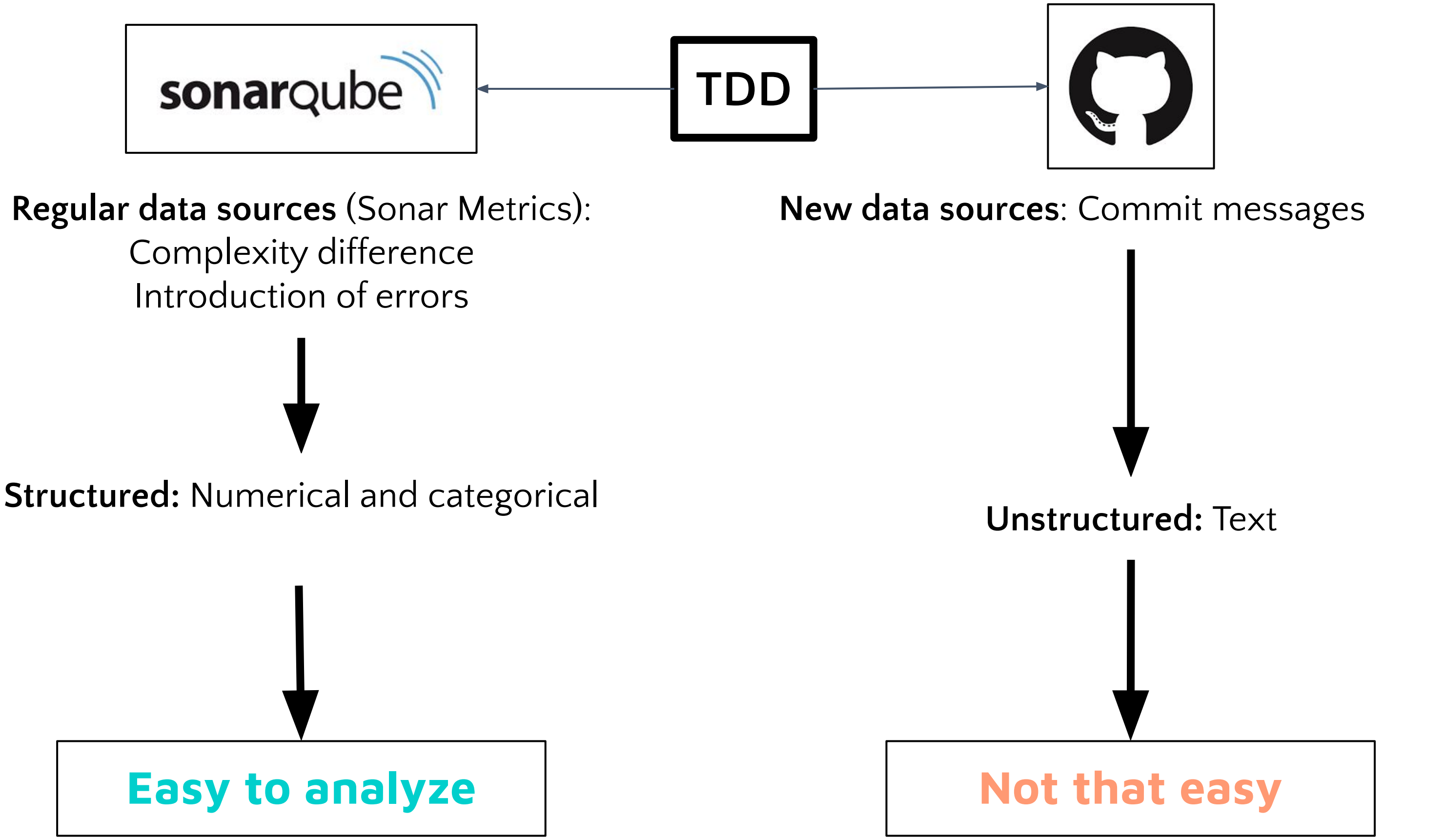
TAED2 Software Analytics Project

EPAA – Using Commit Messages to get insights

BUSINESS UNDERSTANDING



TDD

Regular data sources (Sonar Metrics):
Complexity difference
Introduction of errors

Structured: Numerical and categorical

Easy to analyze

New data sources: Commit messages

Unstructured: Text

Not that easy

BUSINESS UNDERSTANDING AND OBJECTIVES

- **Business objectives:**

  Offer meaningful insights from **git commit messages**

  1. Is there information in the commit message about the bugs in the code?
  2. Can we segment authors in a project based on how they write their commit messages?
  3. Is this segmentation related to the quality of their commits?
  4. Can we detect outlying commit messages or authors based on commit messages?

