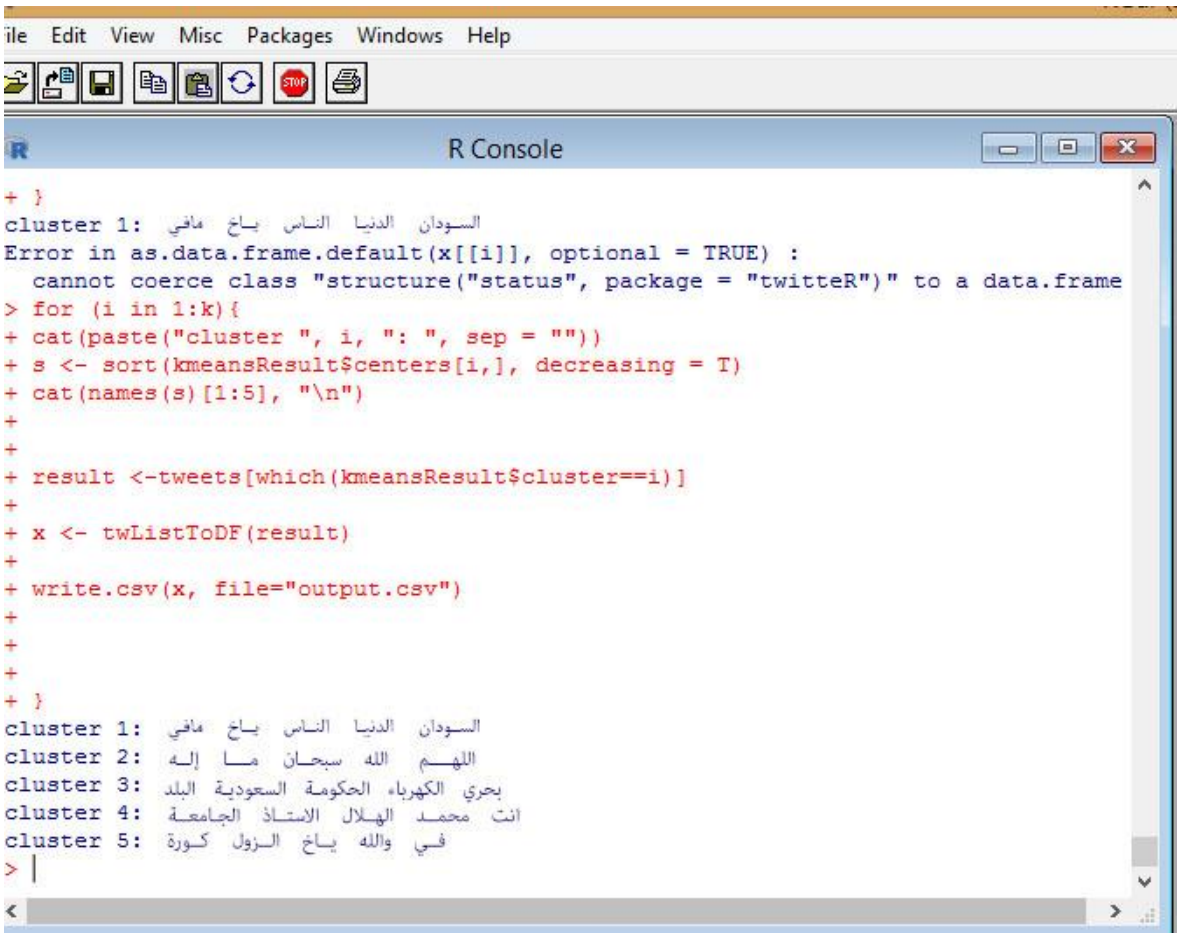
الجدول (4.6.5) يمثل تجميع البيانات في المجموعات اعتماداً على اقرب مركز لها(بالنسبة لمراكز التجمع الجديدة)

وهكذا تتكرر الخطوات السابقة حتى الوصول الى الثبات بعنى عدم إنتقال عنصر من مجموعة الى اخرى .

ملاحظة :

1. يعتمد أداء هذه الخوارزميه على المواقع الأولية لمراكز التجمع, ومن المستحسن تنفيذ هذه الخوارزميه عدة مرات مع إختلاف المراكز في كل مرة عن المرات السابقة.

