Prajit Kumar Das

Advisor: Anupam Joshi, PhD

Co-Advisor: Tim Finin, PhD

App behavioral analysis using system calls

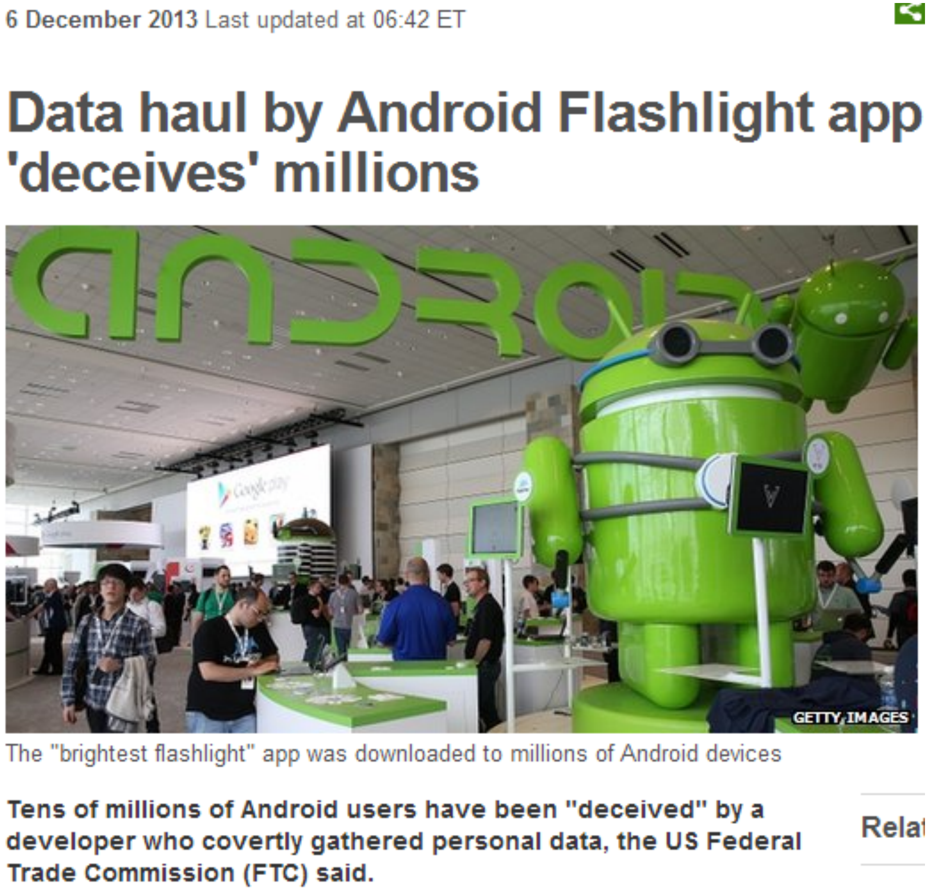
UMBC

AN HONORS UNIVERSITY IN MARYLAND

Problem statement

We present Heimdall, a framework for studying system calls made by mobile apps in order to determine an app’s behavior class and match such behavior to their stated purpose.

Motivation: App issues



6 December 2013 Last updated at 06:42 ET

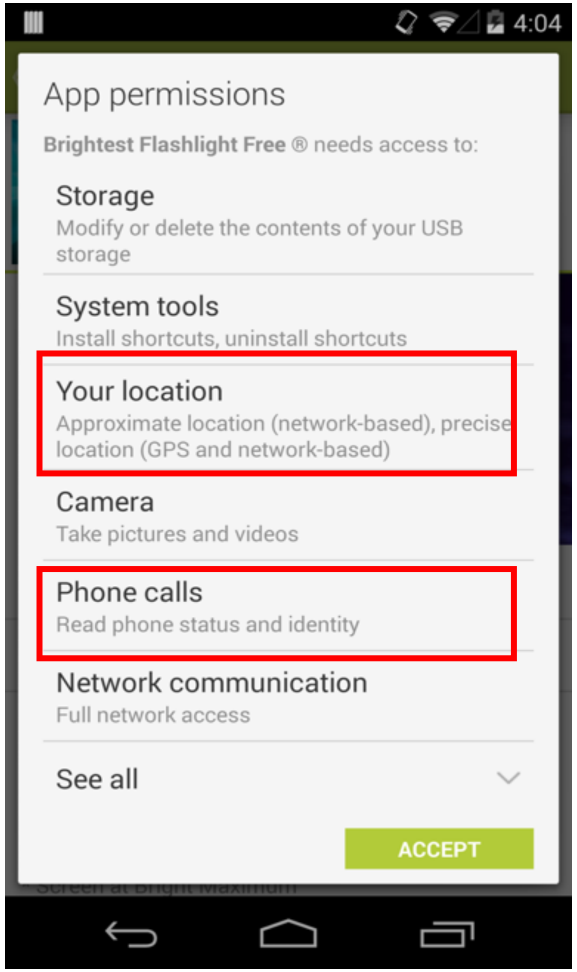
Data haul by Android Flashlight app 'deceives' millions

GETTY IMAGES

The "brightest flashlight" app was downloaded to millions of Android devices

Tens of millions of Android users have been "deceived" by a developer who covertly gathered personal data, the US Federal Trade Commission (FTC) said.

Relat



App permissions

Brightest Flashlight Free ® needs access to:

Storage
Modify or delete the contents of your USB storage

System tools
Install shortcuts, uninstall shortcuts

Your location
Approximate location (network-based), precise location (GPS and network-based)

Camera
Take pictures and videos

Phone calls
Read phone status and identity

Network communication
Full network access

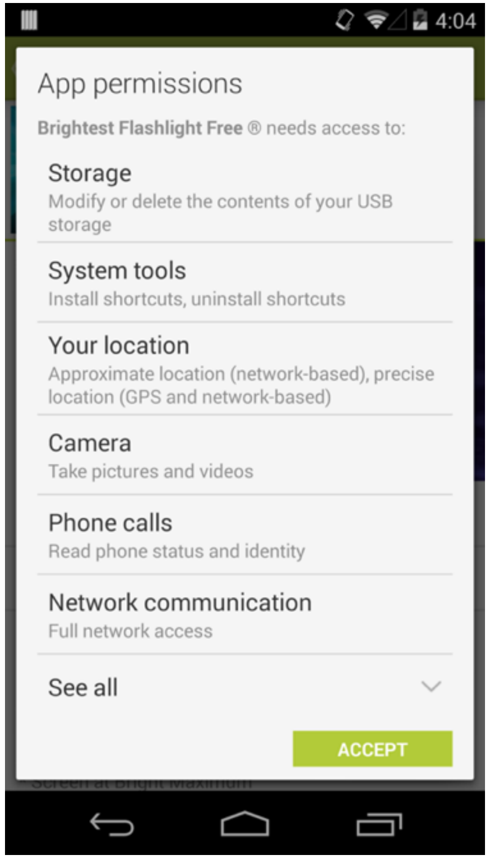
See all

ACCEPT

Image courtesy: Android App Market

Motivation: Permission inadequacy

Pre-Marshmallow



App permissions

Brightest Flashlight Free ® needs access to:

Storage
Modify or delete the contents of your USB storage

System tools
Install shortcuts, uninstall shortcuts

Your location
Approximate location (network-based), precise location (GPS and network-based)

