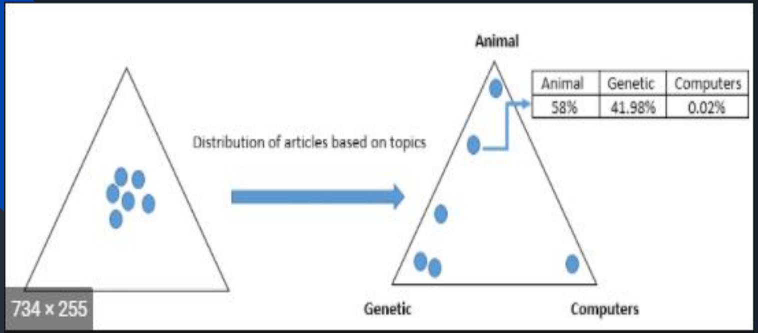
Multidimensional Fingerprinting

By Dennis Sarovski
Project Supervisor PhD. Saed Alrabaee

Background

- 71.81% of mobile users use Android OS
- Mobile & User friendly infrastructure is increasing everyday
- Due to popularity a lot of Iot devices use some distribution of android OS
- Malware developers across the world produce over 10k types of malware per day based on a 2019 study.



What is Multi-Dimensional Fingerprinting

- A way to identify and cluster malware based on unique aspects specific to the malware on different dimensions.
- Like an Police officer looking for evidence (finger prints/ DNA) we look for similar things in the code.

Motivation

- Due to mass innovation in tech industry, many malicious users are flocking to develop technology to take advantage of this rapidly developing industry to cheat their way to success.

- Like in “Big Data” it become increasingly difficult to manually analyze user data. The same is true for malware.

- It is also increasingly difficult to build a one size fits all model due the increasing differences and output of today's malware developers.

Goal

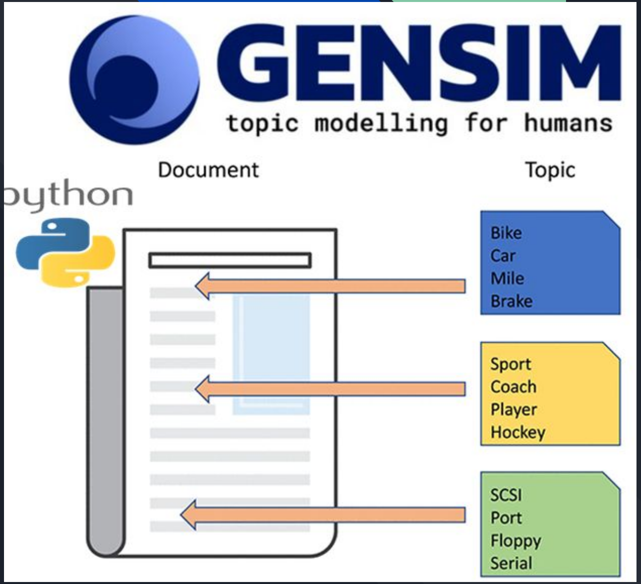
- Build a model that can
  - Reduce the need to manually analyze malware
  - Fingerprint types of malware to allow us to categorize them in specific families.
  - Be able to detect both known and unseen malware

Related Works

- My project is based on 2 papers one being:
- Scalable and robust unsupervised android malware fingerprinting using community-based network partitioning (2020) - by **ElMouatez Billah Karbab, Mourad Debbabi, Abdelouahid Derhabb, Djedjiga Mouhebc**

- MalDozer: Automatic framework for android malware detection using deep learning (2018) - by **ElMouatez Billah Karbab, Mourad Debbabi, Abdelouahid Derhabb, Djedjiga Mouhebc**

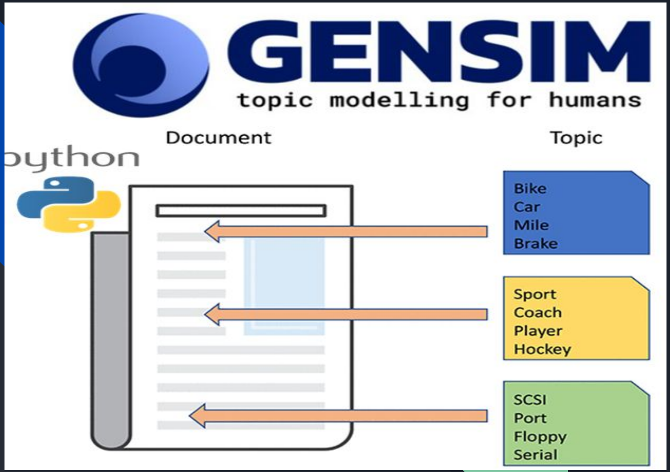
Methodology



In order to tackle our goal I decided that the best approach would be to use some sort of Artificial Intelligence that could help automate the process.

- We ended up using LDA (Latent Dirichlet Allocation)

- LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of data are similar or different.




Landroid/app/AlarmManager_Ljava/lang/CharSequence, Landroid/widget/Button, Ljava/io/DataOutputStream_Landroid/telephony/SmsMessage, Ljava/io/BufferedReader_Landroid/content/ContentResolver, Lorg/json/JSONException, Landroid/net/Proxy, Landroid/app/Activity, Ljava/io/FileOutputStream, Ljava/net/Proxy, Landroid/app/Service

- LDA imagines a fixed set of topics. Each topic represents a set of words. And the goal of LDA is to map all the documents to the topics in a way, such that the words in each document are mostly captured by those imaginary topics.

- We have chosen topic modeling:
    - The reason we decided to choose topic modeling is because I found that it would be easier to cluster different malware under similar families (topics) due to this similarities or differences in the code.

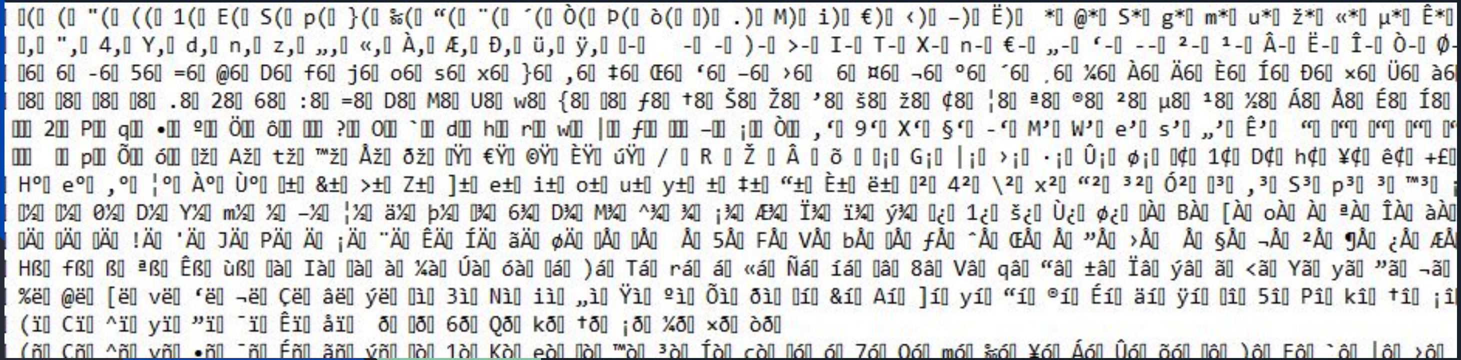
- In reality the topic modeling becomes more complex than the first example due to not being human readable.

