**Web Services and Service Oriented Programming**

**Assignment 4**

**Data Analysis**

Chinmay Kulkarni

This assignment deals with analyzing the API data queried in previous assignment and to get meaningful insights from the data. The data has 45 attributes and 11199 entries, one for each API. Almost all the variables are nominal and few attributes like ‘Rating’ and ‘Updated Year’ are numerical. The data consists mostly of text data, Therefore I have performed Text mining on the Data.

Text mining in this assignment involves, parsing the JSON data of the APIs, this is done using python NLTK package. I have performed Clustering technique to group the APIs with same functionality, which can be used for Web service recommendation. The biggest problem face while clustering is the visualization of the clusters formed. I have tackled this problem by using Principle Component Analysis. The K-means clustering and PCA is performed using Scikit-Learn python machine learning library. *I could cluster 11184 services with 34 clusters*, and the clusters were visualized using MatplotLib python library.

The overall process is explained in the following steps:

**A)- Text mining by forming TF-IDF matrix of the API Data.**

**B)- K-means Clustering of the TF-IDF matrix.**

**C)- Principle Component Analysis of the data.**

**D)- Visualizing clusters using Principle components and cluster prototype.**

**Text Mining:**

* Data is extracted from attributes ‘Title’, ’Category’, ’Summary’, ’Description’, ’Tags’ for each API to get important words describing the data.
* These words were filtered in following ways:
* The words validated using a dictionary of words provided with the code. If the word is not present in the dictionary then it is not considered for text mining.
* It is checked for being a stop word.
* Punctuations are removed.
* Further these files are saved as Text file, with the name of file as the name of the API.
* These files are Parsed Individually, to form a TF-IDF matrix.
* TF-IDF matrix consisted of (11184 X 7797) values, where 7797 is number of words extracted from all services.

**K-Means Clustering:**

K-means clustering is performed using Euclidean distance as distance metric. To choose the value for optimal “K”, to perform Clustering process, K vs Sum of Squared Error (SSE) was plotted and the K was chosen using the elbow technique. The optimal K selected was 34. All the 11184 Web services were clustered. The clusters formed are as follows:

Clusters with prominent keywords and few of its members are displayed below:

Cluster 0 Prototypes(Keywords):

['secur'],['authent'],['servic'],['user'],['ident'],['use'],['verif'],['applic'],['fraud'],['protect'],['valid'],['password'],['we],['data'],['site'],['allow'],['provid'],['develop'],['account'],['login'],

Cluster0 Members:

CheckInOn.Me.txt',Brightcloud .txt',TrustedX.txt',BankCard Central VERePay.txt'.......

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

Cluster 1 Prototypes(Keywords):

['twitter'],['tweet'],['social'],['user'],['follow'],['search'],['servic'],['link'],['share'],['use'],['trend'],['sentiment'],['data'],['post'],['let'],['format'],['photo'],['peopl'],['allow'],['tool'],

Cluster1 Members:

TweetSentiments.txt',JumboURL.txt',Cadmus.txt',TweetBoner.txt',tweethook.txt',Twishort.txt'...

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

Cluster 2 Prototypes(Keywords):

['video'],['stream'],['content'],['media'],['servic'],['platform'],['share'],['user'],['encod'],['upload'],['player'],['allow'],['use'],['applic'],['host'],['manag'],['live'],['access'],['onlin'],['develop'],

Cluster2 Members:

Howcast.txt',Twistage.txt',Sociagram.txt',Videolicious.txt',Brightcove.txt'.....

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

Cluster 3 Prototypes(Keywords):

['trade'],['financi'],['exchang'],['stock'],['currenc'],['market'],['data'],['quot'],['rate'],['servic'],['price'],['user'],['use'],['account'],['provid'],['financ'],['inform'],['histor'],['order'],['requir'],

Cluster3 Members:

Quantopian.txt',PsychSignal.txt',kaChing.txt',InfoValutar Rate.txt',CitiFXpro.txt'....

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

Cluster 4 Prototypes(Keywords):

['scienc'],['databas'],['sequenc'],['biolog'],['data'],['protein'],['genet'],['research'],['structur'],['inform'],['gene'],['servic'],['align'],['access'],['provid'],['search'],['medic'],['refer'],['use'],['method'],

Cluster4 Members:

Conservation Scorer.txt',HMMER.txt',MultAlin.txt',DLESE.txt',CIPF RENATO.txt',Freebase.txt'.....

