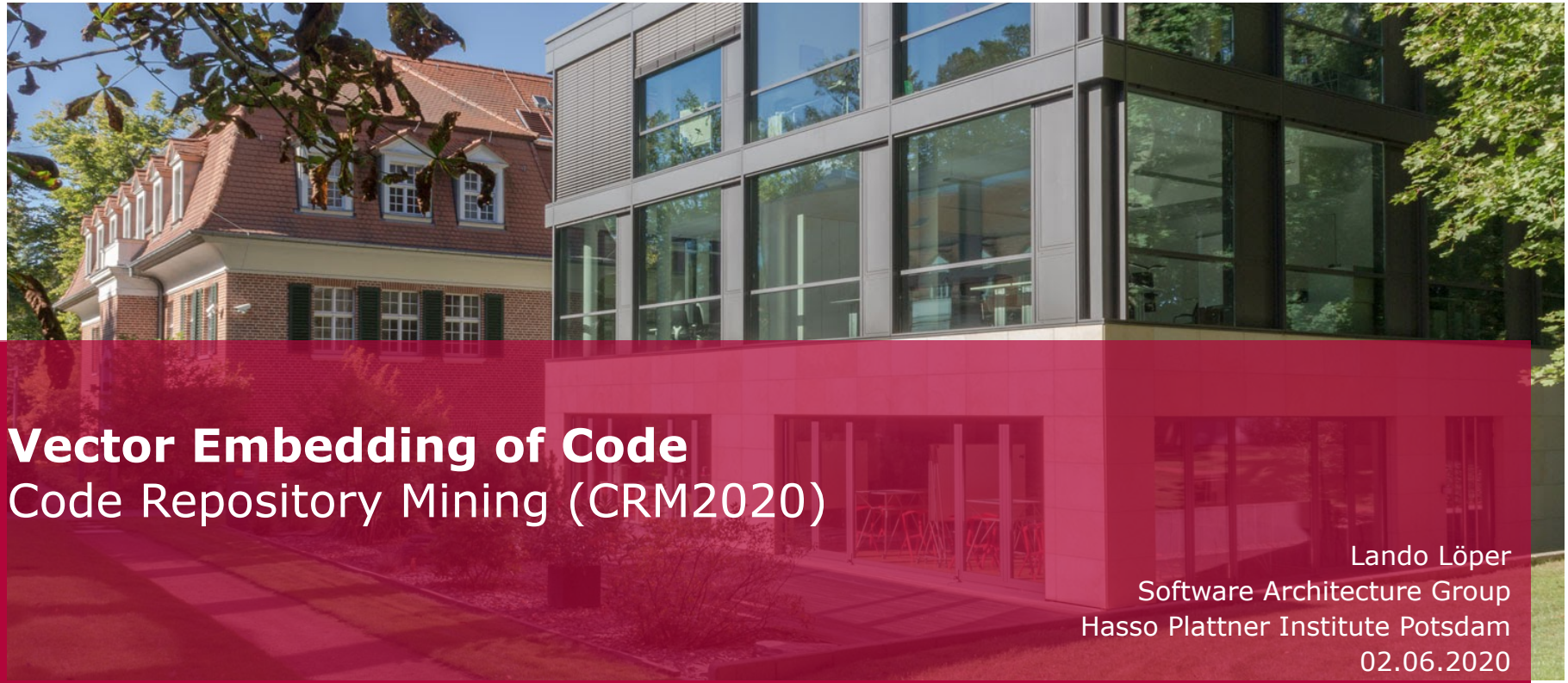


Vector Embedding of Code

Code Repository Mining (CRM2020)

Lando Löper
Software Architecture Group
Hasso Plattner Institute Potsdam
02.06.2020

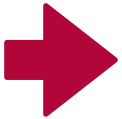
Agenda

1. Motivation
2. Related Work
3. Design Space
4. Data
5. Outlook
6. Discussion

Motivation

What does it do?

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

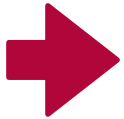


- contains
- count
- indexOf
- reverse

Motivation

What does it do?

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```



- contains
- count
- indexOf
- reverse

- Reading unlabelled code takes time

Motivation

How similar are they?

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

```
function(list, target)
    i := 0
    for  x in list
        if x = target
            i++
    return i
```

Motivation

How similar are they?

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

```
function(list, target)
    i := 0
    for  x in list
        if x = target
            i++
    return i
```

- Similar syntax does not mean similar behaviour

Motivation

How similar are they?