- Once all our data comes together it becomes easier to compare and cluster.

Name	Size	Packed Size	Modified	Created
assets	228 816	123 655		
l	2 273	1 218		
lib	78 612	53 220		
META-INF	30 757	11 266		
res	264 236	167 044		
AndroidManifest.xml	7 880	2 098	2012-05-24 15:15	
classes.dex	825 168	380 232	2012-05-24 15:15	
resources.arsc	98 660	98 660	2012-05-24 15:15	

- First step of pre processing our data would be figuring out what is or isn't important!

- The classes.dex is the most important file we use for our model
  - A Dex file contains code which is executed by the android runtime.
  - Every APK has a single classes.dex file which references any classes or methods used within an app.



```
Processing '0008c4a131a4e6ff6a99ff3795e1b17d.dex'...
Opened '0008c4a131a4e6ff6a99ff3795e1b17d.dex', DEX version '035'
Class #0            -
  Class descriptor  : 'Lcom/androidemu/Cheats$Item;'
  Access flags      : 0x0001 (PUBLIC)
  Superclass        : 'Ljava/lang/Object;'
  Interfaces        -
  Static fields     -
  Instance fields   -
    #0              : (in Lcom/androidemu/Cheats$Item;)
      name          : 'code'
      type          : 'Ljava/lang/String;'
      access        : 0x0001 (PUBLIC)
    #1              : (in Lcom/androidemu/Cheats$Item;)
      name          : 'enabled'
      type          : 'Z'
      access        : 0x0001 (PUBLIC)
    #2              : (in Lcom/androidemu/Cheats$Item;)
      name          : 'name'
      type          : 'Ljava/lang/String;'
      access        : 0x0001 (PUBLIC)
    #3              : (in Lcom/androidemu/Cheats$Item;)
      name          : 'this$0'
      type          : 'Lcom/androidemu/Cheats;'
      access        : 0x1010 (FINAL SYNTHETIC)
  Direct methods    -
    #0              : (in Lcom/androidemu/Cheats$Item;)
      name          : '<init>'
      type          : '(Lcom/androidemu/Cheats;)V'
      access        : 0x10001 (PUBLIC CONSTRUCTOR)
      code          -
      registers     : 2
      ins           : 2
      outs          : 1
      insns size    : 6 16-bit code units
0140f8:                                        |[0140f8] com.androidemu.Cheats$Item.<init>:(Lcom/androidemu/Cheats;)V
014108: 5b01 5500                              |0000: iput-object v1, v0, Lcom/androidemu/Cheats$Item;.this$0:Lcom/androidemu/Cheats; // field@0055
01410c: 7010 280c 0000                         |0002: invoke-direct {v0}, Ljava/lang/Object;.<init>:()V // method@0c28
014112: 0e00                                   |0005: return-void
      catches       : (none)
      positions     :
        0x0000 line=17
      locals        :
        0x0000 - 0x0006 reg=0 this Lcom/androidemu/Cheats$Item;
        0x0000 - 0x0006 reg=1 (null) Lcom/androidemu/Cheats;
```

- When we extract the classes.dex file from each malware it is human unreliable so we need to use a dexdump to make it human readable.

- The dex file to the left is the human readable text

Lbsh/This Lbsh/DelayedEvalBshMethod Lco/lvdou/showshow/ui/FragDiscovery Lu/aly/fs Lorg/json/JSONObject Lco/lvdou/showshow/ui/subject/SubjectUtil Lco/lvdou/showshow/
o/lvdou/showshow/model/f/a/ak Lco/lvdou/showshow/diy/combine/OnCombineChangeListener Landroid/support/v4/os/EnvironmentCompatKitKat Ljavax/servlet/http/HttpServletR
/lvdou/showshow/service/e Lcom/j256/ormlite/stmt/query/SetValue ZZLjava/lang/String org/sax/properties/lexical ILandroid/support/v4/view/PagerAdapter Lorg/cocos2dx/
ite/stmt/query/IsNull Lco/lvdou/showshow/b/h Ljava/lang/Math Lcom/umeng/socialize/controller/impl/k Lco/lvdou/showshow/a/db ILcom/umeng/socialize/common/ResContaine
Exception Lco/lvdou/showshow/g/cd cn/my/wallpaper/shareIt Lco/lvdou/showshow/diy/font/selectbg/ActFontBackground long/2addr cn/app/my/editOtherInfo Landroid/widget/
ssion Lcom/umeng/socialize/net/utils/SocializeNetUtils Lco/lvdou/showshow/diy/font/combine/TxtSizeGalleryAdapter Lcom/tencent/mm/sdk/constants/ConstantsAPI Lco/lvdo
/g/ch Lu/aly/fp cn/trend Lco/lvdou/showshow/util/c/e Landroid/view/View Lcn/zjy/pulltorefreshview/PullToRefreshView Lco/lvdou/showshow/j/c/a/c Lco/lvdou/showshow/mo
oid/support/v4/widget/ScrollerCompatGingerbread ILco/lvdou/showshow/diy/combine/OnUpdateFontListListener Lco/lvdou/showshow/util/usersystem/o Lco/lvdou/showshow/a/b
/showshow/c/c/e Lco/lvdou/extension/OnNativeCallbackListener Lorg/cocos2dx/lib/Cocos2dxGLSurfaceViewManager Lco/lvdou/showshow/c/g Lcom/umeng/socialize/utils/Statis
wshow/files/effect2/ Lco/lvdou/showshow/c/r IIILjava/lang/Object Lco/lvdou/showshow/ui/account/ActRetrieveAccount Lco/lvdou/showshow/c/c/i Lco/lvdou/showshow/model/
odel/e/b Lco/lvdou/showshow/ui/material/ActPicMaterialDetailDelegate Lco/lvdou/showshow/a/av Lco/lvdou/showshow/global/b/d Ljava/io/Closeable Landroid/os/Bundle Lco
va/util/zip/GZIPOutputStream Landroid/support/v4/app/TaskStackBuilderHoneycomb Lcom/j256/ormlite/field/types/IntegerObjectType Lco/lvdou/showshow/e/b/e Lco/lvdou/sh
s2dx/lib/BaseUnlockService Lco/lvdou/showshow/util/h/k Lcom/umeng/analytics/AnalyticsConfig Lorg/jdom2/output/LineSeparator Lco/lvdou/showshow/receiver/TurntableGam
d/support/v4/widget/ContentLoadingProgressBar Lco/lvdou/showshow/view/g Ljava/lang/Object Ljavax/xml/stream/events/EndElement Lco/lvdou/showshow/ui/account/ActRetri
/support/StAXStreamProcessor Lcom/j256/ormlite/field/types/BaseEnumType Ljavax/swing/border/MatteBorder Lcom/umeng/socialize/view/aj Lcom/umeng/socialize/bean/SnsAc
il/AbstractQueue Lcom/viewpagerindicator/v cn/comment/count Lco/lvdou/showshow/util/wallpaper/OnWallpaperInforListener Lco/lvdou/showshow/view/MuliteColorViewGroup
w/SurfaceHolder ILcom/tencent/open/TaskGuide Lco/lvdou/showshow/util/usersystem/k Lu/aly/di Lcom/umeng/socialize/view/s Landroid/os/Parcel Lco/lvdou/showshow/floatw
serzone/ActDiyPickPicHead Lco/lvdou/showshow/util/h/m Lco/lvdou/showshow/g/ca Lco/lvdou/a/c/a/a Lbsh/BlockNameSpace Ljavax/swing/JPanel Ljava/util/regex/Pattern Lan
id/graphics/Xfermode cn/my/wallpaper/sell Lco/lvdou/showshow/a/dc Lco/lvdou/b/a/u Lco/lvdou/showshow/j/d/d/c IILandroid/graphics/Rect Lco/lvdou/showshow/util/c/c Lc
/types/DateStringType Lco/lvdou/showshow/j/av Landroid/widget/ImageView Lco/lvdou/extension/LDResLoader JZLco/lvdou/showshow/model/f/h Lcom/umpay/huafubao/h/a Lbsh/

