Apples, Oranges & Fruits - Understanding Similarity of Software Repositories Through The Lens of Dissimilar Artifacts

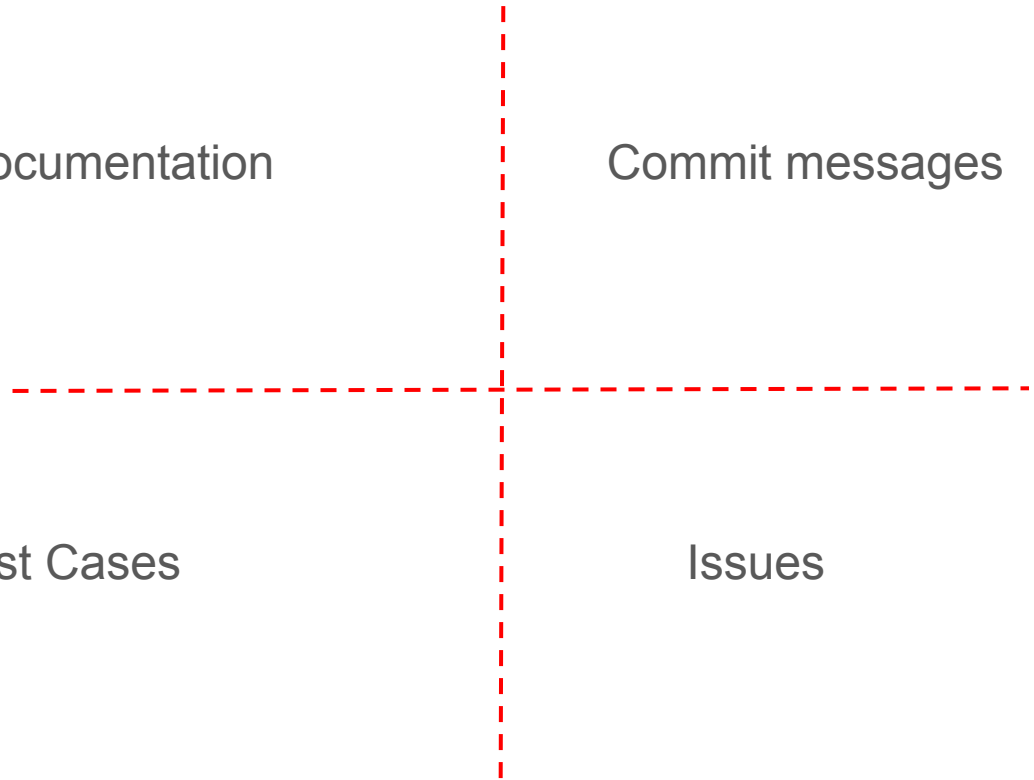
A. E. Rao and S. Chimalakonda, in *2022 IEEE International Conference on Software Maintenance and Evolution (**ICSME**)*, Limassol, Cyprus: IEEE, Oct. 2022, pp. 384–388. doi: 10.1109/ICSME55016.2022.00044.

Similarity of Software Repository

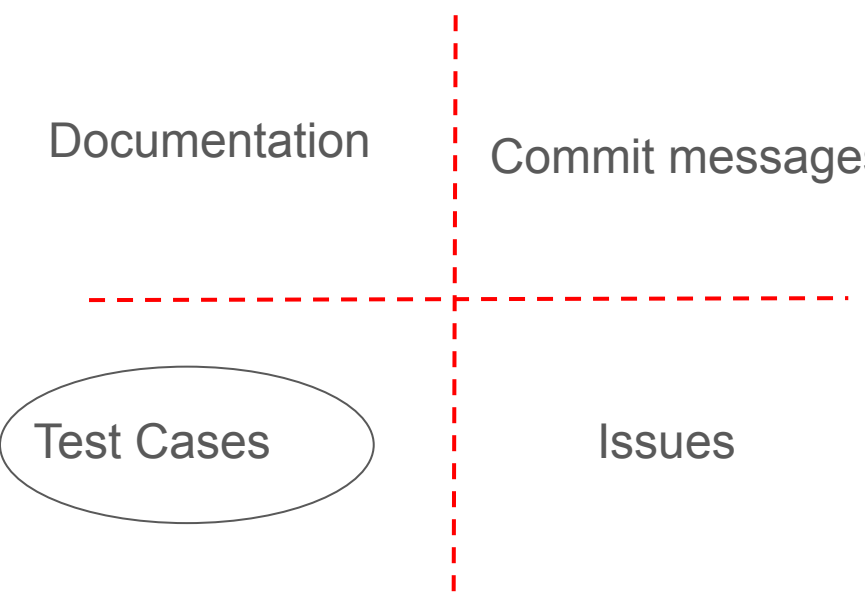
Similarity of Software Repository. But why?

