Duplicate bug report detection through machine learning techniques

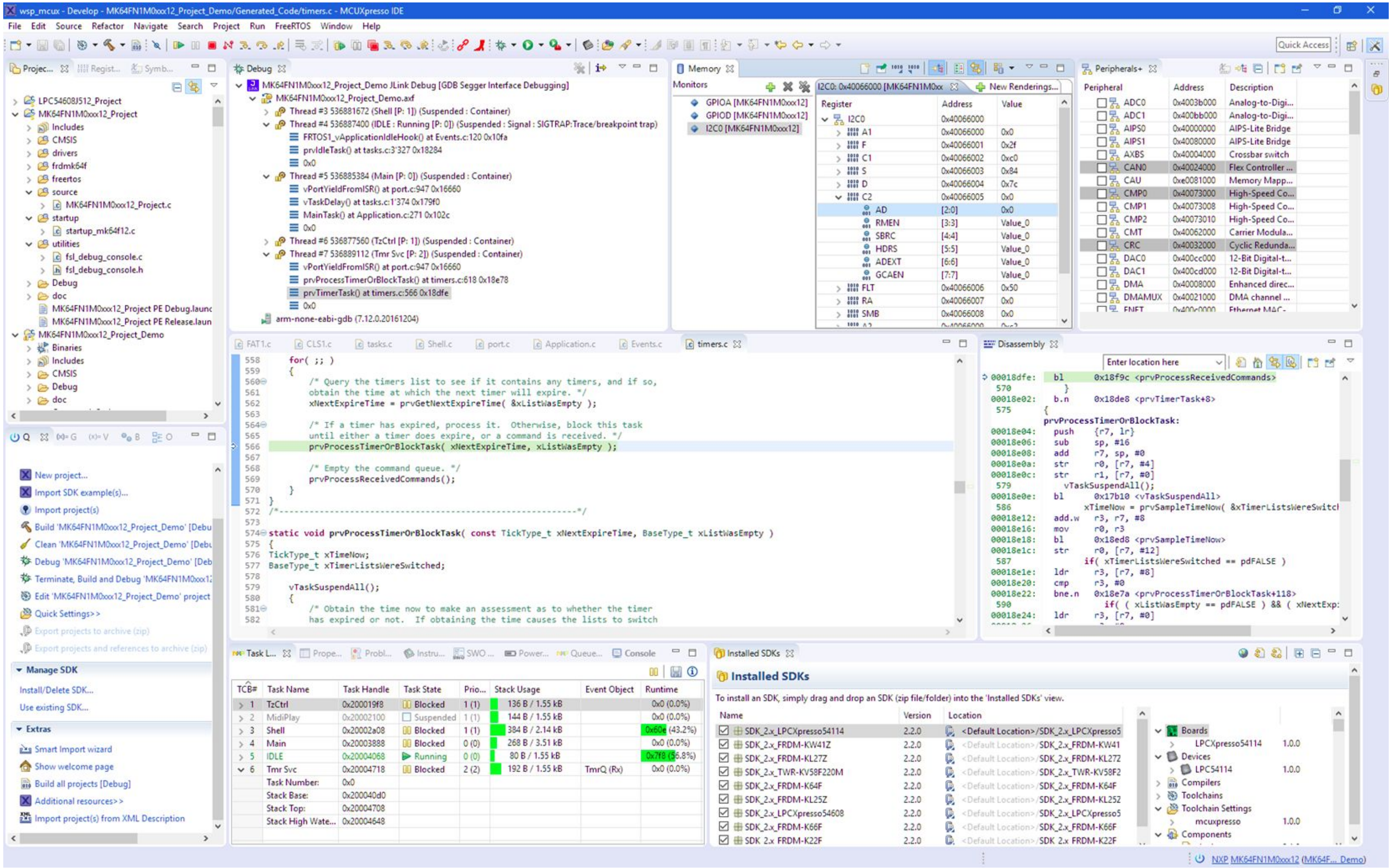
Irving Muller Rodrigues

December 10, 2018

Prof. Daniel Aloise and Prof. Michel Dagenais

Polytechnique Montréal
Laboratoire DORSAL

Introduction


wsp_mcux - Develop - MK64FN1M0xxx12_Project_Demo/Generated_Code/timers.c - MCUXpresso IDE
File Edit Source Refactor Navigate Search Project Run FreeRTOS Window Help
Quick Access
Projec... Regist... Symb...
Debug
Memory
Monitors
Peripherals+
FAT1.c CLS1.c tasks.c Shell.c port.c Application.c Events.c timers.c
Disassembly
Enter location here
Task L... Prope... Probl... Instru... SWO ... Power... Queue... Console
Installed SDKs
New Renderings...
Q G V B O

- LPC54608J512_Project
- MK64FN1M0xxx12_Project
  - Includes
  - CMSIS
  - drivers
  - frdmk64f
  - freertos
  - source
    - MK64FN1M0xxx12_Project.c
  - startup
    - startup_mk64f12.c
  - utilities
    - fsl_debug_console.c
    - fsl_debug_console.h
  - Debug
  - doc
  - MK64FN1M0xxx12_Project PE Debug.launc
  - MK64FN1M0xxx12_Project PE Release.laun
- MK64FN1M0xxx12_Project_Demo
  - Binaries
  - Includes
  - CMSIS
  - Debug
  - doc

New project...
Import SDK example(s)...
Import project(s)
Build 'MK64FN1M0xxx12_Project_Demo' [Debu
Clean 'MK64FN1M0xxx12_Project_Demo' [Debu
Debug 'MK64FN1M0xxx12_Project_Demo' [Deb
Terminate, Build and Debug 'MK64FN1M0xxx12
Edit 'MK64FN1M0xxx12_Project_Demo' project
Quick Settings>>
Export projects to archive (zip)
Export projects and references to archive (zip)

▾ Manage SDK
Install/Delete SDK...
Use existing SDK...

▾ Extras
Smart Import wizard
Show welcome page
Build all projects [Debug]
Additional resources>>
Import project(s) from XML Description

```
MK64FN1M0xxx12_Project_Demo JLink Debug [GDB Segger Interface Debugging]
  MK64FN1M0xxx12_Project_Demo.axf
    Thread #3 536881672 (Shell [P: 1]) (Suspended : Container)
    Thread #4 536887400 (IDLE : Running [P: 0]) (Suspended : Signal : SIGTRAP:Trace/breakpoint trap)
      FRTOS1_vApplicationIdleHook() at Events.c:120 0x10fa
      prvIdleTask() at tasks.c:3'327 0x18284
      0x0
    Thread #5 536885384 (Main [P: 0]) (Suspended : Container)
      vPortYieldFromISR() at port.c:947 0x16660
      vTaskDelay() at tasks.c:1'374 0x179f0
      MainTask() at Application.c:271 0x102c
      0x0
    Thread #6 536877560 (TzCtrl [P: 1]) (Suspended : Container)
    Thread #7 536889112 (Tmr Svc [P: 2]) (Suspended : Container)
      vPortYieldFromISR() at port.c:947 0x16660
      prvProcessTimerOrBlockTask() at timers.c:618 0x18e78
      prvTimerTask() at timers.c:566 0x18dfe
      0x0
  arm-none-eabi-gdb (7.12.0.20161204)
```

- GPIOA [MK64FN1M0xxx12]
- GPIOD [MK64FN1M0xxx12]
- I2C0 [MK64FN1M0xxx12]

I2C0: 0x40066000 [MK64FN1M0xx

Register	Address	Value
I2C0	0x40066000	
A1	0x40066000	0x0
F	0x40066001	0x2f
C1	0x40066002	0xc0
S	0x40066003	0x84
D	0x40066004	0x7c
C2	0x40066005	0x0
AD	[2:0]	0x0
RMEN	[3:3]	Value_0
SBRC	[4:4]	Value_0
HDRS	[5:5]	Value_0
ADEXT	[6:6]	Value_0
GCAEN	[7:7]	Value_0
FLT	0x40066006	0x50
RA	0x40066007	0x0
SMB	0x40066008	0x0

Peripheral	Address	Description
□ ADC0	0x4003b000	Analog-to-Digi...
□ ADC1	0x400bb000	Analog-to-Digi...
□ AIPS0	0x40000000	AIPS-Lite Bridge
□ AIPS1	0x40080000	AIPS-Lite Bridge
□ AXBS	0x40004000	Crossbar switch
□ CAN0	0x40024000	Flex Controller ...
□ CAU	0xe0081000	Memory Mapp...
□ CMP0	0x40073000	High-Speed Co...
□ CMP1	0x40073008	High-Speed Co...
□ CMP2	0x40073010	High-Speed Co...
□ CMT	0x40062000	Carrier Modula...
□ CRC	0x40032000	Cyclic Redunda...
□ DAC0	0x400cc000	12-Bit Digital-t...
□ DAC1	0x400cd000	12-Bit Digital-t...
□ DMA	0x40008000	Enhanced direc...
□ DMAMUX	0x40021000	DMA channel ...

```c
    for( ;; )
    {
        /* Query the timers list to see if it contains any timers, and if so,
        obtain the time at which the next timer will expire. */
        xNextExpireTime = prvGetNextExpireTime( &xListWasEmpty );

        /* If a timer has expired, process it.  Otherwise, block this task
        until either a timer does expire, or a command is received. */
        prvProcessTimerOrBlockTask( xNextExpireTime, xListWasEmpty );

        /* Empty the command queue. */
        prvProcessReceivedCommands();
    }
}
/*-----------------------------------------------------------*/

static void prvProcessTimerOrBlockTask( const TickType_t xNextExpireTime, BaseType_t xListWasEmpty )
{
TickType_t xTimeNow;
BaseType_t xTimerListsWereSwitched;

    vTaskSuspendAll();
    {
        /* Obtain the time now to make an assessment as to whether the timer
        has expired or not.  If obtaining the time causes the lists to switch
```

```
00018dfe:   bl      0x18f9c <prvProcessReceivedCommands>
 570            }
00018e02:   b.n     0x18de8 <prvTimerTask+8>
 575        {
          prvProcessTimerOrBlockTask:
00018e04:   push    {r7, lr}
00018e06:   sub     sp, #16
00018e08:   add     r7, sp, #0
00018e0a:   str     r0, [r7, #4]
00018e0c:   str     r1, [r7, #0]
 579          vTaskSuspendAll();
00018e0e:   bl      0x17b10 <vTaskSuspendAll>
 586                xTimeNow = prvSampleTimeNow( &xTimerListsWereSwitcl
00018e12:   add.w   r3, r7, #8
00018e16:   mov     r0, r3
00018e18:   bl      0x18ed8 <prvSampleTimeNow>
00018e1c:   str     r0, [r7, #12]
 587                if( xTimerListsWereSwitched == pdFALSE )
00018e1e:   ldr     r3, [r7, #8]
00018e20:   cmp     r3, #0
00018e22:   bne.n   0x18e7a <prvProcessTimerOrBlockTask+118>
 590                    if( ( xListWasEmpty == pdFALSE ) && ( xNextExp:
00018e24:   ldr     r3, [r7, #0]
```