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

Cluster 5 Prototypes(Keywords):

['map'],['locat'],['servic'],['data'],['place'],['display'],['geograph'],['applic'],['we],['provid'],['rout'],['gi'],['inform'],['develop'],['use'],['address'],['viewer'],['allow'],['point'],['imag'],

Cluster5 Members:

Strand Map Service.txt',Bing Maps.txt',Google Static Maps.txt',ondi.ro Geocoding.txt'...

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

Cluster 6 Prototypes(Keywords):

['payment'],['card'],['merchant'],['process'],['transact'],['credit'],['servic'],['onlin'],['mobil'],['account'],['allow'],['secur'],['integr'],['user'],['applic'],['shop'],['debit'],['accept'],['method'],['subscript'],

Cluster6 Members:

mygate.txt',Regal Technologies DevConnect.txt',Intellivative.txt',V.me by Visa.txt'

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

Cluster 7 Prototypes(Keywords):

['photo'],['imag'],['share'],['servic'],['user'],['album'],['recognit'],['upload'],['edit'],['print'],['allow'],['applic'],['use'],['develop'],['site'],['social'],['photograph'],['face'],['pictur'],['video'],

Cluster7 Members:

Blipfoto.txt',ZetaPrints.txt',WebPurify Image Moderation.txt',Cortexica.txt',Glosk.txt'

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

Cluster 8 Prototypes(Keywords):

['travel'],['book'],['hotel'],['flight'],['reserv'],['rental'],['search'],['vacat'],['servic'],['tour'],['airlin'],['inform'],['trip'],['accommod'],['provid'],['airport'],['site'],['offer'],['avail'],['tourism'],

Cluster8 Members:

SilverRail Technologies.txt',Serko Online.txt',Alpharooms Affiliate.txt'

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

Cluster 9 Prototypes(Keywords):

['cloud'],['storag'],['host'],['manag'],['server'],['servic'],['internet'],['applic'],['platform'],['user'],['provid'],['comput'],['data'],['develop'],['infrastructur'],['access'],['use'],['enterpris'],['rest'],['allow'],

Cluster9 Members:

SnapCrowd.txt',FME Cloud.txt',Google Drive.txt',CloudPassage.txt',Nimbus Phantom.txt'

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

Cluster 10 Prototypes(Keywords):

['game'],['player'],['social'],['develop'],['video'],['onlin'],['user'],['data'],['inform'],['flash'],['platform'],['use'],['play'],['servic'],['provid'],['allow'],['access'],['server'],['includ'],['statist'],

Cluster10 Members:

PlayCanvas.txt',Ryzom.txt',HabitRPG.txt',Heroes of Newerth Statistics.txt',BF4 Stats.txt'

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

Cluster 11 Prototypes(Keywords):

['govern'],['data'],['open'],['inform'],['dataset'],['state'],['citi'],['servic'],['access'],['provid'],['polit'],['public'],['feder'],['statist'],['avail'],['vote'],['use'],['rest'],['agenc'],['report'],

Cluster11 Members:

DemocracyMap.txt',DataKC.txt',FixMyStreet.txt',DevLeap Codice Fiscale.txt',HealthData.gov catalog.txt'

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

Cluster 12 Prototypes(Keywords):

['shorten'],['internet'],['link'],['short'],['util'],['servic'],['tool'],['use'],['long'],['user'],['expand'],['creat'],['allow'],['return'],['click'],['simpl'],['rest'],['share'],['site'],['applic'],

Cluster12 Members:

Rewd.co.txt',JumboTweet.txt',Qoiob.txt',t0.tv.txt',B54.in .txt',durl.us.txt'

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

Cluster 13 Prototypes(Keywords):

['email'],['market'],['campaign'],['servic'],['manag'],['address'],['valid'],['mail'],['messag'],['newslett'],['send'],['custom'],['user'],['list'],['use'],['deliveri'],['allow'],['provid'],['subscri],['contact'],

Cluster13 Members:

SureVerify.txt',Ongage.txt',CDYNE Email Verification.txt',Campaign Monitor.txt',BlinkCampaign.txt'

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

Cluster 14 Prototypes(Keywords):

['translat'],['languag'],['tool'],['text'],['human'],['servic'],['machin'],['local'],['dictionari'],['project'],['content'],['applic'],['provid'],['use'],['document'],['format'],['we],['develop'],['request'],['profession'],