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

```
function(elements, x)
    i := 0
    n := size of elements
    while i < n
        if elements[i] = x
            return i
        i++
    return -1
```

Motivation

How similar are they?

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

```
function(elements, x)
    i := 0
    n := size of elements
    while i < n
        if elements[i] = x
            return i
        i++
    return -1
```

- Different syntax does not mean different behaviour

Motivation

Direct mapping

```
function(list, target)
    i := 0
    for  x in list
        if x = target
            i++
    return i
```

count

```
function(elements, x)
    i := 0
    n := size of elements
    while i < n
        if elements[i] = x
            return i
        i++
    return -1
```

indexOf

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

indexOf

Motivation

What about a different representation?

```
function(list, target)
    i := 0
    for  x in list
        if x = target
            i++
    return i
```

```
function(elements, x)
    i := 0
    n := size of elements
    while i < n
        if elements[i] = x
            return i
        i++
    return -1
```

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

Motivation

Indirect mapping

```
function(list, target)
    i := 0
    for  x in list
        if x = target
            i++
    return i
```

count

```
function(elements, x)
    i := 0
    n := size of elements
    while i < n
        if elements[i] = x
            return i
        i++
    return -1
```

indexOf

```
function(list, target)
    i := 0
    for x in list
        if x = target
            return i
        i++
    return -1
```

indexOf

Motivation

Embedding

Definition

An embedding is a mapping of a discrete — categorical — variable to a vector of continuous numbers

Benefits

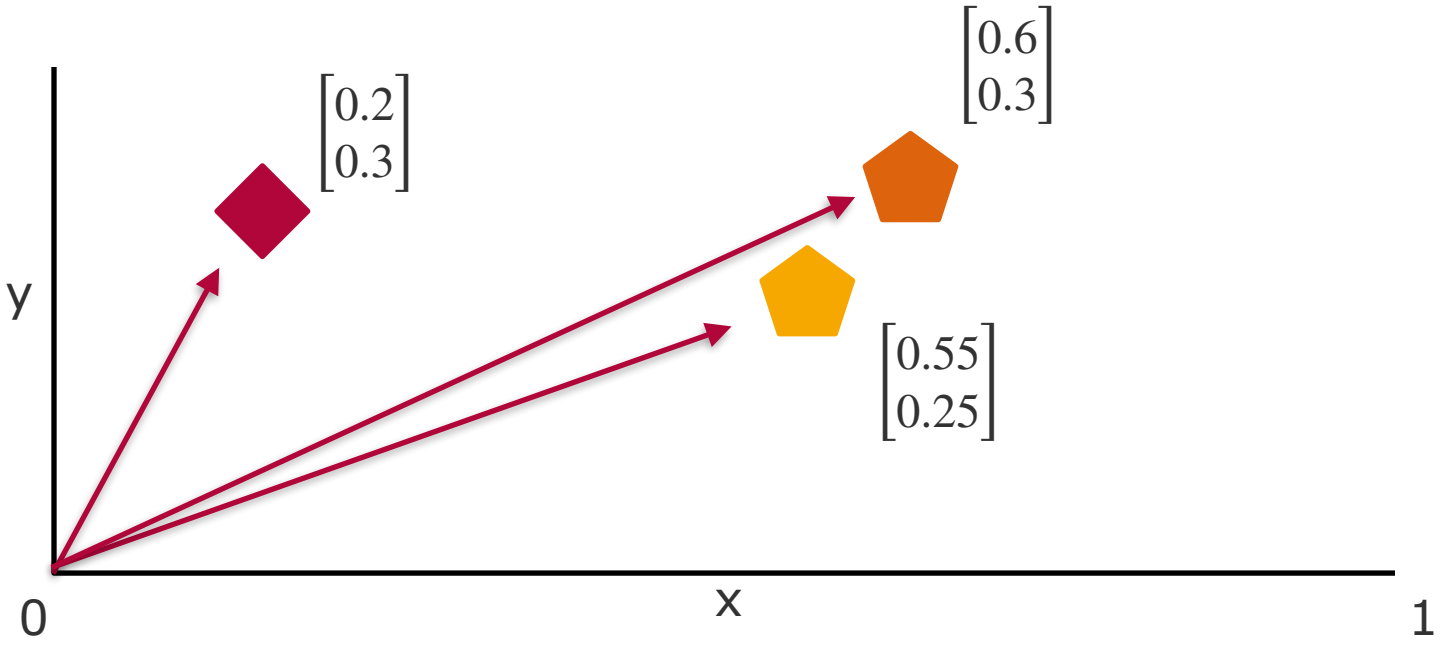
- Finding of nearest neighbours
- Visualisation of relationships

