2nd Workshop of the LIG Axes - 2021

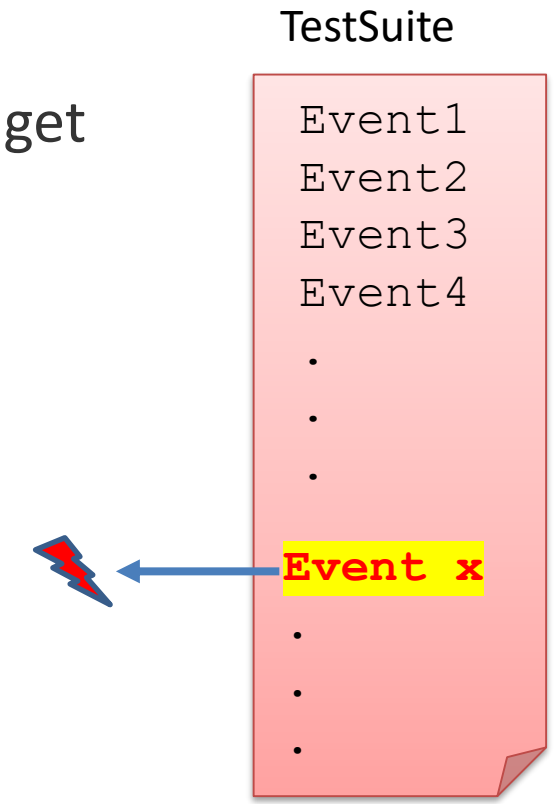
Reducing Regression Test Suites using the Word2Vec Natural Language Processing Tool

Bahareh Afshinpour, Roland Groz, Massih-Reza Amini, Yves Ledru and Catherine Oriat

LIG Lab, Grenoble INP ,University Grenoble Alpes

Motivation

- **Software testing and fault detection**
  - More than 52 percent of total software development budget
  - Time-consuming
  - Highly priced
- **Fault detection by users' log analysis**
  - Processing user log files as a test suite
  - Log files reveal **faults**
  - Log files are sometimes huge
  - Rerunning an entire log file is highly time-consuming



→CAN WE MAKE a LOG FILE SMALLER? (Test Suite Reduction)

→While the small log file still reveals the same faults.

→We must select important events and discard the rest of events

NLP-based approach

- **Other methods :**
  - Generally focus on the combination of the selected events.
  - Do not consider *semantic* of events in the log file.

- **We propose an NLP-based approach:**
  - We process software outputs (logs) using NLP.
  - Consider combination and *semantical similarity* of the events in test traces.
  - *Traces of actions* in log file could be regarded as *sentences of words* in NLP.

Overview

- The Software Under Test
- The proposed method
- Vectorization
- Clustering different sessions
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- Conclusion

Overview

- **The Software Under Test**
- The proposed method
- Vectorization
- Clustering different sessions
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- Conclusion

The Software Under Test (Scanner case study)

- Large files from clients' behavior
- Goal: Select clients that their behavior trigger faults

- Supermarkets scanner device.
  - **Self-scanning items by client.**
  - Actions: Scan, Delete barcodes and Pay, ….
- The Scanner system has a Java implementation.

- We need a buggy software to evaluate the proposed Test Suite Reduction method:
  - To artificially inject faults, the source code of the Scanette software was mutated.
  - Some bugs or ***mutations*** injected.

Example log file :

Client6 **session** from `Unlock` to `Pay`

Index	Time	Session ID	Object	Action	Input	Output
51,	1585070116817,	client6,	scan12,	unlock,	[],	0
52,	1585070116819,	client0,	scan0,	scan,	[3270190022534],	0
53,	1585070116820,	client1,	cashier1,	CloseSession,	[],	0
54,	1585070116820,	client2,	cashier2,	add,	[3570590109324],	0
55,	1585070116824,	client5,	scan5,	scan,	[8718309259938],	0
56,	1585070116825,	client6,	scan12,	scan,	[3560070139675],	0
57,	1585070116837,	client0,	scan0,	scan,	[3560070048786],	0
58,	1585070117030,	client6,	scan12,	scan,	[7640164630021],	-2
59,	1585070117073,	client6,	scan12,	delete,	[7640164630021],	-2
60,	1585070116838,	client1,	cashier1,	pay,	[353.06],	0
61,	1585070116839,	client2,	cashier2,	CloseSession,	[],	0
62,	1585070116840,	client3,	cashier3,	add,	[3570590109324],	0
64,	1585070117687,	client6,	scan12,	transmission,	[caisse6],	0
65,	1585070117687,	client6,	scan12,	abandon,	[],	?
66,	1585070117701,	client6,	cashier4,	OpenSession,	[],	0
67,	1585070116855,	client0,	scan0,	transmission,	[cashier0],	0
68,	1585070116855,	client0,	scan0,	abandon,	[],	?
69,	1585070117716,	client6,	cashier4,	add,	[7640164630021],	0
70,	1585070117731,	client6,	cashier4,	CloseSession,	[],	0
71,	1585070117747,	client6,	cashier4,	Pay,	[260],	9.11