TCB#	Task Name	Task Handle	Task State	Prio...	Stack Usage	Event Object	Runtime
1	TzCtrl	0x200019f8	Blocked	1 (1)	136 B / 1.55 kB		0x0 (0.0%)
2	MidiPlay	0x20002100	□ Suspended	1 (1)	144 B / 1.55 kB		0x0 (0.0%)
3	Shell	0x20002a08	Blocked	1 (1)	384 B / 2.14 kB		0x60e (43.2%)
4	Main	0x20003888	Blocked	0 (0)	268 B / 3.51 kB		0x0 (0.0%)
5	IDLE	0x20004068	▶ Running	0 (0)	80 B / 1.55 kB		0x7f8 (56.8%)
6	Tmr Svc	0x20004718	Blocked	2 (2)	192 B / 1.55 kB	TmrQ (Rx)	0x0 (0.0%)
	Task Number:	0x0					
	Stack Base:	0x200040d0					
	Stack Top:	0x20004708					
	Stack High Wate...	0x20004648					

Installed SDKs

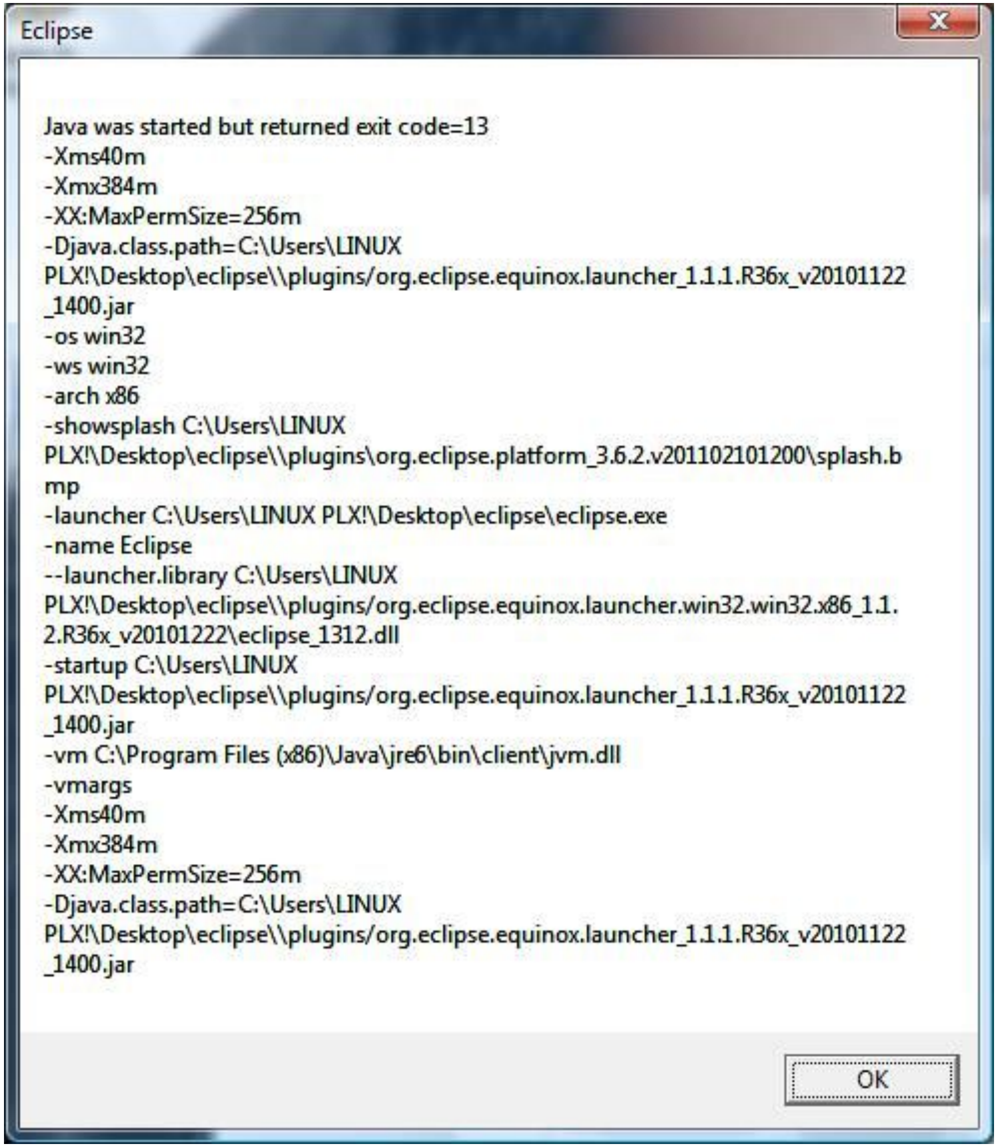
To install an SDK, simply drag and drop an SDK (zip file/folder) into the 'Installed SDKs' view.

Name	Version	Location
☑ SDK_2.x_LPCXpresso54114	2.2.0	<Default Location>/SDK_2.x_LPCXpresso5
☑ SDK_2.x_FRDM-KW41Z	2.2.0	<Default Location>/SDK_2.x_FRDM-KW41
☑ SDK_2.x_FRDM-KL27Z	2.2.0	<Default Location>/SDK_2.x_FRDM-KL27Z
☑ SDK_2.x_TWR-KV58F220M	2.2.0	<Default Location>/SDK_2.x_TWR-KV58F2
☑ SDK_2.x_FRDM-K64F	2.2.0	<Default Location>/SDK_2.x_FRDM-K64F
☑ SDK_2.x_FRDM-KL25Z	2.2.0	<Default Location>/SDK_2.x_FRDM-KL25Z
☑ SDK_2.x_LPCXpresso54608	2.2.0	<Default Location>/SDK_2.x_LPCXpresso5
☑ SDK_2.x_FRDM-K66F	2.2.0	<Default Location>/SDK_2.x_FRDM-K66F
☑ SDK_2.x_FRDM-K22F	2.2.0	<Default Location>/SDK_2.x_FRDM-K22F

Boards	
LPCXpresso54114	1.0.0
Devices	
LPC54114	1.0.0
Compilers	
Toolchains	
Toolchain Settings	
mcuxpresso	1.0.0
Components	

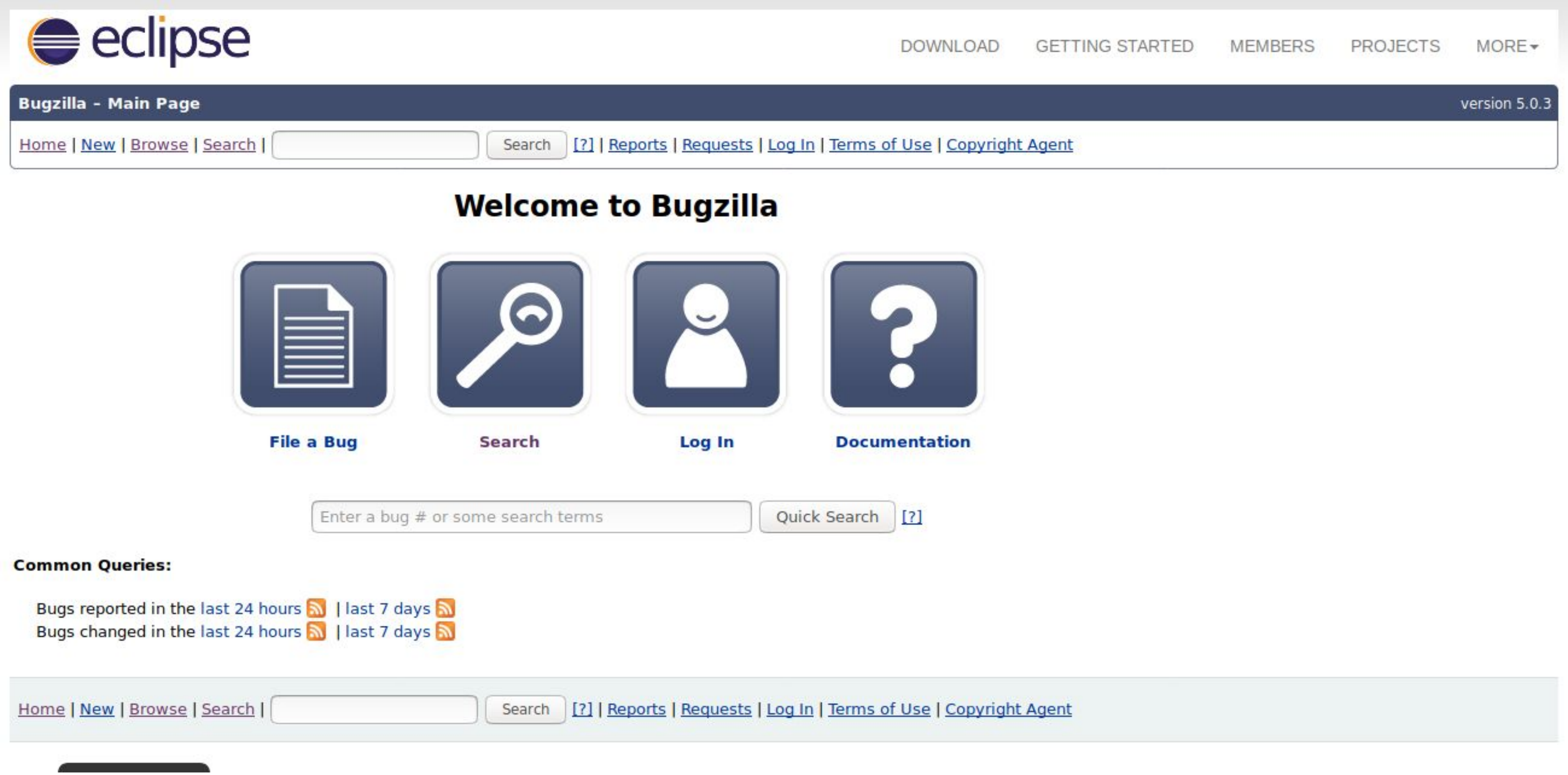
NXP MK64FN1M0xxx12 (MK64F... Demo)

Introduction



```
Java was started but returned exit code=13
-Xms40m
-Xmx384m
-XX:MaxPermSize=256m
-Djava.class.path=C:\Users\LINUX
PLX!\Desktop\eclipse\\plugins/org.eclipse.equinox.launcher_1.1.1.R36x_v20101122
_1400.jar
-os win32
-ws win32
-arch x86
-showsplash C:\Users\LINUX
PLX!\Desktop\eclipse\\plugins\org.eclipse.platform_3.6.2.v201102101200\splash.b
mp
-launcher C:\Users\LINUX PLX!\Desktop\eclipse\eclipse.exe
-name Eclipse
--launcher.library C:\Users\LINUX
PLX!\Desktop\eclipse\\plugins/org.eclipse.equinox.launcher.win32.win32.x86_1.1.
2.R36x_v20101222\eclipse_1312.dll
-startup C:\Users\LINUX
PLX!\Desktop\eclipse\\plugins/org.eclipse.equinox.launcher_1.1.1.R36x_v20101122
_1400.jar
-vm C:\Program Files (x86)\Java\jre6\bin\client\jvm.dll
-vmargs
-Xms40m
-Xmx384m
-XX:MaxPermSize=256m
-Djava.class.path=C:\Users\LINUX
PLX!\Desktop\eclipse\\plugins/org.eclipse.equinox.launcher_1.1.1.R36x_v20101122
_1400.jar
```

Bug Tracking System



Bugzilla – Main Page version 5.0.3

Home | New | Browse | Search | [Search] [?] | Reports | Requests | Log In | Terms of Use | Copyright Agent

Welcome to Bugzilla

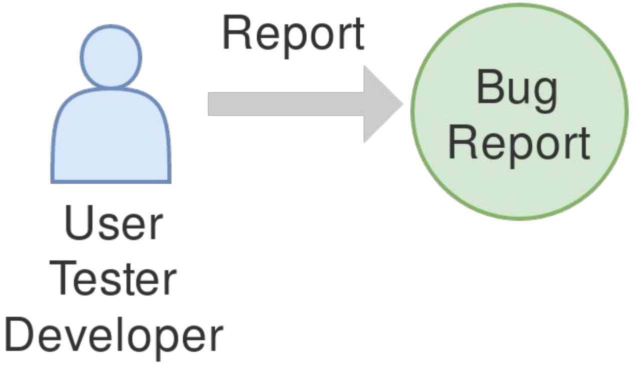
Enter a bug # or some search terms [Quick Search] [?]