2. المثال السابق تم شرحه بشكل مبسط لتوضيح فكرة عمل الخوارزمية.

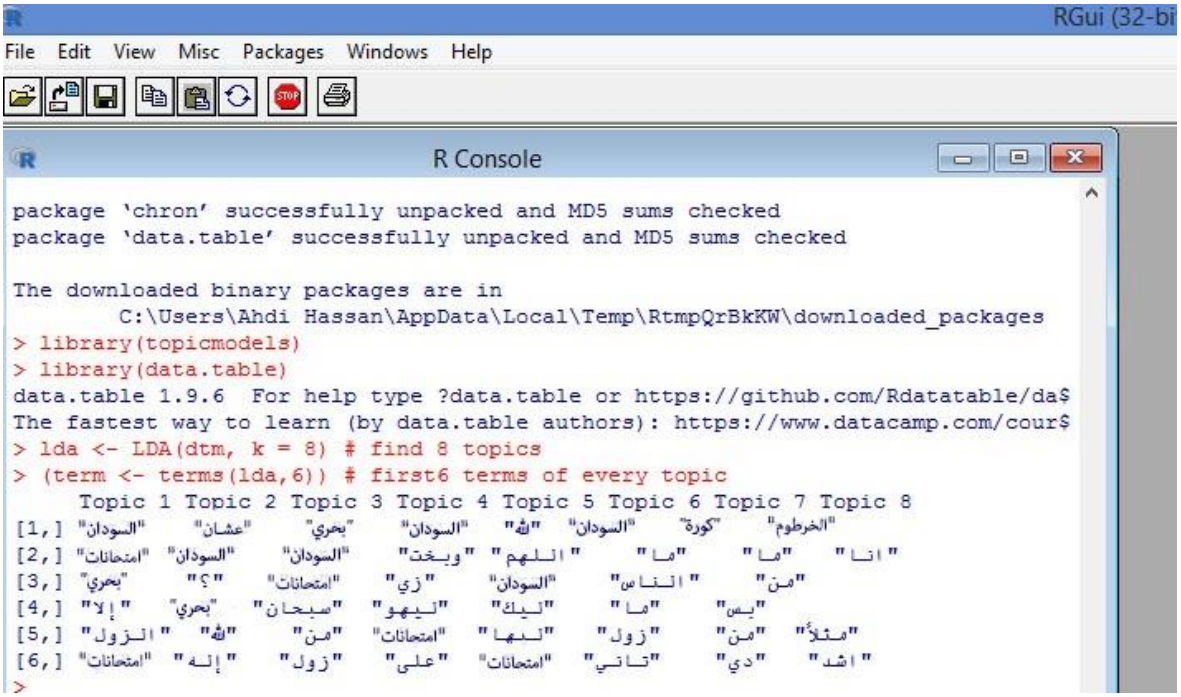


الشكل(2.6.5) نتيجة تطبيق خوارزمية Kmeans على التغريدات

7.5 تطبيق Topic modeling على التغريدات

Topic modeling هو المنهج الإحصائي لإكتشاف "ملخصات" من مجموعة من الوثائق والنصوص بناءَ على إحصاءات من كل كلمة. بعبارات بسيطة، عملية النظر في مجموعة كبيرة من الوثائق وتحديد مجموعات من الكلمات وتجميعها معاً على أساس التشابه وتحديد أنماط في المجموعات التي تظهر في الكثرة.

- تحميل الحزم المطلوبة وهي (topicmodels) وإستدعاء مكتباتها .
- تحديد عدد الموضوعات والمصطلحات لكل موضوع.
- تطبيق خوارزمية Latent Dirichlet Allocation algorithm (LDA) (هو الاسلوب الذي يكتشف الموضوعات التي تحتويها الوثائق تلقائياً).



الشكل (7.5) أكثر الموضوعات تداولا في التغريدات السودانية

الباب السادس

النتائج والتوصيات

Results and Recommendations

1.6 مقدمة

يتناول هذا الفصل النتائج التي تم التوصل اليها والتوصيات التي تخص المشروع.