Camera
Take pictures and videos

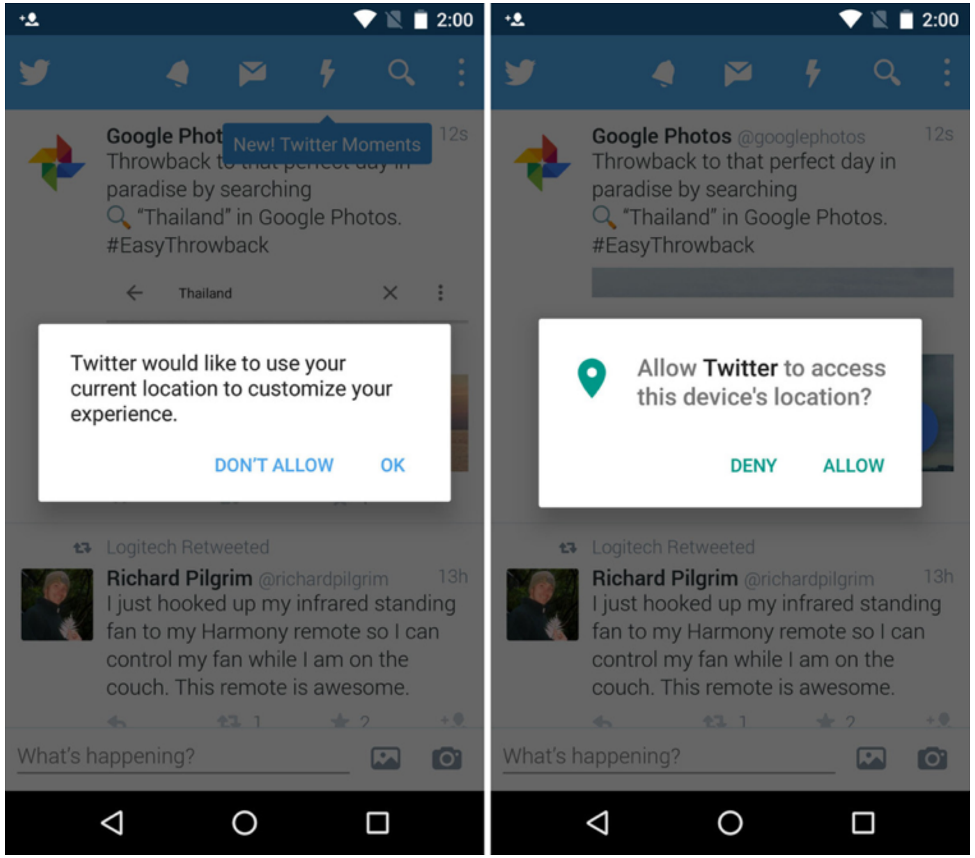
Phone calls
Read phone status and identity

Network communication
Full network access

See all

ACCEPT

Marshmallow



Twitter would like to use your current location to customize your experience.

DON'T ALLOW OK

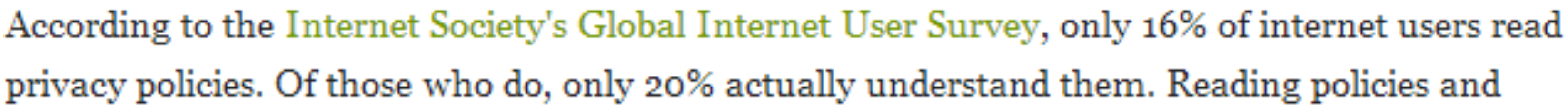
Allow **Twitter** to access this device's location?

DENY ALLOW

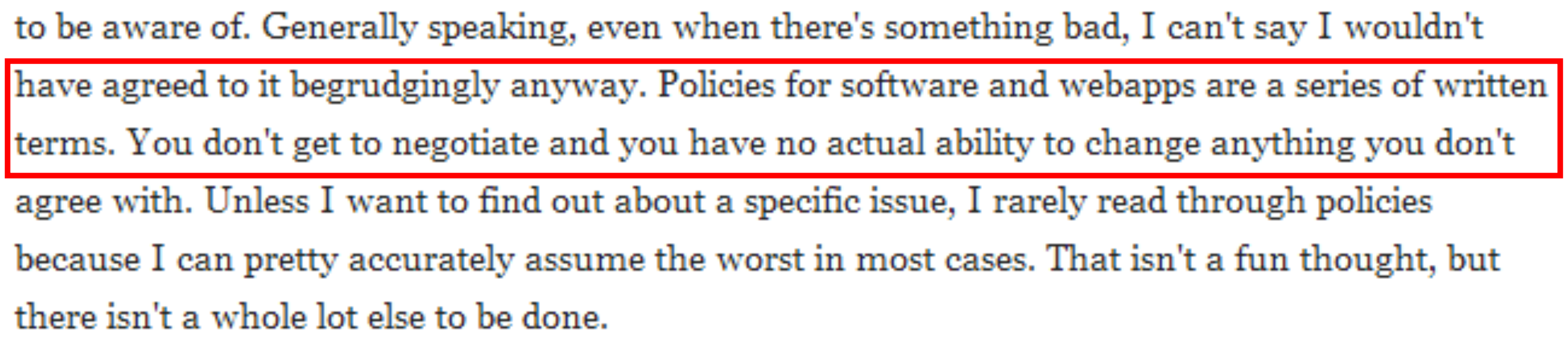
Google Photos @googlephotos 12s
New! Twitter Moments
Throwback to that perfect day in paradise by searching "Thailand" in Google Photos. #EasyThrowback
Thailand
Logitech Retweeted
Richard Pilgrim @richardpilgrim 13h
I just hooked up my infrared standing fan to my Harmony remote so I can control my fan while I am on the couch. This remote is awesome.
What's happening?

Motivation: Software Limitation

Do you read privacy policies and do you understand them?



According to the Internet Society's Global Internet User Survey, only 16% of internet users read privacy policies. Of those who do, only 20% actually understand them. Reading policies and



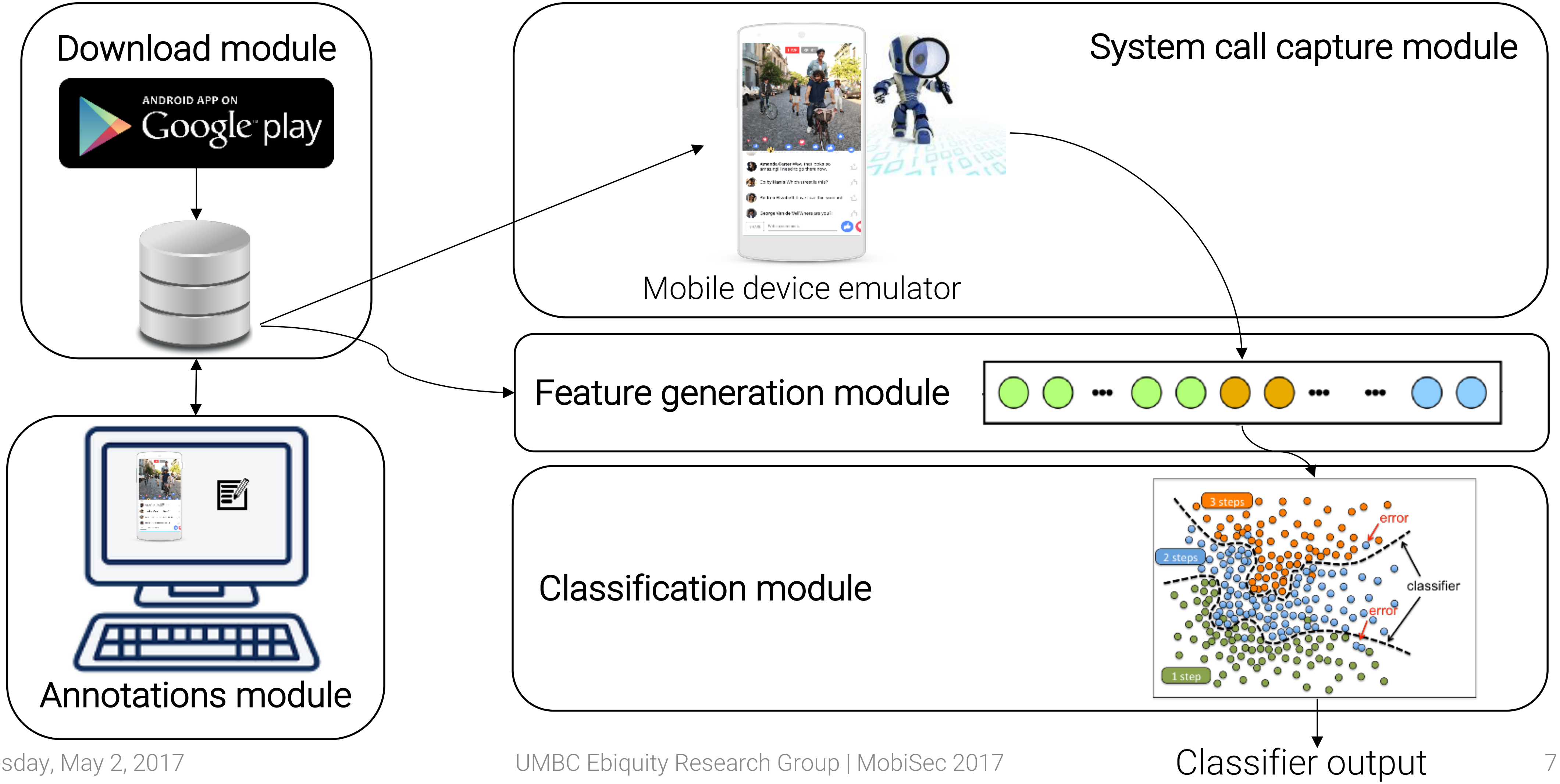
to be aware of. Generally speaking, even when there's something bad, I can't say I wouldn't have agreed to it begrudgingly anyway. Policies for software and webapps are a series of written terms. You don't get to negotiate and you have no actual ability to change anything you don't agree with. Unless I want to find out about a specific issue, I rarely read through policies because I can pretty accurately assume the worst in most cases. That isn't a fun thought, but there isn't a whole lot else to be done.