Common Queries:

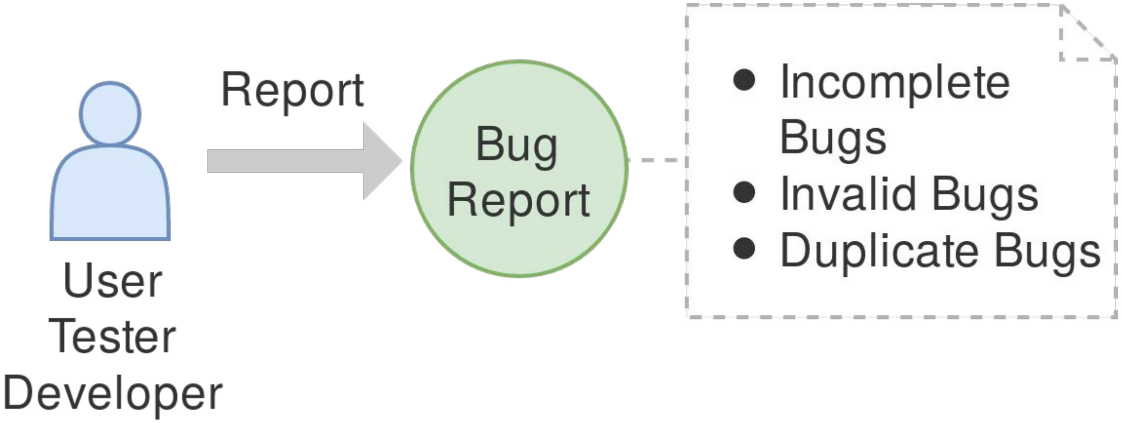
Bugs reported in the last 24 hours | last 7 days
Bugs changed in the last 24 hours | last 7 days

Home | New | Browse | Search | [Search] [?] | Reports | Requests | Log In | Terms of Use | Copyright Agent

Bug Tracking System

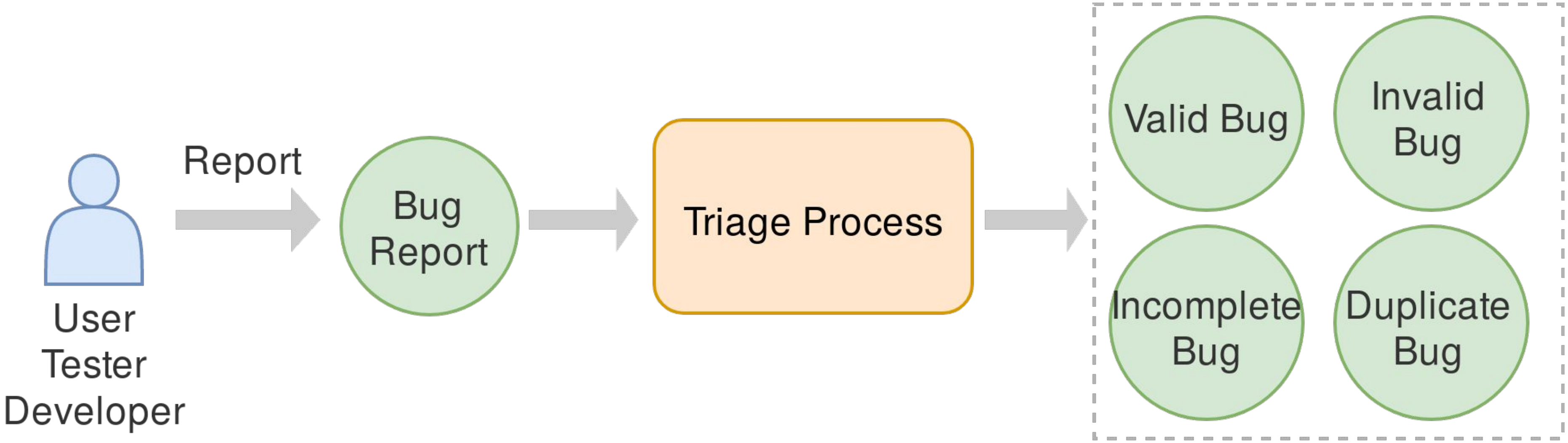


Bug Tracking System



- Incomplete Bugs
- Invalid Bugs
- Duplicate Bugs

Bug Tracking System

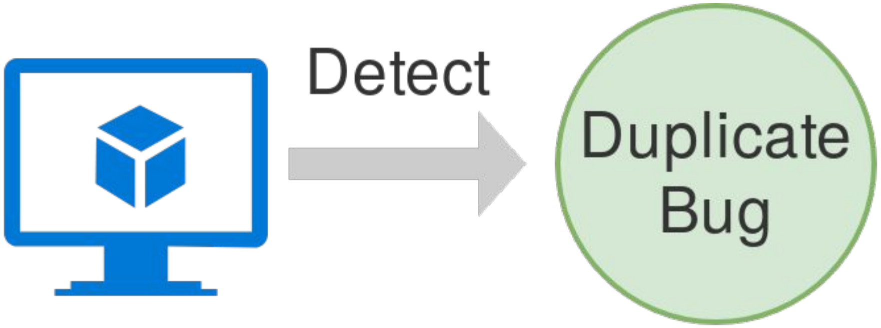


Bug Tracking System

Triage Process

- Manual checking
- Time and money consuming
- Large user base project: Firefox ~300 new reports per day

Objective



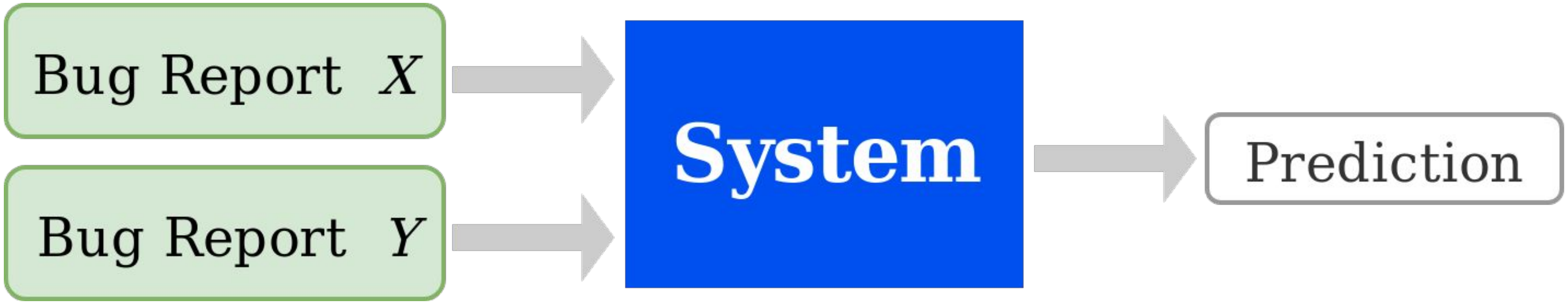
- Increase software quality and save resource
  - Decrease triage team overload
  - Avoid two or more developers fixing the same bug
  - Avoid to fix a bug already solved

Duplicate bug report detection

- Detect whether a bug is duplicate or not
- Master set
  - Master report
  - Duplicate reports
  - Every report is in a master set
- Three approaches
  - Decision-making approach
  - Binary classification approach
  - Ranking approach

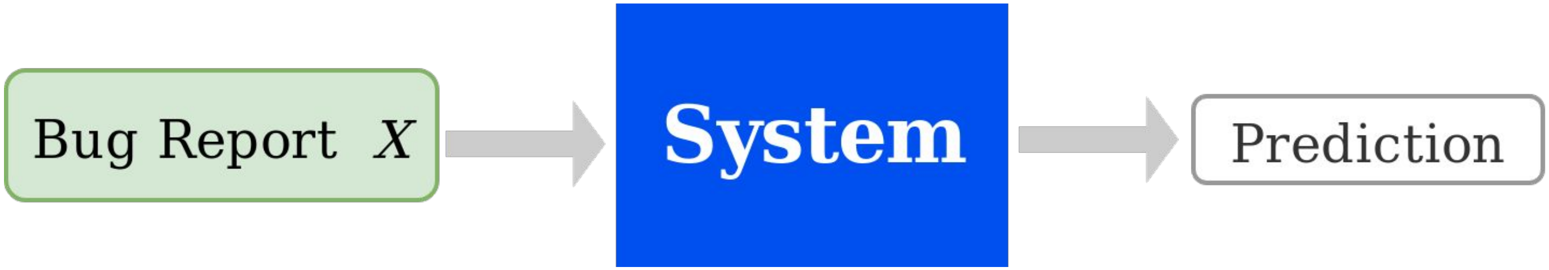
Decision-making approach

- Pairs of bug reports (Training and Evaluation)
- Drawbacks
  - Too Easy
  - High probability to create easy non-duplicate pairs
  - Far from the real scenario
    - Compare new bug with a set of bugs in the dataset