- To preprocess I used regular expressions to collect anything that looked like an API and files that might be specific to the malware and saved it to its own word file.

- Once finished preprocessing, I had to figure out what hyperparameters would best suit the model so I ended up generating multiple models till i found the best coherence.

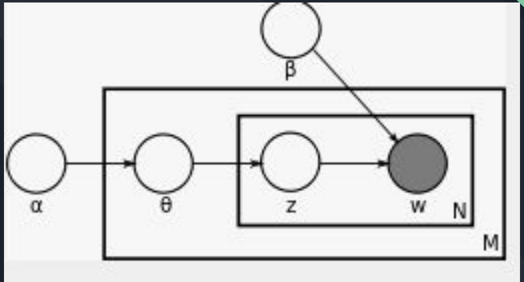
- Coherence was generated based on how many topics we can choose

Hyper-Parameters

- α is the parameter of Dirichlet prior on the per-document topic distribution

- β is the parameter of the Dirichlet prior on the per-topic word distribution

- K is the number of topics



$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi_{Z_{j,t}}),$$

Dominant_Topic	Topic_Perc_Contrib
23.0	0.9656
3.0	0.9590
0.0	0.0312
27.0	0.9530
15.0	0.9580
25.0	0.9169
0.0	0.0312
0.0	0.0312
22.0	0.9667
27.0	0.9209
20.0	0.7875
0.0	0.0312
13.0	0.9402
23.0	0.9054
0.0	0.0312
2.0	0.9135
0.0	0.0312
0.0	0.0312
15.0	0.9705
0.0	0.0312
13.0	0.9560
3.0	0.9641
0.0	0.0312
27.0	0.9559
27.0	0.9606
26.0	0.9511
2.0	0.9431
21.0	0.9451

- Once we finished with choosing which model best suits our needs I processed our data to get the results.

- I used F1 scores to identify and analyze how well and accurate my data was to see if any anomalies occurred in building my model

- Accuracy: Ratio of the correct labeled subjects to the whole pool of subjects

- Precision: Ratio of correctly positive labels by our program to all the positive labels

- Specificity: Ratio of the correctly negative labels by our program to all the documents who are positive in reality

- Misclassification: how often something is wrong

Experimental Setup

- For this project we used the maldozer data set which contains 20090 separate malware of which 20040 was used.

- The data set contained 32 different malware.

- This data uncompressed was ~ 200 GB of human readable dex files.

- The framework that was used was Gensim.
    - This was used due to its superior multithreading capabilities and other tools such as using tf-idf.

- Hardware that was used i7-7700 @ 3.60 GHz
- 32 GB DDR4 @ 4000 MHz
- RTX 2080TI

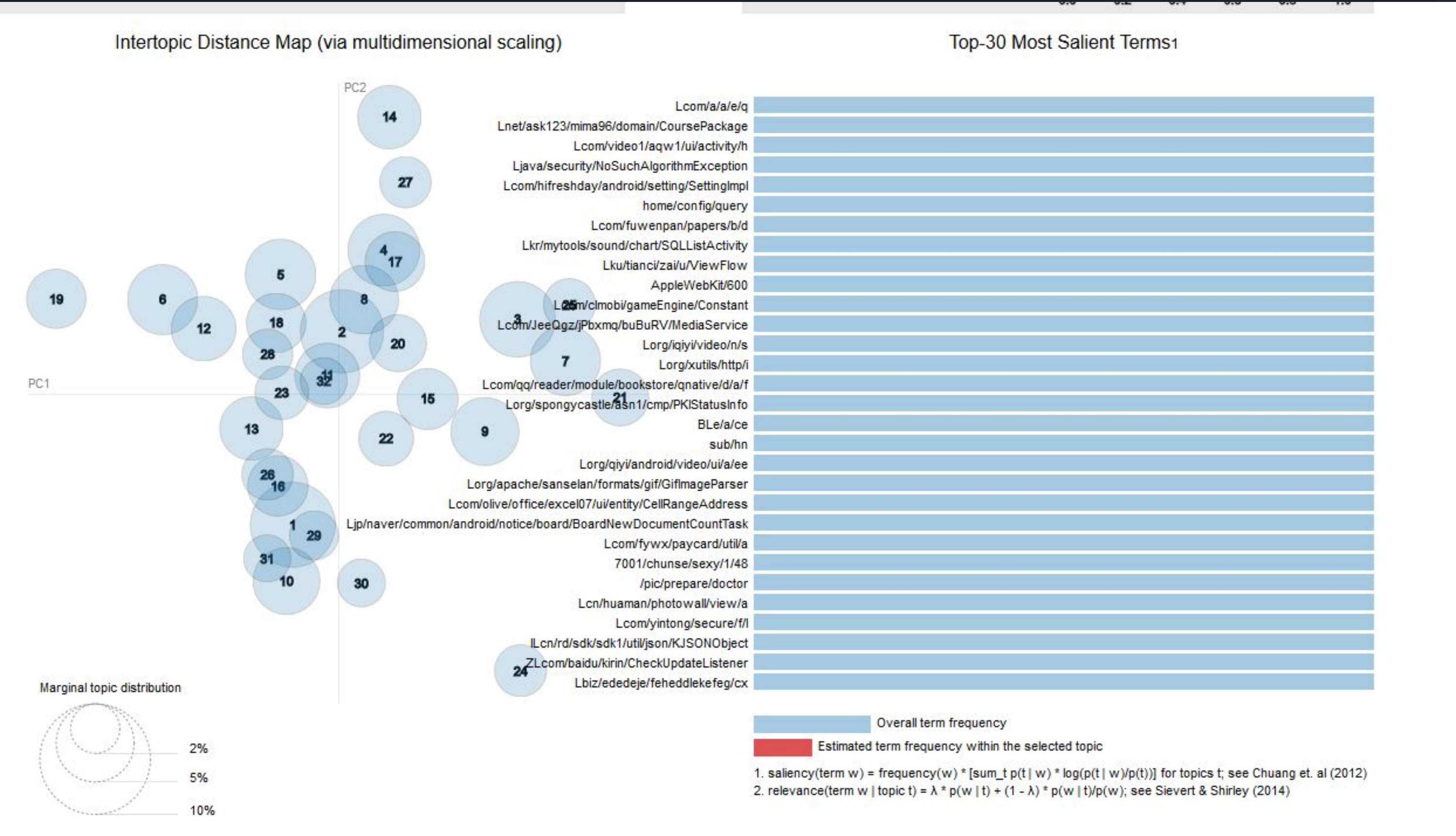
- Based on the hardware it took me ~24h to process my models based on the whole dataset