Cluster14 Members:

Language Studio Asia Online.txt',SDL BeGlobal.txt',LingoTip Translation.txt',Translution.txt',

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

Cluster 15 Prototypes(Keywords):

['search'],['engin'],['result'],['servic'],['provid'],['data'],['user'],['we],['site'],['keyword'],['access'],['allow'],['use'],['develop'],['local'],['semant'],['index'],['queri'],['content'],['applic'],

Cluster15 Members:

NCSU Scholarly Publications Repository.txt',ZEIT ONLINE .txt',Bounce MixedResults.txt'

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

Cluster 16 Prototypes(Keywords):

['sport'],['transport'],['transit'],['data'],['inform'],['rout'],['vehicl'],['bu'],['servic'],['train'],['traffic'],['provid'],['track'],['time'],['footbal'],['locat'],['access'],['schedul'],['leagu'],['includ'],

Cluster16 Members:

RidePSTA.txt',Metro Trip Planner.txt',Global Sports Media.txt',Averitt.txt',GTFS Data Exchange.txt'

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

Cluster 17 Prototypes(Keywords):

['social'],['network'],['user'],['media'],['share'],['platform'],['servic'],['applic'],['friend'],['develop'],['site'],['allow'],['data'],['use'],['commun'],['creat'],['content'],['access'],['event'],['profil'],

Cluster17 Members:

Beevolve Crawler.txt',SlipStream.txt',Brightkite.txt',Zerista.txt',DailyCred.txt',Pinboard.txt'

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

Cluster 18 Prototypes(Keywords):

['project'],['manag'],['task'],['collabor'],['track'],['tool'],['team'],['time'],['develop'],['servic'],['user'],['softwar'],['applic'],['function'],['creat'],['allow'],['enterpris'],['use'],['access'],['platform'],

Cluster18 Members:

LiquidPlanner .txt',AgileZen.txt',The Helioviewer Project.txt',Celoxis.txt',Floorplanner.txt'

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

Cluster 19 Prototypes(Keywords):

['advertis'],['ad'],['market'],['mobil'],['campaign'],['publish'],['platform'],['servic'],['applic'],['manag'],['network'],['develop'],['allow'],['integr'],['access'],['monet'],['user'],['report'],['affili'],['provid'],

Cluster19 Members:

Get Me Listed.txt',Friday-Ad.txt',VigLink.txt',MediaEquals.txt',Amobee.txt',adBrite.txt',Po.st.txt'

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

Cluster 20 Prototypes(Keywords):

['enterpris'],['manag'],['busi'],['custom'],['data'],['sale'],['servic'],['applic'],['platform'],['integr'],['contact'],['market'],['user'],['lead'],['softwar'],['document'],['allow'],['account'],['provid'],['access'],

Cluster20 Members:

SAManage.txt',MINDBODY.txt',Agendize Action.txt',Predixion.txt',rules.io.txt',WORKetc Web Service.txt'

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

Cluster 21 Prototypes(Keywords):

['address'],['code'],['locat'],['postal'],['map'],['servic'],['lookup'],['postcod'],['zip'],['valid'],['citi'],['countri'],['data'],['inform'],['tool'],['refer'],['return'],['use'],['format'],['internet'],

Cluster21 Members:

Geocodio.txt',Accthub.txt',GreatSchools.txt',Singapore Geocoder.txt',FreeGeoIP.txt',SetGetGo IP Geolocation.txt'

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

Cluster 22 Prototypes(Keywords):

['shop'],['product'],['deal'],['retail'],['coupon'],['servic'],['store'],['onlin'],['price'],['cart'],['merchant'],['local'],['affili'],['user'],['custom'],['order'],['offer'],['inform'],['use'],['allow'],

Cluster22 Members:

Slice.txt',Gimahhot Shopping.txt',CleverSet.txt',Ohmydeal.txt',QuarkRank.txt',Best Buy BBYOpen Categories.txt'

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

Cluster 23 Prototypes(Keywords):

['jo],['recruit'],['search'],['employ'],['hire'],['post'],['servic'],['schedul'],['list'],['manag'],['site'],['applic'],['user'],['allow'],['develop'],['seeker'],['hr'],['use'],['compani'],['resum'],

Cluster23 Members:

JobGizmo.txt',Hired in NY.txt',stSoftware Job Track.txt',Google Coordinate.txt',Econz Eservice.txt',Schedulefly.txt'

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

Cluster 24 Prototypes(Keywords):

['weather'],['forecast'],['data'],['meteorolog'],['station'],['condit'],['observ'],['locat'],['provid'],['servic'],['inform'],['report'],['nation'],['climat'],['wind'],['map'],['current'],['method'],['time'],['citi'],

Cluster24 Members:

ServiceObjects DOTS Fast Weather.txt',Meteomedia Weather Development Kit.txt',Weather Source.txt',IMN.txt',WeerAPI.txt'

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

Cluster 25 Prototypes(Keywords):

['mobil'],['applic'],['develop'],['devic'],['platform'],['manag'],['allow'],['build'],['servic'],['access'],['user'],['integr'],['creat'],['function'],['tool'],['app'],['we],['featur'],['telephoni'],['data'],

Cluster25 Members:

iPointer.txt',ChowNow.txt',CloudMine.txt',AppsGeyser.txt',DeployGate.txt',Netatmo.txt'

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

Cluster 26 Prototypes(Keywords):

['telephoni'],['voic'],['phone'],['servic'],['number'],['applic'],['messag'],['mobil'],['commun'],['speech'],['develop'],['provid'],['use'],['user'],['text'],['telephon'],['allow'],['we],['platform'],['custom'],

Cluster26 Members:

MegaPhone Labs.txt',CallZing Voice Broadcast .txt',Sipgate.txt',HummingBytes Q&A.txt'

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

Cluster 27 Prototypes(Keywords):

['internet'],['ship'],['servic'],['we],['monitor'],['tool'],['site'],['user'],['track'],['analyt'],['use'],['applic'],['page'],['data'],['provid'],['develop'],['allow'],['fax'],['host'],['test'],

Cluster27 Members:

AppDirect.txt',SONoMA.txt',PDF24 Javascript PDF.txt',OVH.com.txt',Sentinel Monitoring.txt'

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

Cluster 28 Prototypes(Keywords):

['music'],['song'],['artist'],['radio'],['stream'],['search'],['servic'],['lyric'],['user'],['play'],['event'],['access'],['allow'],['share'],['site'],['concert'],['track'],['applic'],['social'],['data'],

Cluster28 Members:

Nokia Music.txt',setlist.fm.txt',Dubset MixSCAN.txt',DoStuffMedia.txt',Digital Music Engine.txt',iJukebox.txt'

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

Cluster 29 Prototypes(Keywords):

['file'],['storag'],['share'],['upload'],['user'],['cloud'],['servic'],['allow'],['backup'],['store'],['download'],['manag'],['onlin'],['document'],['applic'],['access'],['host'],['sync'],['media'],['use'],

Cluster29 Members:

FileStorm.txt',goo.im.txt',Seafile.txt',Bitcasa.txt',Nirvanix.txt',Mover.io.txt',FileTrek.txt'

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

Cluster 30 Prototypes(Keywords):

['invoic'],['account'],['financi'],['busi'],['onlin'],['payment'],['manag'],['servic'],['enterpris'],['small'],['creat'],['integr'],['user'],['financ'],['access'],['use'],['allow'],['track'],['rest'],['applic'],

Cluster30 Members:

Factura Directa.txt',disdar.txt',PayPal Invoicing.txt',billFLO Seller.txt',Quaderno.txt'

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

Cluster 31 Prototypes(Keywords):

['servic'],['data'],['tool'],['user'],['use'],['inform'],['provid'],['event'],['access'],['allow'],['develop'],['refer'],['applic'],['content'],['educ'],['manag'],['text'],['onlin'],['we],['format'],

Cluster31 Members:

Live Matrix.txt',Adobe Feeds.txt',BetaSignal.txt',UNAVCO Synthetic Aperture Radar.txt'

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

Cluster 32 Prototypes(Keywords):

['domain'],['internet'],['resel'],['host'],['registr'],['servic'],['manag'],['provid'],['regist'],['avail'],['user'],['registrar'],['wholesal'],['allow'],['use'],['we],['lookup'],['email'],['search'],['record'],

Cluster32 Members:

Directi.txt',Subreg.txt',Company Data Trees Parked Domain.txt',http.net DNS.txt',OpenSRS .txt'

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

Cluster 33 Prototypes(Keywords):

