

# **TWITTER ANALYTICS USING SAP HANA AND SAP LUMIRA**

Project Submitted to

The College of Sciences and Engineering

Southern University and A&M College

In Partial Fulfillment of the Requirements for

The Degree of

Master of Science in

Computer Science

By

Mahesh Gottam

Baton Rouge, Louisiana

December, 2016



# **TWITTER ANALYTICS USING SAP HANA AND SAP LUMIRA**

Name: Mahesh Gottam

APPROVED BY

---

**Dr. Mohammad Abdus Salam**  
Committee Chair

---

**Dr. Shizong Yang**  
Committee Member

---

**Dr. Jose Noguera**  
Committee Member

---

**Dr. Ebrahim Khosravi**  
Chair, Department of Computer Science

## TABLE OF CONTENTS

1. INTRODUCTION.....	1
1.1 Background.....	1
1.2 Significance.....	2
1.3 Problem Statement.....	2
1.4 Research Questions.....	3
1.5 Objectives.....	3
1.6 Delimitations.....	4
2. REVIEW OF RELATED LITERATURE.....	5
3. METHODOLOGY.....	8
3.1 Design.....	8
3.2 Requirements.....	11
3.3 Implementation.....	12
3.4 Timeline.....	27
4. CONCLUSION AND FUTURE WORK.....	28
REFERENCES.....	29
APPENDIX.....	30

## **LIST OF FIGURES**

1. Architecture for Development of Application for Text Analysis
2. Create a new application on twitter account
3. Consumer Key and Consumer Secret Key
4. Access Token and Access Token Secret Keys
5. Insertion of Twitter tweets
6. Accessing the HANA database from SAP LUMIRA
7. Tweets stored on the Hana Database
8. Full Index Table for Text Analysis
9. Figure showing the Analysis of data using the measures id by TA\_TYPE
10. Filtering of Tweets using Sentiment
11. Figure showing Positive and Negative Sentiments
12. Figure showing the tag cloud of obtained tweets
13. Stance vs Influence by user
14. Table showing the attributes for stance and influence
15. Table showing the clustering attributes
16. Table showing the Signature of Centers, Results, Data, Params
17. Table showing the Cluster Number and Distance between the nodes
18. Table showing the Centers of the different clusters
19. Table showing the data of users by stance and influence
20. Table showing users by ClusterNumber
21. Table showing the TweepersClusteredSummary
22. Stance and influence by ClusterNumber and user
23. Figure showing the stance and influence by ClusterNumber
24. Timeline for the Project

# **CHAPTER 1**

## **INTRODUCTION**

We are going to get to some continuous online networking data (sample twitter) so we are going to get data on particular tweets that we need to search for and haul those out continuously from twitter API, process those tweets and load that tweets progressively to HANA database.

HANA database will be facilitated on cloud platform which is located in CHICO University and maintain a server for cloud and SAP HANA development for students and faculty or trial HANA cloud platform Developer release for one month on SAP official website and we can register for it and we can use it.

Data Storage is the key and we store data in HANA database and we can do analysis on it say for instance assessment of particular products or persons or anything which can be analyzed using a particular platform such as to comprehend what individuals are stating, enjoying, disliking whatever item or administration tweets that we need to investigate. At that point we might need to fragment or gathering distinctive sorts of persons (twitter clients) taking into account their tweets. There might be individuals who might be stating numerous antagonistic things about the items or administration and there might be some compelling since parcel of individuals might take after and retweets, we then target such individuals take choice in like manner instruct them and so on. So we do some prescient examination to fragment bunches sending comparable sorts of tweets together by utilizing SAP Predictive Analytics Library. We can interface with SAP HANA Database facilitated in cloud through shroud IDE where we introduce and design SAP HANA studio. At that point utilizing this HANA studio we get to HANA database and do operations required, for example, SQL operation or displaying and so forth utilizing SQL editor from HANA studio.

### **1.1 BACKGROUND**

Numerous organizations have been precisely and thoroughly putting away information for a considerable length of time. The information is in assortment of structures like online networking posts, email, websites, news, input, tweets, and Business reports and so on. It is vital to remove important data without reading each and every sentence. For instance on the

off chance that we consider an organization needs to think about what clients are saying in regards to their brands or items? For that we perform Text Analysis utilizing SAP Hana and SAP Predictive Analysis.

The objective of this undertaking is to furnish organizations with an extra, capable intends to extend their business sectors by centering their development impacts on particular districts and demographics. Information Visualization and Statistical strategies will be utilized to decide designs among the relationships between different measure of information moving through online networking and sorts of constant information.

## **1.2 SIGNIFICANCE**

These days, enormous collections of data which is in unstructured format and huge amount of information are being caught in CRM, blogs, and call centers and additionally online social networking, websites, discussions, e-mails, reports, and so forth. Organizations are attempting to extricate significant, organized data from the immense volume information, attempting to discover and examine the substance to offer their business methodology some assistance with running and tackle basic issues. For all this issues we use Text Analysis, an innovation to structure, change, or advance unstructured information with the end goal of revelation or investigation. Content Analysis in SAP HANA can extricate the significant data from writings, apply proper phonetic principles for the specific dialect and after that semantically translate the information into a visual statistic.

## **1.3 PROBLEM STATEMENT**

Concentrating on the information streaming in this present reality and online networking, we are going to build up an application to perform the content examination and concentrate the information, sustaining to cloud database like SAP Hana and after that performing the information representation utilizing SAP Predictive Analysis. In this task it mainly concentrates on the most proficient method to get to the constant online networking data utilizing Twitter and interact with Twitter API and collect the information into SAP Hana and perform the supposition examination and in addition Segmentation utilizing Hana Clustering. Once the information has been stacked we will discover whether individuals are tweeting

positive or negative proclamations and afterward distinctive sorts of tweeters can be gathered and sectioned by evaluating the impact of tweeter utilizing the Predictive Analysis Library.

## **1.4 RESEARCH QUESTIONS**

- Which platform is needed and efficient to develop an application?
- Is it necessary to add web services for login and registration of end users?
- How to fetch the old data from the twitter if needed?
- Can we fetch the data of local organization for Text Analysis?

## **1.6 OBJECTIVES**

The main objective of this task is to store the ongoing information from social network (twitter) and this information is used for text analysis and also sentiment analysis and additionally perform Segmentation utilizing SAP HANA clustering. In this contextual analysis we are going to concentrate on the most proficient method to get to constant online networking data utilizing Twitter. In light of our intrigued pattern that we need to break down we haul out the particular tweets progressively utilizing Twitter API, then process and stacked this information into a SAP HANA database. To interact with the Twitter API and stack the information into SAP HANA, we can make utilize node.js which is a free open source. Task can be finished by utilizing SAP HANA (in memory database) and Its Predictive Analysis Library calculations in a down to earth approach. Once the information has been stacked we will perform data analysis. Sentiment Analysis is performed by discovering whether individuals are tweeting positive or negative articulations. At that point distinctive sorts of tweeters can then be assembled and portioned by assessing the impact of tweeter utilizing the Predictive Analysis Library. For instance, we can figure out who is powerful because of a high number of adherents and re-tweets and afterward amass the persuasive tweeters who are communicating negative contemplations around an item together. We can then focus on that gathering with an instructive effort program. Be that as it may, as a contextual analysis here I am not going to concentrate on a specific item since it's up to the client necessity who need to concentrate on which sort of information from online networking. In this we are simply concentrating on the most proficient method to bolster the real time information nourishes from social networking site (twitter) and how this information, then can be utilized to perform the

content analysis for slant investigation what's more, and also Segmentation utilizing SAP HANA clustering.

## **1.7 DELIMITATIONS**

There will be some limitations to this project:

- We use a particular word to extract the tweets related to particular word.
- Application is only for real time data feeds.
- Every tweets cannot be retrieved at once and also the fetched data depends on the keyword we gave at the time of interacting with the Twitter API.



## CHAPTER 2

### REVIEW OF RELATED LITERATURE

Twitter is a well-known social networking blog and long range informal communication administration that introduces numerous open doors for analysis in Natural Language Processing (NLP) and machine learning. Since its initiation in 2006 [1], Twitter has developed to the point where <http://twitter.com> is the eleventh most visited site in the world, and the eighth most visited site in the United States, and more than 100 million Twitter accounts have been made [1].

Clients of Twitter post messages, called "tweets," on an assortment of themes, extending from news occasions and popular society, to unremarkable day by day occasions and spam postings. As of February 2015, clients of Twitter were creating 140 million tweets for each day, a normal of 1680 tweets per second [1].