Three log files (test-suites)

- 1026-event , 100,043-event and 200,035-event

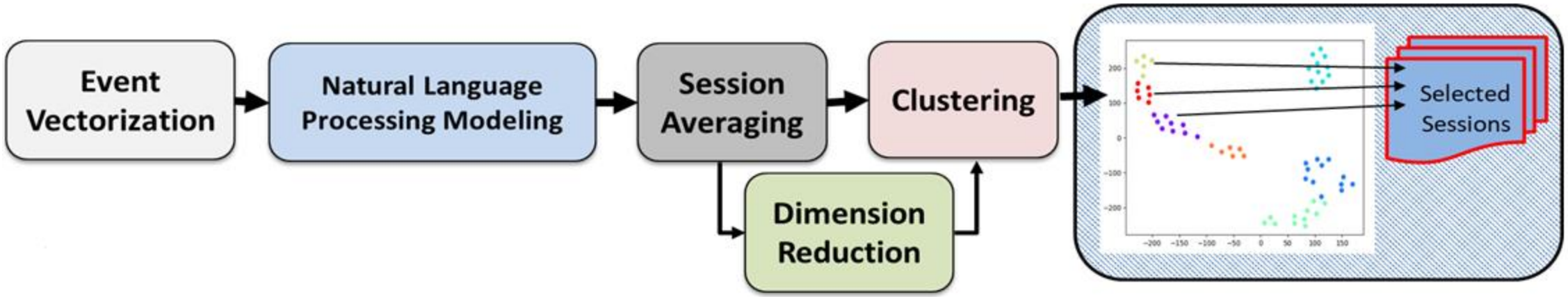
- If a test-suite can find the injected fault, we say it ***kills*** the mutant.

Testsuite	Number of clients	Number of events	Number of killed mutants
1026-event	61	1026	19
100043-event	7079	100043	22
200035-event	14443	200035	23

Overview

- The Software Under Test
- **The proposed method**
- Vectorization
- Clustering different sessions
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- Conclusion

The Proposed Method



Overview

- The Software Under Test
- The proposed method
- **Vectorization**
- Clustering different sessions
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- Conclusion

Step #1: Vectorization

- In order to process the events, we need to vectorize them.
- We have a set of actions
  - $A=\{A_1,A_2,A_3,\ldots,A_n\}$
    - For example A={unlock, abandon, transmit, scan, pay}
- And set of input parameters
  - $P=\{P_1, P_2, \ldots,P_m\}$
    - For example P={checkout0, 5410188006711, 68.27, 640164630021 , [], …..}
- And a set of outputs
  - $O=\{O_1,O_2, \ldots, O_k\}$
    - For example O={0,-2, …}

Vectorizing the actions

```
# timestamp, sessID, object, action, inputs, output
1570573649196, 41, scan3, abandon, [], 0
1570573649191, 42, scan1, transmit, [], 0
1570573649197, 42, scan1, abandon, [], 0
1570573649355, 43, scan1, unlock, [], 0
1570573649358, 42, checkout0, openSession, [], 0
1570573649996, 43, scan1, scan, [5410188006711], 0
1570573650366, 43, scan1, scan, [5410188006711], 0
1570573650366, 42, checkout0, add, [3570590109324], 0
1570573650389, 44, scan2, scan, [3046920010856], 0
1570573651369, 42, checkout0, closeSession, [], 0
1570573652376, 42, checkout0, pay, [68.27], 0
1570573655132, 40, scan0, scan, [7640164630021], -2
1570573656245, 44, scan2, scan, [3270190022534], 0
1570573656633, 43, scan1, scan, [3474377910724], 0
```