['messag'],['send'],['bulk'],['servic'],['text'],['mobil'],['gateway'],['user'],['receiv'],['telephoni'],['applic'],['global'],['provid'],['market'],['use'],['deliveri'],['allow'],['integr'],['sent'],['group'],

Cluster33 Members:

youRoom.txt',Large Bulk SMS.txt',Pulse.to.txt',GenesisBulkSMS.txt',Wholesale SMS.txt',Upside Wireless.txt'

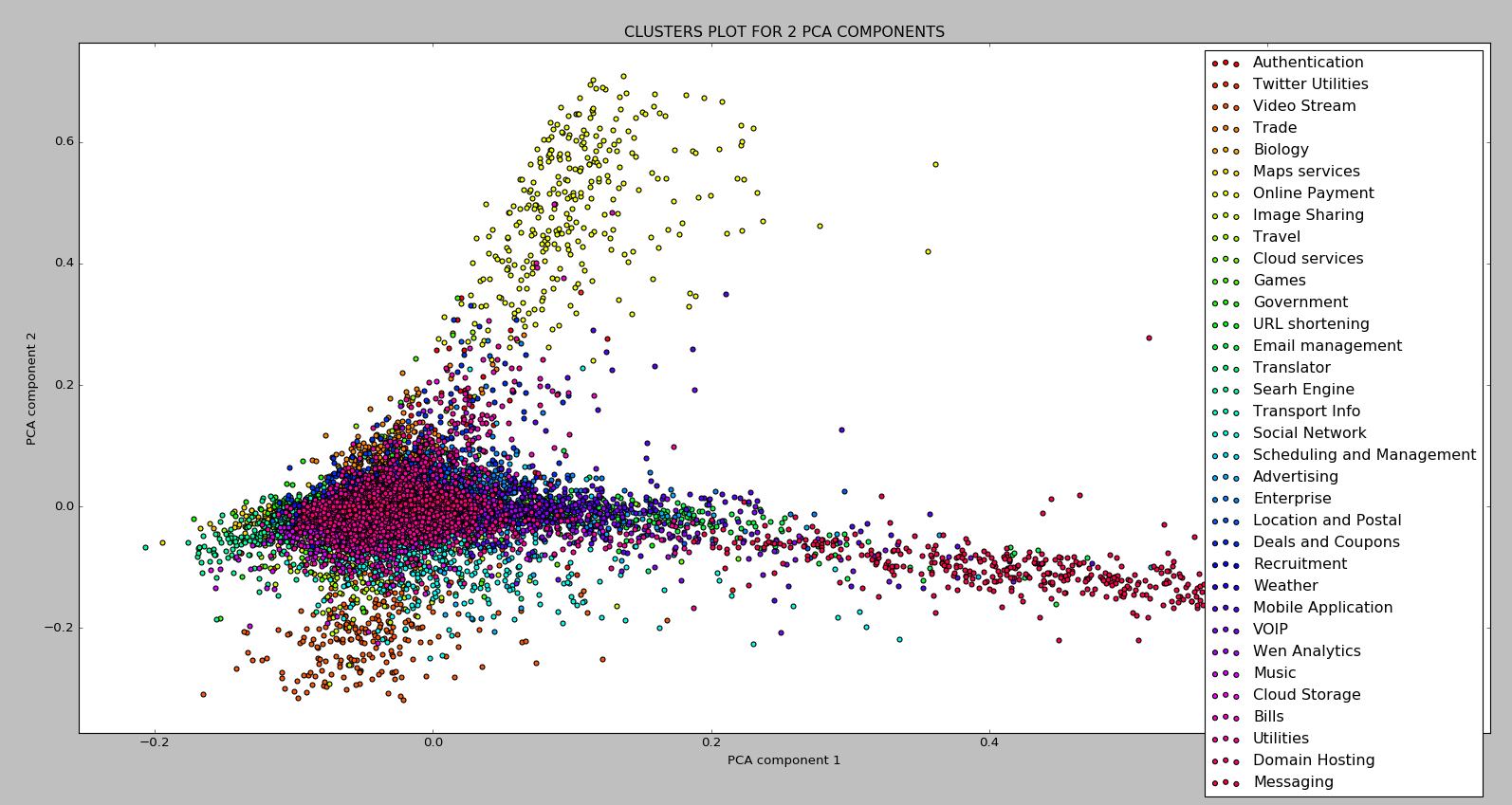
---------------------------------------------

**Principle Component Analysis of the data:**

Principle component is performed on the data to visualize the clusters. PCA is performed as both 2D and 3D, to get both the perspectives of the clusters formed. It is performed using Sickit-learn python library.