Twitter introduces some captivating opportunities for applications of NLP and machine learning. One such part of Twitter that gives opportunities is slanting points - words and phrases, highlighted on the primary page of Twitter, that are as of now famous in clients' tweets. The fame and development of Twitter exhibits a few difficulties for uses of NLP and machine adapting, be that as it may. The length limitations of the messages make linguistic and basic traditions that are not seen in more conventional corpora, and the extent of the Twitter system produces a consistently changing, dynamic corpus [1]. In expansion, there is a considerable amount of substance on Twitter that would be named irrelevant to an outside onlooker, comprising of individual data or spam, which should be sifted through in request to precisely recognize the components of the corpus that are important to the Twitter group all in all, and could in this way be thought to be potential inclining subjects. The challenge of Twitter's prominence is that to distinguish and recognize inclining subjects, one must specimen and analyze an extensive volume of spilling information [2].

A significant part of the data driving your business originates from online sources like email, tweets, what's more, sites – and the quantity of pages on the Web duplicates every day. Your foundation applications produce always information as well, as they grow to handle new ongoing information from business forms. To catch understanding from online information,

you should distil client sentiments, recognitions, and encounters and at that point consolidate that data with information from SAP and non-SAP back-office applications to nourish examination arrangements. What's more, to keep pace with the data blast, you require an in-memory figuring motor to control the coordinated procedure. It saddles in-memory investigation to bring you a new business esteem by offering you, for instance, some assistance with monitoring and enhance advertising effort, plan fruitful item improvements, and oversee valuing also, marketing taking into account quick request [2].

The arrangement of data is preconfigured to help you obtain unstructured information from Facebook, Twitter, and Google+. It likewise underpins coordination with social information frameworks from DataSift furthermore, Gnip and with the SAP Jam social programming platform [2]. A setup guide contains proposals for collecting unstructured information from an assortment of other online sources. You can perform semantic extraction and information load from the sources you indicate onto the SAP HANA platform, nourishing concoction relationship investigations in the middle of campaign and administration information. You can coordinate unstructured online networking and other content information straightforwardly with crusade and administration devices in the SAP Customer Relationship Management application. Preconfigured scientific substance gives end clients quick understanding through staggering graphical investigation of client conclusion in an interface predictable over their favored devices [2].

The SAP HANA Sentiment Intelligence rapid deployment arrangement utilizes content information handling usefulness in SAP Data Services programming to parse and disambiguate unstructured information [3]. In impact, it gives you a chance to extricate usable significance from HTML, content, and XML reports from the Web, record frameworks, spreadsheets, and other vaults. It orders feelings recorded in these reports from "strong positive" through "impartial or neutral" to "strong negative." The arrangement then changes separated substance into organized information for institutionalizing, coordinating, furthermore, incorporating into the SAP HANA database [3].

Understanding to-activity sees bolster inquiry, investigation, also, reporting over the whole database, so data gathered from unstructured content assembles a solitary strong establishment on which you can move from information to choice [3]. Preconfigured diagnostic substance utilizing a local HTML5 toolbox for SAP HANA gives clients a steady, natural assumption cockpit on desktop what's more, cell phones. They can take advantage of

discretionary SAP Lumira programming to fabricate what's more, share profound and expansive diagnostic reports. What's more, they can apply prescient investigation on information given by supposition knowledge to contribute to item and campaign arranging [4].

The consumerization of IT has hoisted client self-support of new levels. Client connections today depend on sub second reactions as clients communicate with big business frameworks. Applications that were initially made as frameworks of record now need to end up frameworks of engagement. The SAP HANA stage is architected to change existing frameworks to convey new administration levels without interruption. SAP HANA additionally gives a bound together platform for the coordination of big business applications with customer applications. While giving ongoing following by buyers, it can at the same time give ongoing bits of knowledge to the undertaking. This capable cutting edge application empowers organizations to react with exactness at the point when making continuous offers to customers [4].

The term Mobile Application Development Platform (MADP) appeared two or three years later when MEAP (Mobile Enterprise Access Platform) included cross-stage improvement from a solitary code base [4]. MADPs tended to complex versatile issues however needed openness. Their way to deal with settling logged off information provisioning furthermore, synchronization issues was regularly exclusive and muddled. That was then and this is presently, where the world is a great deal more versatile, and portable open source system ventures are excessively various, making it impossible to count [4].

Engineers use coordinated advancement situations and systems promptly accessible to offer them some assistance with creating applications all the more rapidly. With improvement devices and packs offered at a negligible expense, does it bode well to put resources into an MEAP or MADP platform? In a word: Yes. "Purchase VERSUS BUILD" BECOMING "Purchase VERSUS BUILD AND BORROW" [4].

Numerous open source versatile segments are promptly accessible; however, the watchword is "numerous." You generally require more than one. The great buy versus build choice now incorporates acclimatizing one or more open source bundles [5]. "Purchase versus build" has ended up "purchase versus manufacture and get."

For a few applications, that equation is sufficient. Suppose you have to assemble an online application that keeps running on numerous gadget working frameworks. You could

utilize a "compose once, run numerous" open source application holder, for example, Apache Cordova and fabricate the client interface from an open source UI system, for example, Sencha Touch or OpenUI5 [5]. "Extraordinary," you say. In any case, hold on. That covers the gadget systems, and you need Representational State Transfer (REST) furthermore, JavaScript Object Notation (JSON) administrations systems to bolster it. On the off chance that you are a Windows shop, you may utilize Microsoft .NET administrations to assemble Open Data Protocol (OData) Web administrations. In a SAP world, you may depend on SAP Gateway innovation. Java open tooling additionally gives numerous alternatives [5].

## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 DESIGN**

##### **3.1.1 Integrating and Mining Data**

This paper has integrated and mined data from multiple sources. It has divided this data into smaller pieces of information. We have coordinated and mined big data from various sources to interpret and use the structure of natural systems to shed new bits of knowledge on the elements of organic frameworks. We address the hypothetical underpinnings and present and future empowering advancements for incorporating and mining natural systems. We have extended and incorporated the strategies and routines in data procurement, transmission, and handling for data systems. We have created techniques for semantic-based information integration, robotized theory era from mined information, and mechanized adaptable scientific instruments to assess reenactment results and refine models.

##### **3.1.2 Big Data Fast Response**

Here a big data framework has been built for fast response. This adapts modules which change and adjust according to the data. We propose to fabricate a stream-based Big Data systematic structure for quick reaction and constant choice making.

- Designing Big Data examining instruments to diminish Big Data volumes to a sensible size for handling
- Building forecast models from Big Data streams. Such models can adaptively acclimate to the element changing of the information, and also precisely foresee the information's pattern later on; and
- A learning indexing structure to guarantee ongoing information observing and arrangement for Big Data application.

##### **3.1.3 Pattern matching and mining**

This paper performs pattern matching using wild cards. We perform a systematic investigation on pattern matching, pattern mining with wildcards, and application problems as follows:

- Exploration of the NP-hard complexity of the matching and mining problems,
- Multiple patterns matching with wildcards,
- Approximate pattern matching and mining, and

- Application of our research onto personalized information processing and bioinformatics

### **3.1.4 Key technologies for integration and mining**

Here combining the data which was searched using pattern matching is done. We have performed an examination on the accessibility and measurable regularities of multisource, gigantic and element data, including cross-media hunt in light of data extraction, testing, unverifiable information questioning, and cross-area and cross-stage data polymerization. To get through the impediments of conventional information mining routines, we have examined heterogeneous data disclosure and mining in complex inline information, mining in information streams, multi-granularity learning revelation from monstrous multisource information, dispersion regularities of gigantic learning, quality combination of huge learning.

### **3.1.5 Procedure**

#### **Step1: We shall install all the tools required**

To perform text analysis, we should install the SAP HANA studio and SAP Lumira with a username and password to the Workbench we are going to use to develop the application.

#### **Step2: We shall have the Twitter Account**

To interact with the Twitter API and to establish a connection with real-time data flowing through it we should have an account in that and should register as an app developer in that.

#### **Step3: We shall develop a node.js application to establish connection between the database and the twitter.**

The methods used will be developed based on the modules of the JavaScript application and server establishing techniques.

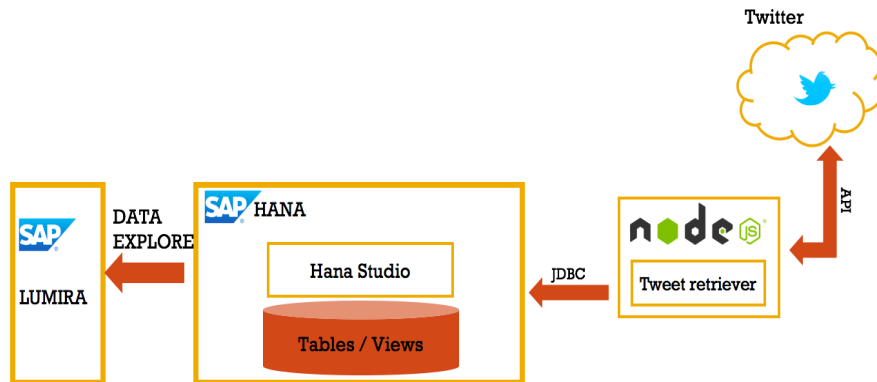
#### **Step4: To program the queries to extract the data from the twitter and load the data into the database.**

We should write the queries to extract the useful data from the twitter and load that data to SAP HANA databases.