Results

	Document_No	Dominant_Topic	Topic_Perc_Contrib
0	0	23.0	0.9656
1	1	3.0	0.9590
2	2	0.0	0.0312
3	3	27.0	0.9530
4	4	15.0	0.9580
5	5	25.0	0.9169
6	6	0.0	0.0312
7	7	0.0	0.0312
8	8	22.0	0.9667
9	9	27.0	0.9209

18750d3a30a52e508aa4a03fdada630e-Adwo
1874ed85ba7160d4b8002f682d8b3102-SmsPay
1873ebb0538fc2656ea67aa93815dba6-FakeInst
187347fd95d5675ac183f86f8409789b-Adwo
18732dd40714f304cfa8647fd4b2b020-SMSReg
18648135b9a314e35920ddfe28db4186-Dowgin
185f1e54de7b139f56a65c4a8134b39a-FakeInst
18546fd960ae37cb2d335a908699c94c-FakeInst
184ce6255faf70d13575784c02ed8f52-Plankton
184c22371a693f0906f87474101023a1-Dowgin

Lcom/calendar/example/fmt/f/al, ILcom/googlea/callback/a, Landroid/os/Looper, Lname/rgmipn/jjhpku/juprvtnh/f8, Lcom/gsapp/wifitransport/network/transport/socket/nio/NioDatagramConnector, Lcom/mango/sanguo/view/battleNet/BattleNetPreliminaryViewController, cn/MobileGameCenter/gh/TaskCenter, Lcom/bodong/smartad/sdk/other/br, Lfahrbot/apps/rootcallblocker/ui/ProfilePreferenceActivity, Linfo/eeeg
Lcom/rosefloweri/kkgidyhoService, FLandroid/view/animation/Transformation, Leu/nordeus/topeleven/android/modules/clubshop/a/f, Lcom/edeeeeee/ececdlekefeg/cm, Landroid/support/v4/app/DialogFragment, Lcom/qihoo/video/widget/PhotoPlayView, Lcom/baidu/mapapi/MKOfflineMapListener, Ljava/util/Collections, Ljavax/crypto/Cipher, Lcom/cmread/bplusc/presenter/nativerequest/batchSubscribeFascicle
Landroid/provider/BaseColumns, Lcom/google/hfapservice/a/l, Lcom/tencent/karaoke/common/reporter/click/AppStartReporter, Lcom/sy/ishow/ui/show/a, Lcom/malangstudio/alarmmon/AlarmMonApplication, Lcom/gamelist/packs/Ad, smallgame2/icon16, Lgame/model/arrow/WeaponsLazer, IIZZZLjava/lang/String, Lcom/echo/holographlibrary/PieGraph
com/maps, Lnet/nmntntn/tnqoknmnlnr/au, Lcom/adwo/adsdk/f, Ljava/io/DataInputStream, Landroid/webkit/DownloadListener, Llogic/bean/VipRes, Lcom/adwo/adsdk/v, Lcom/waps/DisplayAdNotifier, Ljava/lang/Throwable, receiver/uninstall
Lcom/memezhibo/android/activity/mobile/show/CashRecordActivity, Lcom/pplive/android/data/b/n, Lcom/kaikai/live/juese/ui/Vip3Activity, Lcom/aligo/calculations/regressions/Regressions, Ljava/io/ByteArrayOutputStream, Lcom/youku/crazytogether/app/components/db/roomad/a, Lcom/mv/util/PhoneInfoUtil, Lcom/BHBRMILsm/b, Landroid/content/res/Configuration, Lvn/me/engine/network/CoreServices
Lcom/sds/android/ttpod/app/online/OnlineMusicActivity, IJLjava/lang/String, Lcom/dd725/comic/UI/ShowComic, Lxiaoying/utils/QSize, Lco/lvdou/showshow/ui/web/ActMyPost, Lcom/xiequ/c0010/an, Lcom/cutt/zhiyue/android/view/activity/chatting/ChattingTaskListActivityController, Lcom/wonlourev/kmageviua148343/SetPreferences, Lcom/android/volley/u, JJLjava/io/PrintWriter
Landroid/provider/BaseColumns, Lcom/google/hfapservice/a/l, Lcom/tencent/karaoke/common/reporter/click/AppStartReporter, Lcom/sy/ishow/ui/show/a, Lcom/malangstudio/alarmmon/AlarmMonApplication, Lcom/gamelist/packs/Ad, smallgame2/icon16, Lgame/model/arrow/WeaponsLazer, IIZZZLjava/lang/String, Lcom/echo/holographlibrary/PieGraph
Landroid/provider/BaseColumns, Lcom/google/hfapservice/a/l, Lcom/tencent/karaoke/common/reporter/click/AppStartReporter, Lcom/sy/ishow/ui/show/a, Lcom/malangstudio/alarmmon/AlarmMonApplication, Lcom/gamelist/packs/Ad, smallgame2/icon16, Lgame/model/arrow/WeaponsLazer, IIZZZLjava/lang/String, Lcom/echo/holographlibrary/PieGraph
Landroid/graphics/Camera, Landroid/support/v4/view/GestureDetectorCompat, Ljava/lang/Enum, Lorg/edekebe/eegegdlekefeg/k, Lcom/mmpay/happylordmatch/a/a, Ljava/io/DataInputStream, Lcom/tbu/a/d/a, Lcom/showselfu/ar, Landroid/view/animation/Animation, //details
com/maps, Lnet/nmntntn/tnqoknmnlnr/au, Lcom/adwo/adsdk/f, Ljava/io/DataInputStream, Landroid/webkit/DownloadListener, Llogic/bean/VipRes, Lcom/adwo/adsdk/v, Lcom/waps/DisplayAdNotifier, Ljava/lang/Throwable, receiver/uninstall
Landroid/widget/ListAdapter, IILandroid/os/Bundle, s3/eu, ILcom/duomi/superdj/logic/ay, Ligudi/com/ergushi/ap, Lframework/net/DownloadImgEventListener, /data/libr9umd, Lcom/tiqiaa/b/ah, ZCLjava/lang/String, Lbiz/edeiehe/befedeleieiei/bx