Success criteria → Define if git commit message contain **meaningful data**

DATA MINING GOALS

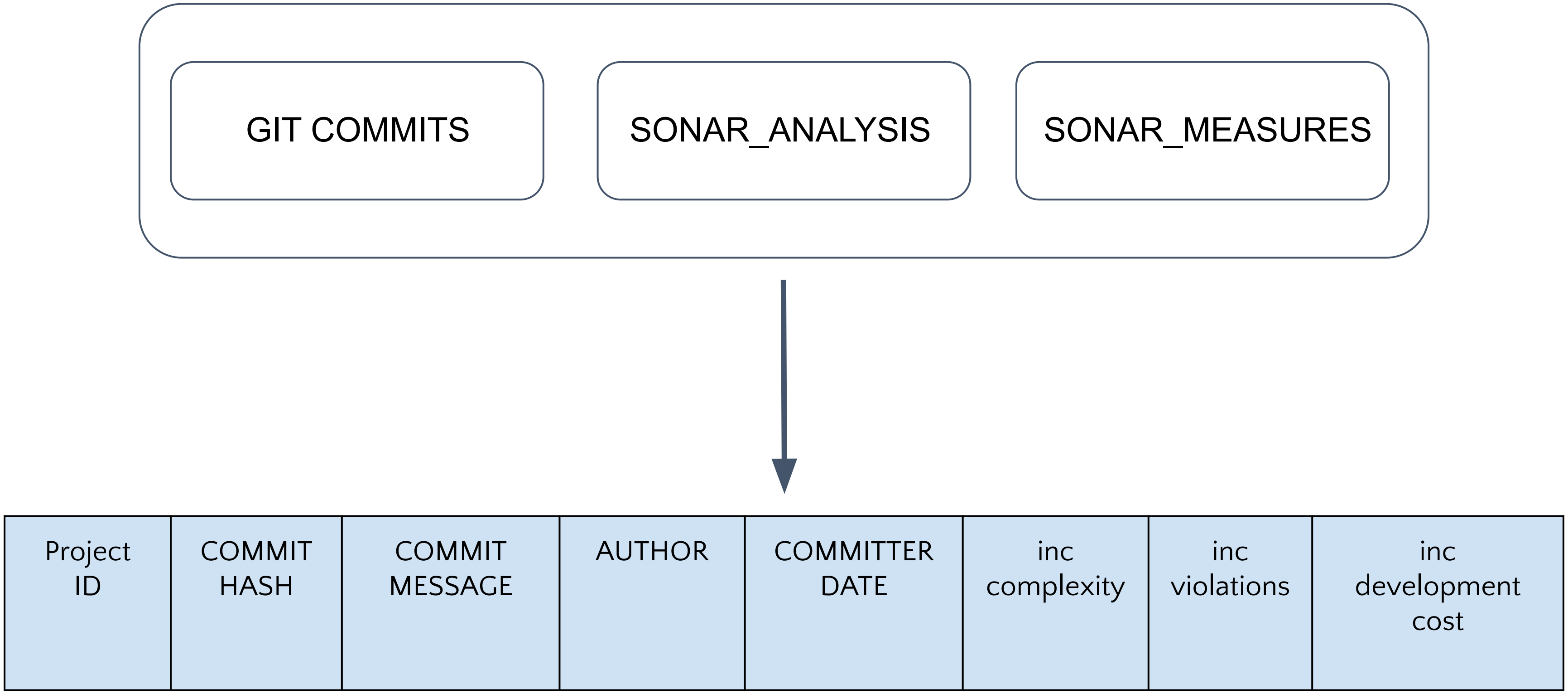
Our main goal is to obtain **valuable insights**:

1. Finding relation between COMMIT_TEXT and complexity measures
2. Find relationship between developers based on COMMIT_TEXT
3. Analyze distinguished and misleading commit text authors (clustering).