#### **Step5: Loading data from the SAP Hana Database to SAP Predictive Analysis for Data Visualization**

We shall connect the database to the SAP Predictive Analysis and load the data into it to perform the required analysis.

As shown below in Figure 1, architecture explains the steps we are following



**Figure 1: Architecture for Development of Application for Text Analysis**

## 3.2 REQUIREMENTS

### Hardware Requirements:

- Processor - Intel Core I3
- RAM - 2 GB
- Hard Disk - 1 TB

### Software Requirements:

- Operating System - Linux, Windows
- Framework - Node.js
- Database - SAP HANA Database
- Scripting Languages - JavaScript
- Platforms - SAP HANA Studio, SAP Predictive Analysis or SAP Lumira



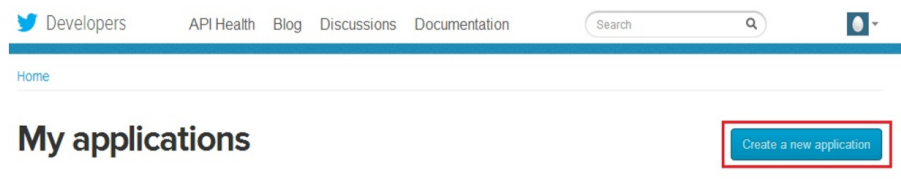
## 3.3 IMPLEMENTATION

### 3.3.1 Registration at Twitter Developers:

To develop a twitter analysis application and to extract the twitter data in real time, it is mandatory to create an application with Twitter Developer and to invoke the APIs later we need the information regarding to the authentication of the application.

We can register and create an OAuth Tokens at Twitter by logging into your twitter account and we followed the below steps.

1. Login into your twitter account and click on the Twitter Developers and create the application as shown in the Figure 2.



**Figure 2: Create a new application on twitter account**

2. Create the twitter application with the details of the project and get the access tokens by creating my access token. After that we are able to see the OAuth settings as shown below in Figure 3. and copy the Consumer Key, Consumer Secret Key, Access token secret and Access token.

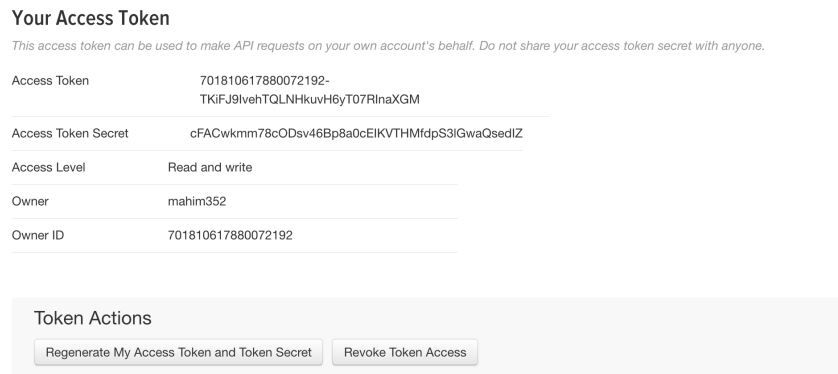
#### Application Settings

*Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.*

Consumer Key (API Key)	uaGIHlnVJDLNO7UyPbX3jYZSW
Consumer Secret (API Secret)	BUk7MIQjiCGQdOxsYACHlojahLewU0Xn13vkcYrQadeuouE6qK
Access Level	Read and write ( <a href="#">modify app permissions</a> )
Owner	mahim352
Owner ID	701810617880072192

**Figure 3: Consumer Key and Consumer Secret Key**

Access token and Access token secret by clicking on the generate access token and token secret, we get the tokens as shown below in the Figure 4.



**Figure 4: Access Token and Access Token Secret Keys**

### 3.3.2 Install Node.JS

Install Node.JS and also the node modules. These node modules are used to facilitate the application development. We are using six modules in this application and these modules are controlled by using node package manager(npm). We created a package.json file to install the required modules at once and we create a Node application. We installed the following modules.

- express – It includes the basic functionality such as handling requests and sending responses.
- oauth – It implements the authentication required by the twitter api.
- Node-tweet-stream – implements the api which is required to interact with applications of twitter.
- hdb – it allows to execute the SQL commands and interaction with SAP HANA.
- emoji-strip – It strips emojis from the gathered tweets because HANA doesn't support the emojis.

After creating the node modules we create the node server which consists of config.js and app.js files in which config.js contains the information to configure with Twitter api and HANA server.

We run the node server in the terminal by using the following command:

“ node app”

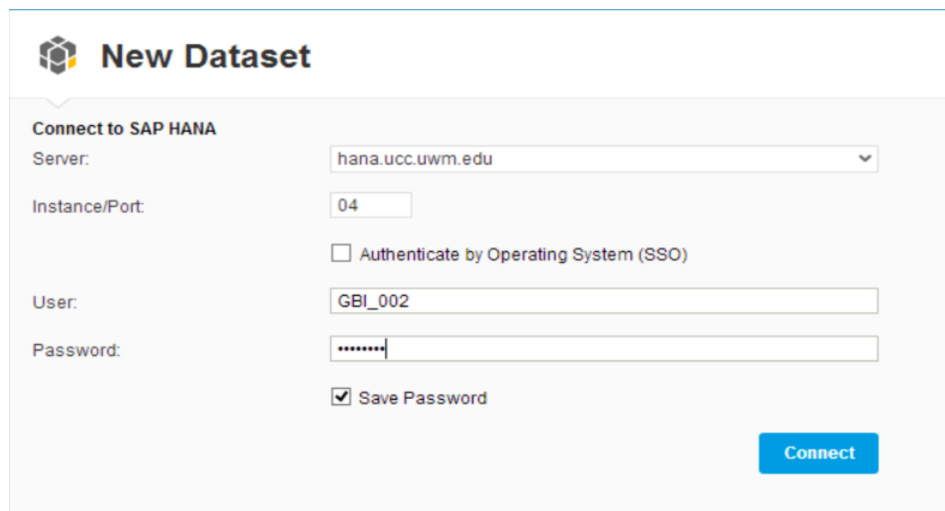
Tweets are inserted from the twitter to the Hana and we can see it in the console as shown below in the Figure 5.

```
Maheshs-MBP:Sap Twitter Project Mahesh$ node app
Listening on port 8888
Press Ctrl-C to terminate
Start tracking amazon
Tweet inserted: 793698795213066240 2016-11-2 1:18:33 1
Tweet inserted: 793698795695538177 2016-11-2 1:18:33 1
Tweet inserted: 793698796777668609 2016-11-2 1:18:34 1
Tweet inserted: 793698797918392320 2016-11-2 1:18:34 1
Tweet inserted: 793698799076073473 2016-11-2 1:18:34 1
Tweet inserted: 793698799298367489 2016-11-2 1:18:34 1
Tweet inserted: 793698800548315136 2016-11-2 1:18:34 1
Tweet inserted: 793698802276245504 2016-11-2 1:18:35 1
Tweet inserted: 793698802100080640 2016-11-2 1:18:35 1
Tweet inserted: 793698804725747712 2016-11-2 1:18:35 1
Tweet inserted: 793698805795348481 2016-11-2 1:18:36 1
```

**Figure 5: Insertion of Twitter tweets**

### **3.3.3 Installing SAP HANA studio and SAP LUMIRA**

Sap Hana is a database server located in the California State University Chico. They gave us account with password and created a space. We install Sap Lumira to do sentiment analysis and Text Analysis. In sap lumira, we connect to the host by giving the hostname and port as shown below in the Figure 6.



**New Dataset**

**Connect to SAP HANA**

Server: hana.ucc.uwm.edu

Instance/Port: 04

☐ Authenticate by Operating System (SSO)

User: GBI\_002

Password: .....