Vector Embedding of Code

Lando Löper, Code Repository Mining, 02.06.2020

Motivation

Code embedding



Vector Embedding of Code

Lando Löper, Code Repository Mining, 02.06.2020

Related Work

Word2Vec

Idea

Words that appear in the same context are similar

Example

- “The grey cat is sitting on the porch.”
- “The white dog is sitting in the garage.”

https://arxiv.org/pdf/1301.3781.pdf

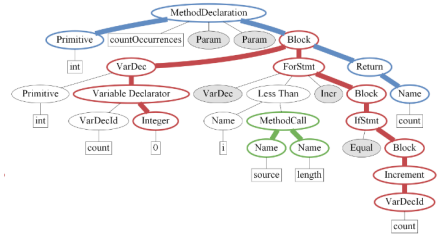
Related Work

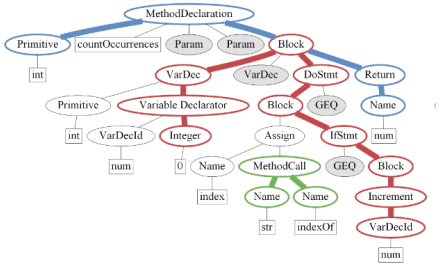
Code2Seq

Idea

Functions with a similar AST are similar

Example





https://openreview.net/pdf?id=H1gKYo09tX

Vector Embedding of Code

Lando Löper, Code Repository Mining, 02.06.2020

Related Work

DYPRO

Idea

Functions with a similar runtime behaviour are similar

Example

- Sorting Algorithm

https://arxiv.org/pdf/1711.07163.pdf

Related Work

DYPRO

Idea

Functions with a similar runtime behaviour are similar

Example

- Sorting Algorithm

Bubble	Insertion
[5,5,1,4,3]	[5,5,1,4,3]
[5,8,1,4,3]	[5,8,1,4,3]
[5,1,1,4,3]	[5,1,1,4,3]
[5,1,8,4,3]	[5,1,8,4,3]
[1,1,8,4,3]	[5,1,4,4,3]
[1,5,8,4,3]	[5,1,4,8,3]
[1,5,4,4,3]	[5,1,4,3,3]
[1,5,4,8,3]	[5,1,4,3,8]
[1,4,4,8,3]	[1,1,4,3,8]
[1,4,5,8,3]	[1,5,4,3,8]
[1,4,5,3,3]	[1,4,4,3,8]
[1,4,5,3,8]	[1,4,5,3,8]
[1,4,3,3,8]	[1,4,3,3,8]
[1,4,3,5,8]	[1,4,3,5,8]
[1,3,3,5,8]	[1,3,3,5,8]
[1,3,4,5,8]	[1,3,4,5,8]

https://arxiv.org/pdf/1711.07163.pdf

Design Space

	Idea	Assumption	Method
Code as sequence of tokens	▪ Learn embeddings from token context	▪ Code can be tokenised	▪ Word2Vec
Code as abstract syntax tree	▪ Learn embeddings from syntax	▪ Code can be parsed	▪ Code2Seq
Code as execution traces	▪ Learn embeddings from runtime behaviour	▪ Code can be executed ▪ Program state can be logged	▪ DYPRO

Data

Source Code

- Java (https://s3.amazonaws.com/code2seq/datasets/java-large.tar.gz)
- Python (https://eth-sri.github.io/py150)
- COSET (https://arxiv.org/abs/1905.11445)

Metric

- Use-Case (Classification)
  - Accuracy
  - Precision / Recall / F1

Outlook

Problem

Predicting function names from their body.

Current State

- Python dataset
- Training code embeddings with Word2Vec

Next Steps

- Training code embeddings including AST information

Discussion

Questions?

- Can you think of other applications for code embeddings?
- Can you think of methods to evaluate code embeddings?