['abandon','Nothing','0']

['transmit','Nothing','0']

Example : 1026_event

Client 2:

- [['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['ouvrirSession', 'Nothing', '0'], ['ajouter', 'Barcode', '0'], ['fermerSession', 'Nothing', '0'], ['payer', 'Price-float', '0']]

Client 42:

- [['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['ouvrirSession', 'Nothing', '0'], ['ajouter', 'Barcode', '0'], ['fermerSession', 'Nothing', '0'], ['payer', 'Price-integer', 'Float Number']]

Overview

- The Software Under Test
- The proposed method
- Vectorization
- **Clustering different sessions (clients)**
  - **Create Word2Vec Model**
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- Conclusion

The goal of clustering

Client 2:

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['ouvrirSession', 'Nothing', '0'], ['ajouter', 'Barcode', '0'], ['fermerSession', 'Nothing', '0'], ['payer', 'Price-float', '0']]

Client 9:

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['ouvrirSession', 'Nothing', '0'], ['ajouter', 'Barcode', '0'], ['fermerSession', 'Nothing', '0'], ['payer', 'Price-integer', 'Float Number']]

Client 20:

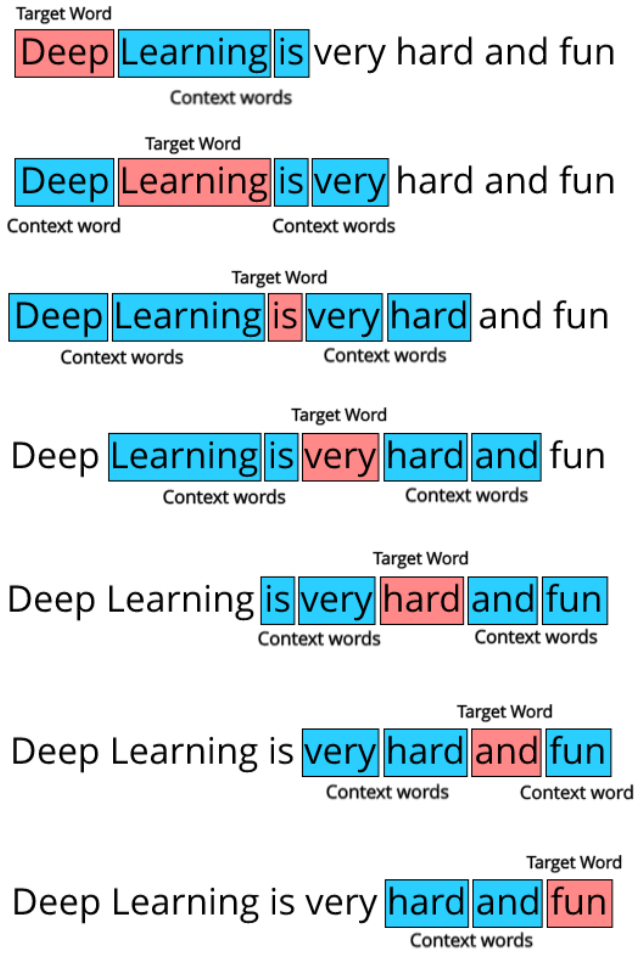
[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-float', '0']]

A clustering method may tell us that some sessions are equivalent or very close to each other.

We need a method to say if these clients are in the same cluster?

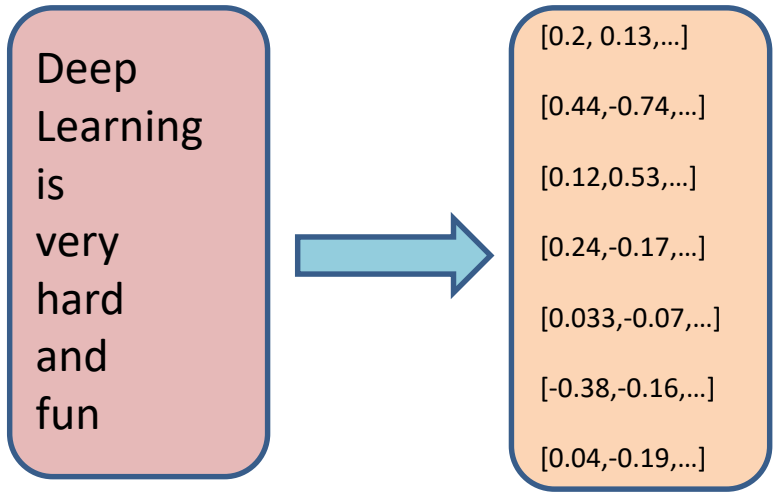
16

Step #2: Word2Vec[1] Model Construction



https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73

- Word embedding technique
- Uses deep learning and neural networks-based techniques
- Converts words into vectors
- Finally semantically similar words(*vocabs*) have close vectors.