☒ Save Password

**Connect**

**Figure 6: Accessing the HANA database from SAP LUMIRA**

We created a database table and named as “Tweets” in our schema. We used the following SQL code to create it

SQL:

```
CREATE COLUMN TABLE "Tweets"  
( "id" VARCHAR(256) NOT NULL,  
  "created" TIMESTAMP,  
  "text" NVARCHAR(256),  
  "lang" VARCHAR(256),  
  "user" VARCHAR(256),  
  "replyUser" VARCHAR(256),  
  "retweetedUser" VARCHAR(256),  
  "lat" DOUBLE,  
  "lon" DOUBLE,  
  PRIMARY KEY ("id")  
);
```

SELECT TOP 1000 \* FROM "GBI\_361"."Tweets"

	id	created	text	lang	user	replyUser	retweetedUser
1	7778910692455274434	2016-09-24 16:0 AM	/Se acabó el "boom" de las apps?: Cuando salió el primer iPhone, la novedad real -además de...	es	UberSocialVzia		
2	777891063625240577	2016-09-24 16:0 AM	そういえば2ヶ月くらい前にアイフォンのケース当たったんだけど、シャンパンゴールドで、	ja	shimamuu498		
3	777891064975720454	2016-09-24 17:0 AM	RT @erodouaadeonani: セフレに異性ある人必見!!	ja	mutturisukeb...		erodouaadeo...
4	777891064527122432	2016-09-24 16:0 AM	Un iPhone 7 GRATIS. particio del sorteo de @TecnonautaTV https://t.co/FnixDnSAec #iPHO...	it	JaredPeralta...		
5	777891065038725121	2016-09-24 17:0 AM	@ouou nino	ja	muamua aral...	ouou ...	
6	77789106505547392	2016-09-24 17:0 AM	@FashionInsight we design iPhone cases that hold headphones. Hapov to send a sample.	en	dome8cases	Fashion...	
7	777891065147846656	2016-09-24 17:0 AM	RT @FrancescoPupa: Ma dià rutt u cazz s aodiornamento, per sbloccare l'iPhone mi devo imbr...	it	LuddTesoriere		FrancescoPupa
8	777891065613262848	2016-09-24 17:0 AM	/Se acabó el "boom" de las apps?: Cuando salió el primer iPhone, la novedad real -además de...	es	AztkDeVzia		
9	777891066410184704	2016-09-24 17:0 AM	モイ！ iPhoneからキャス配信中 - 画やーべス https://t.co/xSxCorrvD	ja	Xx0692		
10	777891066510921729	2016-09-24 17:0 AM	@tibbsaf which iPhone	en	KvleRancourt	tibbsaf	
11	777891066934534145	2016-09-24 17:0 AM	モイ！ iPhoneからキャス配信 - 暇だ、ねれない https://t.co/XdF8Vv65EV	en	RabbitAMNO...		
12	777891067731517440	2016-09-24 17:0 AM	RT @toetaoTav: https://t.co/i7Hou18CER	und	MvName IsF...	toetaoTav	
13	777891068339687424	2016-09-24 17:0 AM	RT @HesJustKidrauh: hacer cola para comprar un Iphone es lo mas normal del mundo, pero h...	es	ashtOnrth4rv	HesJustKidr...	
14	777891069983731716	2016-09-24 18:0 AM	持ち傷を自己修復するiPhone 7用ハードケース発売 - LINE NEWS https://t.co/LsczmoSVF #line...	ja	S Mmax		
15	777891070579470337	2016-09-24 18:0 AM	As Per Casper - Goodniht ft. Fatiniza	en	WhiteCubeUAE		
16	777891070994636800	2016-09-24 18:0 AM	RT @iPhone News: Samsung's exploding phones the result of a rush to beat the 'dull' iPhone...	en	MariaKatieba	iPhone News	
17	777891071636365312	2016-09-24 18:0 AM	RT @tendenza io: 【動画あり】 「iPhone 7」をへりから落とした結果！ そうiPhoneならね http...	en	tateNobita	tendenza io	
18	777891072420646916	2016-09-24 18:0 AM	RT @cultofmac: iPhone 7 destroys Samsung's Galaxy Note 7 in speed test https://t.co/7ivsT9...	en	thecultcast	cultofmac	
19	777891072244457473	2016-09-24 18:0 AM	RT @khaiochi: ทดสอบ iPhone 7 ด้วยภาพทดสอบ ภาพว่าจริง และเมื่อเห็น ด้วยคำ ไม่ค่อยมีอะไร...	th	mmminchv	khaiochi	
20	777891073490181664	2016-09-24 19:0 AM	Un iPhone 7 GRATIS. particio del sorteo de @TecnonautaTV https://t.co/6Rh1LSBmVA #IPH...	it	17Benvides		
21	777891073058336768	2016-09-24 18:0 AM	Trvino a new android outlet in Talon Pro...last decent third party app I found for Twitter was @t...	en	dothebruce		
22	777891073234834400	2016-09-24 19:0 AM	Koratis PRO Screen (gratis + IAP - iPhone/iPad, iOS 7.0+) - (was €6.99) Mask ie - Lovendar; T...	nl	Media krant		
23	777891074786398209	2016-09-24 19:0 AM	Я coбoяи 12 800 зoлoтux вoвeи! Kтo вa cлoжoкeт 6oнeуe? https://t.co/gIH75FLWm #...	ru	33WHesOoQ...		
24	777891075260383232	2016-09-24 19:0 AM	Featured: Swappa listina for Apple iPhone 6 (AT&amo:T): \$399 https://t.co/DuRrJSEMzI	en	SwappaMucho		
25	777891075637719040	2016-09-24 19:0 AM	I got a reward: Architectural Treasure in BioBusiness Deluxe for iOS https://t.co/USa3iSsTGo #...	en	FunNfiber		

Figure 7: Tweets stored on the Hana Database

### 3.3.3.1 Text Analysis

As indicated by SAP Help, content investigation “linguistic, statistical and machine-learning calculations that model and structure the data substance are three noteworthy parts of Text Analysis which are printed sources in different dialects.” One of the important concept of text analysis is sentiment analysis. Sentiment analysis tries to calculate the sentiment of the extracted entities and to allocate the positive and negative sentiment to them.

#### Perform the Text Analysis:

To create the text analysis, we have to create a full text index on the tweets table. EXTRACTION\_CORE\_VOICEOFCUSTOMER is dictionary which consists of set of rules and entities to extract the sentiments and requests and so on of the customers. We use this content to get specific information about our needs and perceptions of our customer when processing and analyzing the data.

The content of this index are shown below in the Figure 8.

SELECT TOP 1000 \* FROM "GBI\_361"."\$TA\_tweets"

id	TA_RULE	TA_COUNTER	TA_TOKEN	TA_LANGUAGE	TA_TYPE	TA_NORMALIZED	TA_STEM	TA_PARAGRAPH	TA_SENTENCE	TA_CREATED_AT	TA_OFFSET	TA_PARENT
777890603606634496	Entity Extraction	1	iPhone 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890606131605504	Entity Extraction	2	iPhone	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	62	?
777890606448063616	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890606102278144	Entity Extraction	1	@C4ETe...	en	SOCIAL...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890605334691840	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890606458822656	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890606458822656	Entity Extraction	4	iPHONE 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	58	?
777890607679303680	Entity Extraction	1	@WuTan...	en	SOCIAL...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890607851241473	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890607851241473	Entity Extraction	3	iPhone 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	64	?
777890608962822144	Entity Extraction	1	iOS 9	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890612582486020	Entity Extraction	2	iPhone 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	31	?
777890612238516224	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890612238516224	Entity Extraction	3	iPhone 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	15	?
777890612896894977	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890612896894977	Entity Extraction	4	iPhone	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	54	?
777890612821520384	Entity Extraction	1	@cinem...	en	SOCIAL...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890613320646656	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890613320646656	Entity Extraction	3	iPhone 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	46	?
777890613744263168	Entity Extraction	6	iPhone 7	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	63	?
777890614134341632	Entity Extraction	2	iPhone 6s	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	32	?
777890614176198656	Entity Extraction	1	RT	en	ORGANI...	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	0	?
777890614176198656	Entity Extraction	3	iPhone	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	30	?
777890616990511104	Entity Extraction	1	El iPhon...	en	PRODUCT	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	12	?
777890616990511104	Entity Extraction	5	https://t...	en	URL	?	?	1	1	Sep 19. 2016 3:32:55.758 PM	77	?

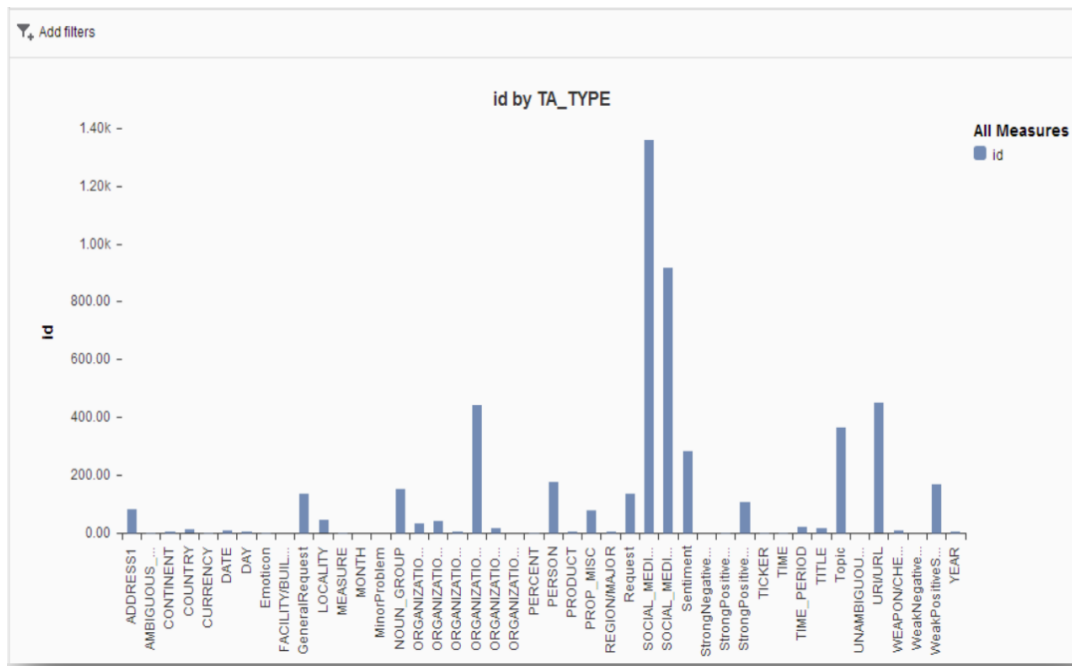
**Figure 8: Full Index Table for Text Analysis**

This table consists of columns id, TA\_RULE, TA\_COUNTER, TA\_TOKEN, TA\_LANGUAGE, TA\_TYPE. Tweets are added to the table in live index such a way that they are analyzed and added to the index.