2.6 النتائج :

بعد تطبيق خوارزمية (K means) على التغريدات السودانية تم تقسيم الآراء كالتالي:

المجموعة	النسبة
الآراء الدينية	%36.41
الآراء السياسية	%27.20
الآراء الترفيهيه	%13.03
الآراء الرياضية	%11.51
اخرى	%11.85

3.6 التوصيات :

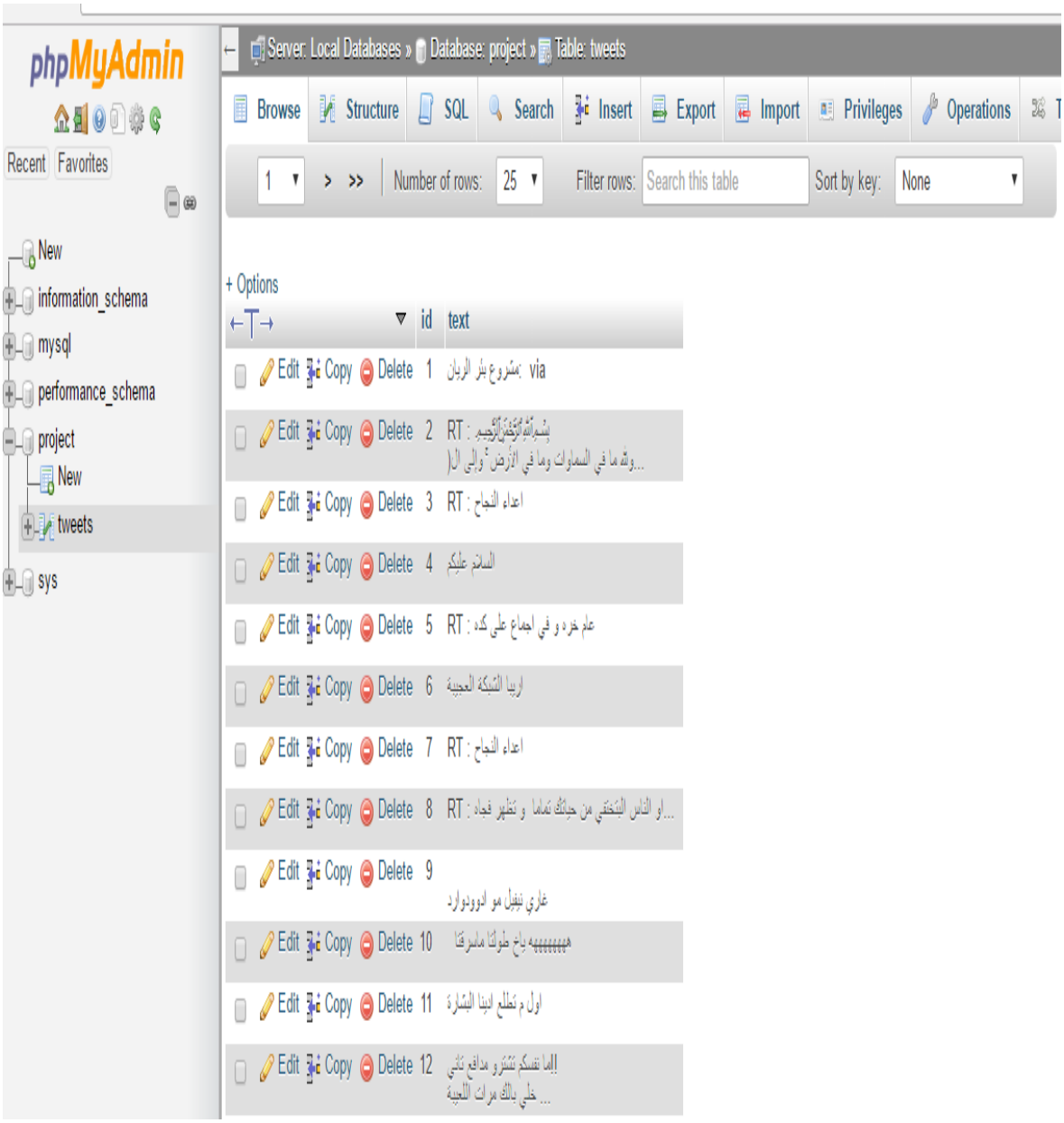
1. يمكن تطبيق خوارزميات التصنيف Classification algorithms وخوارزميات تجميع اخرى و مقارنة النتائج بنتائج المتحصلة عليها.
2. إضافة بيانات من وسائل تواصل إجتماعية اخرى مثل الفيسبوك .
3. تحليل اعمق للبيانات.