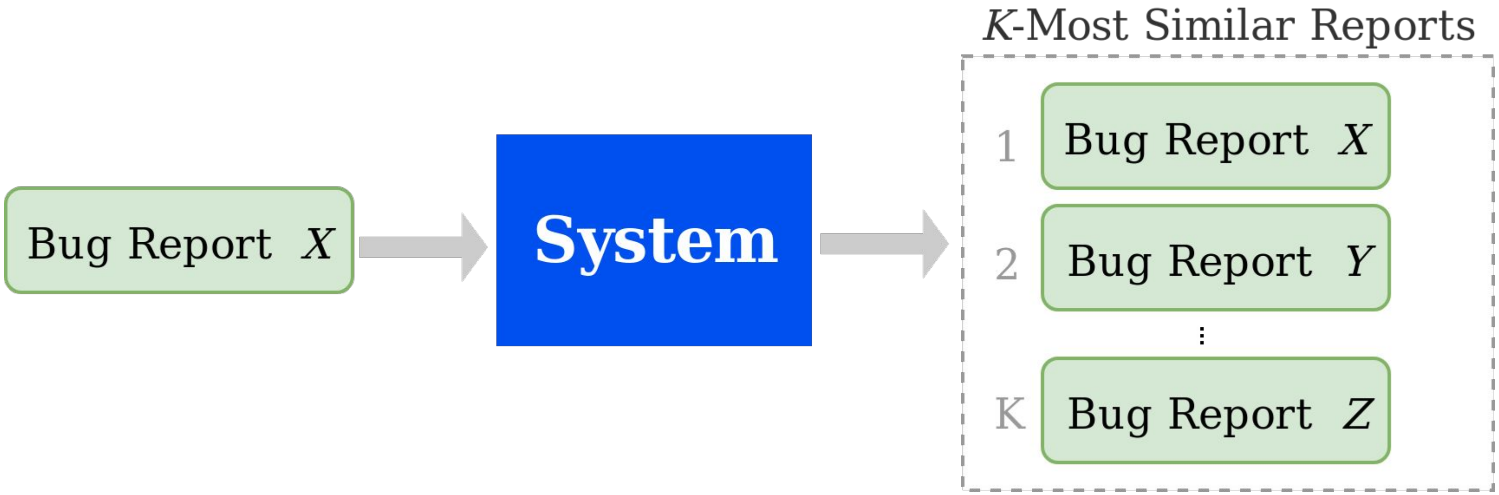
Binary classification approach

- Automatic prediction of the report as duplicate or not
  - General information extracted from the database and the new bug reports
- False negative can have a great impact
- Really difficult task



Ranking approach

- Recommend a similarity list
- A person check the list and label the report as duplicate or not
  - Decrease the decision time
- The most used approach in the literature
- Metric: Recall Rate
  - Rate of reports whose the lists have at least one bug report from the same master set



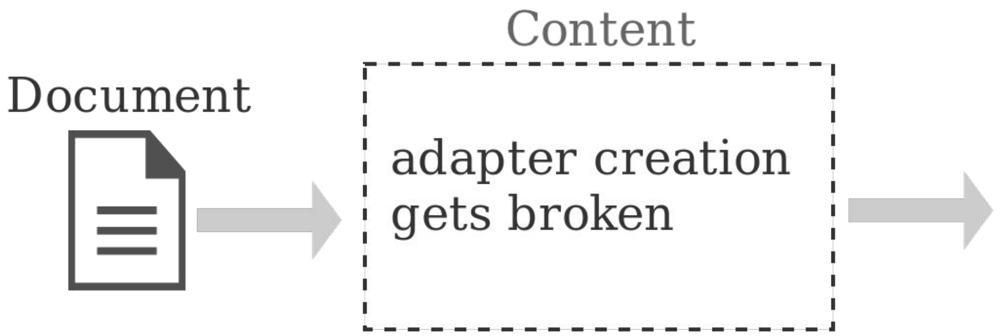
Ranking approach

- Two methodologies: Deshmukh et al. 2017 and Sun et al. 2011
- Deshmukh et al. 2017
  - Training, validation and test datasets are randomly generated
  - Evaluation: similarity list are created using bug from the test dataset
  - Unrealistic scenario
  - It makes the problem easier
    - Decrease number of comparisons
    - Concept Drift mitigation
- Sun et al. 2011
  - Reports are sorted by creation date
  - Training, validation and test are generate by period of time
  - New bug report is compared with all previous bug reports
  - More realistic scenario

Our Solution

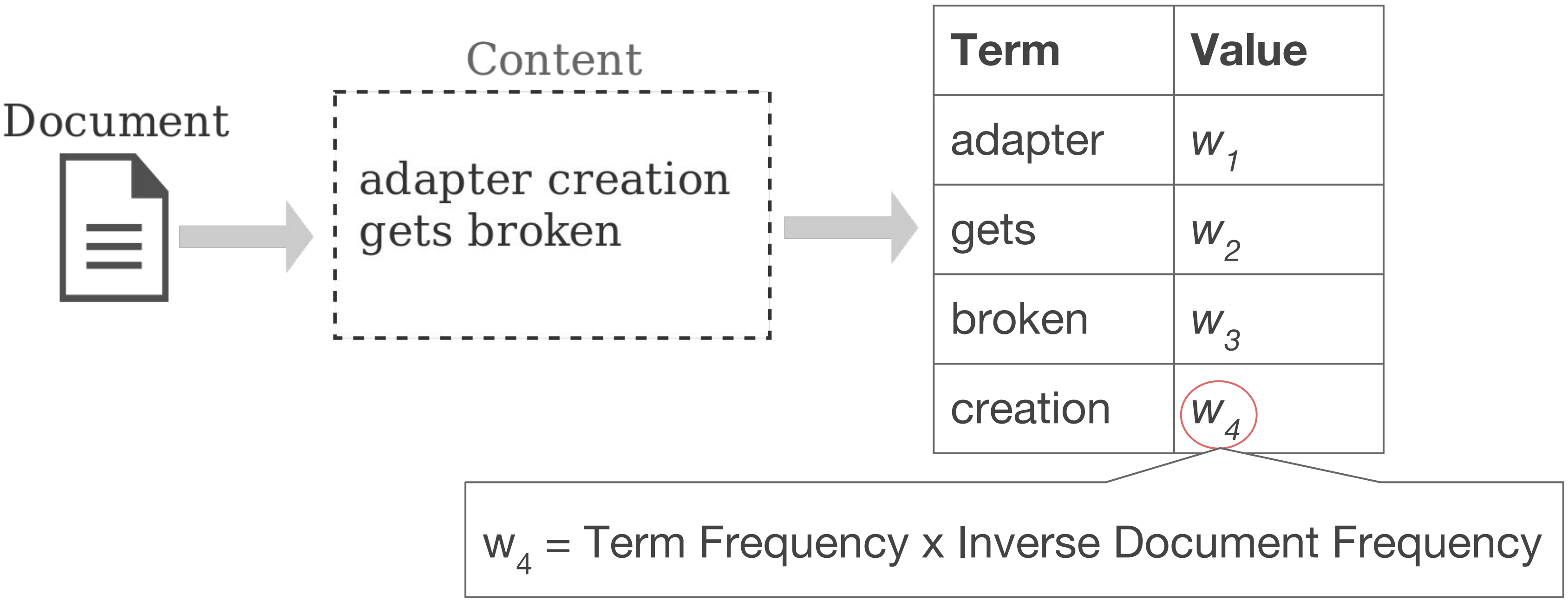
- Ranking approach + Sun’s Methodology
- Only textual data
    - Summary and description
- Baseline: TF-IDF
- Model: Word Embeddings + Convolution Neural Network