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
14
27
4
17
5
19
6
12
18
8
2
20
28
11
32
23
13
22
15
9
3
25
7
21
26
16
1
29
31
10
30
24
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
Lcom/a/a/e/q
Lnet/ask123/mima96/domain/CoursePackage
Lcom/video1/aqw1/ui/activity/h
Ljava/security/NoSuchAlgorithmException
Lcom/hifreshday/android/setting/SettingImpl
home/config/query
Lcom/fuwenpan/papers/b/d
Lkr/mytools/sound/chart/SQLListActivity
Lku/tianci/zai/u/ViewFlow
AppleWebKit/600
Lcom/clmobi/gameEngine/Constant
Lcom/JeeQgz/jPbxmq/buBuRV/MediaService
Lorg/iqiyi/video/n/s
Lorg/xutils/http/i
Lcom/qq/reader/module/bookstore/qnative/d/a/f
Lorg/spongycastle/asn1/cmp/PKIStatusInfo
BLe/a/ce
sub/hn
Lorg/qiyi/android/video/ui/a/ee
Lorg/apache/sanselan/formats/gif/GifImageParser
Lcom/olive/office/excel07/ui/entity/CellRangeAddress
Ljp/naver/common/android/notice/board/BoardNewDocumentCountTask
Lcom/fywx/paycard/util/a
7001/chunse/sexy/1/48
/pic/prepare/doctor
Lcn/huaman/photowall/view/a
Lcom/yintong/secure/f/l
lLcn/rd/sdk/sdk1/util/json/KJSONObject
ZLcom/baidu/kirin/CheckUpdateListener
Lbiz/ededeje/feheddlekefeg/cx
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

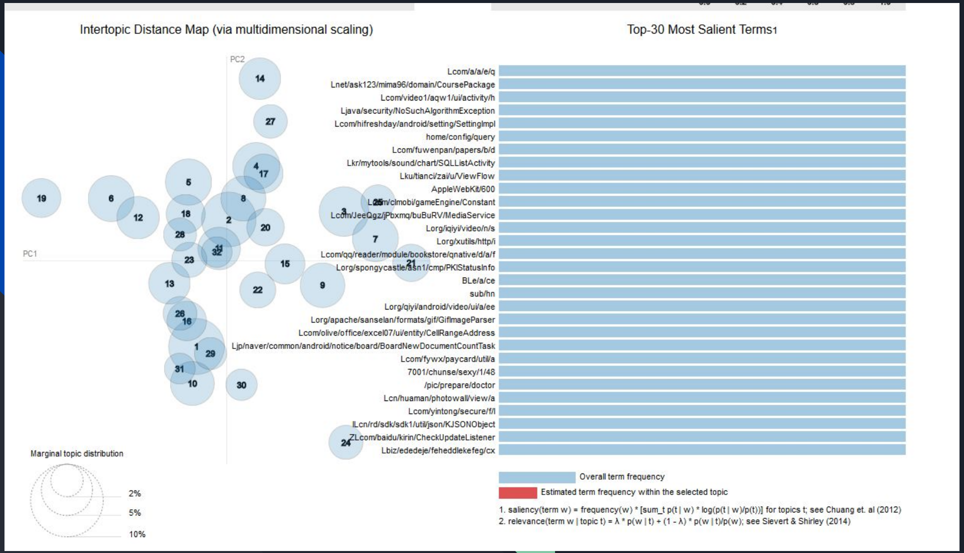
	Adwo	SmsPay	FakeInst	SMSReg	Dowgin	Plankton	AppQuanta	SmsSpy	HiddenAds	Youmi	FakeApp	Wapsx	Utchi	GingerMaster	Geinimi	Kuguo	DroidKungFu	Kmin	FakeDoc	RATC	SMSSend	SMSKey	Agent	Mseg	BaseBridge	Iop	InfoStealer	HiddenApp	GingerBreak	Dropper	DDLight	MobilePay
0	35	106	4468	73	365	55	3	157	33	14	170	45	39	131	16	14	91	30	2	16	612	49	156	8	38	339	176	23	32	132	10	31
1	18	39	8	14	48	35	5	0	12	14	10	23	8	28	1	9	4	5	4	2	11	2	18	2	3	0	3	8	4	3	4	9
2	47	44	13	16	68	11	19	3	5	14	6	37	10	16	4	20	10	4	4	2	33	1	22	5	3	0	7	5	5	2	3	9
3	8	42	9	11	62	3	0	10	7	2	9	15	11	13	7	14	3	4	3	4	9	0	16	6	36	0	5	8	8	2	3	0
4	24	112	15	31	101	55	32	3	8	5	6	22	13	24	1	16	11	2	14	6	11	6	25	2	7	0	5	4	0	1	5	7
5	19	23	5	10	65	9	4	2	18	5	2	23	17	18	2	6	6	6	16	4	11	4	19	4	3	0	2	6	1	1	5	2
6	52	39	13	19	49	6	1	2	101	10	2	17	13	9	4	7	5	16	0	0	18	4	20	2	9	0	1	20	2	3	2	3
7	9	22	5	5	45	2	0	1	3	10	2	15	8	12	2	5	3	3	4	1	4	0	12	1	4	0	0	2	0	3	2	3
8	69	33	11	29	54	15	0	3	12	10	2	23	13	9	3	23	8	11	3	4	11	0	14	4	8	1	1	9	5	3	7	1
9	32	46	13	15	82	9	3	10	5	12	6	38	19	10	4	22	6	8	0	5	20	3	30	3	2	0	2	5	3	4	5	11
10	63	21	13	17	55	35	12	4	7	7	8	25	11	7	2	8	7	0	2	3	5	4	11	5	1	0	0	6	2	3	4	0
11	104	35	11	6	49	4	2	0	6	12	0	36	9	13	2	7	12	4	16	10	14	0	10	1	7	0	2	9	2	0	2	3
12	19	25	8	13	42	6	3	1	3	9	4	27	14	8	1	13	1	1	8	2	10	2	14	3	2	0	0	1	1	1	1	3
13	19	28	11	5	41	6	9	2	2	10	1	24	9	10	0	12	6	6	3	2	13	0	14	1	16	0	2	2	1	2	4	0
14	77	35	13	34	84	15	1	2	13	12	1	25	7	14	3	22	4	16	4	5	25	3	20	3	2	1	2	5	3	0	4	3
15	15	26	9	9	37	5	1	1	2	15	0	30	14	12	1	7	6	3	5	2	8	2	18	2	17	0	1	1	1	0	2	3
16	58	92	35	38	198	13	15	6	28	25	1	46	17	27	4	42	6	6	5	4	58	3	48	2	12	1	4	28	3	3	2	3
17	40	42	5	14	43	1	0	3	18	4	1	19	9	9	2	6	6	2	2	2	14	4	18	3	1	0	0	5	0	1	0	5
18	9	22	8	12	33	12	18	0	1	5	1	14	11	7	3	6	4	2	4	4	13	4	18	2	1	0	1	0	1	4	1	6
19	34	77	17	16	82	14	1	4	1	10	2	34	12	10	1	8	8	8	9	1	14	1	16	7	5	0	12	5	2	0	3	7
20	13	39	11	15	47	8	0	6	3	7	0	22	14	17	0	17	6	5	1	0	15	3	18	3	3	0	2	1	2	1	1	2
21	120	35	13	13	60	11	60	9	3	18	12	27	16	15	3	25	19	8	2	4	13	4	16	8	47	0	0	4	6	3	2	1
22	36	127	21	21	69	28	4	6	6	12	4	26	13	16	7	13	10	5	3	1	26	2	40	7	2	1	1	10	2	1	3	4
23	91	26	9	17	48	5	0	2	0	22	4	28	8	10	4	21	1	8	2	5	12	7	21	3	5	0	8	6	0	3	7	9
24	6	111	13	21	43	1	0	7	99	3	4	17	5	7	3	10	7	7	3	0	18	1	34	1	2	1	4	15	0	3	1	1
25	29	19	7	6	54	10	0	1	1	4	1	17	11	12	2	6	5	4	5	2	6	2	17	5	2	0	4	3	2	0	6	3
26	44	60	10	19	92	37	0	3	17	8	5	24	10	16	3	18	6	3	2	2	23	6	20	1	50	0	5	12	2	3	5	7
27	309	40	10	43	54	60	3	3	5	42	6	49	12	17	6	28	34	25	3	4	7	4	18	11	6	0	2	0	1	4	2	2
28	30	17	10	28	47	7	3	8	5	4	1	18	11	13	5	11	2	4	1	3	12	1	19	4	3	0	0	6	3	5	1	2
29	19	94	8	37	40	256	0	5	0	13	6	27	7	10	5	9	4	8	3	8	11	4	18	3	5	0	1	0	5	1	2	2
30	22	29	13	67	47	79	2	4	0	6	0	23	13	8	4	10	7	4	3	2	18	1	16	4	18	0	0	3	3	0	4	2
31	22	38	7	12	42	4	1	1	2	11	1	11	13	5	1	7	5	3	4	0	9	3	22	0	21	0	0	2	1	3	1	0