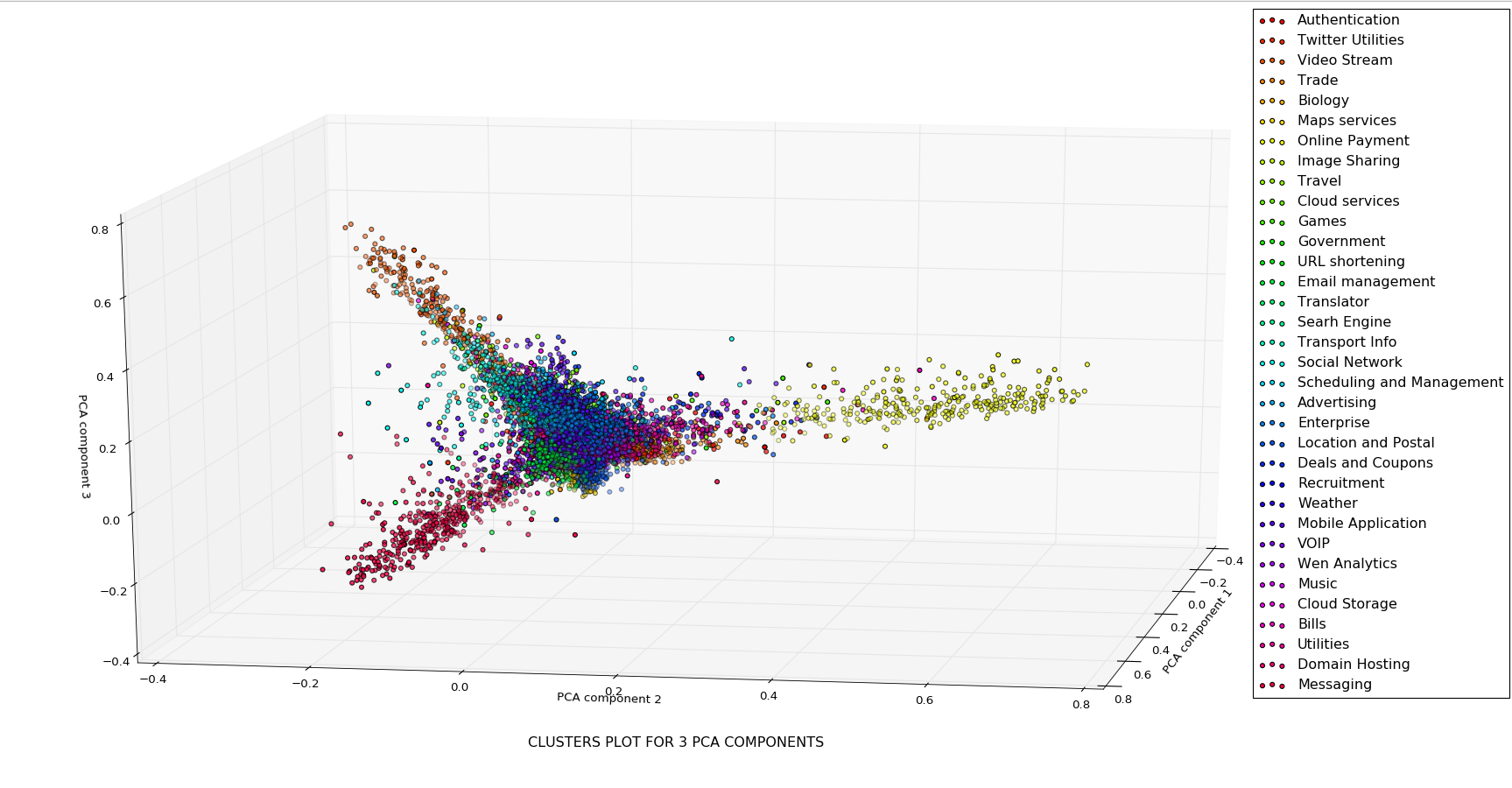
**Visualization:**

**2D Plot**

****

In this we can see clusters, but to get a better view, we can have a look at the 3D plot of same clusters.

**3D PLOT:**

****