Source: lifehacker

Related Work

- System calls used for software analysis *Kosoresow'97*
- Three areas of research for mobile app analysis
  - Malware classification *Zhou'12*
  - NLP techniques *Pandtia'13, Gorla'14*
  - Taint tracking *Enck'10*
- Google PHA taxonomy *Google Android Security'16*
- PrivacyGrade: Grading The Privacy Of Smartphone Apps
  - Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. Lin'12
  - Modeling users' mobile app privacy preferences, Lin'14

Architecture



Download module

Annotations module

System call capture module

Mobile device emulator

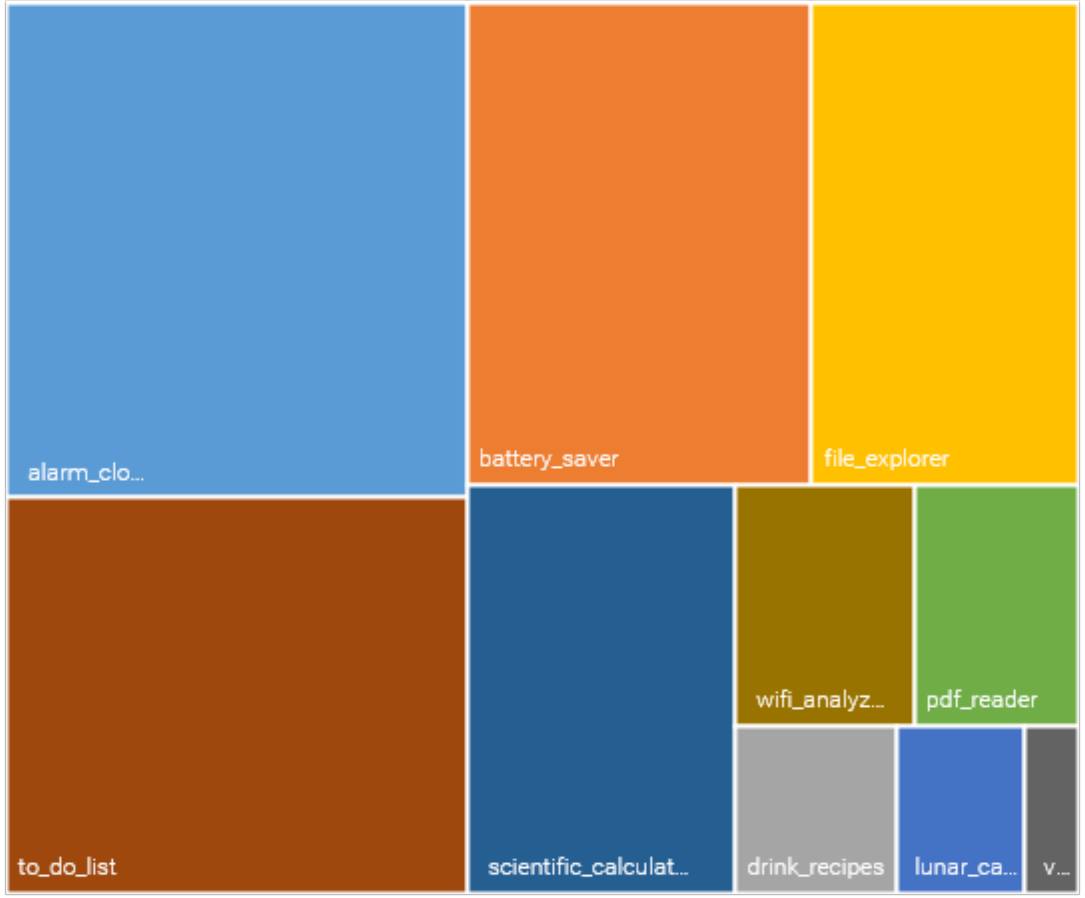
Feature generation module

Classification module

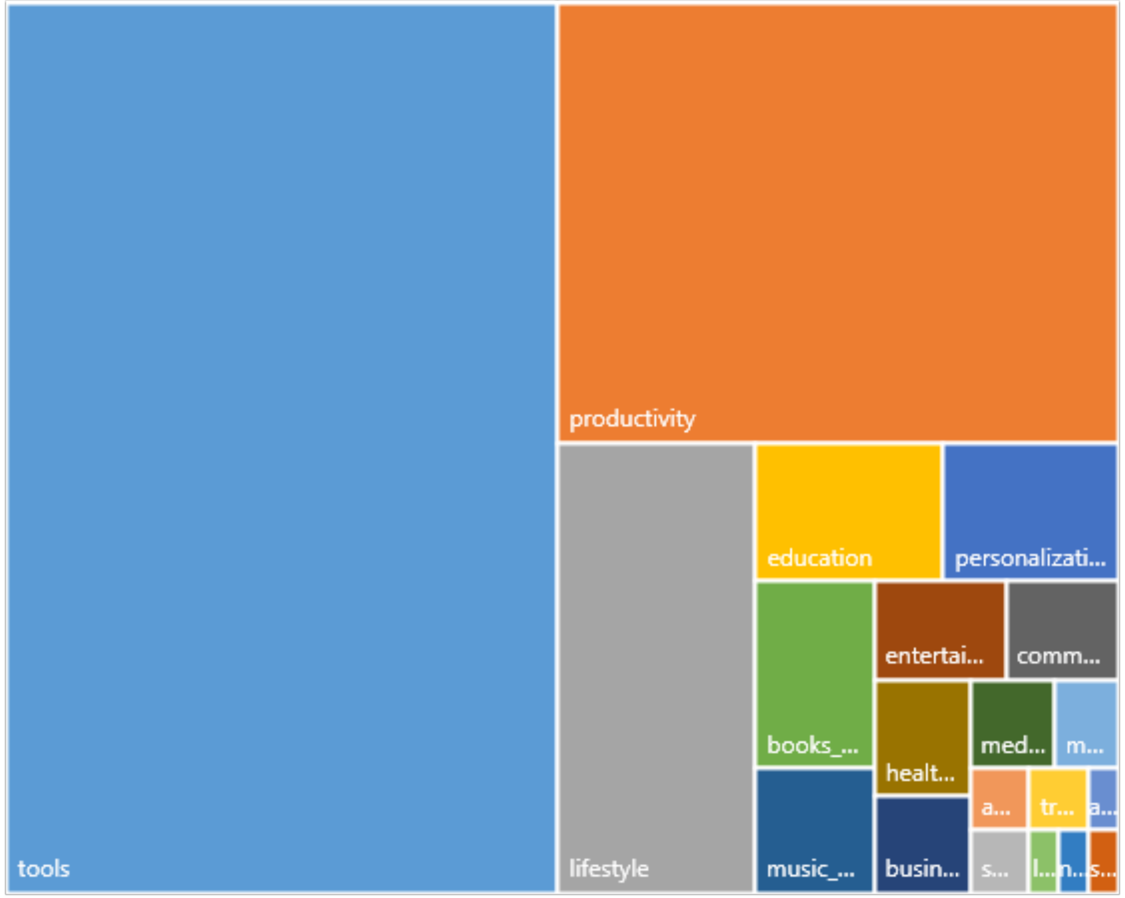
Classifier output

Dataset distribution

10 annotated categories, 20 Google Play categories – 75% tool and productivity



Annotated class distribution



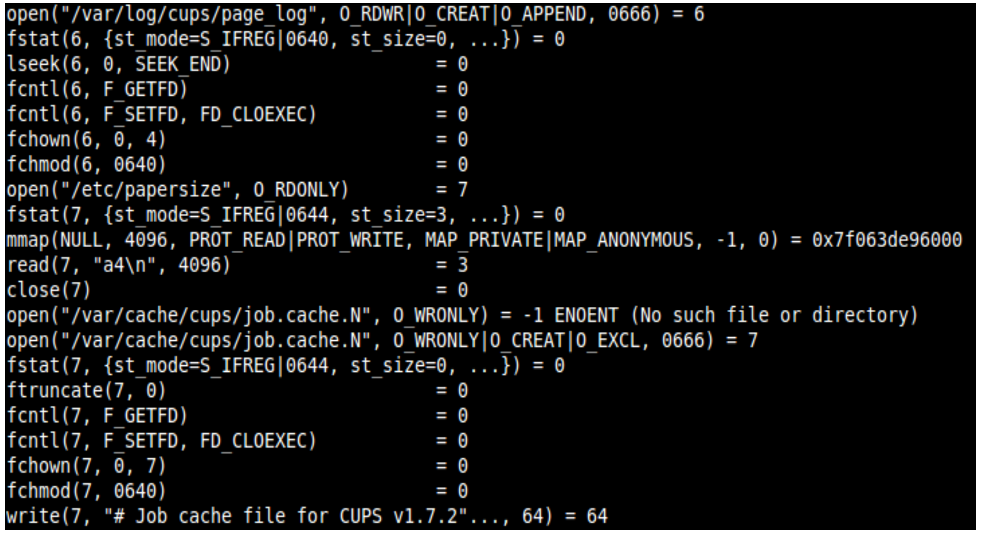
Google categories distribution

Experimental setup

- 1560 apps
- 534 successfully executed

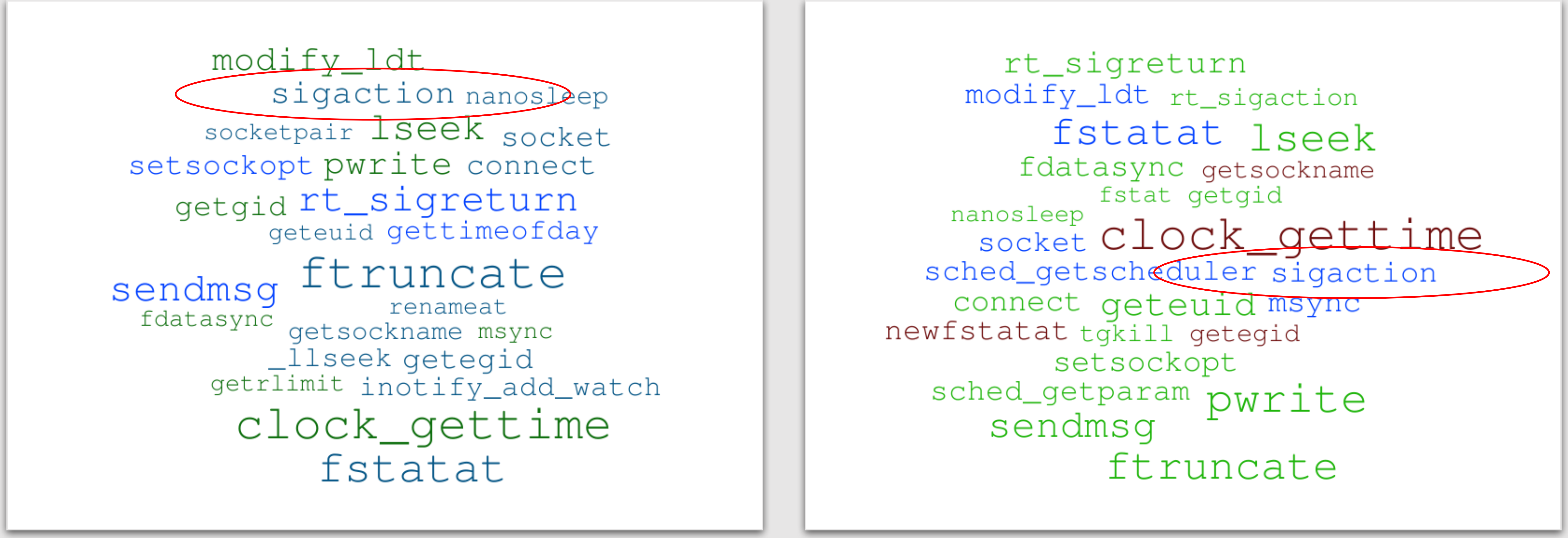
- Android 6.0.1 December 2015 build used
- SVM, MLP, J48 and NB used
- Best F1-score from MLP at 0.44

- **`strace`** can only be used on emulator