TF-IDF



Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

TF-IDF

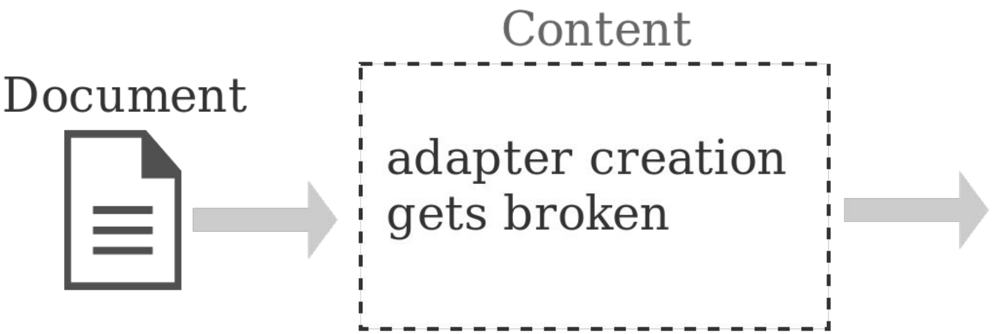


adapter creation gets broken

Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

w_4 = Term Frequency x Inverse Document Frequency

TF-IDF

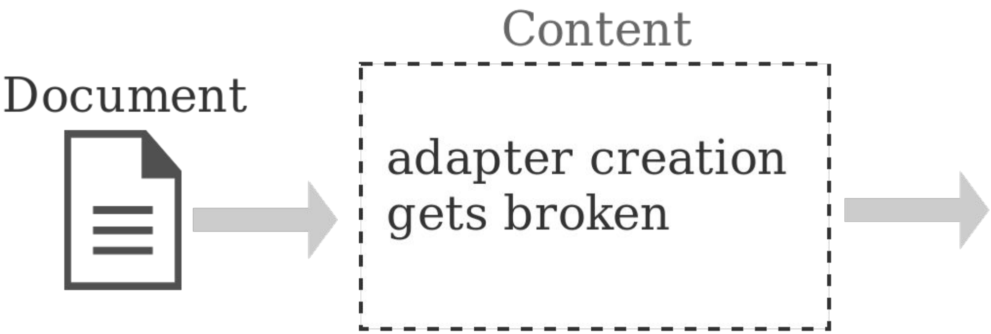


adapter creation
gets broken

Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

w_4 = Term Frequency x Inverse Document Frequency

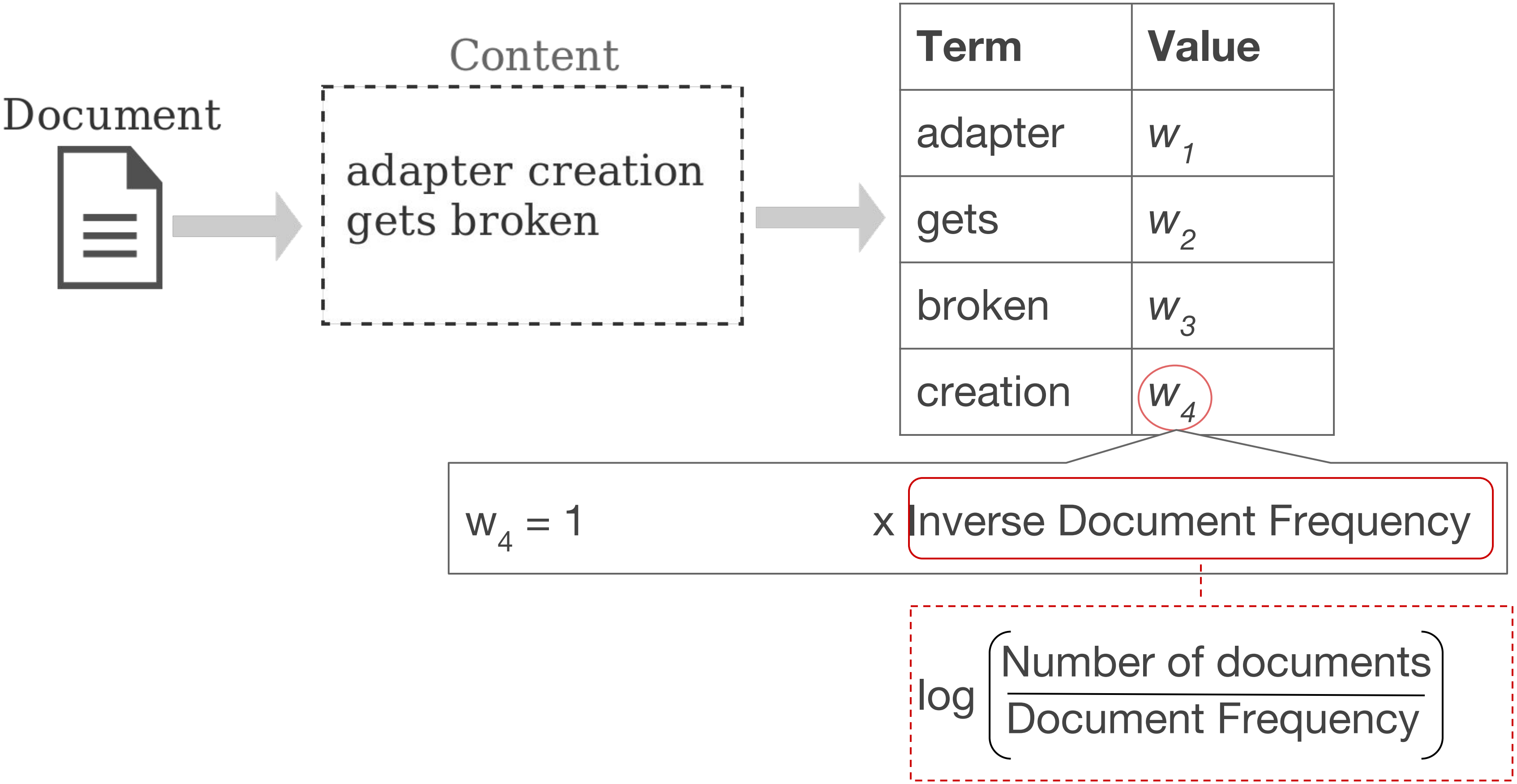
TF-IDF



Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

$w_4 = 1$ x Inverse Document Frequency

TF-IDF



Document

Content

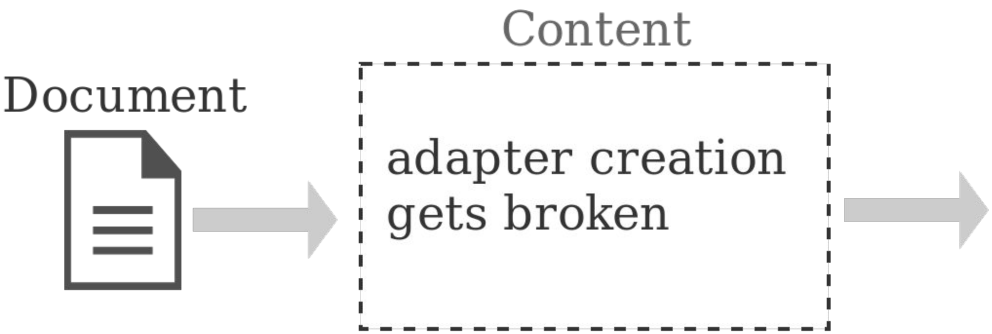
adapter creation gets broken

Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

$w_4 = 1 \times$ Inverse Document Frequency

$$\log\left(\frac{\text{Number of documents}}{\text{Document Frequency}}\right)$$

TF-IDF



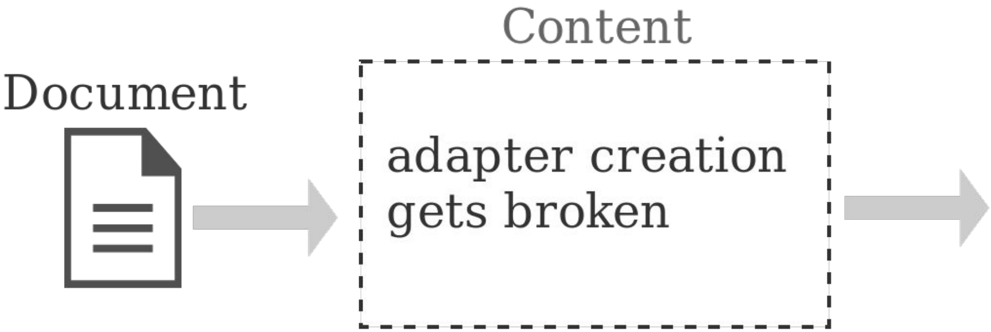
adapter creation gets broken

Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

$w_4 = 1$ x Inverse Document Frequency

$$\log\left(\frac{10}{8}\right)$$

TF-IDF



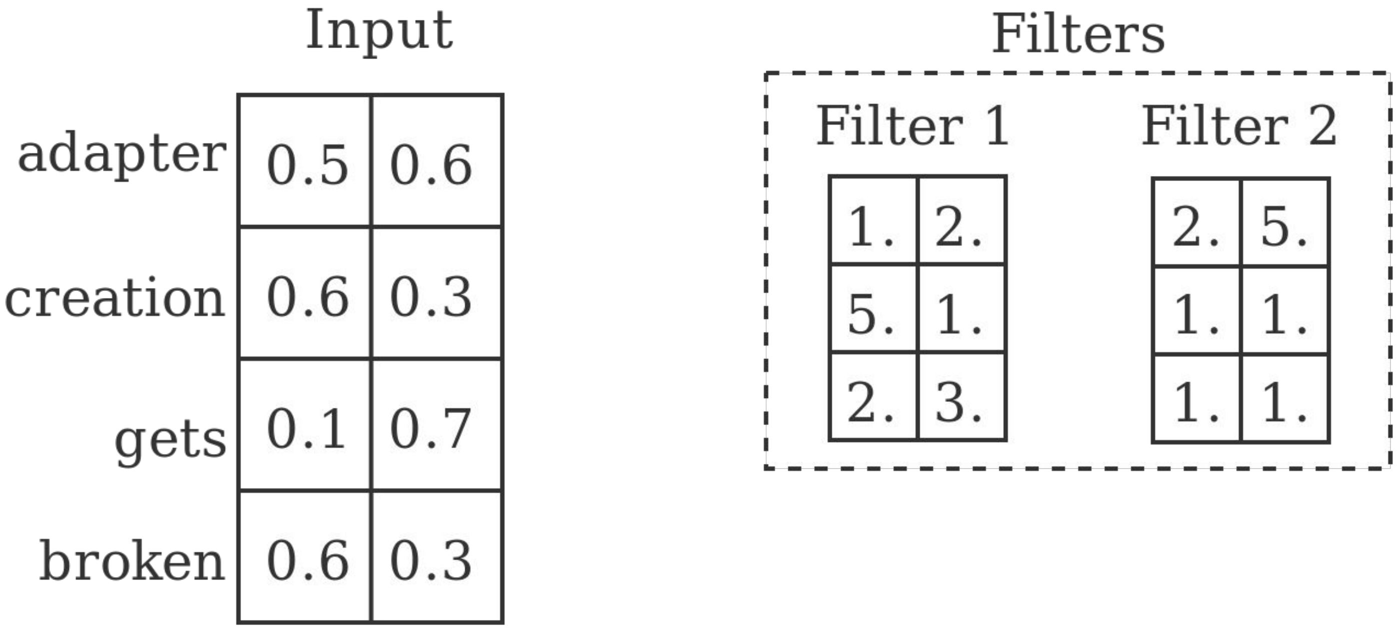
Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	0.09

Represent word as vector

- Word Embedding
  - Dense vectors with real numbers
  - More compact representation
  - Semantic and syntactic information

Word	Representation
adapter	[0.5, 0.6]
broken	[0.3, 0.2]
gets	[0.1, 0.7]
creation	[0.6, 0.3]

Convolution Neural Network for NLP



Input

adapter	0.5	0.6
creation	0.6	0.3
gets	0.1	0.7
broken	0.6	0.3

Filters

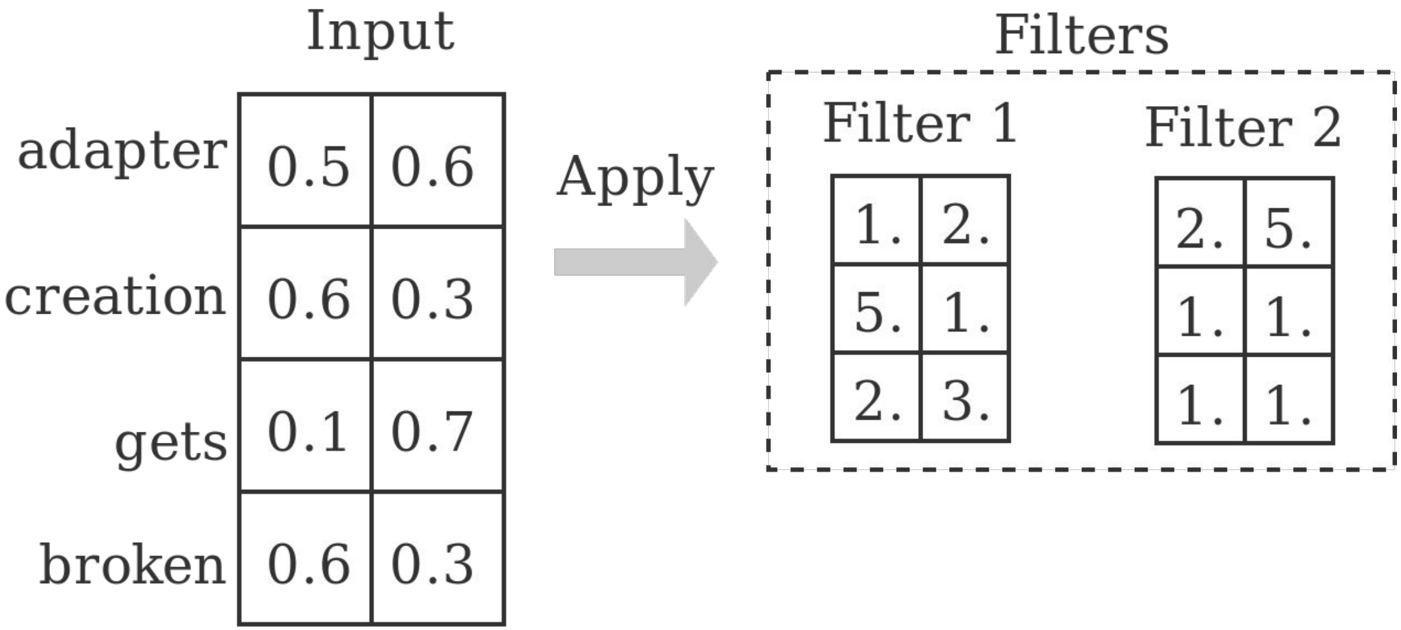
Filter 1

1.	2.
5.	1.
2.	3.

Filter 2

2.	5.
1.	1.
1.	1.