	Adwo	SmsPay	FakeInst	SMSReg	Dowgin	Plankton	AppQuanta	SmsSpy	HiddenAds	Youmi	FakeApp	Wapsx	Utchi	GingerMaster	Geinimi	Kuguo	DroidKungFu	Kmin	FakeDoc	RATC	SMSSend	SMSKey	Agent	Mseg	BaseBridge	Iop	InfoStealer	HiddenApp	GingerBreak	Dropper	DDLight	MobilePay
0	2.346	6.865	92.659	10.641	16.251	6.732	1.485	58.364	7.746	3.944	61.151	5.441	9.824	24.578	15.094	3.167	29.073	13.575	1.429	14.545	56.458	37.692	20.051	6.897	11.144	98.547	69.565	10.748	31.068	67.692	9.615	21.528
1	1.206	2.526	0.166	2.041	2.137	4.284	2.475	NaN	2.817	3.944	3.597	2.781	2.015	5.253	0.943	2.036	1.278	2.262	2.857	1.818	1.015	1.538	2.314	1.724	0.880	NaN	1.186	3.738	3.883	1.538	3.846	6.250
2	3.150	2.850	0.270	2.332	3.028	1.346	9.406	1.115	1.174	3.944	2.158	4.474	2.519	3.002	3.774	4.525	3.195	1.810	2.857	1.818	3.044	0.769	2.828	4.310	0.880	NaN	2.767	2.336	4.854	1.026	2.885	6.250
3	0.536	2.720	0.187	1.603	2.760	0.367	NaN	3.717	1.643	0.563	3.237	1.814	2.771	2.439	6.604	3.167	0.958	1.810	2.143	3.636	0.830	NaN	2.057	5.172	10.557	NaN	1.976	3.738	7.767	1.026	2.885	NaN
4	1.609	7.254	0.311	4.519	4.497	6.732	15.842	1.115	1.878	1.408	2.158	2.660	3.275	4.503	0.943	3.620	3.514	0.905	10.000	5.455	1.015	4.615	3.213	1.724	2.053	NaN	1.976	1.869	NaN	0.513	4.808	4.861
5	1.273	1.490	0.104	1.458	2.894	1.102	1.980	0.743	4.225	1.408	0.719	2.781	4.282	3.377	1.887	1.357	1.917	2.715	11.429	3.636	1.015	3.077	2.442	3.448	0.880	NaN	0.791	2.804	0.971	0.513	4.808	1.389
6	3.485	2.526	0.270	2.770	2.182	0.734	0.495	0.743	23.709	2.817	0.719	2.056	3.275	1.689	3.774	1.584	1.597	7.240	NaN	NaN	1.661	3.077	2.571	1.724	2.639	NaN	0.395	9.346	1.942	1.538	1.923	2.083
7	0.603	1.425	0.104	0.729	2.004	0.245	NaN	0.372	0.704	2.817	0.719	1.814	2.015	2.251	1.887	1.131	0.958	1.357	2.857	0.909	0.369	NaN	1.542	0.862	1.173	NaN	NaN	0.935	NaN	1.538	1.923	2.083
8	4.625	2.137	0.228	4.227	2.404	1.836	NaN	1.115	2.817	2.817	0.719	2.781	3.275	1.689	2.830	5.204	2.556	4.977	2.143	3.636	1.015	NaN	1.799	3.448	2.346	0.291	0.395	4.206	4.854	1.538	6.731	0.694
9	2.145	2.979	0.270	2.187	3.651	1.102	1.485	3.717	1.174	3.380	2.158	4.595	4.786	1.876	3.774	4.977	1.917	3.620	NaN	4.545	1.845	2.308	3.856	2.586	0.587	NaN	0.791	2.336	2.913	2.051	4.808	7.639
10	4.223	1.360	0.270	2.478	2.449	4.284	5.941	1.487	1.643	1.972	2.878	3.023	2.771	1.313	1.887	1.810	2.236	NaN	1.429	2.727	0.461	3.077	1.414	4.310	0.293	NaN	NaN	2.804	1.942	1.538	3.846	NaN
11	6.971	2.267	0.228	0.875	2.182	0.490	0.990	NaN	1.408	3.380	NaN	4.353	2.267	2.439	1.887	1.584	3.834	1.810	11.429	9.091	1.292	NaN	1.285	0.862	2.053	NaN	0.791	4.206	1.942	NaN	1.923	2.083
12	1.273	1.619	0.166	1.895	1.870	0.734	1.485	0.372	0.704	2.535	1.439	3.265	3.526	1.501	0.943	2.941	0.319	0.452	5.714	1.818	0.923	1.538	1.799	2.586	0.587	NaN	NaN	0.467	0.971	0.513	0.962	2.083
13	1.273	1.813	0.228	0.729	1.825	0.734	4.455	0.743	0.469	2.817	0.360	2.902	2.267	1.876	NaN	2.715	1.917	2.715	2.143	1.818	1.199	NaN	1.799	0.862	4.692	NaN	0.791	0.935	0.971	1.026	3.846	NaN
14	5.161	2.267	0.270	4.956	3.740	1.836	0.495	0.743	3.052	3.380	0.360	3.023	1.763	2.627	2.830	4.977	1.278	7.240	2.857	4.545	2.306	2.308	2.571	2.586	0.587	0.291	0.791	2.336	2.913	NaN	3.846	2.083