Success criteria:

1. Predicting the modification in the metrics with 30% error (at most).
2. Differentiate authors between 2 well defined groups.
3. Identify some misleading and distinguishable commit text authors.

Data Preparation



GIT COMMITS

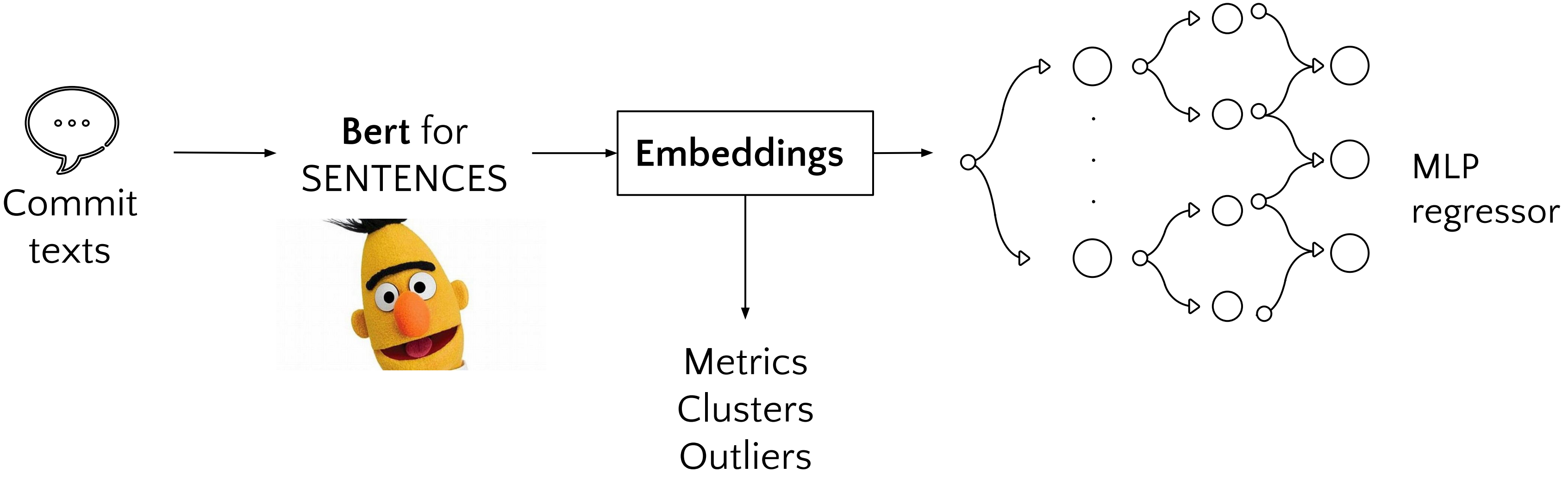
SONAR_ANALYSIS

SONAR_MEASURES

Project ID	COMMIT HASH	COMMIT MESSAGE	AUTHOR	COMMITTER DATE	inc complexity	inc violations	inc development cost