Convolution Neural Network for NLP



Input

adapter	0.5	0.6
creation	0.6	0.3
gets	0.1	0.7
broken	0.6	0.3

Filters

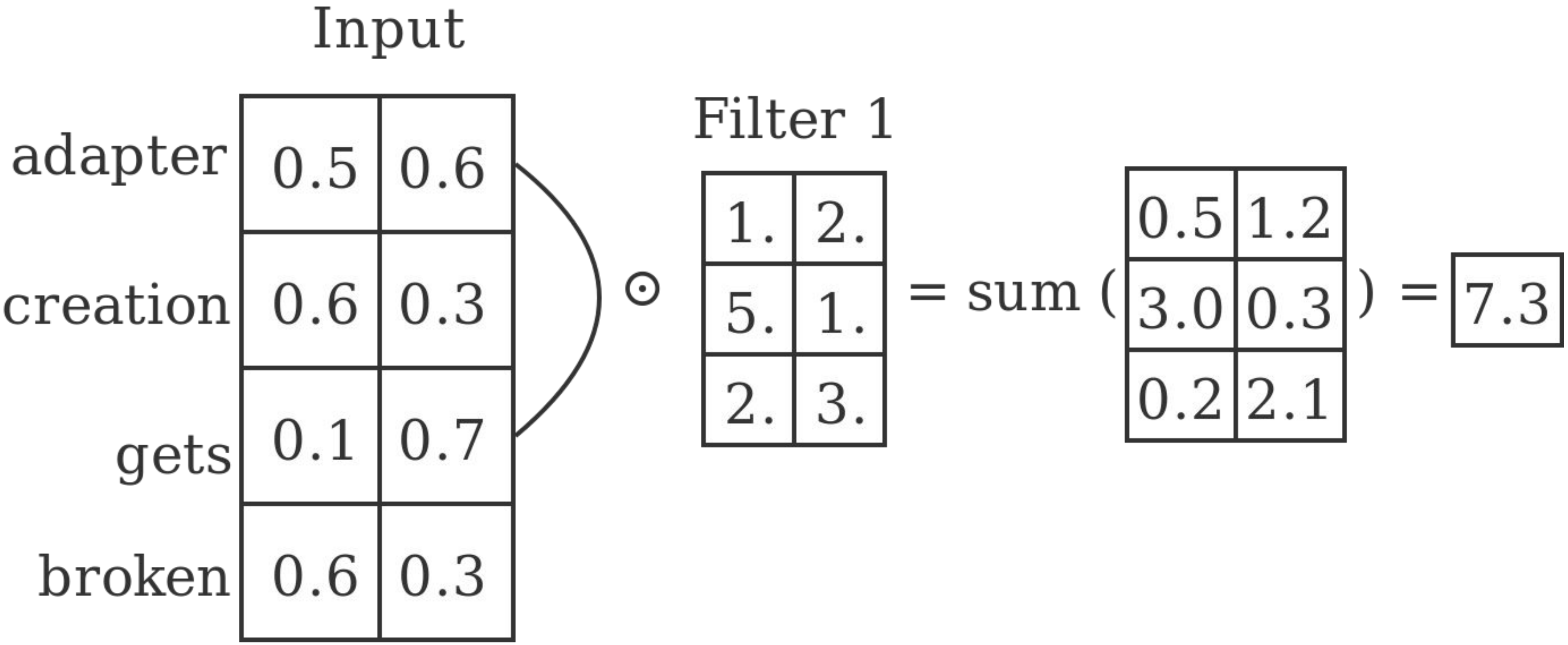
Filter 1

1.	2.
5.	1.
2.	3.

Filter 2

2.	5.
1.	1.
1.	1.

Convolution Neural Network for NLP



Input

adapter	0.5	0.6
creation	0.6	0.3
gets	0.1	0.7
broken	0.6	0.3

$\odot$

Filter 1

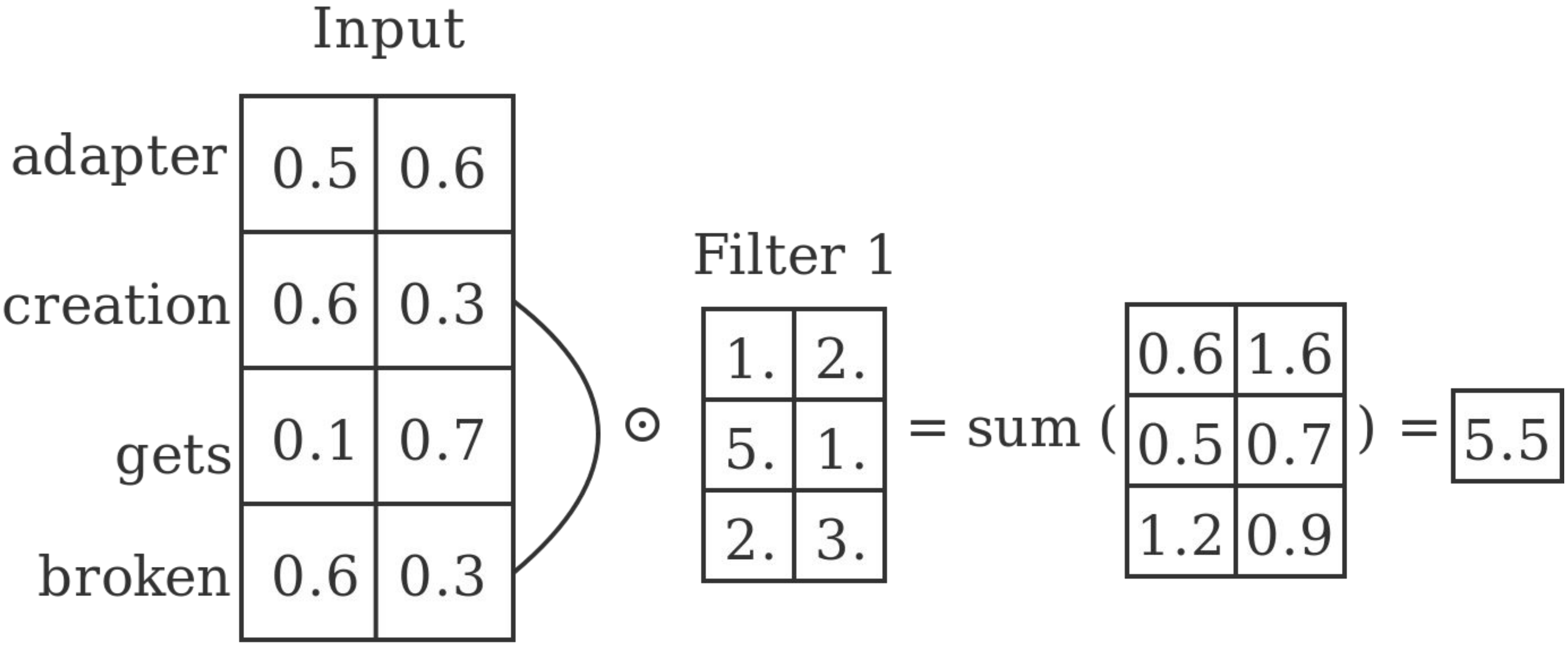
1.	2.
5.	1.
2.	3.

= sum (

0.5	1.2
3.0	0.3
0.2	2.1

) = $\boxed{7.3}$

Convolution Neural Network for NLP



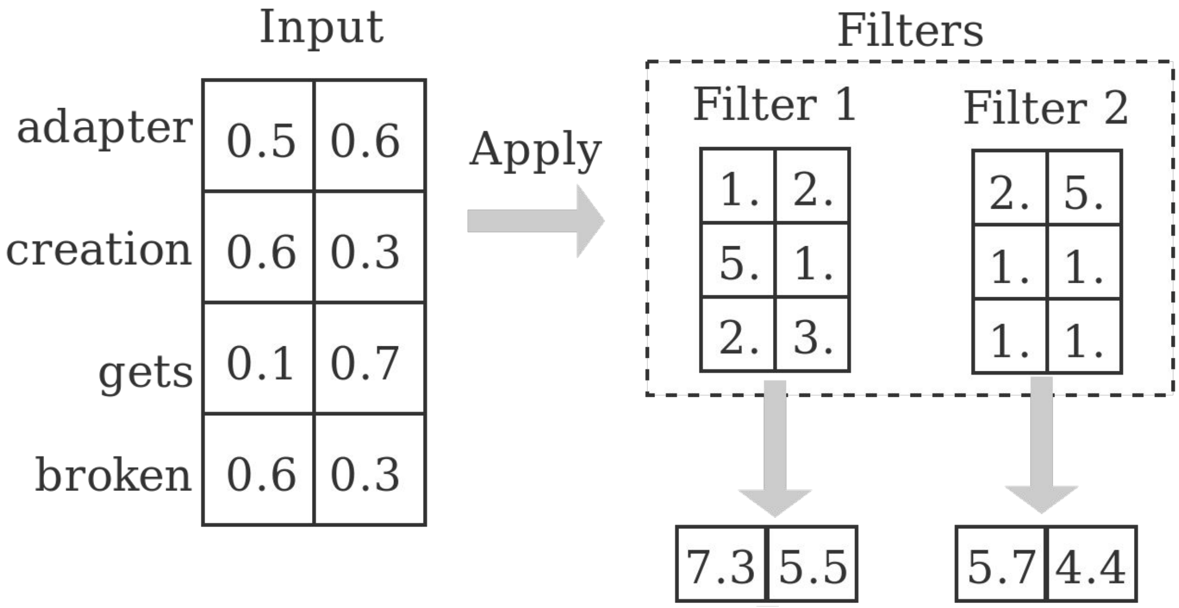
Input

$$\begin{array}{r|c|c|} \text{adapter} & 0.5 & 0.6 \\ \text{creation} & 0.6 & 0.3 \\ \text{gets} & 0.1 & 0.7 \\ \text{broken} & 0.6 & 0.3 \end{array}$$

Filter 1

$$\odot \begin{bmatrix} 1. & 2. \\ 5. & 1. \\ 2. & 3. \end{bmatrix} = \text{sum}\left(\begin{bmatrix} 0.6 & 1.6 \\ 0.5 & 0.7 \\ 1.2 & 0.9 \end{bmatrix}\right) = \boxed{5.5}$$

Convolution Neural Network for NLP



Input

adapter	0.5	0.6
creation	0.6	0.3
gets	0.1	0.7
broken	0.6	0.3

Apply

Filters

Filter 1

1.	2.
5.	1.
2.	3.