15	1.005	1.684	0.187	1.312	1.647	0.612	0.495	0.372	0.469	4.225	NaN	3.628	3.526	2.251	0.943	1.584	1.917	1.357	3.571	1.818	0.738	1.538	2.314	1.724	4.985	NaN	0.395	0.467	0.971	NaN	1.923	2.083
16	3.887	5.959	0.726	5.539	8.816	1.591	7.426	2.230	6.573	7.042	0.360	5.562	4.282	5.066	3.774	9.502	1.917	2.715	3.571	3.636	5.351	2.308	6.170	1.724	3.519	0.291	1.581	13.084	2.913	1.538	1.923	2.083
17	2.681	2.720	0.104	2.041	1.915	0.122	NaN	1.115	4.225	1.127	0.360	2.297	2.267	1.689	1.887	1.357	1.917	0.905	1.429	1.818	1.292	3.077	2.314	2.586	0.293	NaN	NaN	2.336	NaN	0.513	NaN	3.472
18	0.603	1.425	0.166	1.749	1.469	1.469	8.911	NaN	0.235	1.408	0.360	1.693	2.771	1.313	2.830	1.357	1.278	0.905	2.857	3.636	1.199	3.077	2.314	1.724	0.293	NaN	0.395	NaN	0.971	2.051	0.962	4.167
19	2.279	4.987	0.353	2.332	3.651	1.714	0.495	1.487	0.235	2.817	0.719	4.111	3.023	1.876	0.943	1.810	2.556	3.620	6.429	0.909	1.292	0.769	2.057	6.034	1.466	NaN	4.743	2.336	1.942	NaN	2.885	4.861
20	0.871	2.526	0.228	2.187	2.093	0.979	NaN	2.230	0.704	1.972	NaN	2.660	3.526	3.189	NaN	3.846	1.917	2.262	0.714	NaN	1.384	2.308	2.314	2.586	0.880	NaN	0.791	0.467	1.942	0.513	0.962	1.389
21	8.043	2.267	0.270	1.895	2.671	1.346	29.703	3.346	0.704	5.070	4.317	3.265	4.030	2.814	2.830	5.656	6.070	3.620	1.429	3.636	1.199	3.077	2.057	6.897	13.783	NaN	NaN	1.869	5.825	1.538	1.923	0.694
22	2.413	8.225	0.436	3.061	3.072	3.427	1.980	2.230	1.408	3.380	1.439	3.144	3.275	3.002	6.604	2.941	3.195	2.262	2.143	0.909	2.399	1.538	5.141	6.034	0.587	0.291	0.395	4.673	1.942	0.513	2.885	2.778
23	6.099	1.684	0.187	2.478	2.137	0.612	NaN	0.743	NaN	6.197	1.439	3.386	2.015	1.876	3.774	4.751	0.319	3.620	1.429	4.545	1.107	5.385	2.699	2.586	1.466	NaN	3.162	2.804	NaN	1.538	6.731	6.250
24	0.402	7.189	0.270	3.061	1.915	0.122	NaN	2.602	23.239	0.845	1.439	2.056	1.259	1.313	2.830	2.262	2.236	3.167	2.143	NaN	1.661	0.769	4.370	0.862	0.587	0.291	1.581	7.009	NaN	1.538	0.962	0.694
25	1.944	1.231	0.145	0.875	2.404	1.224	NaN	0.372	0.235	1.127	0.360	2.056	2.771	2.251	1.887	1.357	1.597	1.810	3.571	1.818	0.554	1.538	2.185	4.310	0.587	NaN	1.581	1.402	1.942	NaN	5.769	2.083
26	2.949	3.886	0.207	2.770	4.096	4.529	NaN	1.115	3.991	2.254	1.799	2.902	2.519	3.002	2.830	4.072	1.917	1.357	1.429	1.818	2.122	4.615	2.571	0.862	14.663	NaN	1.976	5.607	1.942	1.538	4.808	4.861
27	20.710	2.591	0.207	6.268	2.404	7.344	1.485	1.115	1.174	11.831	2.158	5.925	3.023	3.189	5.660	6.335	10.863	11.312	2.143	3.636	0.646	3.077	2.314	9.483	1.760	NaN	0.791	NaN	0.971	2.051	1.923	1.389
28	2.011	1.101	0.207	4.082	2.093	0.857	1.485	2.974	1.174	1.127	0.360	2.177	2.771	2.439	4.717	2.489	0.639	1.810	0.714	2.727	1.107	0.769	2.442	3.448	0.880	NaN	NaN	2.804	2.913	2.564	0.962	1.389
29	1.273	6.088	0.166	5.394	1.781	31.334	NaN	1.859	NaN	3.662	2.158	3.265	1.763	1.876	4.717	2.036	1.278	3.620	2.143	7.273	1.015	3.077	2.314	2.586	1.466	NaN	0.395	NaN	4.854	0.513	1.923	1.389
30	1.475	1.878	0.270	9.767	2.093	9.670	0.990	1.487	NaN	1.690	NaN	2.781	3.275	1.501	3.774	2.262	2.236	1.810	2.143	1.818	1.661	0.769	2.057	3.448	5.279	NaN	NaN	1.402	2.913	NaN	3.846	1.389
31	1.475	2.461	0.145	1.749	1.870	0.490	0.495	0.372	0.469	3.099	0.360	1.330	3.275	0.938	0.943	1.584	1.597	1.357	2.857	NaN	0.830	2.308	2.828	NaN	6.158	NaN	NaN	0.935	0.971	1.538	0.962	NaN