62917 rows in total

Embeddings



Commit texts

Bert for SENTENCES

Embeddings

Metrics
Clusters
Outliers

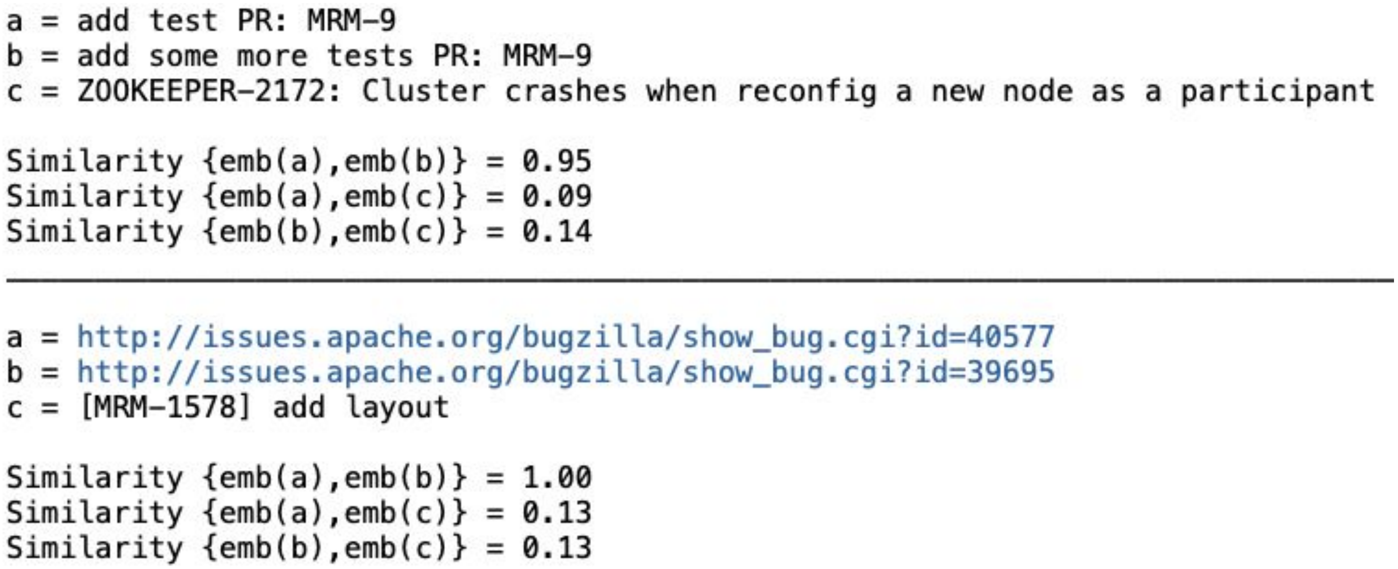
MLP regressor

Used the pretrained model 'all-MiniLM-L6-v2' from sentence_transformers

Sentence embeddings of size **384**

Embeddings

Commit message Embeddings → powerful representation



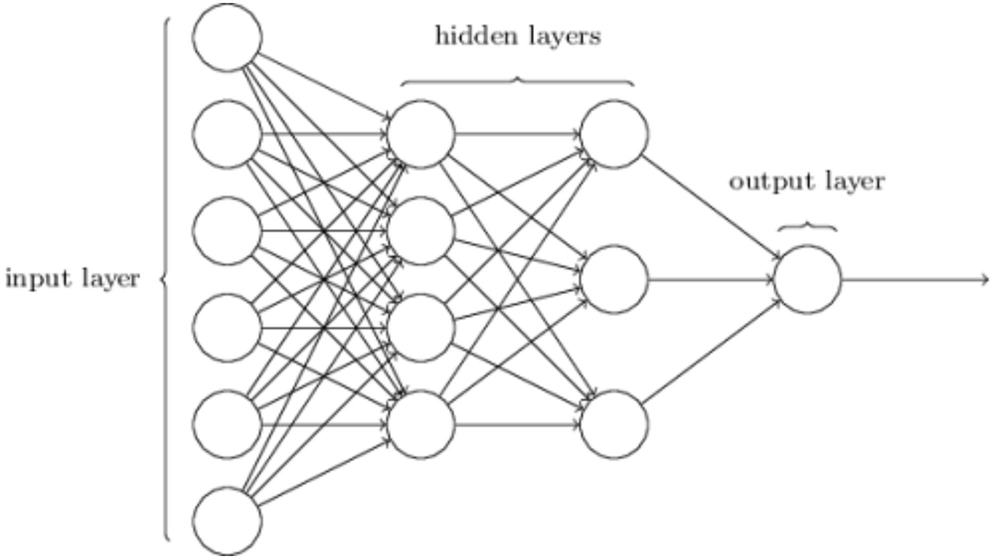
```
a = add test PR: MRM-9
b = add some more tests PR: MRM-9
c = ZOOKEEPER-2172: Cluster crashes when reconfig a new node as a participant

Similarity {emb(a),emb(b)} = 0.95
Similarity {emb(a),emb(c)} = 0.09
Similarity {emb(b),emb(c)} = 0.14
________________________________________________________________________________

a = http://issues.apache.org/bugzilla/show_bug.cgi?id=40577
b = http://issues.apache.org/bugzilla/show_bug.cgi?id=39695
c = [MRM-1578] add layout

Similarity {emb(a),emb(b)} = 1.00
Similarity {emb(a),emb(c)} = 0.13
Similarity {emb(b),emb(c)} = 0.13
```