Overview of TA\_tweets index table is newly made content list called \$TA\_tweets will show up in rundown of tables. To see the information right tap on the \$TA\_tweets table and select information and we will see some vital sections. These incorporate TA\_RULE, a segment that affirms that substance extraction is happening. The TA\_TOKEN records the bit of data that the content investigation is performed on. TA\_TYPE indicates what sort of content investigation is it and what has been extricated. Cases incorporate area, online networking and estimation. The content record decides the five sorts of estimation a Tweet can have. For instance, "Great" is resolved to be strong positive sentiment and "Disappointment" is resolved to be weak negative sentiment. A few words are named equivocal and could be considered debase, for example, "loser."

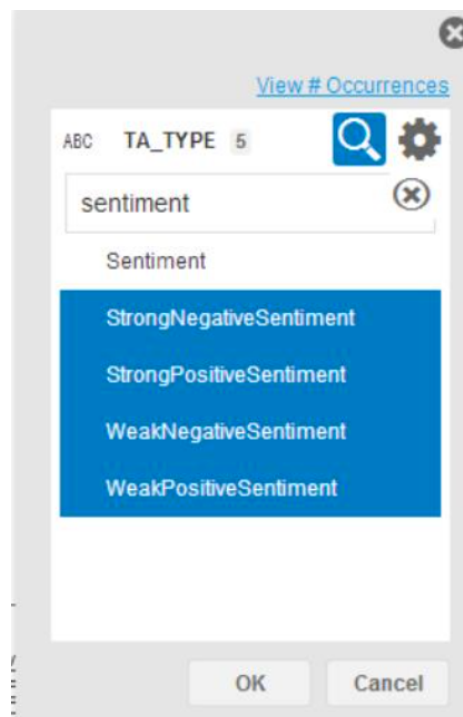
### 3.3.3.2 Exploring the Data:

We add the data to SAP LUMIRA by selecting the table TA\_tweets index and we create the index measure by selecting the x-axis and y-axis. Some of the analyzed data is shown below in the Figure 9:



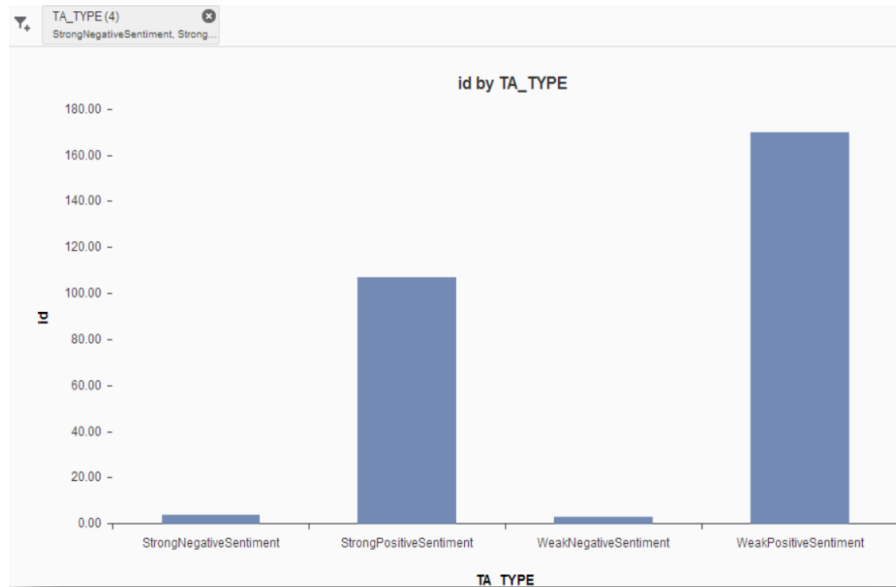
**Figure 9: Figure showing the Analysis of data using the measures id by TA\_TYPE**

We are doing the sentiment analysis for the extracted tweets and we do so by selecting the sentiment which are both positive and negative sentiments as shown below in the Figure 10.

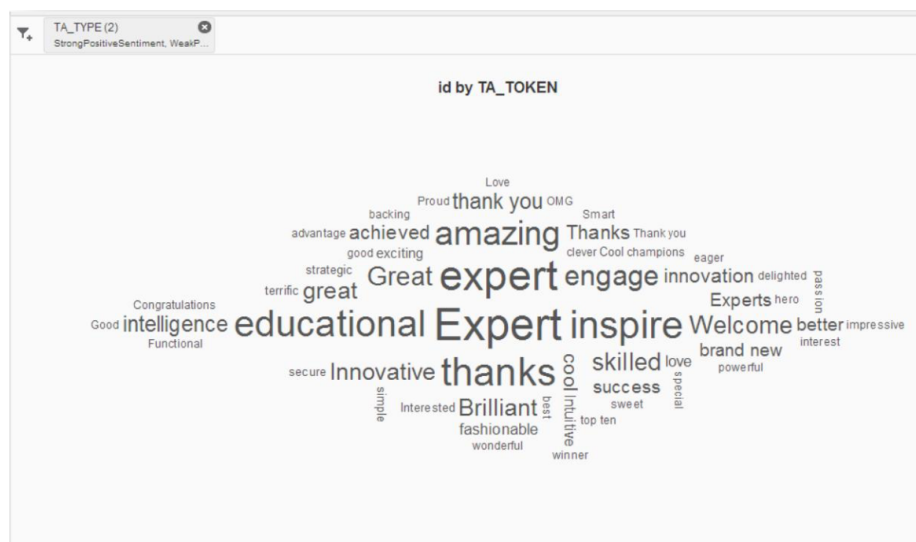


**Figure 10: Filtering of Tweets using Sentiment**

Below charts are much more informative and we observe a much more preponderance of positive sentiments in the respective twitter stream we stored on our database as shown below in the Figure 11 and Figure 12.



**Figure 11: Figure showing Positive and Negative Sentiments**



**Figure 12: Figure showing the tag cloud of obtained tweets**



### 3.3.3.3 Stance and Influence

The initial phase in the cluster analysis is to make a quality view that contains information on the most powerful tweeters. The position of a tweeter is dictated by the relative number of positive and negative sentiment tokens separated from their tweets. Stance can be calculated as following:

$$\text{Stance} = 5 * \text{SP} + 2 * \text{WP} - 5 * \text{SN} - 2 * \text{WN}$$

Where:

SP – Strong Positive

WP – Weak Positive

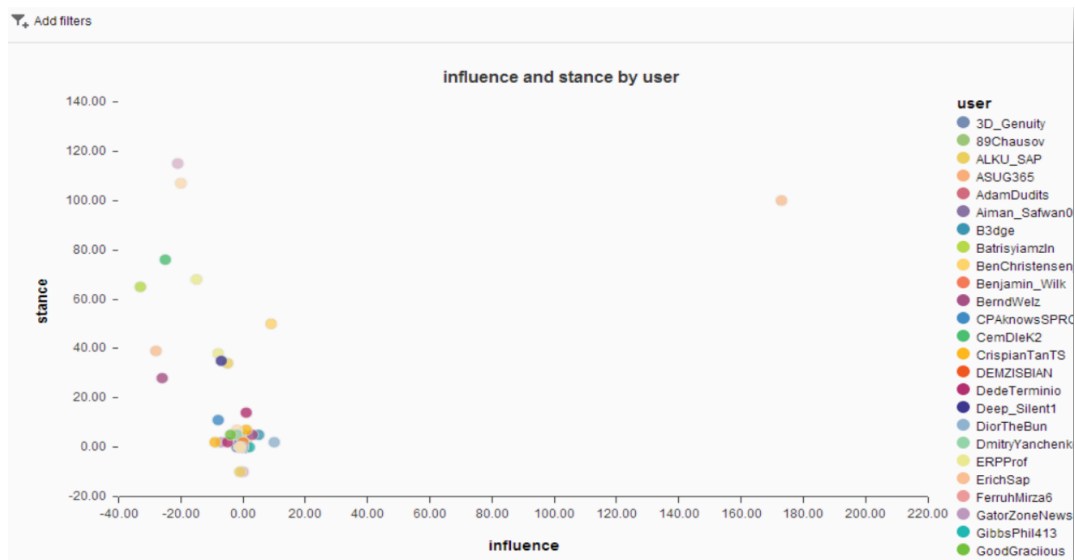
SN – Strong Negative

WN – Weak Negative

Influence is determined by the quantity of times a client is retweeted and answered to. The replyUser and retweetedUser segments in the Tweets table is utilized to ascertain this. What we are keen on is clients who have strong stances with high impact.