- UI/Application exerciser tool used

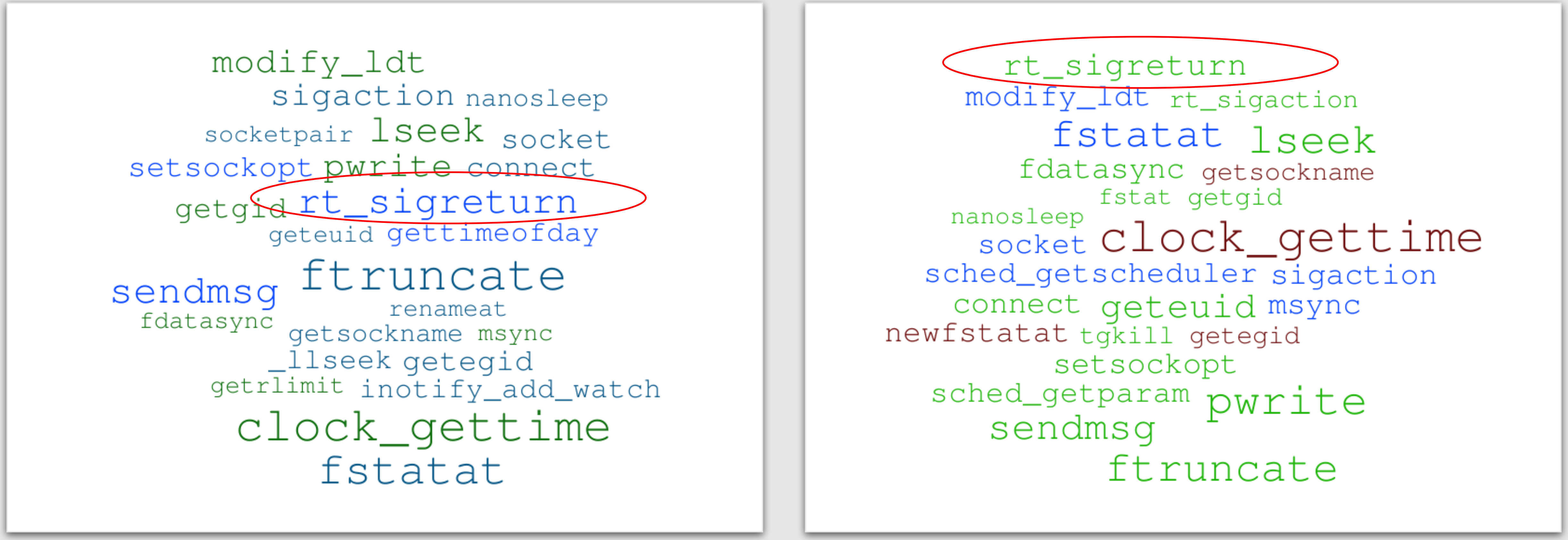


```
open("/var/log/cups/page_log", O_RDWR|O_CREAT|O_APPEND, 0666) = 6
fstat(6, {st_mode=S_IFREG|0640, st_size=0, ...}) = 0
lseek(6, 0, SEEK_END)                   = 0
fcntl(6, F_GETFD)                       = 0
fcntl(6, F_SETFD, FD_CLOEXEC)           = 0
fchown(6, 0, 4)                         = 0
fchmod(6, 0640)                         = 0
open("/etc/papersize", O_RDONLY)        = 7
fstat(7, {st_mode=S_IFREG|0644, st_size=3, ...}) = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x7f063de96000
read(7, "a4\n", 4096)                   = 3
close(7)                                = 0
open("/var/cache/cups/job.cache.N", O_WRONLY) = -1 ENOENT (No such file or directory)
open("/var/cache/cups/job.cache.N", O_WRONLY|O_CREAT|O_EXCL, 0666) = 7
fstat(7, {st_mode=S_IFREG|0644, st_size=0, ...}) = 0
ftruncate(7, 0)                         = 0
fcntl(7, F_GETFD)                       = 0
fcntl(7, F_SETFD, FD_CLOEXEC)           = 0
fchown(7, 0, 7)                         = 0
fchmod(7, 0640)                         = 0
write(7, "# Job cache file for CUPS v1.7.2"..., 64) = 64
```