Complexity Prediction

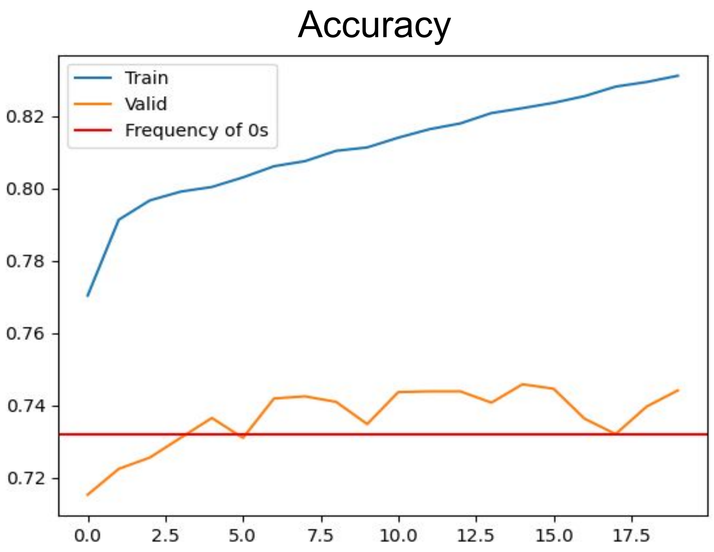
Two classes:

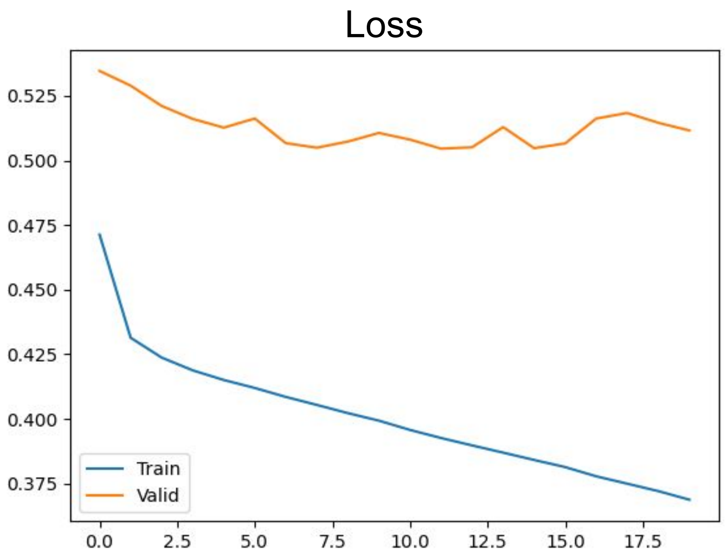
- 1, Complexity grows after that commit
- 0, Complexity doesn't grow after the commit

Architecture: MLP with 384-1024-120-1 neuron layers and ReLu activations



Complexity Prediction





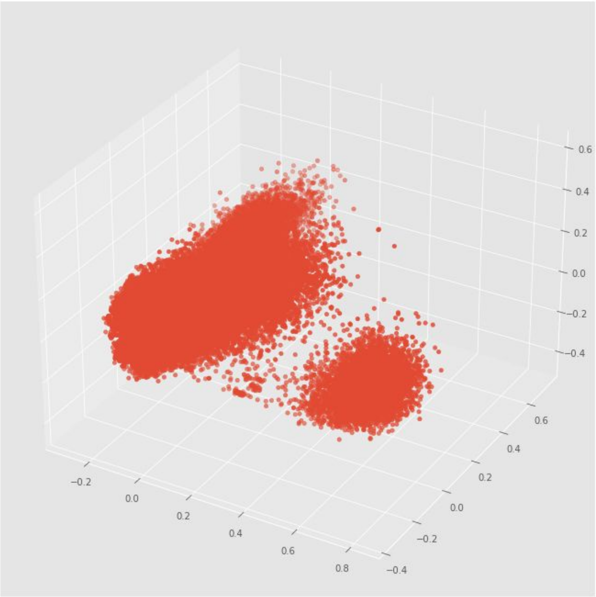
No Learning, just remembering training observations