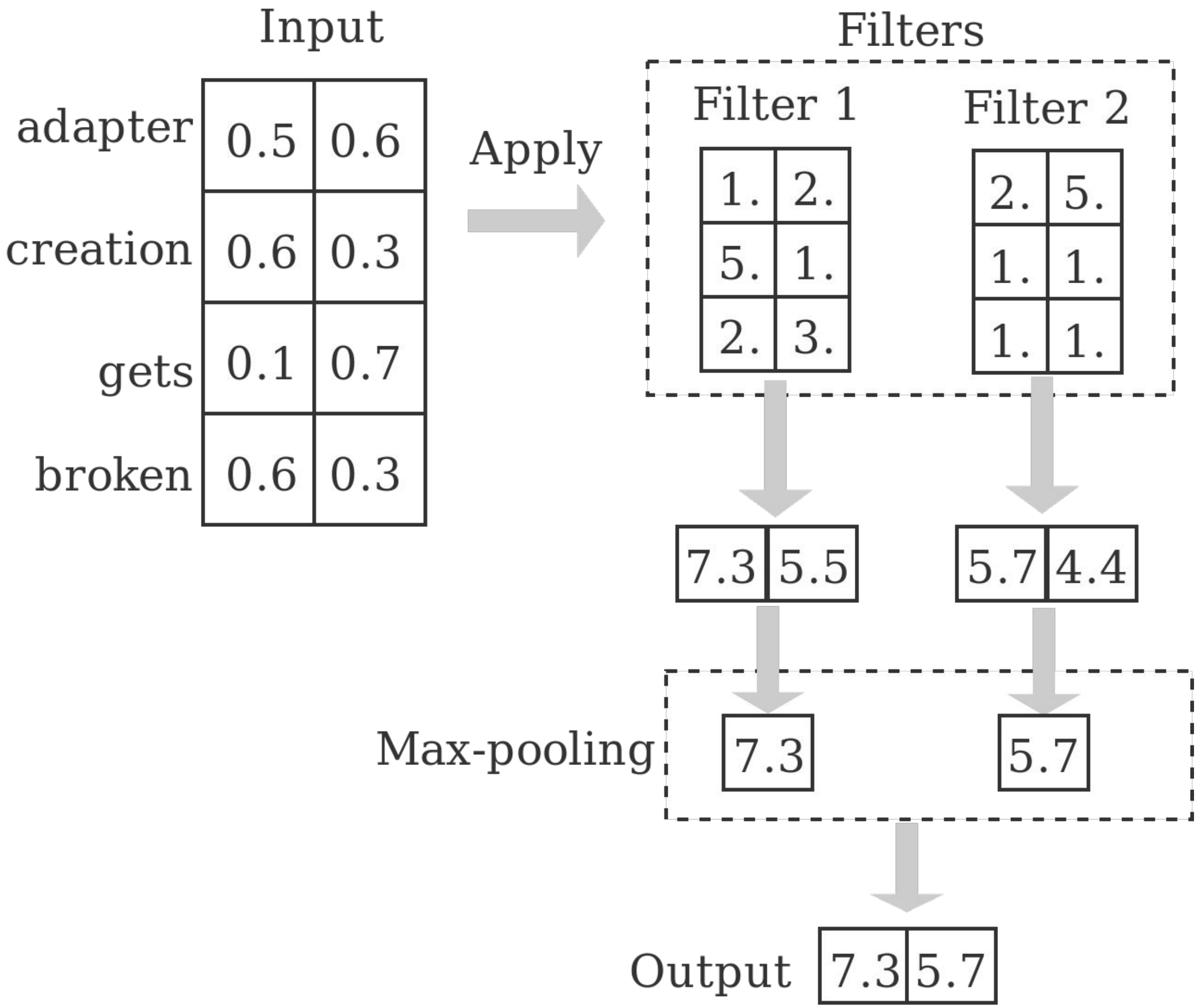
Filter 2

2.	5.
1.	1.
1.	1.

7.3	5.5

5.7	4.4

Convolution Neural Network for NLP



Input

adapter	0.5	0.6
creation	0.6	0.3
gets	0.1	0.7
broken	0.6	0.3

Apply

Filters

Filter 1

1.	2.
5.	1.
2.	3.

Filter 2

2.	5.
1.	1.
1.	1.