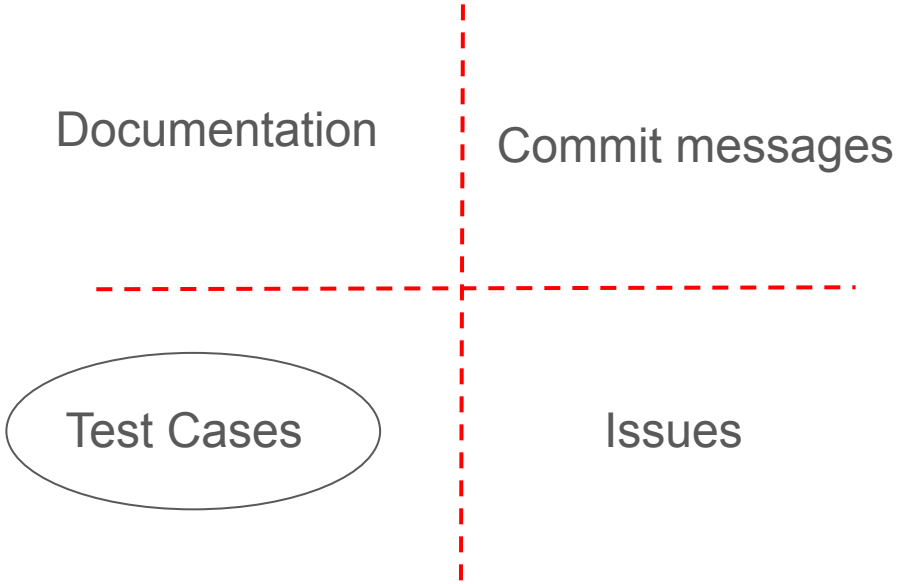
- Alternate implementations
- Best practices
- Rapid prototyping
- Code reuse
- Detect plagiarism

Dissimilar Artifacts

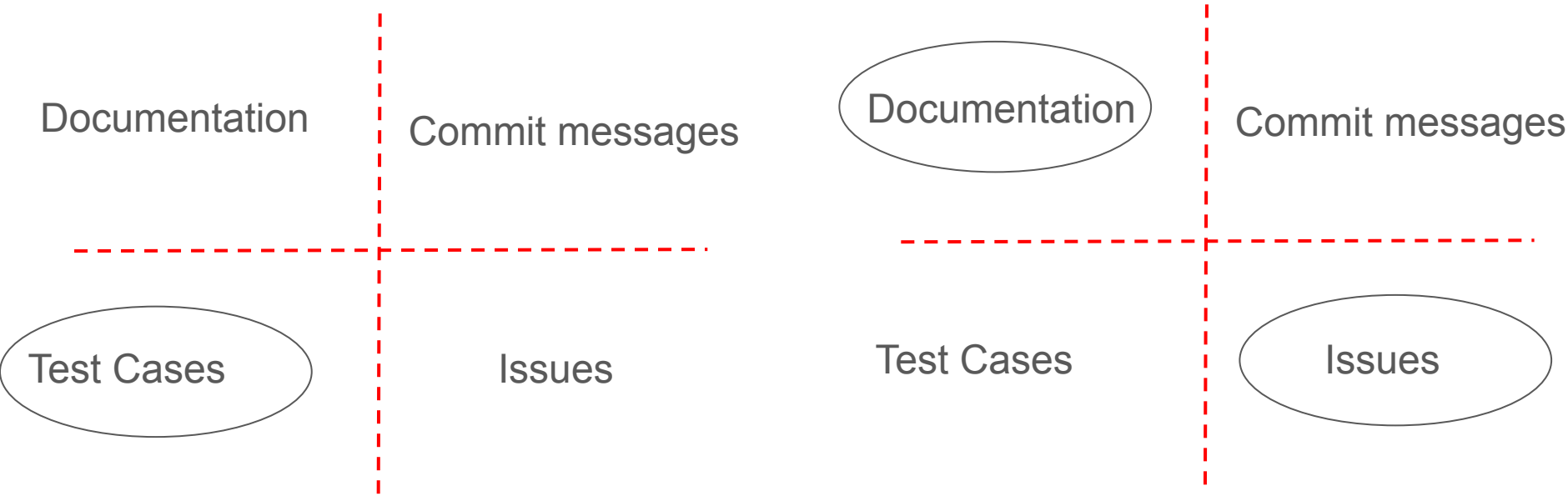


Works so far used… only similar artifacts





This work proposes… using dissimilar artifacts



An example.