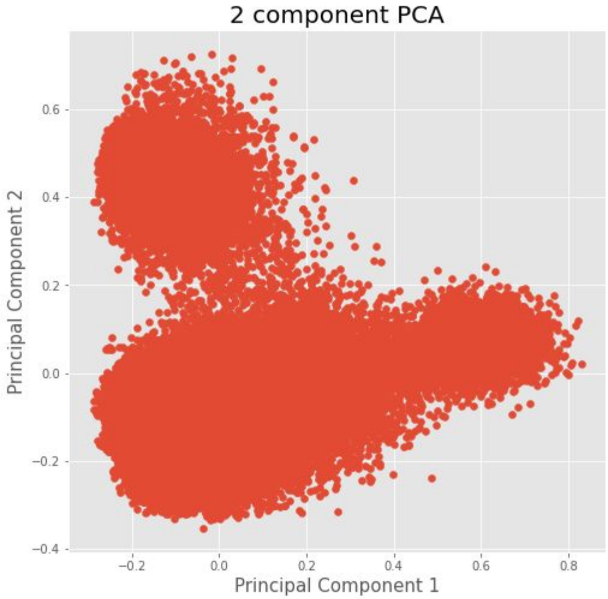
Data inherent problem

Clustering – K-MEANS

⟶Mini Batches Kmeans (1024 batches)

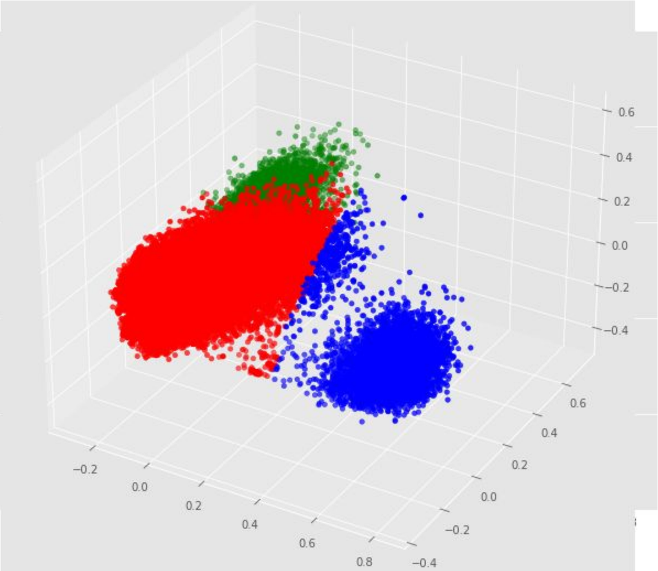
⟶ **Input**:   Principal Component Analysis



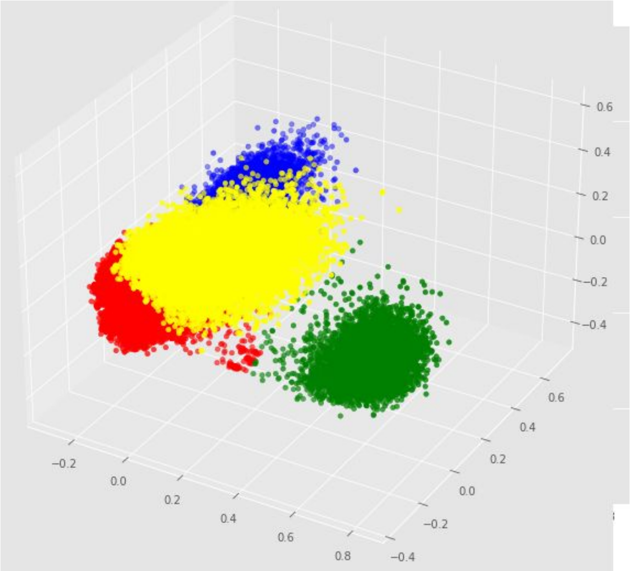


Clustering – K-MEANS

K = 3



K = 4



Calinski-Harabasz	4879.38	4121.24
Davies-Bouldies	2.2	3.48

We select **3 clusters** for the analysis

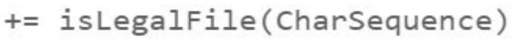
Clustering - K-MEANS

	Cluster 0	Cluster 1	Cluster 2
COMMITS	48308	8811	5798
AUTHORS	342	13	3

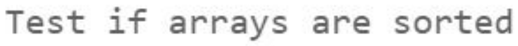
⟶ One big cluster

⟶ Cluster of authors: each author are assign to the cluster with its maximum number of commits