After executing the respective SQL statements, we get the stance and influence of tweets and we calculate the sentiments and then by exploring it into SAP Lumira we get the visualization content. In this SAP Lumira, after the data exported we create the measure by selecting the scatter chart with x-axis influence and y-axis stance. We can see that in the below Figure 13.

In this analysis, we are using the k-means algorithm of the predictive analysis library (PAL) of HANA. At first we create the AFL (Advanced Function Library) stored procedure which acts like a wrapper and intermediary to the algorithm. This defines outputs, inputs and parameters to the k-means algorithm.



**Figure 13: Stance vs Influence by user**

We create the table types such as PAL\_T\_DATA, PAL\_T\_PARAM, PAL\_T\_RESULTS and PAL\_T\_CENTERS to configure the behavior of k-means algorithm and they can be used to create the output and input tables as shown below in Figure 14 and Figure 15.

Table Name:		Schema:						
PAL_T_CENTERS		GBL_361						
Columns	Indexes	Further Properties	Runtime Information					
	Name	SQL Data Type	Dimens	Column Store	Data Type	Key	Not Null	Default
1	ClusterNumber	INTEGER		INT				
2	stance	INTEGER		INT				
3	influence	INTEGER		INT				

**Figure 14: Table showing the attributes for stance and influence**

SELECT TOP 1000 * FROM "GBI_361"."PAL_PARAMS"				
	NAME	INTARGS	DOUBLEARGS	STRINGARGS
1	THREAD_NUMBER	2		? ?
2	GROUP_NUMBER	3		? ?
3	GROUP_NUMBE...	5		? ?
4	GROUP_NUMBE...	10		? ?
5	INIT_TYPE	4		? ?
6	DISTANCE_LEVEL	2		? ?
7	MAX_ITERATION	100		? ?
8	EXIT_THRESHOLD	?	0.000001	?
9	NORMALIZATION	0		? ?

**Figure 15: Table showing the clustering attributes**

In the next step we create the PAL\_SIGNATURE table as shown below in Figure 16 by including with all the other table types we created. We create a parameter table which controls the execution of algorithm. Next we create the output tables such as PAL\_RESULTS and PAL\_CENTERS as shown in Figure 17 and Figure 18.

SELECT TOP 1000 * FROM "GBI_361"."PAL_SIGNATURE"			
	ID	TYPENAME	DIRECTION
1	1	GBI_361.PAL_T_DATA	in
2	2	GBI_361.PAL_T_PARAMS	in
3	3	GBI_361.PAL_T_RESULTS	out
4	4	GBI_361.PAL_T_CENTERS	out
5	1	GBI_361.PAL_T_DATA	in
6	2	GBI_361.PAL_T_PARAMS	in
7	3	GBI_361.PAL_T_RESULTS	out
8	4	GBI_361.PAL_T_CENTERS	out

**Figure 16: Table showing the Signature of Centers, Results, Data, Params.**

SELECT TOP 1000 * FROM "GBI_361"."PAL_RESULTS"			
	user	ClusterNumber	distance
1	narsgoddess	2	1.338331523697228
2	equipment_bot	2	1.1397665441915075
3	SocialNBroo...	2	3.8971729066269485
4	aneobitoyone	2	1.1397665441915075
5	bladeit	2	1.338331523697228
6	Infinitebook1	2	3.8971729066269485
7	ebstt	2	1.1397665441915075
8	techforcepo...	2	1.1397665441915075
9	SalesEd1	2	1.1397665441915075
10	Persisplanner	1	0.7071067811865476
11	daffyllove_06...	2	1.1397665441915075
12	gadget_ande...	2	1.1397665441915075
13	rtogai	2	1.1397665441915075
14	o_gjx	2	1.1397665441915075
15	drarvindkumar	2	3.9598005197497805
16	SocialNColu	2	3.8971729066269485

**Figure 17: Table showing the Cluster Number and Distance between the nodes**

SELECT TOP 1000 * FROM "GBI_361"."PAL_CENTERS"			
	ClusterNumber	stance	influence
1	0	-5	-1
2	1	14	-1
3	2	1	0

**Figure 18: Table showing the Centers of the different clusters**

Views are necessary in creating the tables and we create two views for the data collected. They are `TweetersClustered` which combines both `Tweeters` view and cluster numbers (located in the users) and the second one is `TweetersClusteredSummary` as shown below in Figure 21 which combines both `Tweeters` data and also groups by `ClusterNumber` in stance and influence as shown in Figure 19 and Figure 20.

SELECT TOP 1000 * FROM "GBI_361"."Tweeters"			
	user	stance	influence
1	narsgoddess	0	-1
2	equipment_bot	0	0
3	SocialNBroo...	5	0
4	aneobitoyone	0	0
5	bladeit	0	-1
6	Infinitebook1	5	0
7	ebstt	0	0
8	techforcepo...	0	0
9	SalesEd1	0	0
10	Persisplanner	14	-2
11	daffyllove_06...	0	0
12	gadget_ande...	0	0
13	rtogai	0	0
14	o_gjx	0	0
15	drarvindkumar	5	-1
16	SocialNColu...	5	0
17	Info_DE	2	0
18	SocialNDenver	5	0
19	utahfoodfree...	0	-1
20	saibaba2016	0	0
21	home0077	0	0
22	58QJ9WJrX...	0	0
23	nora_suzu	0	-1

Figure 19: Table showing the data of users by stance and influence

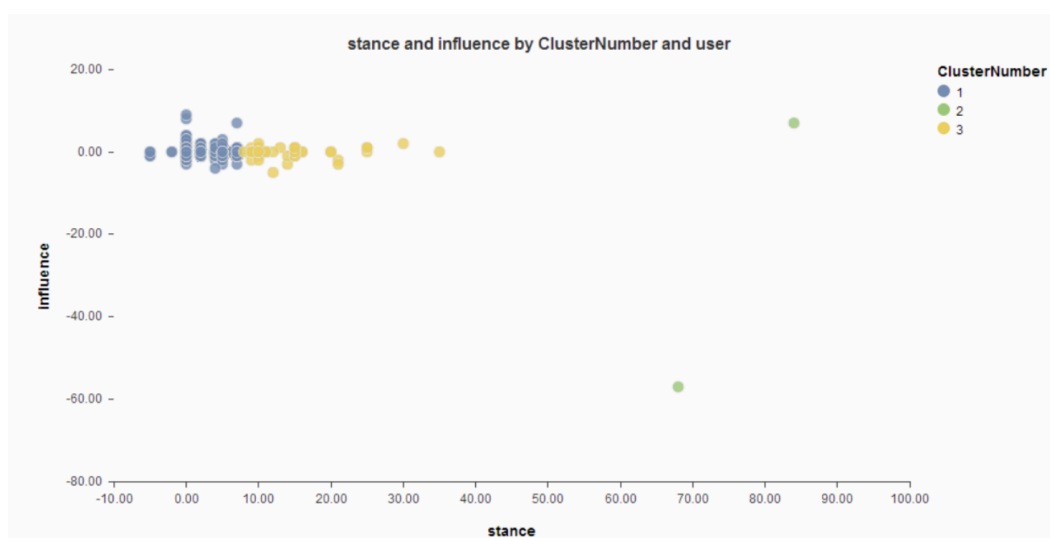
SELECT TOP 1000 * FROM "GBI_361"."TweetersClustered"				
	user	stance	influence	ClusterNumber
1	narsgoddess	0	-1	3
2	equipment_bot	0	0	3
3	SocialNBroo...	5	0	3
4	aneobitoyone	0	0	3
5	bladeit	0	-1	3
6	Infinitebook1	5	0	3
7	ebstt	0	0	3
8	techforcepo...	0	0	3
9	SalesEd1	0	0	3
10	Persisplanner	14	-2	2
11	daffyllove_06...	0	0	3
12	gadget_ande...	0	0	3
13	rtogai	0	0	3
14	o_gjx	0	0	3
15	drarvindkumar	5	-1	3
16	SocialNColu...	5	0	3
17	Info_DE	2	0	3
18	SocialNDenver	5	0	3
19	utahfoodfree...	0	-1	3

Figure 20: Table showing users by ClusterNumber

SELECT TOP 1000 * FROM "GBI_361"."TweetersClusteredSummary"				
	ClusterNumber	stance	influence	users
1	3	0	-1	13
2	3	0	0	35
3	3	5	0	9
4	2	14	-2	1
5	3	5	-1	2
6	3	2	0	2
7	1	-5	-1	1
8	3	7	0	1
9	3	4	-1	1
10	2	15	-1	1