I have a ***pytorch*** project. But no ***tests***. How do I write tests?

An example.

I have a ***pytorch*** project. But no ***tests***. How do I write tests?

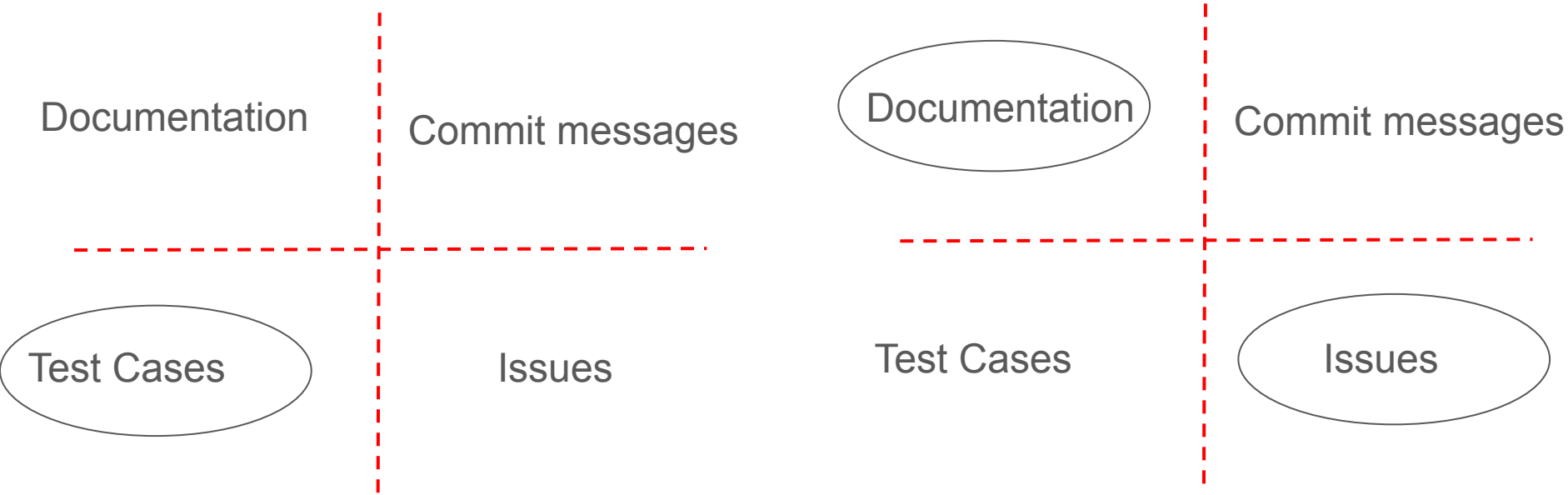
Look for ***test files***.

An example.