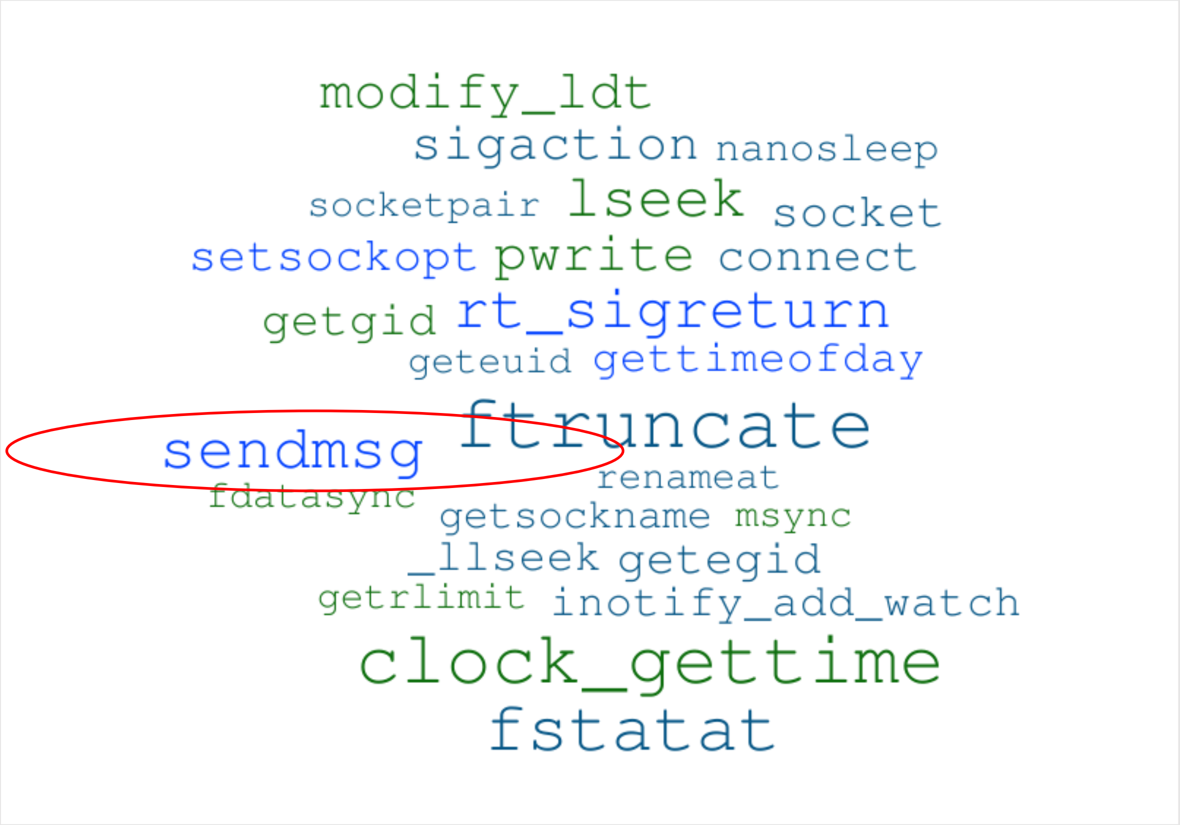
- 1-hot vectors – Call present or absent
- TF-IDF weight vectors – Uniqueness and significance of system calls for app