⟶ Cluster 2 having a clear outlier author (no links after text)



```
+= isLegalFile(CharSequence)
```

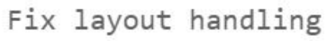


```
Test if arrays are sorted
```

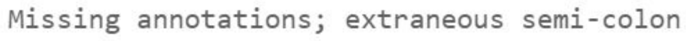
Clustering - KMEANS

Characteristics found per cluster:

- Cluster 0: Keywords “fix” and “add” but lots of variety

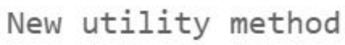


```
Fix layout handling
```

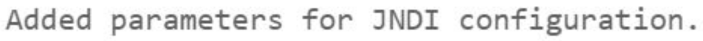


```
Missing annotations; extraneous semi-colon
```

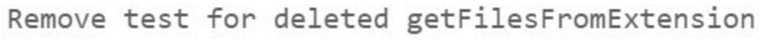
- Cluster 1: No found patterns, a lot of variety



```
New utility method
```

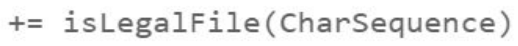


```
Added parameters for JNDI configuration.
```

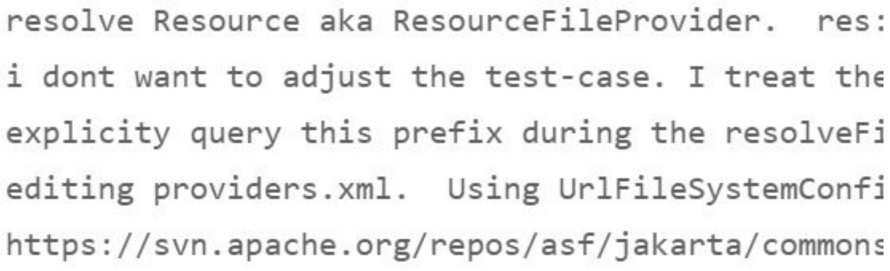


```
Remove test for deleted getFilesFromExtension
```

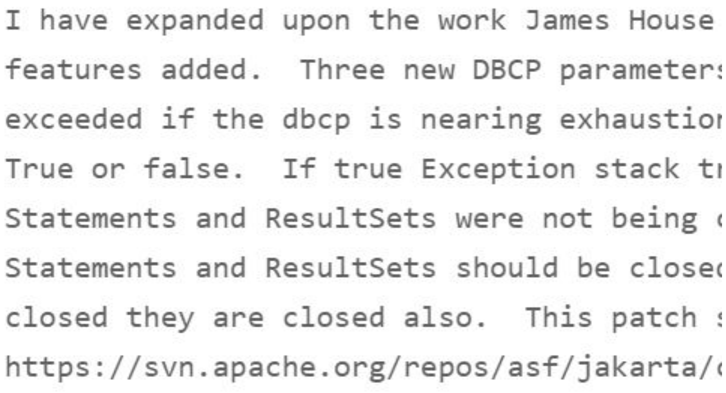
- Cluster 2: Differentiated messages and variety



```
+= isLegalFile(CharSequence)
```