[1] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Step #2: Word2Vec Model Construction

- Finding similar vector representation for actions that occur in the same event.
- For our case:
  - Each triplet in a session is a *vocab* for the Word2Vec model.
  - We store each triplet as a string (*e.g*: '['scanner', 'Barcode', '0']')
  - We treat triplets like words in natural language processing.

client0:

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-float', '0']]

One vocab

Word2Vec on all the Sessions

The Word2Vec package from the Gensim libray

Example: the vectorized representation of word2vec

```
['ajouter', 'Barcode', '0']
[0.244 0.257 0.528 2.099 -0.873 0.674 0.641 0.587 -1.204 0.873 -2.233
 -0.225 0.383 -0.687 0.515]
^^^^^^^^^^^^^^^^^^^^^^^^^^
['fermerSession', 'Nothing', '0']
[0.133 0.214 0.162 1.278 -0.390 0.337 0.326 0.385 -0.778 0.524 -1.136
 -0.235 0.164 -0.275 0.236]
^^^^^^^^^^^^^^^^^^^^^^^^^^
['supprimer', 'Barcode', '0']
[0.007 -0.324 0.222 -0.645 -0.170 -0.164 -0.327 -0.414 0.603 -0.213 0.450
 0.138 0.071 0.055 0.190]
^^^^^^^^^^^^^^^^^^^^^^^^^^
```

File name	Number of different triplets
1026-event	15
100043-event	18
200035-event	20

The dimension of the W2V vectors are equal to the number of vocabs

Overview

- The Software Under Test
- The proposed method
- Vectorization
- **Clustering different sessions**
  - Create Word2Vec Model
  - **Averaging of the Word2Vec vectors**
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- Conclusion

Step #3: Comparing two Sessions

- Need a *measure* for each session …
  - To compare two sessions and find their similarity.

- *Measure:*
  - An average of the Word2Vec vectors in each session [2][3].

[2]Tomas Mikolov. "Distributed representations of sentences and documents." In *International conference on machine learning*, pp. 1188-1196. 2014.
[3]*Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. "Towards universal paraphrastic sentence embeddings." arXiv preprint arXiv:1511.08198 (2015).*

Averaging a session

- **client60:**

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-integer', 'Float Number']]

([0.015 -0.505
0.285 -0.973
-0.233 -0.270
-0.448 -0.670
0.952 -0.364
0.848 0.159
0.001 0.111
0.170]

+ [0.183 -1.838
1.492 -2.503
-1.275 -0.563
-1.398 -2.107
2.791 -0.838
1.838 0.596
0.519 0.122
1.077]

+ …) = […..
…..
……
……..
…….
…….
]

/ 10 = **[0.694 0.449
-0.182 0.304
0.306 -0.590
0.221 -0.607
0.850 -0.789
-0.843 -0.996
0.506 -0.587
-0.097]**

Averaging a session

client60:

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-integer', 'Float Number']]

[0.694 0.449
-0.182 0.304
0.306 -0.590
0.221 -0.607
0.850 -0.789
-0.843 -0.996
0.506 -0.587
-0.097]

client40:

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-integer', 'Float Number']]

[0.399 -0.829
-0.209 -1.239
1.379 1.404
-0.528 0.081
-0.763 -0.999
1.066 -1.038
-0.355 0.963
-0.973]

client17:

[['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-integer', 'Float Number']]

[-1.437 0.207
-0.594 - 0.754
1.211 1.220
0.543 -0.852
0.897 -0.951
0.577 1.473 -
1.082 -0.819
0.887]