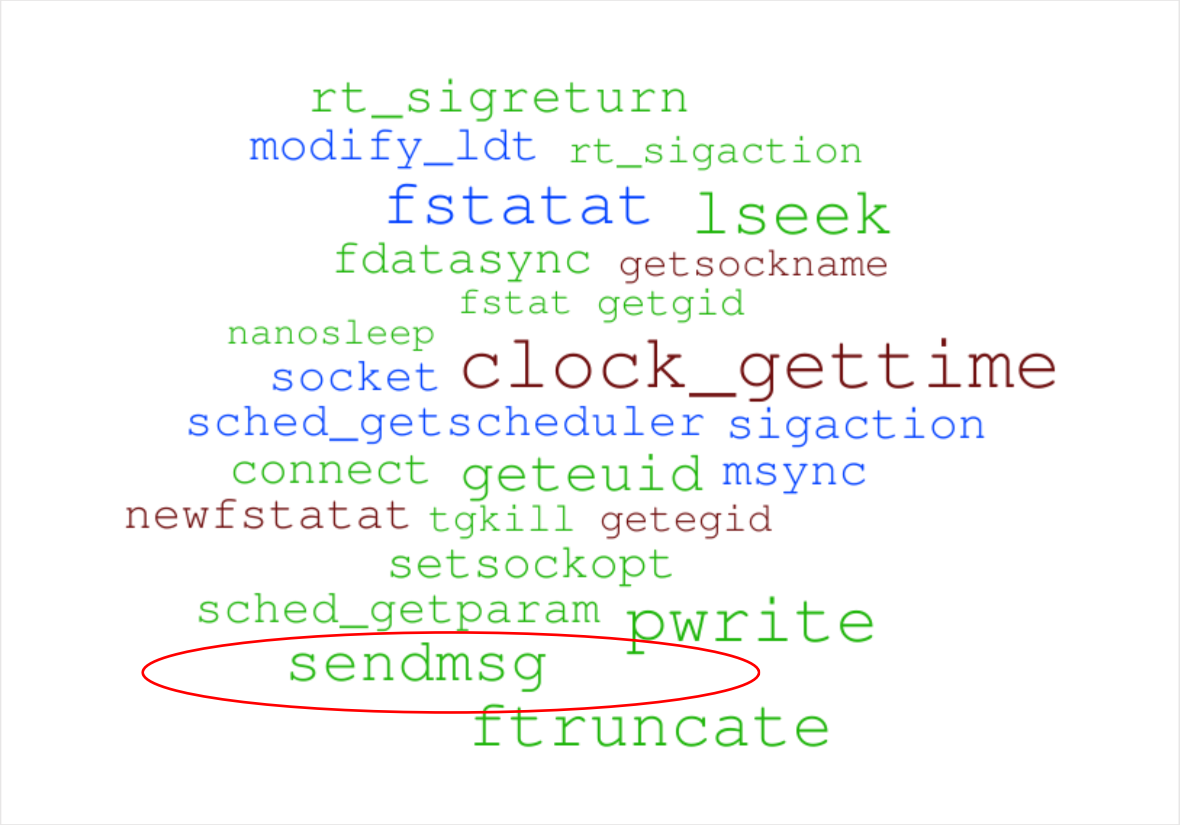


Similar TF-IDF scores - Scientific calculator vs To-do list

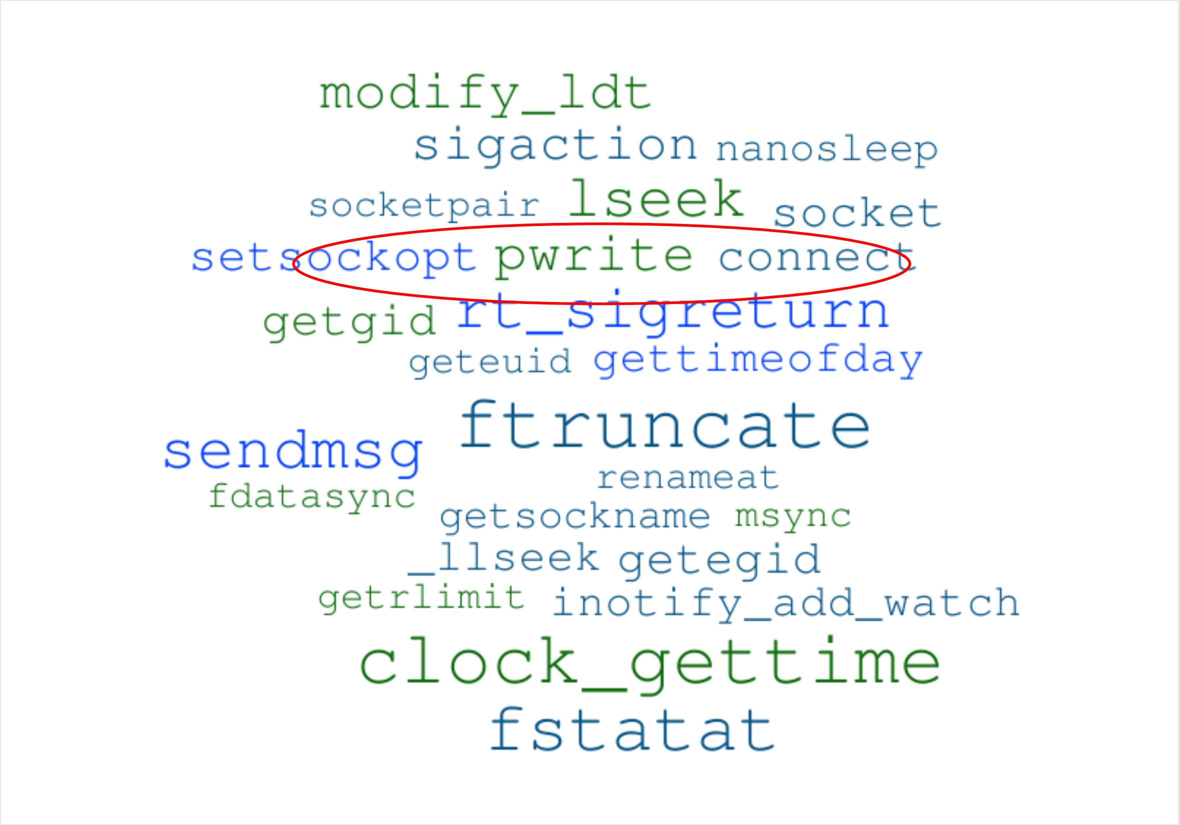


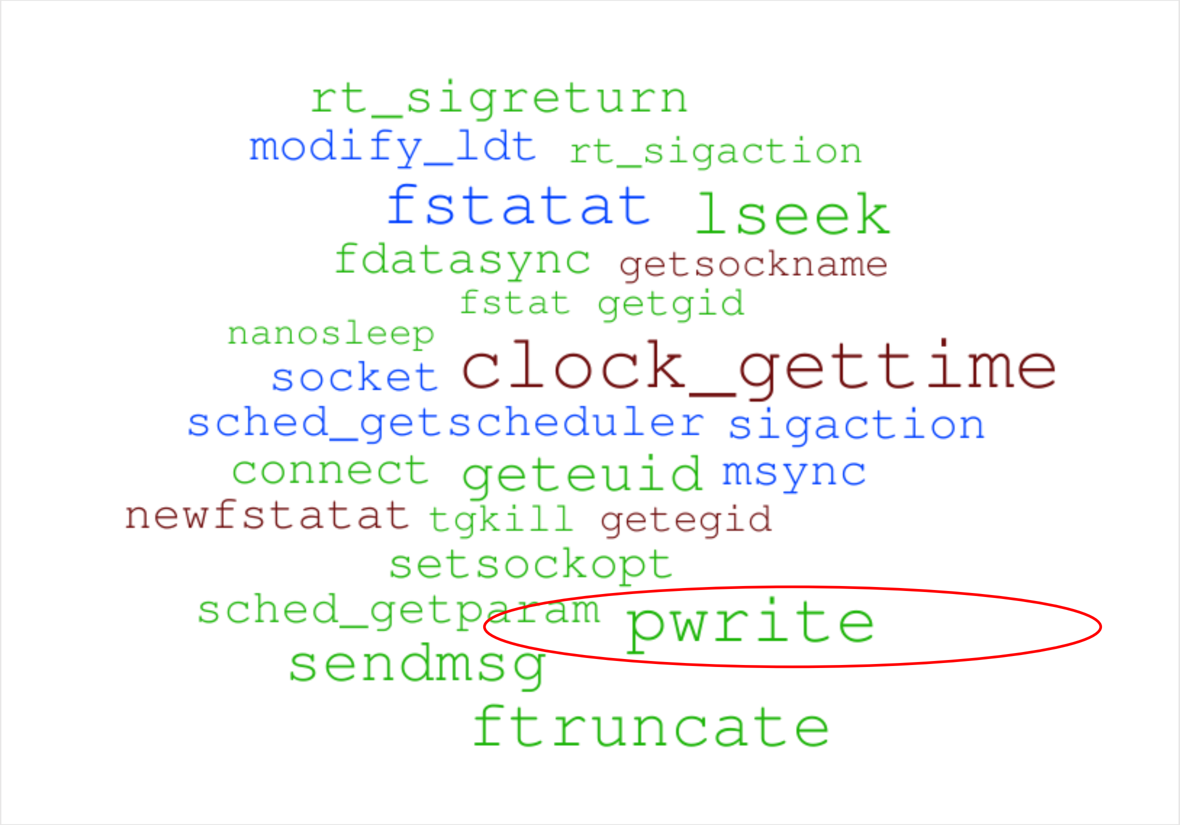
Similar TF-IDF scores - Scientific calculator vs To-do list