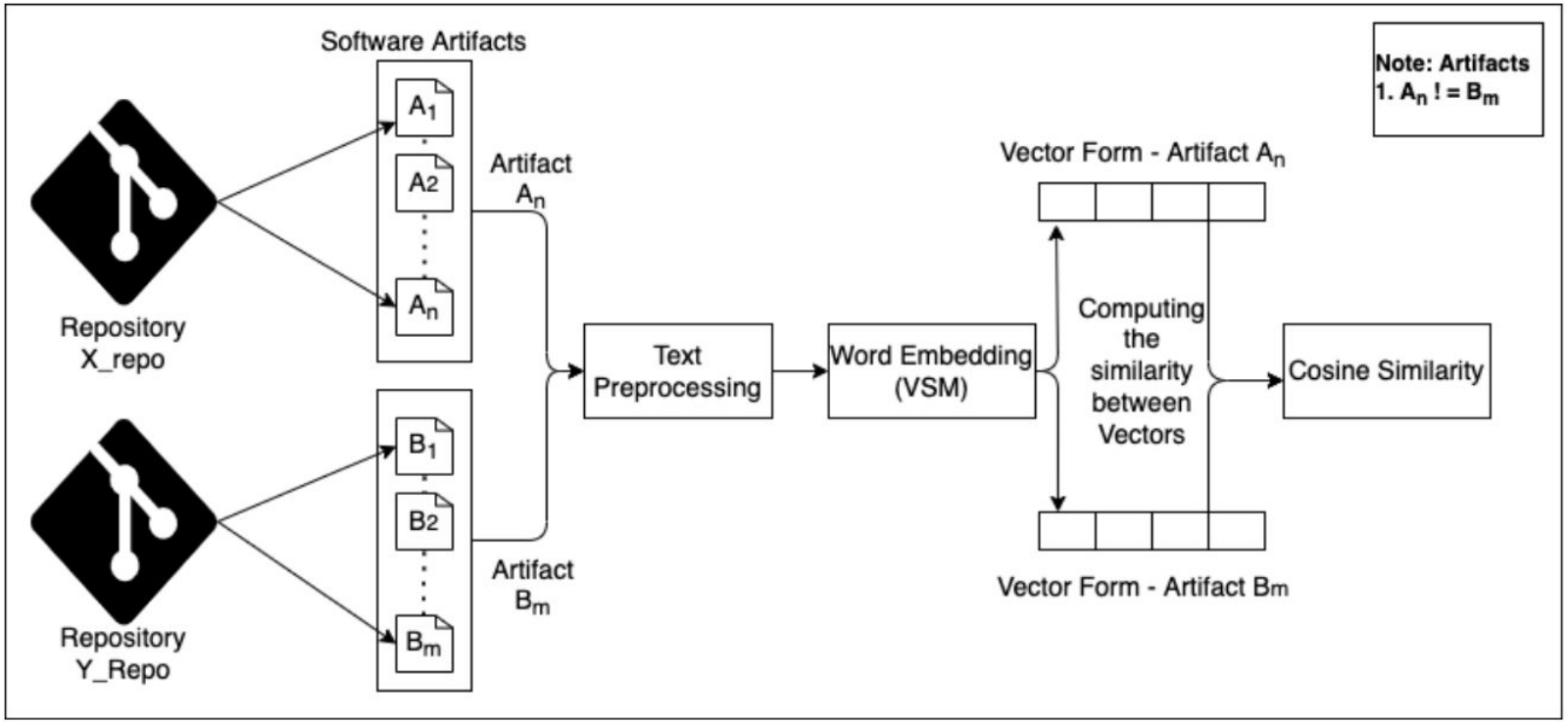
I have a ***pytorch*** project. But no ***tests***. How do I write tests?

1. Look for ***pytorch*** projects
2. Check ***documentation***
3. Look for ***test cases***
4. Ctrl + C, Ctrl + V

This work proposes… using dissimilar artifacts ✅



Approach


Software Artifacts
Note: Artifacts
1. An ! = Bm
A1
A2
An
Artifact An
Repository X_repo
Vector Form - Artifact An
Text Preprocessing
Word Embedding (VSM)
Computing the similarity between Vectors
Cosine Similarity
B1
B2
Bm
Artifact Bm
Repository Y_Repo
Vector Form - Artifact Bm

Evaluation

Android app vs Android app

Dissimilar Artifacts Pairs

Dissimilar Artifacts Pair	Highest	Average	Lowest
Commits (A) vs Pull Requests (B)	0.34	0.10	0
Commits (A) vs Readme Files (B)	0.21	0.05	0
Issues (A) vs Pull Requests (B)	0.25	0.04	0
Issues (A) vs Readme Files (B)	0.26	0.06	0
Pull Requests (A) vs Readme Files (B)	0.24	0.05	0

Similar Artifacts Pairs

Similar Artifacts Pair	Highest	Average	Lowest
Commits (A) vs Commits (B)	1.00	0.10	0
Pull Requests (A) vs Pull Requests (B)	0.52	0.04	0
Issues (A) vs Issues (B)	0.42	0.08	0
Readme Files (A) vs Readme Files (B)	0.74	0.04	0

Android app vs Music app

Dissimilar Artifacts Pairs

Dissimilar Artifacts Pair	Highest	Average	Lowest
Commits (A) vs Pull Requests (B)	0.17	0.03	0
Commits (A) vs Readme Files (B)	0.13	0.02	0
Issues (A) vs Pull Requests (B)	0.14	0.03	0
Issues (A) vs Readme Files (B)	0.12	0.02	0
Pull Requests (A) vs Readme Files (B)	0.15	0.01	0

Similar Artifacts Pairs

Similar Artifacts Pair	Highest	Average	Lowest
Commits (A) vs Commits (B)	0.27	0.03	0
Pull Requests (A) vs Pull Requests (B)	0.28	0.05	0
Issues (A) vs Issues (B)	0.24	0.05	0
Readme Files (A) vs Readme Files (B)	0.31	0.04	0

Discussion Points

1. Limited to text based artifacts.
2. Bad Evaluation. Choice of dataset. Presentation of results.
3. Future relevance.