Android challenge - data minining

Computational linguistics

Ondrej Platek

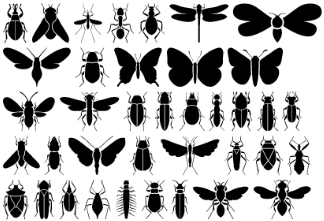
UNIBZ

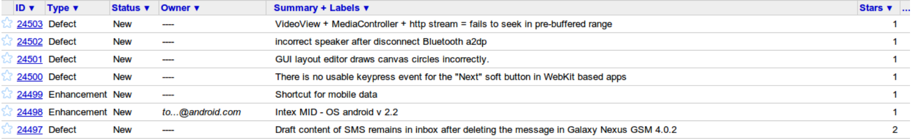
January 27, 2012

Content

1. Key features
2. Tools
3. Data statistics
4. Results
5. Summary

Key features





ID ▾	Type ▾	Status ▾	Owner ▾	Summary + Labels ▾	Stars ▾	...
☆ 24503	Defect	New	----	VideoView + MediaController + http stream = fails to seek in pre-buffered range	1	
☆ 24502	Defect	New	----	incorrect speaker after disconnect Bluetooth a2dp	1	
☆ 24501	Defect	New	----	GUI layout editor draws canvas circles incorrectly.	1	
☆ 24500	Defect	New	----	There is no usable keypress event for the "Next" soft button in WebKit based apps	1	
☆ 24499	Enhancement	New	----	Shortcut for mobile data	1	
☆ 24498	Enhancement	New	*to...@android.com*	Intex MID - OS android v 2.2	1	
☆ 24497	Defect	New	----	Draft content of SMS remains in inbox after deleting the message in Galaxy Nexus GSM 4.0.2	2	

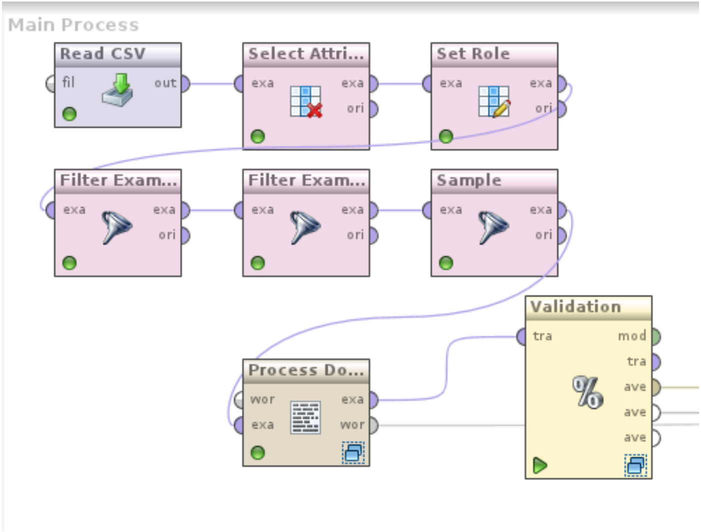
Bugs reports on Google Code

Interesting properties manually inserted

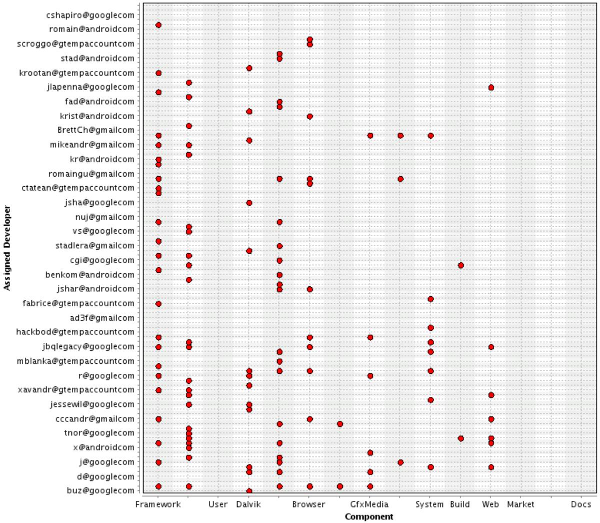
Some properties added automatically

Tools used

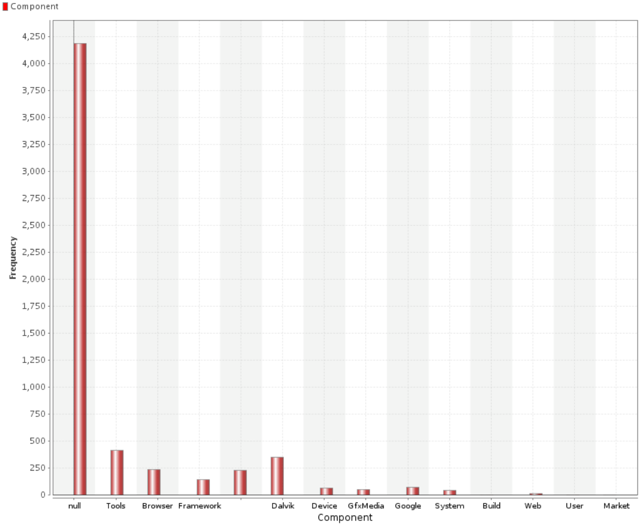
Process


Main Process
Read CSV
fil
out
Select Attri...
exa
exa
ori
Set Role
exa
exa
ori
Filter Exam...
exa
exa
ori
Filter Exam...
exa
exa
ori
Sample
exa
exa
ori
Validation
tra
mod
tra
ave
ave
ave
Process Do...
wor
exa
exa
wor

Developers by component



Histogram of components



Classification results

Type	Algorithm	Sample Size	Time	Accuracy
Decision Trees	Weka J48	500	2:19	44.80%
		1000	10:25	45.50%
Decision Trees	Weka LAD-Tree	500	13:40	39.80%
Lazy Classifiers	15-NN	500	0:11	45.40%
		1000	0:38	49.70%
		2000	3:13	55.90%
Lazy Classifiers	30-NN	500	0:21	45.10%
		1000	0:56	49.00%
		2000	3:31	54.40%
SVM	Weka SMO	500	0:53	41.40%
		1000	4:15	47.20%

Classification results 2

Type	Algorithm	Size	Accuracy 1	Accuracy 2
Decision Trees	Weka J48	500	44.80%	46.5%
		1000	45.50%	47.8%
Decision Trees	Weka LAD-Tree	500	39.80%	41.6%
Lazy Classifiers	15-NN	500	45.40%	43.4%
		1000	49.70%	50.7%
		2000	55.90%	54.8%
Lazy Classifiers	30-NN	500	45.10%	46.2%
		1000	49.00%	49.9%
		2000	54.40%	55.3%
SVM	Weka SMO	500	41.40%	42.5%
		1000	47.20%	48.2%

Summary

- Technologies: (Weka), RapidMiner
- Classification on real data
- Improved tokenizer
- Compared algorithms and document importance (TF-IDF vs Binary)

Questions?