الخاتمة

تم بحمد الله ما اردنا جمعه وكتابته عن تحليل التغريدات السودانية وتصنيفها الى الفئات المحددة سلفاً حيث قمنا بإستخدام خوارميات التجميع في الذكاء الإصطناعي وتطبيقها على التغريدات السودانية المخزنة في قاعدة البيانات .

هذا واسأل الله أن ينفع به إخواني المسلمين ، كما أساله جل شأنه أن يجعله في موازين أعمالي وأن يكون حجه لي لا علي يوم أن ألقاه إنه جواد كريم و بالإجابه قدير . فإن أصبت فمن الله تعالى وإن أخطأت فمن نفسي ومن الشيطان وكما قال الله تعالى : (" لا يكلف الله نفسا إلا وسعها ").

الملاحق



قاعدة البيانات

المراجع

[1] Dr. Dinesh Chand Sharma , Analysis on new Avenues in Opinion mining and Sentiment Analysis , International JournalOf Engineering Research and General ScienceVolume 3, Issue 1,January-February, 2015 ISSN 2091-2730 ,page 856,857

[2] سميحة ناصر خليف, أنواع مواقع التواصل الإجتماعي, آخر تحديث :٠٢:٥٨ ، ٢٨ يناير ٢٠١٦, تاريخ الإطلاع 20 مايو 2016

http://mawdoo3.com/%D8%A3%D9%86%D9%88%D8%A7%D8%B9_%D9%85%D9%88%D8%A7%D9%82%D8%B9_%D8%A7%D9%84%D8%AA%D9%88%D8%A7%D8%B5%D9%84_%D8%A7%D9%84%D8%A7%D8%AC%D8%AA%D9%85%D8%A7%D8%B9%D9%8A

محمود العراقي ,الموضوع الشامل عن مواقع الشبكات الإجتماعية , تاريخ الإطلاع 20 يناير 2016

/http://www.traidnt.net/vb/traidnt1921726

السرحاني, عطا الله بن فهد , شبكات التواصل الاجتماعي ,2014-12-07T16:01:33Z , جامعة نايف العربية للعلوم الأمنية, تاريخ الاطلاع 22 يناير 2016

http://repository.nauss.edu.sa/handle/123456789/59614

[3] خالد عمر, 10 معلومات لا تعلمها عن موقع التواصل الاجتماعي "تويتر " , أخر تحديث الاثنين 16 نوفمبر 2015 - 20:49 م القاهرة, تاريخ الإطلاع 3 فبراير 2016

http://www.rassd.com/165482.htm
مصطفى ثابت , تويتر (Twitter) , السبت 21/مايو/2016 - 08:43 م , تاريخ الإطلاع 7 فبراير 2016

http://www.elfagr.org/2146146

[4] أحمد الأحمدي, ما هي ال APIs و علاقتها بتطبيقات الطرف الثالث (third party apps) ؟ ,تاريخ الإطلاع 10 فبراير 2016

https://io.hsoub.com/webdev/3299-%D9%85%D8%A7-%D9%87%D9%8A-%D8%A7%D9%84-apis

https://dev.twitter.com/rest/public/search

https://dev.twitter.com/rest/public

https://dev.twitter.com/streaming/overview

https://dev.twitter.com/overview/documentation

[5] Syed Akib Anwar Hridoy, M. Tahmid Ekram, Mohammad Samiul Islam, Faysal Ahmed and Rashedur M. Rahman, Localized twitter opinion mining using sentiment analysis ,Decision Analytics, Volume 2, Number 1, Page 1, 22 October 2015, ISSN: 2193-8636
https://decisionanalyticsjournal.springeropen.com/articles/10.1186/s40165-015-0016-4