7.3 5.5

5.7 4.4

Max-pooling

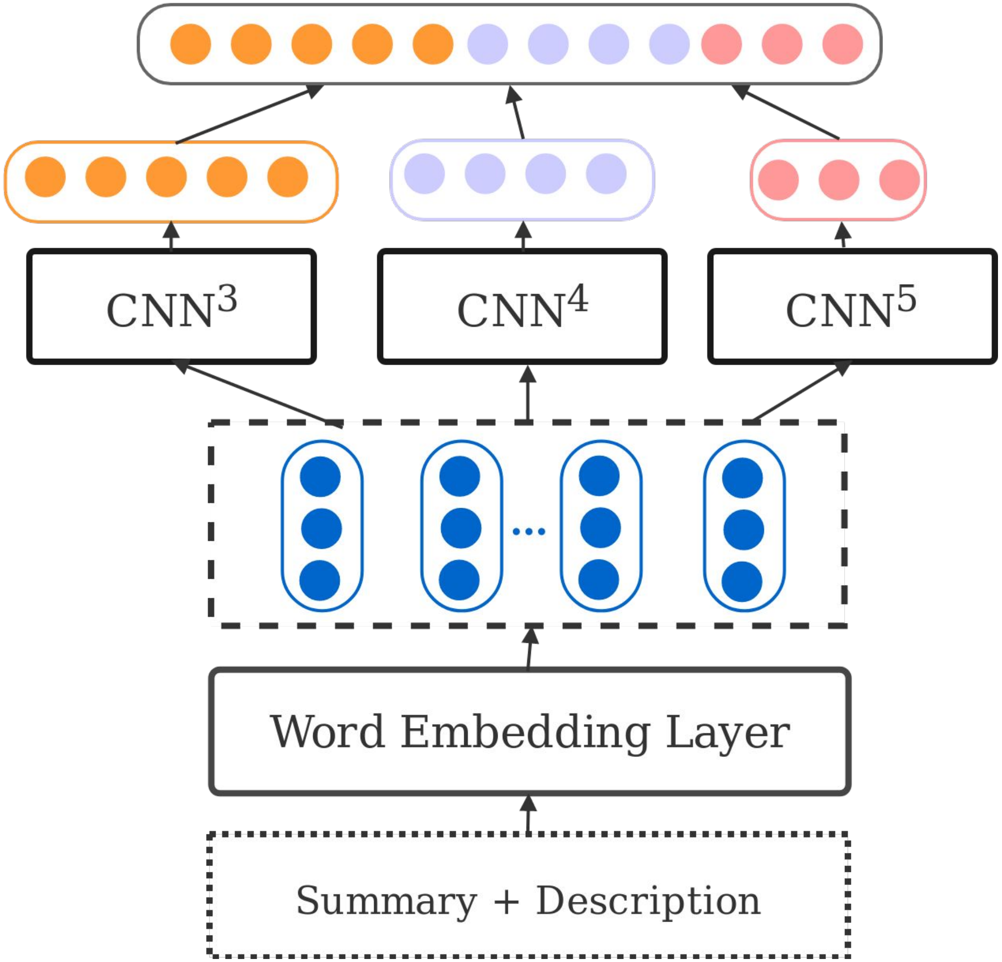
7.3

5.7

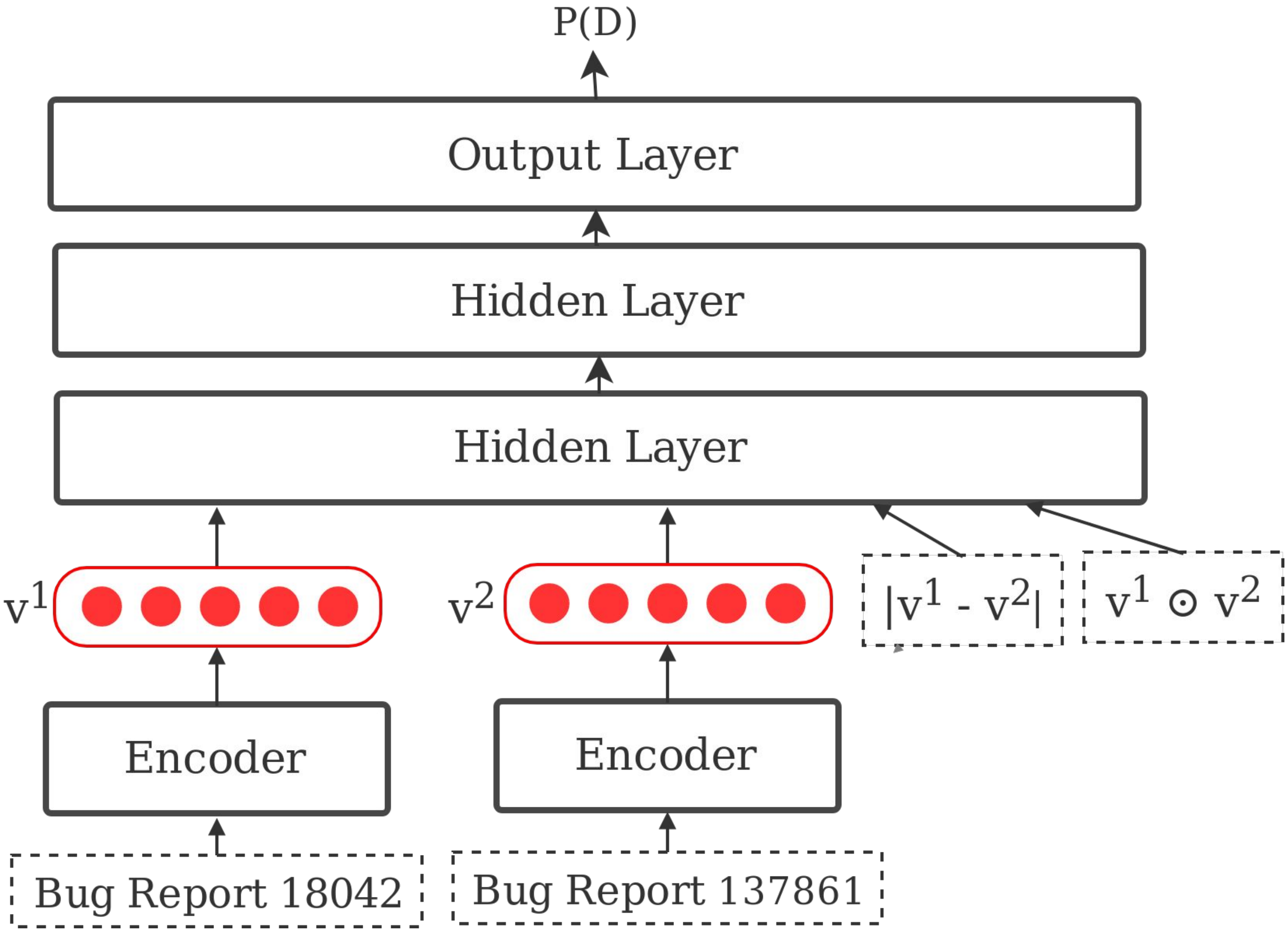
Output 7.3 5.7

Our Deep Learning Model

- Encoder
  - Represent the report as vector



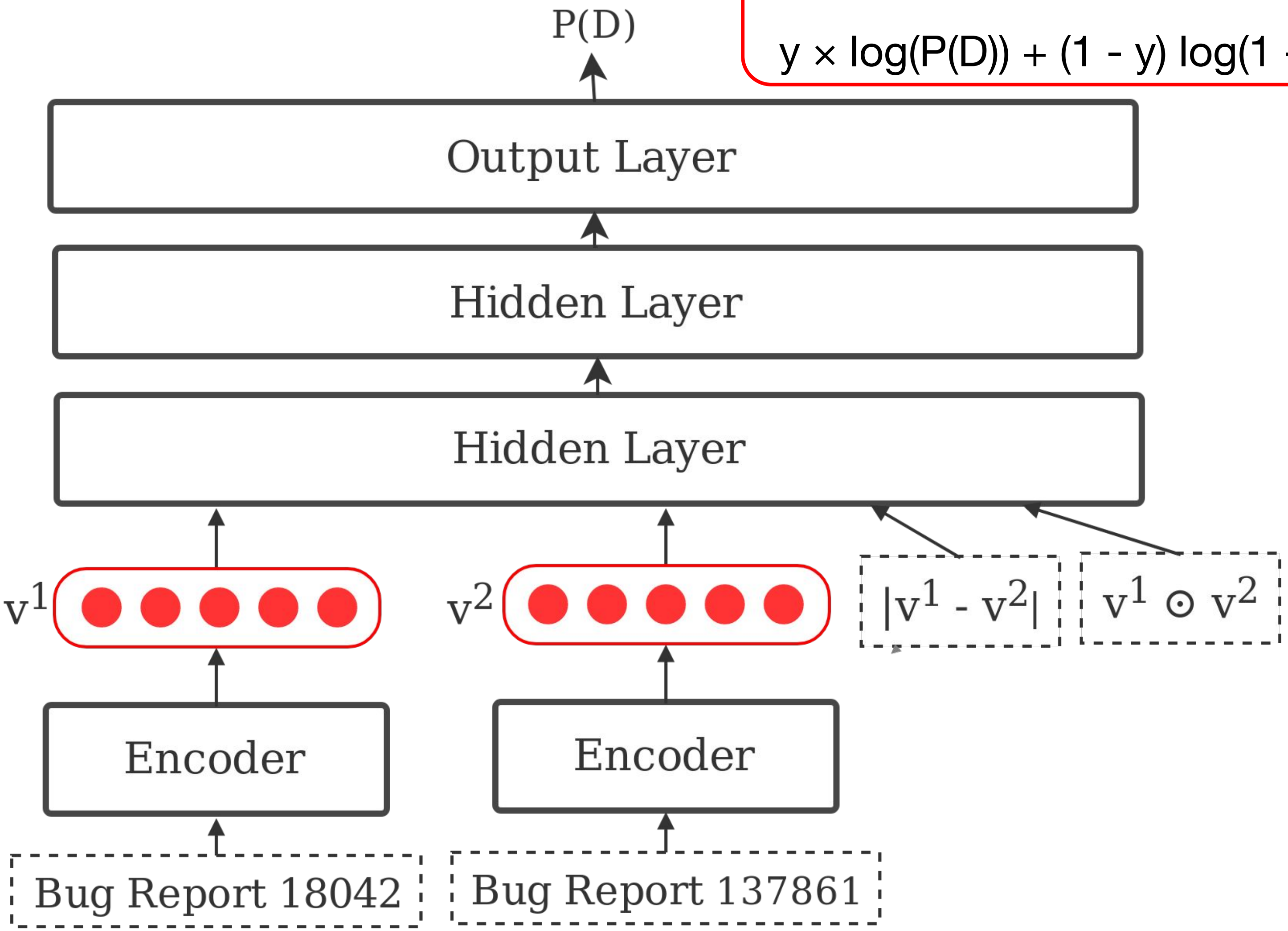
Our Deep Learning Model



Our Deep Learning Model

Cross Entropy

$y \times \log(P(D)) + (1 - y) \log(1 - P(D))$



Preliminar Results

Model	Top-5	Top-10	Top-15	Top-20
TF-IDF	44.80%	51.27%	54.97%	57.88%
DL Model	37.11%	43.95%	48.61%	52.03%

Our Deep Learning Model

- Challenge:
  - Generate relevant non-duplicate pairs (negative) can be difficult
  - Most non-duplicate pairs are easy
  - ~ n^2 different combinations
  - $n = 174{,}002 \Rightarrow n^2 \cong 30 \times 10^9$
- Solution: Random subsample negative examples each epoch
  - Constraint: loss has to be greater than 0
  - Keep rate between positive and negative examples

Preliminar Results

Model	**Top-5**	**Top-10**	**Top-15**	**Top-20**
TF-IDF	44.80%	51.27%	54.97%	57.88%
DL Model	37.11%	43.95%	48.61%	52.03%
DL Model - subsampling by epoch	44.02%	51.03%	55.49%	58.43%

Preliminar Results

Model	**Top-5**	**Top-10**	**Top-15**	**Top-20**
TF-IDF	44.80%	51.27%	54.97%	57.88%
DL Model	37.11%	43.95%	48.61%	52.03%
DL Model - subsampling by epoch	44.02%	51.03%	55.49%	58.43%

6.40%

Future Work

- Bottleneck: select negative pairs
  - Try different approaches
- Encoder receives information from the first bug
  - Attention
- Combine different information sources
  - Categorical information, stack trace, tracing
- Use our solution to help our partners
  - Partner data

Thank you for your attention!

Questions?

Irving Muller Rodrigues
irving.muller-rodrigues@polymtl.ca

References

- Deshmukh, J., M, A. K., Podder, S., Sengupta, S., & Dubash, N. (2017). Towards Accurate Duplicate Bug Retrieval Using Deep Learning Techniques. 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), 115–124. http://doi.org/10.1109/ICSME.2017.69
- Lazar, A., Ritchey, S., & Sharif, B. (2014). Generating duplicate bug datasets. Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014, 392–395. http://doi.org/10.1145/2597073.2597128
- Sabor, K. K., Hamou-Lhadj, A., & Larsson, A. (2017). DURFEX: A feature extraction technique for efficient detection of duplicate bug reports. Proceedings - 2017 IEEE International Conference on Software Quality, Reliability and Security, QRS 2017, 240–250. http://doi.org/10.1109/QRS.2017.35

References

- Anh Tuan Nguyen, Tung Thanh Nguyen, Tien N Nguyen, David Lo, and Chengnian Sun. Duplicate bug report detection with a combination of information retrieval and topic modeling. In Automated Software Engineering (ASE), 2012 Proceedings of the 27th IEEE/ACM International Conference on, pages 70–79. IEEE, 2012.
- Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber. LSTM: A Search Space Odyssey. CoRR abs/1503.04069 (2015)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

References

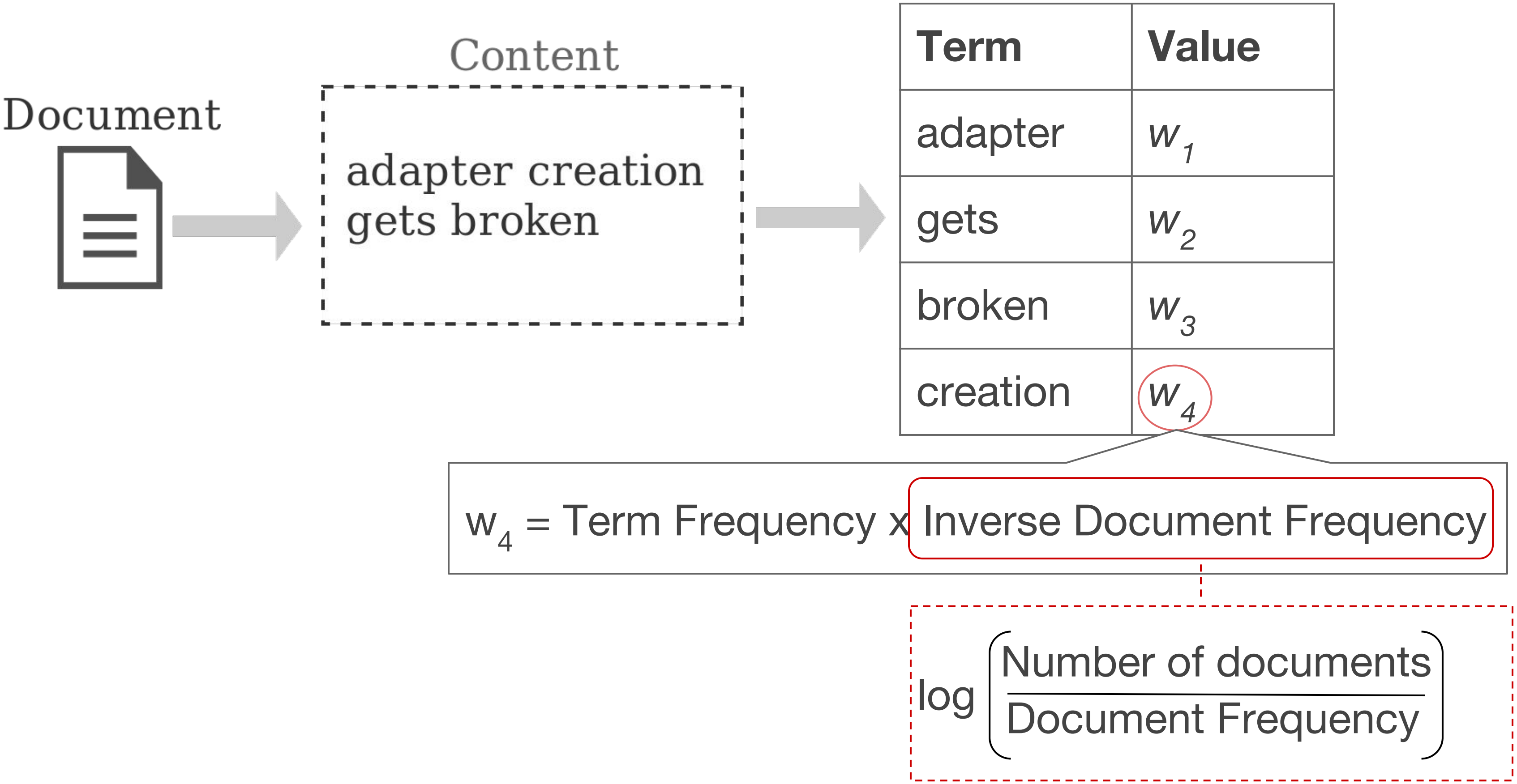
- Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." Paper presented at the meeting of the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 2014.
- C. Sun, D. Lo, S. Khoo and J. Jiang, "Towards more accurate retrieval of duplicate bug reports," 2011 26th IEEE/ACM International Conference on Automated Software Engineering (ASE 2011), Lawrence, KS, 2011, pp. 253-262.

Represent word as vector

- One hot encoding
  - Binary Vectors
  - Vector Size = Vocabulary Size
  - Curse of Dimensionality

Word	Representation
adapter	[1,0,0,0]
broken	[0,1,0,0]
gets	[0,0,1,0]
creation	[0,0,0,1]

TF-IDF



Document

Content

adapter creation gets broken

Term	Value
adapter	w_1
gets	w_2
broken	w_3
creation	w_4

w_4 = Term Frequency x Inverse Document Frequency

$$\log\left(\frac{\text{Number of documents}}{\text{Document Frequency}}\right)$$