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	GingerMaster		SmsSpy		FakeDoc	HiddenAds		Iop	Utchi		RATC					FakeInst			InfoStealer		AppQuanta	SmsPay	SMSKey			BaseBridge	Adwo	Dropper	Plankton	SMSReg
			Geinimi					DDLight	MobilePay							Dowgin					FakeApp						Youmi			
			GingerBreak													Kuguo											Wapsx			
																SMSSend											DroidKungFu			
																Agent											Kmin			
																HiddenApp											Mseg			

Dominant_Topic	Topic_Perc_Contrib
23.0	0.9656
3.0	0.9590
0.0	0.0312
27.0	0.9530
15.0	0.9580
25.0	0.9169
0.0	0.0312
0.0	0.0312
22.0	0.9667
27.0	0.9209
20.0	0.7875
0.0	0.0312
13.0	0.9402
23.0	0.9054
0.0	0.0312
2.0	0.9135
0.0	0.0312
0.0	0.0312
15.0	0.9705
0.0	0.0312
13.0	0.9560
3.0	0.9641
0.0	0.0312
27.0	0.9559
27.0	0.9606
26.0	0.9511
2.0	0.9431
21.0	0.9451

	Precision	Recall	F1_Score	Accuracy	Specificity	Misclassification
Adwo	0.9871	0.7702	0.8652	0.9363	0.9964	0.0637
SmsPay	0.8898	0.2534	0.3944	0.7753	0.9872	0.2247
FakeInst	0.8000	0.0061	0.0121	0.0525	0.9698	0.9475
SMSReg	0.9403	0.3580	0.5185	0.8294	0.9922	0.1706
Dowgin	0.9646	0.3032	0.4614	0.8014	0.9957	0.1986
Plankton	0.9844	0.8182	0.8936	0.9266	0.9921	0.0734
AppQuanta	1.0000	0.9524	0.9756	0.9851	1.0000	0.0149
SmsSpy	1.0000	0.0602	0.1136	0.4201	1.0000	0.5799
HiddenAds	0.9802	0.6644	0.7920	0.8779	0.9928	0.1221
Youmi	0.9048	0.4270	0.5802	0.8451	0.9850	0.1549
FakeApp	0.9167	0.0576	0.1084	0.3489	0.9885	0.6511
Wapsx	0.9184	0.3261	0.4813	0.8827	0.9942	0.1173
Utchi	0.9474	0.2769	0.4286	0.8791	0.9970	0.1209
GingerMaster	1.0000	0.1407	0.2467	0.6792	1.0000	0.3208
Geinimi	0.8571	0.2222	0.3529	0.7925	0.9873	0.2075
Kuguo	0.9524	0.5970	0.7339	0.9344	0.9947	0.0656
DroidKungFu	1.0000	0.2297	0.3736	0.6358	1.0000	0.3642
Kmin	0.9600	0.4800	0.6400	0.8778	0.9942	0.1222
FakeDoc	1.0000	0.6667	0.8000	0.9429	1.0000	0.0571
RATC	1.0000	0.3571	0.5263	0.8364	1.0000	0.1636
SMSSend	0.9483	0.0788	0.1455	0.4041	0.9922	0.5959
SMSKey	0.8571	0.0896	0.1622	0.5231	0.9841	0.4769
Agent	0.8333	0.1342	0.2312	0.6581	0.9833	0.3419
Mseg	1.0000	0.6111	0.7586	0.9397	1.0000	0.0603
BaseBridge	0.9800	0.5213	0.6806	0.8651	0.9960	0.1349
Iop	1.0000	0.0029	0.0059	0.0145	1.0000	0.9855
InfoStealer	1.0000	0.0606	0.1143	0.2648	1.0000	0.7352
HiddenApp	1.0000	0.4912	0.6588	0.8645	1.0000	0.1355
GingerBreak	0.8750	0.1556	0.2642	0.6214	0.9828	0.3786
Dropper	0.6000	0.0191	0.0370	0.2000	0.9474	0.8000
DDLight	1.0000	0.4118	0.5833	0.9038	1.0000	0.0962
MobilePay	0.6364	0.1094	0.1867	0.5764	0.9500	0.4236



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	GingerMaster		SmsSpy		FakeDoc	HiddenAds		Iop	Utchi		RATC					FakeInst			InfoStealer		AppQuanta	SmsPay	SMSKey			BaseBridge	Adwo	Dropper	Plankton	SMSReg
			Geinimi					DDLight	MobilePay							Dowgin					FakeApp						Youmi			
			GingerBreak													Kuguo											Wapsx			
																SMSSend											DroidKungFu			
																Agent											Kmin			
																HiddenApp											Mseg			

1,29,16,26		3,21		23,28		6,19		8,11		9,22				23		5		27		30	
GingerMaster	67%	SmsSpy	42%	SMSKey	52%	HiddenAds	87%	Iop	1.00%	Utchi	87%			SMSKey	52%	FakeDoc	94%	Adwo	94%	SMSReg	83%
Plankton	93%	Geinimi	79%	Dropper	20%	InfoStealer	26%	DDLight	90%	MobilePay	57%							Youmi	84%		
FakeInst	5%	GingerBreak	62%					RATC	84%	SmsPay	77%							Wapsx	88%		
Dowgin	80%	AppQuanta	98%															DroidKungFu	64%		
Kuguo	93%	FakeApp	35%															Kmin	87%		
SMSSend	40%																	Mseg	93%		
Agent	66%																				
HiddenApp	86%																				
BaseBridge	87%																				
Plankton		AppQuanta				HiddenAds		DDLight		Utchi						FakeDoc		Adwo		SMSReg	
Dowgin		Geinimi						RATC		SmsPay								Youmi			
Kuguo		GingerBreak																Wapsx			
HiddenApp																		Kmin			
BaseBridge																		Mseg			

Limitations/Future Work

In the end i do believe that my models has shown for some of the malware success in clustering different types into families but overall i did feel like there was some limitation,

- Thing I could improve is using a much larger data set where types of malware are equally distributed.

- Using different types of N-Grams to try to form a more holistic model.

- Using different aspects of the Dex file such as the assembly or meta data.

- For the future I can take this model as a precursor for a larger scale framework which is designed to identify malware