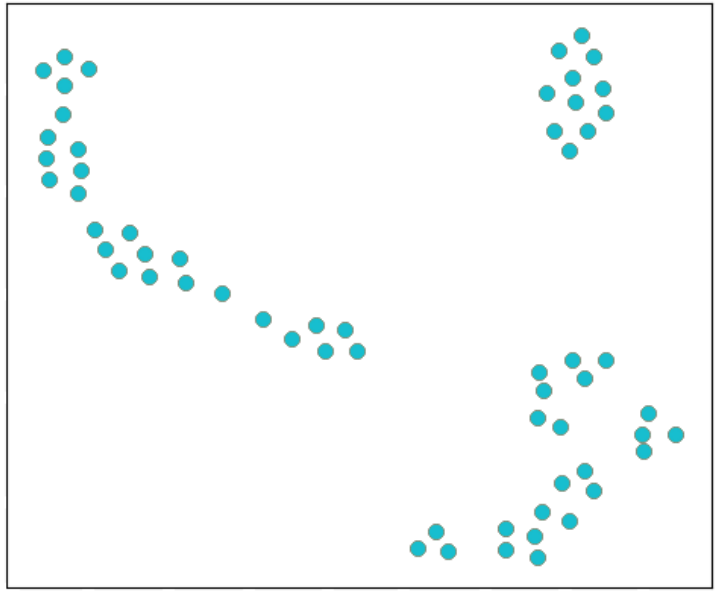
Now we have a measure for each session.

Overview

- The Software Under Test
- The proposed method
- Vectorization
- **Clustering different sessions**
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - **Use t-SNE (dimensionality reduction technique)**
  - **K-means clustering**
- Results
- Conclusion

Step #4: t-SNE (dimensionality reduction technique)

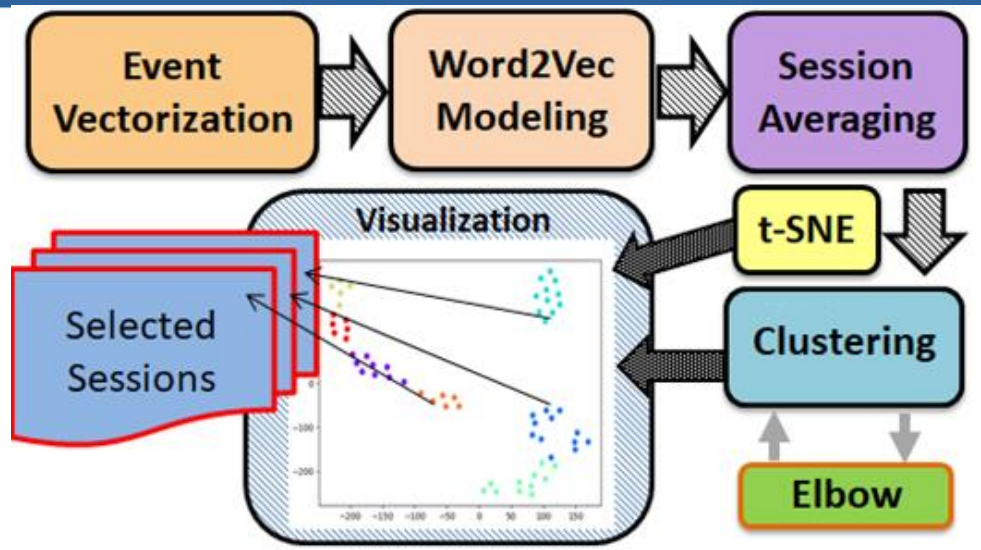
- Dimension of Word2Vec output is high.
- t-SNE is a nonlinear dimensionality reduction technique
- It is extensively applied in image processing, NLP, genomic data and speech processing.
- It is a technique for visualizing high-dimensional space (2- or 3-D).