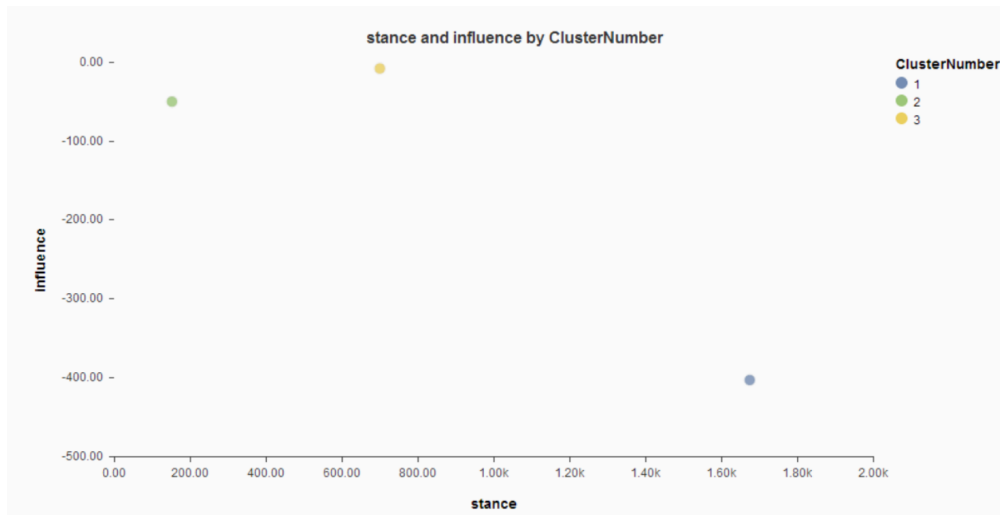
**Figure 21: Table showing the TweetersClusteredSummary**

We explore the above data into SAP LUMIRA by expanding the TweetersClustered view and we create the measure with stance and influence as shown below in Figure 22. The data can be seen as below in visual format.



**Figure 22: Stance and influence by ClusterNumber and user**

There three different types of clusters can be seen the above diagram and they are grouped different users. If we remove the users and we can see the clusters as center as below in Figure 23.



**Figure 23: Figure showing the stance and influence by ClusterNumber**

### 3.4 TIMELINE

Task /Year 2016	Jan	Feb	Mar	Apr	Aug	Sept	Oct	Nov	Dec
Researching a particular area of interest	X								
Committee Selection	X								
Discussing with committee for improvement in the area of interest	X								
Doing research in the area of interest		X							
Analyzing the problem to be solved		X							
Highlighting the objectives of the research		X							
Literature review		X	X						
Developing a procedure			X						
Getting approval of the proposal			X						
Listing out the H\W and S\W components				X					
Developing the proposal by writing code					X	X			
Making the required changes and enhancing the Research							X		
Testing Phase								X	
Final project presentation								X	X

**Figure 24: Timeline for the Project**



## **CHAPTER 4**

### **CONCLUSION AND FUTURE WORK**

- Analyzed the twitter data and generated the calculation views by taking the tweet counts and with different index types through the aggregation of time dimension and TA\_tweets using SAP HANA.
- We can see the data in a visual format which is easily understandable by using the SAP LUMIRA.
- To facilitate the work, we would like to develop a mobile application and web application for the organizations using OData services in SAP HANA.
- We would like to improve the work by taking multiple keywords at a time such as considering AMAZON, TARGET, STARBUST and other organization and to perform sentiment analysis by the tweets from the twitter.
- To perform analysis on other social media networks such as Facebook, Pinterest etc....

## REFERENCES

- [1] J. Allan, R. Papka, and V. Lavrenko, “On-line New Event Detection and Tracking,” Proc. The 21st ACM International Conference on Information Retrieval (SIGIR), pp. 37–45, 1998.
- [2] Y. Arakawa, S. Tagashira, and A. Fukuda, “Relationship Analysis between User Contexts and Input Word with Twitter,” Transactions of Information Processing Society of Japan, Vol. 52, No. 7, pp. 2268–2276, 2011. (in Japanese).
- [3] T. Joachims, “Text Categorization with Support Vector Machines,” Proc. the 10th European Conference on Machine Learning (ECML), pp. 137–142, 1998.
- [4] B. O’Connor, M. Krieger, and D. Ahn, “TweetMotif: Exploratory search and topic summarization for twitter,” Proc. 4th International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.
- [5] T. Yamanaka, Y. Tanaka, Y. Hijikata, and S. Nishida, “A Supporting System for Situation Assessment using Text Data with Spatio-temporal Information,” Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 22, No. 6. pp. 691–706, 2010.

## APPENDIX

### Detail Code

#### Tweets Table SQL

```
CREATE COLUMN TABLE "Tweets"  
( "id" VARCHAR(256) NOT NULL,  
  "created" TIMESTAMP,  
  "text" NVARCHAR(256),  
  "lang" VARCHAR(256),  
  "user" VARCHAR(256),  
  "replyUser" VARCHAR(256),  
  "retweetedUser" VARCHAR(256),  
  "lat" DOUBLE,  
  "lon" DOUBLE,  
  PRIMARY KEY ("id")  
);
```

## **package.json**

```
{ "name": "live3",  
  "description": "SAP HANA Academy - live3 - load Twitter stream to SAP HANA",  
  "version": "0.0.1",  
  "private": true,  
  "dependencies": {  
    "emoji-strip": "0.0.3",  
    "express": "4.7.*",  
    "hdb": "0.3.*",  
    "node-tweet-stream": "^1.8.1",  
    "oauth": "0.9.*",  
    "request": "2.39.*" } }
```

## Config.js

```
var config = {  
  twitter: {  
    consumer_key : 'uQheiLKCNI27wHsQiwBnFpRUH',  
    consumer_secret : 'F1H0TdqvFMsXAcTefWbfWIFM1aLwghNp7qKsYfopqi2WJAmKyi',  
    token : '162070434-UTSE0SPdtuAmRQt8lv5HRvpzRY1u1OjtGxYQ8wdC',  
    token_secret : 'oScSIchneAbVqy8geVhhkwCW6mT1FDqsYZSWIC4kmw4n4'  
  },  
  hana: {  
    host : 'trinity.cob.csuchico.edu',  
    port : 30059,  
    user : 'GBI_361',  
    password : 'cob123',  
    hdb_schema : 'GBI_361'  
  },  
};  
  
module.exports = config;
```

## App.js

```
//Load the node modules

var express = require('express')

util = require('util'),

url = require('url'),

hdb = require('hdb')

emojiStrip = require('emoji-strip'),

tweets = require('node-tweet-stream');

//Load the config.js file

var config = require('./config.js');

//Configure the HANA connection information using data from config.js

var hdb = hdb.createClient({

    host : config.hana.host,

    port : config.hana.port,

    user : config.hana.user,

    password : config.hana.password

});

var hdb_schema = config.hana.hdb_schema;

//Connect to HANA. If an error occurs it will be written to the console. hdb.connect(function

(err) {

    if (err) {

        return console.error('HDB connect error:', err);

    } });

//Configure the Twitter connection object var stream = new tweets({

    consumer_key: config.twitter.consumer_key,

    consumer_secret: config.twitter.consumer_secret,

    token: config.twitter.token,
```

```

token_secret: config.twitter.token_secret

});

//Create the application object using the express.js node module

var app = express();

//Add paramters to the response header that allows cross-domain requests app.all('*',
function(req, res, next) {

res.header("Content-Type", "text/plain");

res.header("Access-Control-Allow-Origin", "*");

next();

});

//Route that matches http://localhost:8888 used to test if server is functioning app.get('/',
function(req, res, next){

res.send("Listening...");

});

//Route that starts tracking twitter stream

//It takes to parameters: table (the table in HANA) and track (the keyword to track)
app.get('/do/start', function(req, res, next) {

var table = req.param('table');

var track = req.param("track");

//If the user doesn't provide the table name, use Tweets

if(table === undefined) {

table = "Tweets";

}

//If a keyword is provided, write it to the log, send it to the browser

//and start tracking

if (track !== undefined) {

console.log('Start tracking ' + track);

res.send('Start Tracking ' + track);

```

```

//Each time a tweet is captured, the callback function will execute inserting
//the data into the table in HANA
stream.on('tweet', function(data){

var myDate = new Date(Date.parse(data.created_at.replace(/( +)/, 'UTC$1')));

var createdAt = myDate.getFullYear() + '-' + eval(myDate.getMonth() + 1) + '-' +
myDate.getDate() + ' ' + myDate.getHours() + ':' + myDate.getMinutes() + ':' +
myDate.getSeconds();

var replyUser = "";

if (data.in_reply_to_screen_name !== null) {

replyUser = data.in_reply_to_screen_name

}

var retweetedUser = "";

if (typeof data.retweeted_status !== 'undefined') {

retweetedUser = data.retweeted_status.user.screen_name;

}

var lat = null;

var lon = null;

if (data.geo !== null) {

    lat = data.geo.coordinates[0];

    lon = data.geo.coordinates[1];

}

//console.log('Tweet:', data.id_str, data.lang, createdAt, data.user.screen_name, data.text,
replyUser, retweetedUser, lat, lon);

var sql = 'INSERT INTO ' + hdb_schema + '.' + table + ' '
('id','created','text','lang','user','replyUser','retweetedUser';

if (data.geo !== null)

{

sql += ', "lat", "lon"';

}