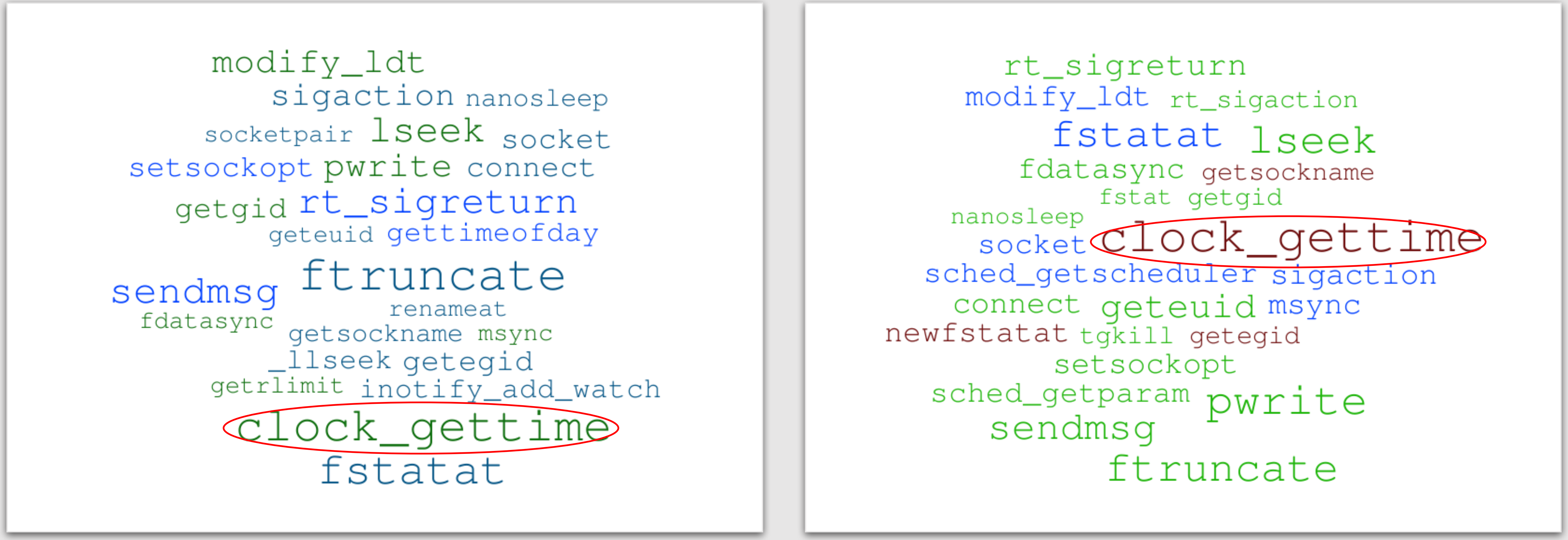


Similar TF-IDF scores - Scientific calculator vs To-do list

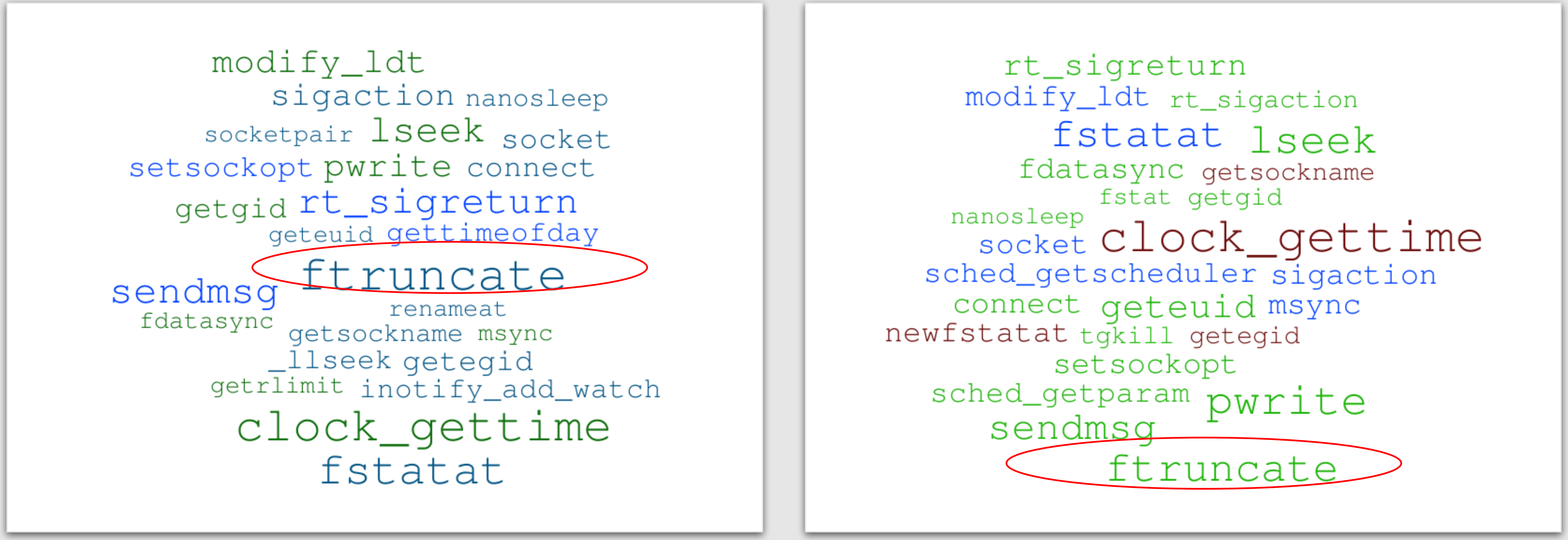




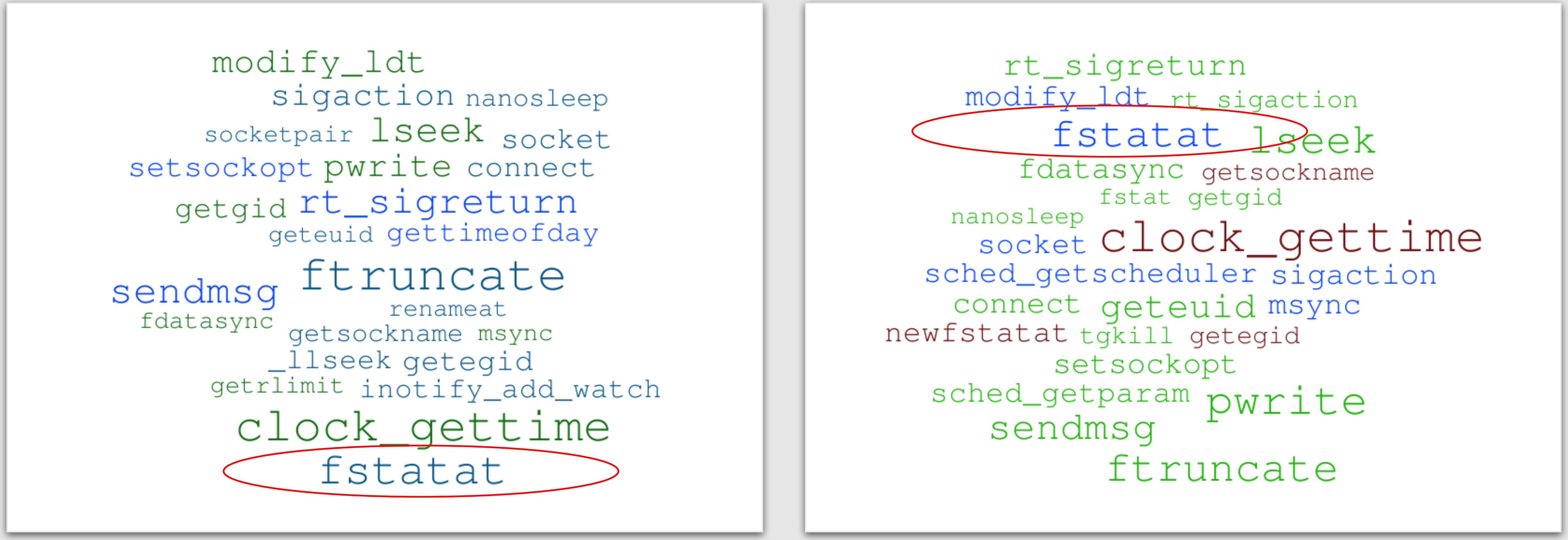
Similar TF-IDF scores - Scientific calculator vs To-do list



Similar TF-IDF scores - Scientific calculator vs To-do list



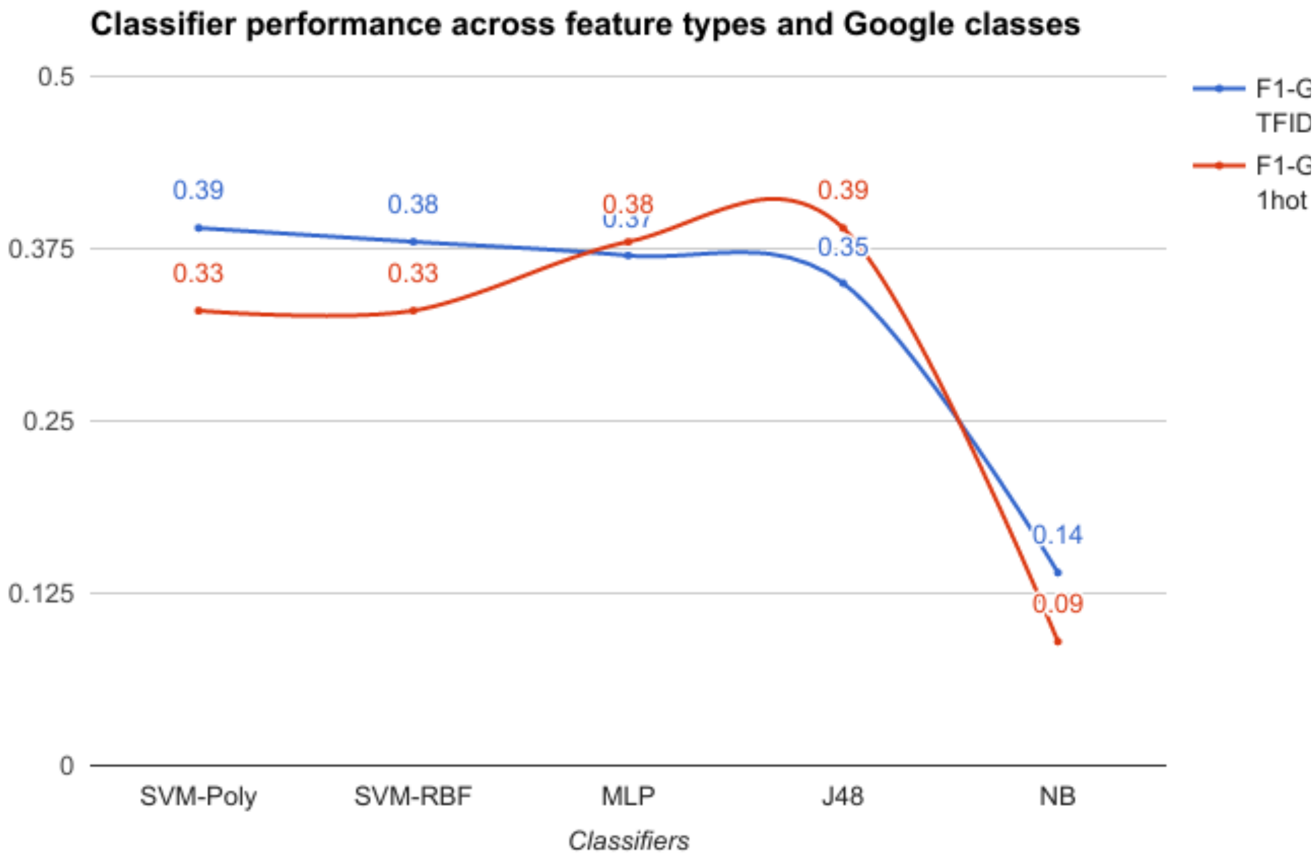
Similar TF-IDF scores - Scientific calculator vs To-do list