[6] Jong Woo Kim and Dong Sung Kim ,Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power

Advanced Science and Technology Letters

Vol.51 (CES-CUBE 2014), pp.224-228

http://dx.doi.org/10.14257/astl.2014.51.5
http://onlinepresent.org/proceedings/vol51_2014/51.pdf

International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 11 (2014), pp. 373-384
http://dx.doi.org/10.14257/ijmue.2014.9.11.3
http://www.sersc.org/journals/IJMUE/vol9_no11_2014/36.pdf

[7]رزان نجار, دراسة: تغريدات الغضب تعد مؤشرا للإصابة بأمراض القلب ! الإثنين 26 يناير 2015

آخر تحديث: الاثنين 6 ربيع الثاني 1436هـ - 26 يناير 2015م02:12 GMT - 05:12 KSA
https://news.webteb.com/%D8%AF%D8%B1%D8%A7%D8%B3%D8%A9-%D8%AA%D8%BA%D8%B1%D9%8A%D8%AF%D8%A7%D8%AA-%D8%A7%D9%84%D8%BA%D8%B6%D8%A8-%D8%AA%D8%B9%D8%AF-%D9%85%D9%88%D8%B4%D8%B1%D8%A7-%D9%84%D9%84%D8%A7%D8%B5%D8%A7%D8%A8%D8%A9-%D8%A8%D8%A7%D9%85%D8%B1%D8%A7%D8%B6-%D8%A7%D9%84%D9%82%D9%84%D8%A8_3804

نيويورك ـ أ.ش.أ, الأربعاء / يناير 2015 GMT 06:11 , دراسة تؤكد أن التغريدات السلبية مؤشر للإصابة بأمراض القلب, تاريخ الإطلاع 4 مارس 2016

http://www.emiratesvoice.com/207/%D8%AF%D8%B1%D8%A7%D8%B3%D8%A9-%D8%AA%D8%A4%D9%83%D8%AF-%D8%A3%D9%86-%D8%A7%D9%84%D8%AA%D8%BA%D8%B1%D9%8A%D8%AF%D8%A7%D8%AA-%D8%A7%D9%84%D8%B3%D9%84%D8%A8%D9%8A%D8%A9-%D9%85%D8%A4%D8%B4%D8%B1-%D9%84%D9%84%D8%A5%D8%B5%D8%A7%D8%A8%D8%A9-%D8%A8%D8%A3%D9%85%D8%B1%D8%A7%D8%B6-%D8%A7%D9%84%D9%82%D9%84%D8%A8

[8] محمد علام, 25Trends أول أداة مصرية لتحليل المشاركات الإجتماعية, تاريخ الإطلاع, 17 يوليو 2016

https://aitnews.com/2013/04/16/25trends-%D8%A3%D9%88%D9%84-%D8%A3%D8%AF%D8%A7%D8%A9-%D9%85%D8%B5%D8%B1%D9%8A%D8%A9-%D9%84%D8%AA%D8%AD%D9%84%D9%8A%D9%84-%D8%A7%D9%84%D9%85%D8%B4%D8%A7%D8%B1%D9%83%D8%A7%D8%AA-%D8%A7%D9%84%D8%A7%D8%AC/

[9] أحمد عبدالقادر, موقع Twitalyzer خدمة لتحليل محتوي الشبكات الإجتماعية, تاريخ الإطلاع 17 يوليو, 2016

https://aitnews.com/2010/05/17/12791/

[10] أحمد عبدالقادر,AsTweeted أول أداة مصرية لأرشفة التغريدات,تاريخ الإطلاع 17 يوليو, 2016

https://aitnews.com/2013/04/15/astweeted-%D8%A3%D9%88%D9%84-%D8%A3%D8%AF%D8%A7%D8%A9-%D9%85%D8%B5%D8%B1%D9%8A%D8%A9-%D9%84%D8%A3%D8%B1%D8%B4%D9%81%D8%A9-%D8%A7%D9%84%D8%AA%D8%BA%D8%B1%D9%8A%D8%AF%D8%A7%D8%AA/

