Comparing Apples and Bananas

Gerrit Gruben

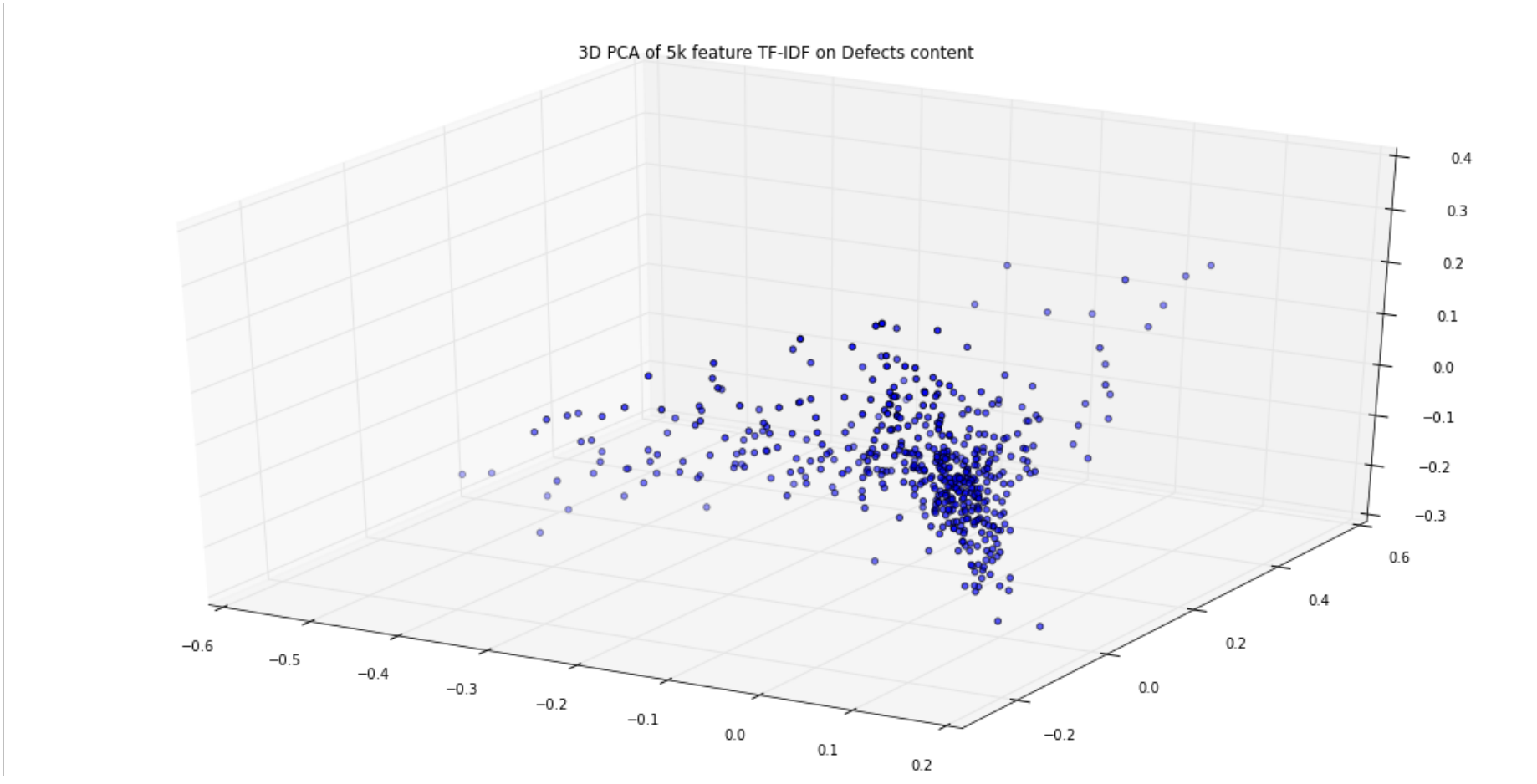
November 8, 2015

Bug reports (Apples) coming in.

Bug reports (Apples) coming in.
Similar bug reports? What about tech articles (Bananas)?

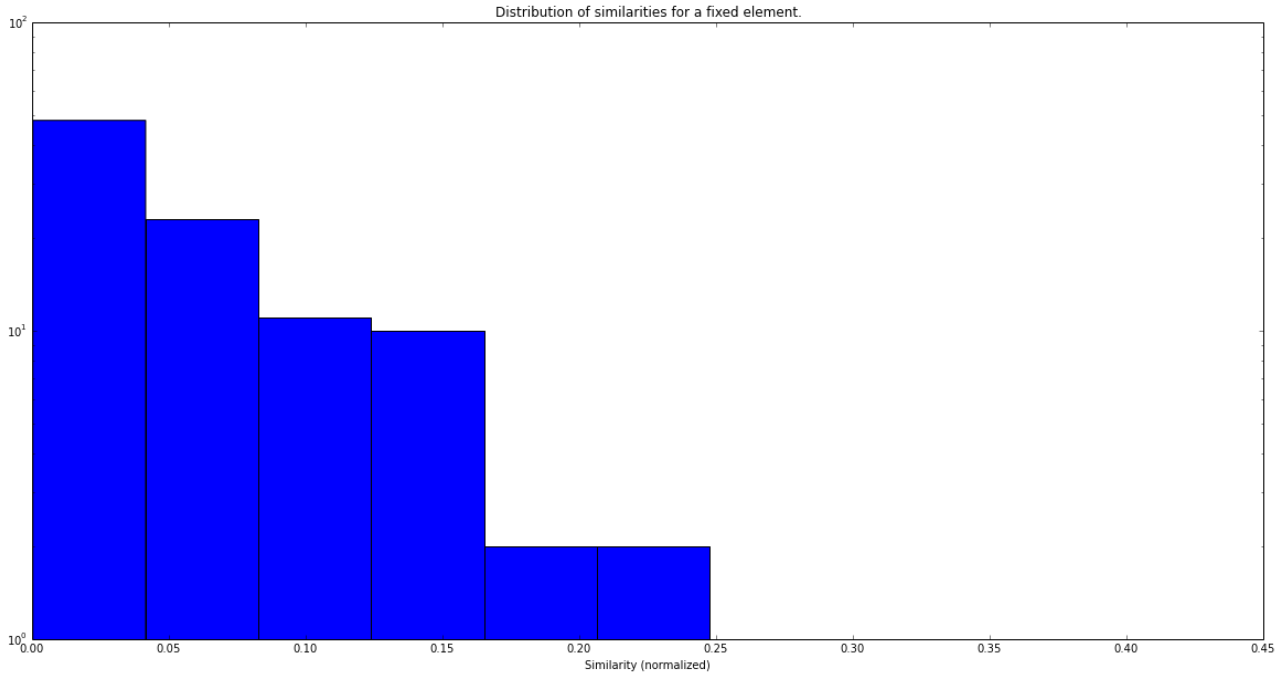
Bug reports (Apples) coming in.
Similar bug reports? What about tech articles (Bananas)?
Need it ASAP!

Defect


3D PCA of 5k feature TF-IDF on Defects content
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.6
0.4
0.2
0.0
−0.2
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2

Both have same structure: content $\rightarrow$ string similarity kernel (Lohdi, 2002)

Both have same structure: content $\rightarrow$ string similarity kernel (Lohdi, 2002)



- Model can incorporate more information (e.g. if both are apples, prior knowledge)
- Possible to use in an online setting (cf. online SVM)
- Free algorithms.

- Slow to compute all pairwise similarities (be clever i.e. clustering)
- Kernel methods not really supported in Big Data.

Thanks for your attention!
Questions?