```



```

}

sql += ' ) VALUES(' + data.id_str + '\,' + createdAt + '\,' + emojiStrip(data.text.replace(/\g,
" ")) + '\,' + data.lang + '\,' + data.user.screen_name + '\,' + replyUser + '\,' +
retweetedUser + '\';

if (data.geo !== null)

{

    sql += ',' + lat + ',' + lon;

}

sql += ')';

hdb.exec(sql, function (err, affectedRows) {

if (err) {

console.log('Error:', err);

console.log('SQL:', sql);

return console.error('Error:', err);

}

//When a tweet is inserted, write to the console console.log('Tweet inserted:', data.id_str,
createdAt, affectedRows);

});

});

//Start tracking the keyword

stream.track(track);

}

else {

res.send('Nothing to track');

}

});

```

```
//Route to stop tracking
app.get('/do/stop', function(req, res, next){
  if(stream)
  {
    stream.abort();
    console.log('Stop');
    res.send('Stop');
  }
});

//Starts the server and waits
var server = app.listen(8888, function() {
  console.log('Listening on port %d', server.address().port);
  console.log('Press Ctrl-C to terminate');
});
```

### **Create Text Analysis Index on Tweets**

```
CREATE FULLTEXT INDEX "tweets" ON "Tweets"("text")  
CONFIGURATION 'EXTRACTION_CORE_VOICEOFCUSTOMER'  
LANGUAGE COLUMN "lang"  
LANGUAGE DETECTION ('EN','FR','DE','ES','ZH')  
TEXT ANALYSIS ON  
;
```

### Create Tweeters View with stance and influence scores

```
CREATE VIEW "Tweeters" AS

SELECT t."user",

       CAST(

           (CASE WHEN SP IS NULL THEN 0 ELSE SP * 5 END)

           + (CASE WHEN WP IS NULL THEN 0 ELSE WP * 2 END)

           - (CASE WHEN WN IS NULL THEN 0 ELSE WN * 2 END)

           - (CASE WHEN SN IS NULL THEN 0 ELSE SN * 5 END)

           AS INT) AS "stance",

       CAST(

           ((CASE WHEN RRC IS NULL THEN 0 ELSE RRC END) + (CASE

WHEN RTRC IS NULL THEN 0 ELSE RTRC END))

           - ((CASE WHEN RSC IS NULL THEN 0 ELSE RSC END) + (CASE WHEN

RTSC IS NULL THEN 0 ELSE RTRC END))

           AS INT) AS "influence"

       FROM (SELECT DISTINCT "user" FROM "Tweets") t

       LEFT JOIN ( SELECT "user", SUM(SP) AS SP, SUM(WP) AS WP,

SUM(WN) AS WN, SUM(SN) AS SN

FROM "Tweets" t

LEFT JOIN (

SELECT "id", SUM(CASE TA_TYPE WHEN 'StrongPositiveSentiment' THEN "total"

END) AS SP,

SUM(CASE TA_TYPE WHEN 'WeakPositiveSentiment' THEN "total" END) AS

WP,

SUM(CASE TA_TYPE WHEN 'WeakNegativeSentiment' THEN "total" END) AS

WN,

SUM(CASE TA_TYPE WHEN 'StrongNegativeSentiment' THEN "total" END) AS
```

SN

```
FROM (

SELECT "id", TA_TYPE, COUNT(*) AS "total"

FROM "$TA_tweets"

WHERE TA_TYPE = 'StrongPositiveSentiment' OR

      TA_TYPE = 'WeakPositiveSentiment' OR

      TA_TYPE = 'WeakNegativeSentiment' OR

      TA_TYPE = 'StrongNegativeSentiment'

GROUP BY "id", TA_TYPE

)

GROUP BY "id"

) i ON t."id" = i."id"

GROUP BY "user"

) s ON s."user" = t."user"

LEFT JOIN (

SELECT "replyUser", COUNT(*) AS RRC

FROM "Tweets"

WHERE "replyUser" != "

GROUP BY "replyUser"

) rrc ON rrc."replyUser" = t."user"

LEFT JOIN (

SELECT "retweetedUser",

COUNT(*) AS RTRC FROM "Tweets"

WHERE "retweetedUser" != "

GROUP BY "retweetedUser"
```

```

) rtrc ON rtrc."retweetedUser" = t."user"

LEFT JOIN (

    SELECT "user",

    COUNT(*) AS RSC FROM "Tweets"

    WHERE "replyUser" != "

    GROUP BY "user"

) rsc ON rsc."user" = t."user"

LEFT JOIN (

    SELECT "user",

    COUNT(*) AS RTSC FROM "Tweets"

    WHERE "retweetedUser" != "

    GROUP BY "user"

) rtsc ON rtsc."user" = t."user";

```

## Create the Table Types

```
CREATE TYPE PAL_T_DATA AS TABLE ("user" VARCHAR(256), "stance"  
INTEGER, "influence" DOUBLE);
```

```
CREATE TYPE PAL_T_PARAMS AS TABLE (NAME VARCHAR(50), INTARGS  
INTEGER, DOUBLEARGS DOUBLE, STRINGARGS VARCHAR(100));
```

```
CREATE TYPE PAL_T_RESULTS AS TABLE ("user" VARCHAR(256),  
"ClusterNumber" INTEGER, "distance" DOUBLE);
```

```
CREATE TYPE PAL_T_CENTERS AS TABLE ("ClusterNumber" INTEGER, "stance"  
INTEGER, "influence" INTEGER);
```

## **Create PAL\_SIGNATURE TABLE**

```
CREATE COLUMN TABLE PAL_SIGNATURE (ID INTEGER, TYPENAME  
VARCHAR(100), DIRECTION VARCHAR(100));  
  
INSERT INTO PAL_SIGNATURE VALUES (1, 'GBI_361.PAL_T_DATA', 'in')  
  
INSERT INTO PAL_SIGNATURE VALUES (2, 'GBI_361.PAL_T_PARAMS', 'in')  
  
INSERT INTO PAL_SIGNATURE VALUES (3, 'GBI_361.PAL_T_RESULTS', 'out')  
  
INSERT INTO PAL_SIGNATURE VALUES (4, 'GBI_361.PAL_T_CENTERS', 'out');  
  
GRANT SELECT ON PAL_SIGNATURE TO SYSTEM;  
  
CALL SYSTEM.AFL_WRAPPER_GENERATOR ('GBI_361_TweetersClustering',  
'AFLPAL', 'KMEANS', PAL_SIGNATURE);
```



## Create Parameter Table

```
CREATE COLUMN TABLE PAL_PARAMS LIKE PAL_T_PARAMS;  
INSERT INTO PAL_PARAMS VALUES ('THREAD_NUMBER', 2, null, null);  
INSERT INTO PAL_PARAMS VALUES ('GROUP_NUMBER', 3, null, null);  
INSERT INTO PAL_PARAMS VALUES ('GROUP_NUMBER_MIN', 5, null, null);  
INSERT INTO PAL_PARAMS VALUES ('GROUP_NUMBER_MAX', 10, null, null);  
INSERT INTO PAL_PARAMS VALUES ('INIT_TYPE', 4, null, null);  
INSERT INTO PAL_PARAMS VALUES ('DISTANCE_LEVEL', 2, null, null);  
INSERT INTO PAL_PARAMS VALUES ('MAX_ITERATION', 100, null, null);  
INSERT INTO PAL_PARAMS VALUES ('EXIT_THRESHOLD', null, 1.0E-6, null);  
INSERT INTO PAL_PARAMS VALUES ('NORMALIZATION', 0, null, null);
```

### **Create Output Tables**

```
CREATE COLUMN TABLE PAL_RESULTS LIKE PAL_T_RESULTS;
```

```
CREATE COLUMN TABLE PAL_CENTERS LIKE PAL_T_CENTERS;
```

## Create Views for Odata

```
CREATE VIEW "TweetersClustered" AS
```

```
    SELECT s.*, c."ClusterNumber" + 1 AS "ClusterNumber"
```

```
    FROM "Tweeters" s
```

```
    INNER JOIN "PAL_RESULTS" c ON c."user" = s."user" ;
```

```
CREATE VIEW "TweetersClusteredSummary" AS
```

```
    SELECT "ClusterNumber", c."stance", c."influence", COUNT(*) AS "users"
```

```
    FROM "Tweeters" t
```

```
    INNER JOIN "TweetersClustered" c ON c."user" = t."user"
```

```
    GROUP BY "ClusterNumber", c."stance", c."influence" ;
```