Appendix

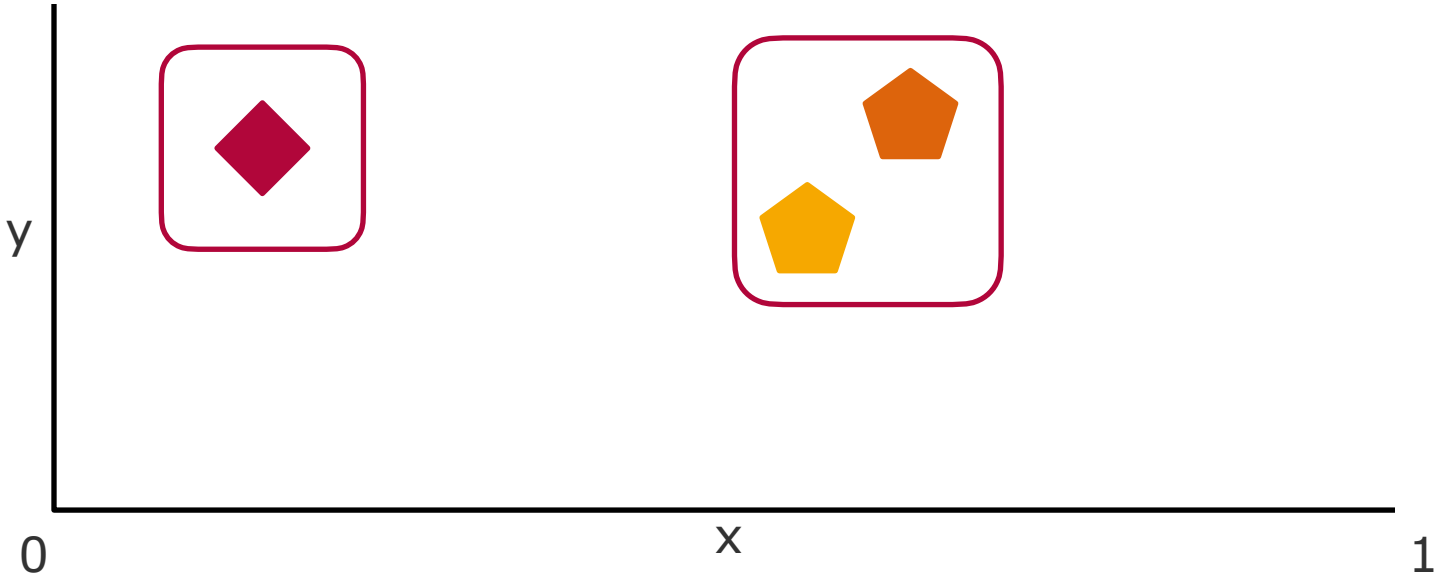
- Code Embedding Examples
- Word Embedding Examples
- Visualisations of current state

Thank you
for your attention!

Lando Löper
Software Architecture Group
Hasso Plattner Institute Potsdam
02.06.2020