Similar TF-IDF scores - Scientific calculator vs To-do list

Results: Google categories

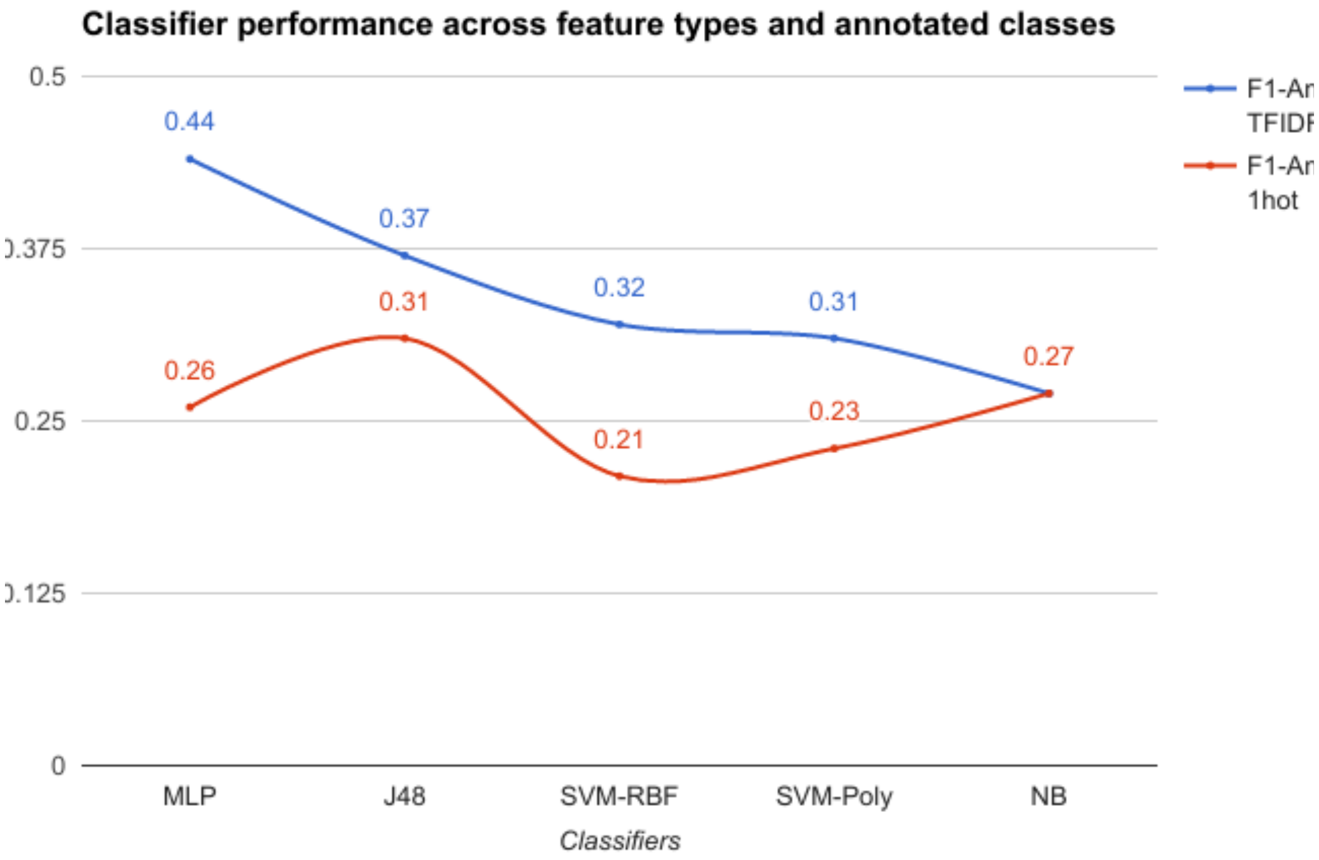
- Google's categories are developer provided
- Could be misleading
- System calls could classify behavior better
- Additional features required

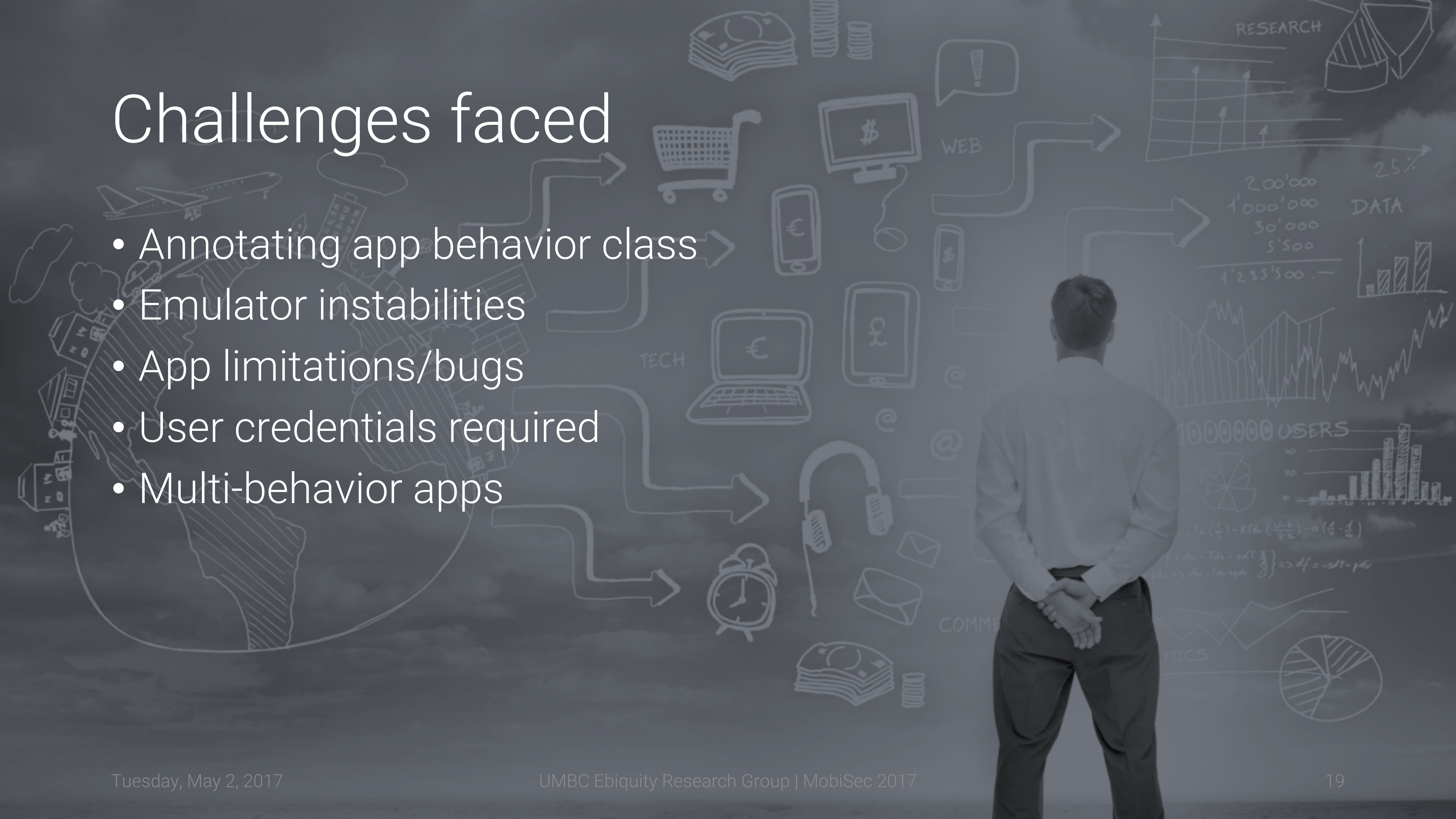


Results: Annotated classes

- System calls not *totally* useless
- 1-hot vs TF-IDF

Classifier	TF-IDF	1-hot
MLP	0.44	0.26
SVM-Poly	0.31	0.23
SVM-RBF	0.32	0.21
J48	0.27	0.31
NB	0.27	0.27





Challenges faced

- Annotating app behavior class
- Emulator instabilities
- App limitations/bugs
- User credentials required
- Multi-behavior apps

Conclusions

Can system calls be used to distinguish between how an app "behaves" and it's perceived/stated purpose?

- System calls – insufficient as features
- Emulator – better ones required
- Coarser behavior classes required

Future work

- System call bi-grams, tri-grams – for capturing call sequences
- Better emulator – Genymotion
- Malware classification – less scope (state-of-the-art is 96.7%)
- Generating contextual policies from behavior estimation
- Features planned for future:
    - App descriptions
    - App ratings
    - PHA app analysis – require samples for Mobile Unwanted Software

Questions?