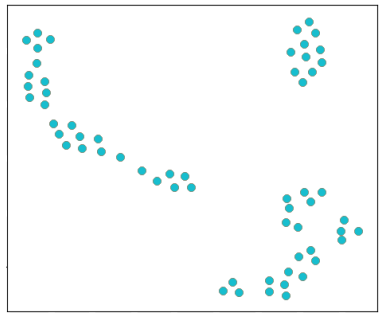


t-SNE output of 1026-event

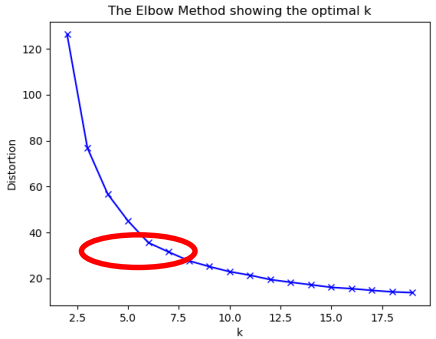
Step #5: Clustering

- K-Means
- Elbow



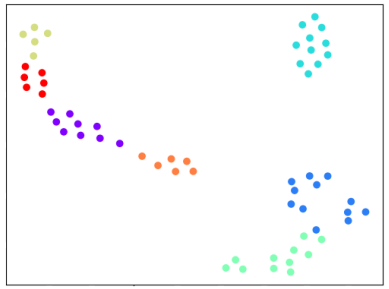


1026-event



Elbow method

K=7



[3]The distortion is the sum of square errors (SSE) : $\sum(\text{data } x_i - \text{centroid } x)^2 + \sum(\text{data } y_i - \text{centroid } y)^2 + \ldots$

Final Step :Select a cluster representative

- We choose the longest session from each cluster
- We will have K selected sessions

Select one session(client) from each cluster

Cluster 2:

client 39 : [['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['scanner', 'Barcode', '0'], ['transmission', 'CaisseNumber', '0'], ['abandon', 'Nothing', 'Error'], ['payer', 'Price-float', '0']] ……..

client 19 : [['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'] ……..

client 30 : [['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '-2'], ['scanner', 'ErrorBarcode', '0'], ……

client 33 : [['debloquer', 'Nothing', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['scanner', 'ErrorBarcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ['scanner', 'Barcode', '0'], ….

SessionID	Set of mutants
client 46	[7, 17, 21, 25, 26]
client 51	[7, 17, 21, 25, 26]
client 60	[7, 17, 21, 25, 26]
client 39	[7, 12, 17, 21, 25, 26]
client 33	[7, 12, 17, 21, 25, 26]
client 19	[0, 7, 9, 10, 12, 17, 21, 25, 26, 34, 35, 37, 41, 44]
client 30	[0, 1, 7, 9, 10, 12, 17, 21, 25, 26, 34, 37, 41, 44]
client 16	[7, 17, 20, 21, 25, 26, 42]

(Ctrl)

Selected session(client)

Overview

- The Software Under Test
- The proposed method
- Vectorization
- Clustering different sessions
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- **Results**
- Conclusion

Result on 1026-event test suite

K	#Events	Sessions Numbers	Killed mutants	Killed Mutant IDs
2	36	[30, 24]	9	[7, 12, 17, 19, 20, 21, 25, 26, 42]
3	67	[30, 23, 35]	17	[0, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44]
4	81	[30, 6, 10, 24]	10	[7, 12, 17, 19, 20, 21, 25, 26, 42, 48]
5	104	[30, 6, 10, 24, 23]	18	[0, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]
6	117	[30, 28, 23, 6, 24, 10]	18	[0, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]
7	128	[6, 27, 23, 10, 24, 30, 28]	18	[0, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]
8	147	[27, 52, 23, 10, 24, 6, 30, 28]	18	[0, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]
9	171	[10, 28, 19, 52, 27, 24, 6, 30, 23]	18	[0, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]
10	194	[28, 22, 23, 26, 35, 24, 30, 10, 6, 33]	19	[0, 1, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]
11	214	[23, 52, 22, 35, 30, 24, 6, 27, 33, 28, 10]	19	[0, 1, 7, 9, 10, 12, 17, 19, 20, 21, 25, 26, 34, 35, 37, 41, 42, 44, 48]

Result of 100043-event and 200035-event

100043-event			200035-event		
K	#Events	Killed mutants	K	#Events	Killed mutants
16	424	**20**	15	234	**10**
18	467	**22**	30	443	**18**
22	551	**22**	60	912	**21**
30	709	**22**	73	1111	**23**

Overview

- The Software Under Test
- The proposed method
- Vectorization
- Clustering different sessions
  - Create Word2Vec Model
  - Averaging of the Word2Vec vectors
  - Use t-SNE (dimensionality reduction technique)
  - K-means clustering
- Results
- **Conclusion**

Conclusion

- Applying word2vec on vectorized client's sessions is effective on test feature extractions
- word2vec clustering separates sessions that kill similar mutants
- By choosing the longest session from each cluster we succeed to kill the same mutants killed by the entire test suite

Future work

- Find a concept space for events in order to have a more meaningful and interpretable result.

- There can be many redundant events that could be removed. We are developing an analysis of the relation between events that can trigger a fault.

Thank you