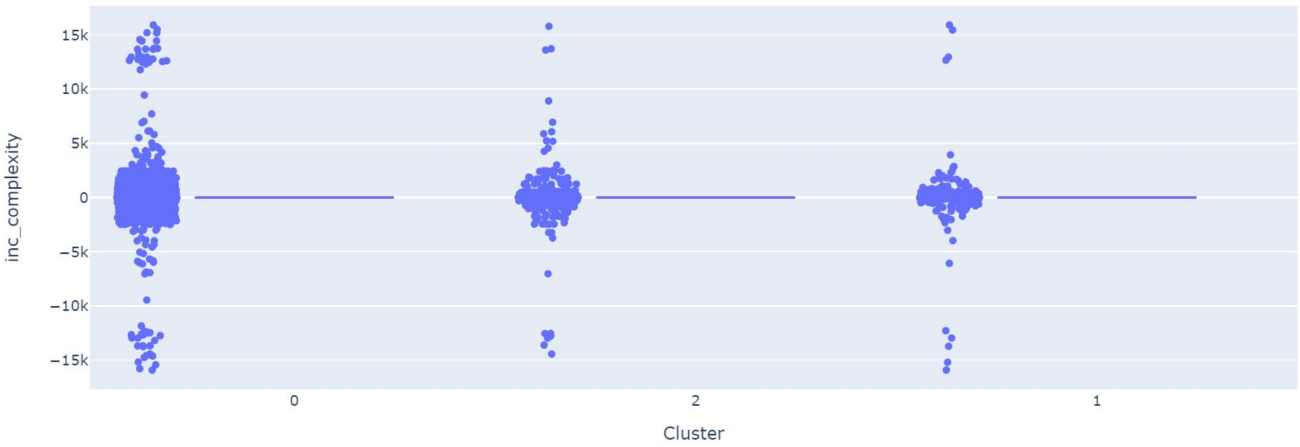


```
resolve Resource aka ResourceFileProvider.  res:
i dont want to adjust the test-case. I treat the
explicity query this prefix during the resolveFi
editing providers.xml.  Using UrlFileSystemConfi
https://svn.apache.org/repos/asf/jakarta/commons
```

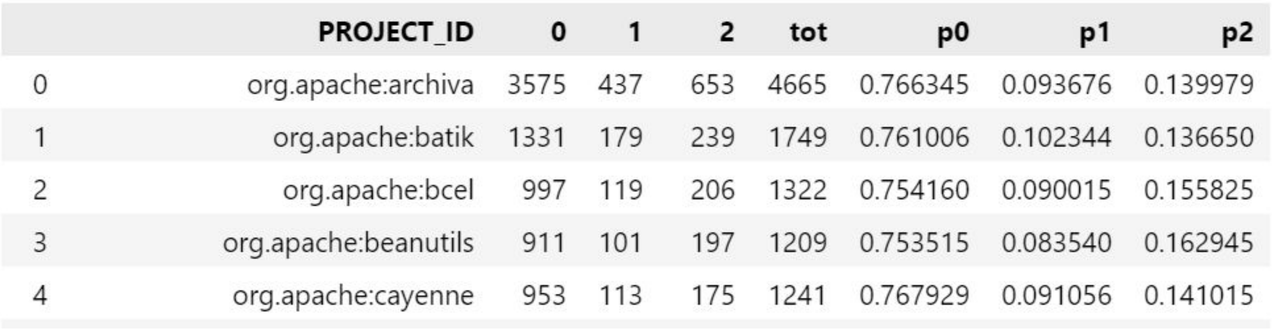


```
I have expanded upon the work James House
features added.  Three new DBCP parameter
exceeded if the dbcp is nearing exhaustio
True or false.  If true Exception stack t
Statements and ResultSets were not being
Statements and ResultSets should be close
closed they are closed also.  This patch
https://svn.apache.org/repos/asf/jakarta/
```

Clustering – K-MEANS



→ Big variance, no clear difference between clusters



	PROJECT_ID	0	1	2	tot	p0	p1	p2
0	org.apache:archiva	3575	437	653	4665	0.766345	0.093676	0.139979
1	org.apache:batik	1331	179	239	1749	0.761006	0.102344	0.136650
2	org.apache:bcel	997	119	206	1322	0.754160	0.090015	0.155825
3	org.apache:beanutils	911	101	197	1209	0.753515	0.083540	0.162945
4	org.apache:cayenne	953	113	175	1241	0.767929	0.091056	0.141015

→ Same proportion of commits clustering per project

Conclusions

- In **Business terms**:

Provided **meaningful** embeddings & clustering insights for a potential project (for example for Github)

There seems to be no relationship that can be modeled between commit messages and increase of complexity

- In **Data Mining terms**:

Able to detect outlier authors and created usable embeddings

Created efficient and **repeatable** process to merge and clean the tables to create the final curated database and **reproduce results**

Conclusions for future data mining

Further Research & Improvements

- Try standardizing data
- Creating End-To-End Embeddings: Fine tune them for the Commits Messages processed dataset
- Explore LSTM methods using word embeddings instead of sentence embeddings

Thanks for watching!

If you have any doubt, contact us!
https://github.com/megaelius/EPAA

Elias Abad

Andrea Garcia

Àlex Martí

Pau Bernat