[11] د.فاطمة السالم, تحليل للرأي العام الإلكتروني عقب استهداف الأماكن المقدسة: مؤامرة ضد السعودية... وإيران متهمة, تاريخ الإطلاع 20 أغسطس, 2016

/http://alqabas.com/163631

[12] طة مختار البابا , تعيين المراكز الإبتدائية بشكل مدروس في خوارزمية K- Medoids, مجلة جامعة البعث ,المجلد 36 , العدد 2 , 2014 , الصفحات 63 ,64 ,65 , 67 , 68 , 70 , 71 .

file:///C:/Users/user/Downloads/document_1b20255d2fdbb42436671e482f3bacd4(1).pdf

[13] جامعة حلب, كلية الهندسة الكهربائية والإلكترونية, قسم هندسة الحواسيب, دراسة حول: تجميع البيانات Data clustering, تاريخ الإطلاع 4 يوليو 2016

موسوخة علوم الحاسب , تجميع البيانات (Data clustering),تاريخ الإطلاع

https://compuscienpedia.wordpress.com/2010/07/05/%d8%aa%d8%ac%d9%85%d9%8a%d8%b9-%d8%a7%d9%84%d8%a8%d9%8a%d8%a7%d9%86%d8%a7%d8%aa-data-clustering/

[14] خالد الشمعة, تلخيص البيانات واختصارها عبر تحليل المكونات الرئيسية (PCA) في لغة R , 12 يونيو 2015 ,تاريخ الإطلاع 10 يوليو 2016

/https://academy.hsoub.com/programming/r-language

[16] Valli Pandy, How to Insert JSON Data into MySQL using PHP
http://www.kodingmadesimple.com/2014/12/how-to-insert-json-data-into-mysql-php.html

[15] عبدالله الحامدي, مقدمة لـMySQL , تاريخ الإطلاع 1 أغسطس 2016

http://blog.w3teacher.net/%D9%85%D9%82%D8%AF%D9%85%D8%A9-%D9%84%D9%80-mysql/#.WAHwGDXQMzU

محمد أحمد العيل, مدخل إلى برنامج إدارة قواعد البيانات MySQL

https://academy.hsoub.com/devops/servers/databases/mysql/%D9%85%D8%AF%D8%AE%D9%84-%D8%A5%D9%84%D9%89-%D8%A8%D8%B1%D9%86%D8%A7%D9%85%D8%AC-%D8%A5%D8%AF%D8%A7%D8%B1%D8%A9-%D9%82%D9%88%D8%A7%D8%B9%D8%AF-%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA-/mysql-r28

زايد السيد ,_ما هي MySQL , تاريخ الإطلاع 7 أغسطس 2016

http://itwadi.com/what-is-mysql

[16] **ما هو PHP و لماذا أنا بحاجة إليه؟ , تاريخ الإطلاع 20 أغسطس 2016**

http://www.webbloog.com/2013/03/Part-1-What-is-PHP.html

ما هي لغة البرمجة بي اتش بيphp , تاريخ الإطلاع 1 سبتمبر 2016 , تاريخ النشر 2013/03/14

http://mafhome.com/%D9%85%D8%A7-%D9%87%D9%8A-%D9%84%D8%BA%D8%A9-%D8%A7%D9%84%D8%A8%D8%B1%D9%85%D8%AC%D8%A9-/%D8%A8%D9%8A-%D8%A7%D8%AA%D8%B4-%D8%A8%D9%8A-php

[17] أحمد الزوكي, مزايا لغة الـ" PHP“ , تاريخ الإطلاع 20 سبتمبر 2016,

تاريخ النشر الخميس, 19/فبراير 2004 10:45

http://tibanews.com/ss/tibanet/index.php/php/intro-php/basic-php

1 Yassine Bajdou, شرح كامل ومبسط لتثبيت السيرفر المحلي WAMP Server , تاريخ الإطلاع اكتوبر 2016

[18] Posted by Valli Pandy On 12/08/2014,How to Insert JSON Data into MySQL using PHP

http://www.kodingmadesimple.com/2014/12/how-to-insert-json-data-into-mysql-php.html

الشكل (7.3.2) نتيجة تحليل الرأي العام الإلكتروني عقب إستهداف الأماكن المقدسة