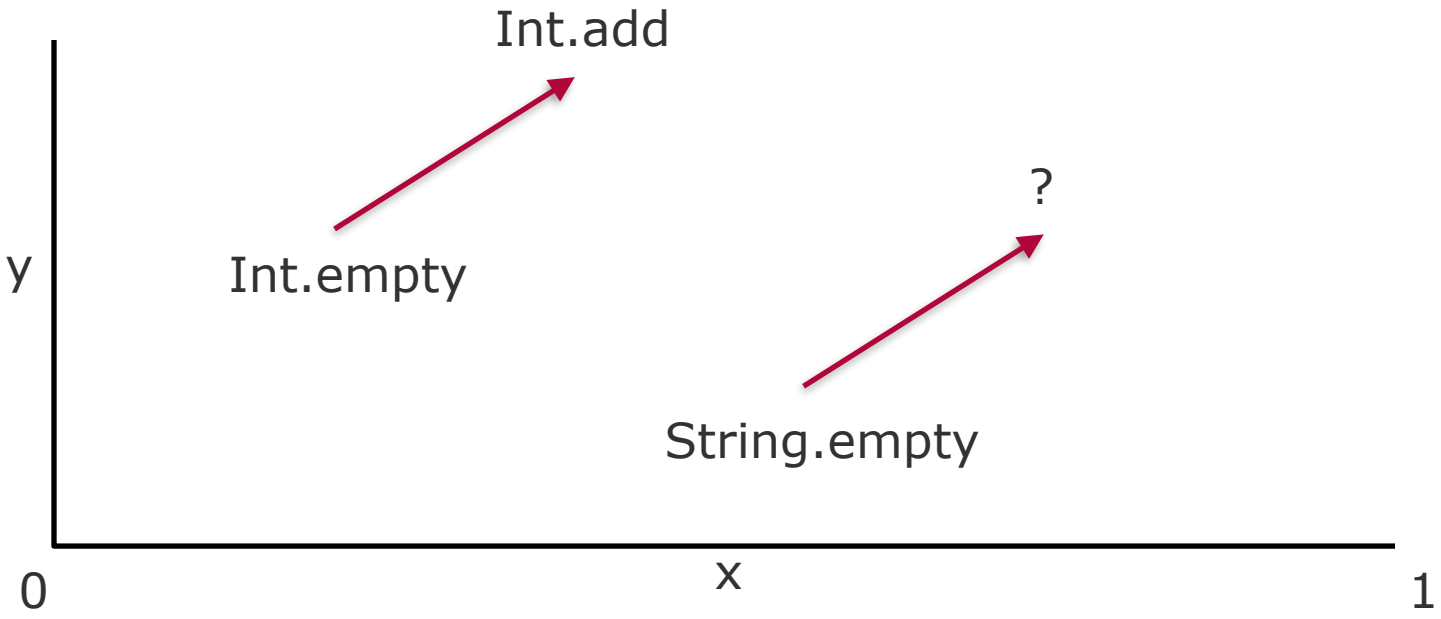
Appendix A

Nearest Neighbour



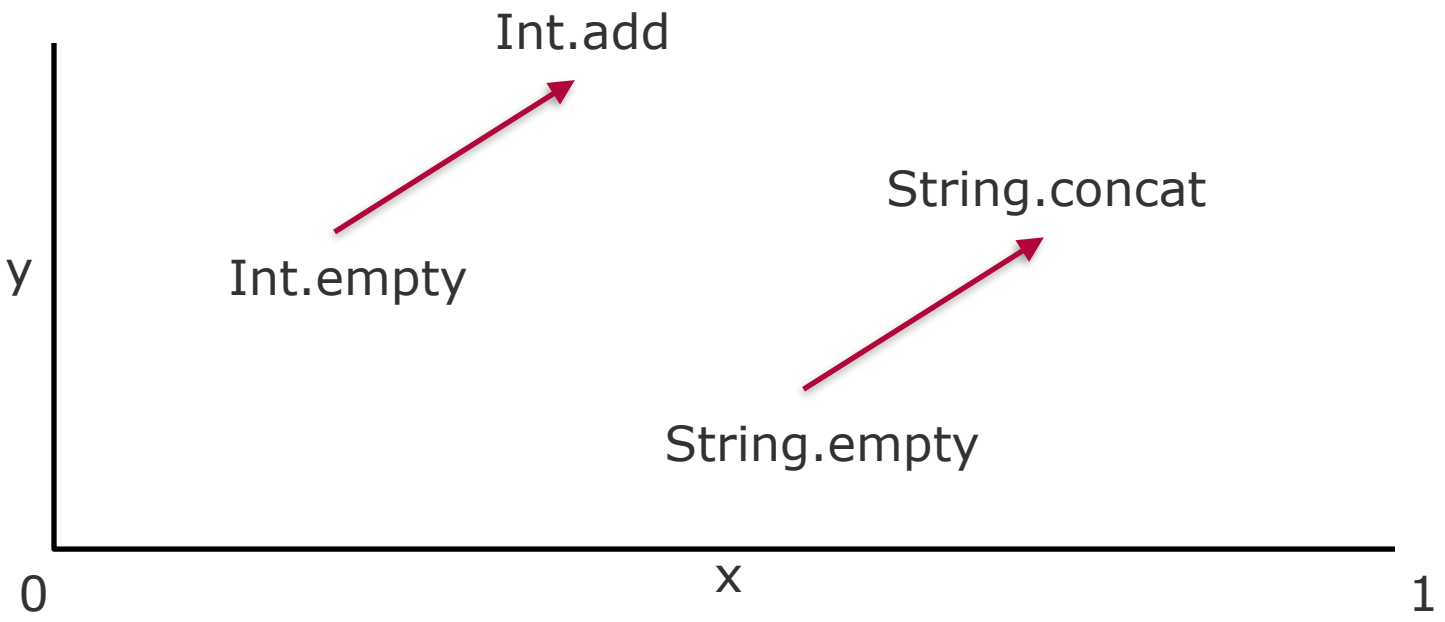
Appendix A

Relationships



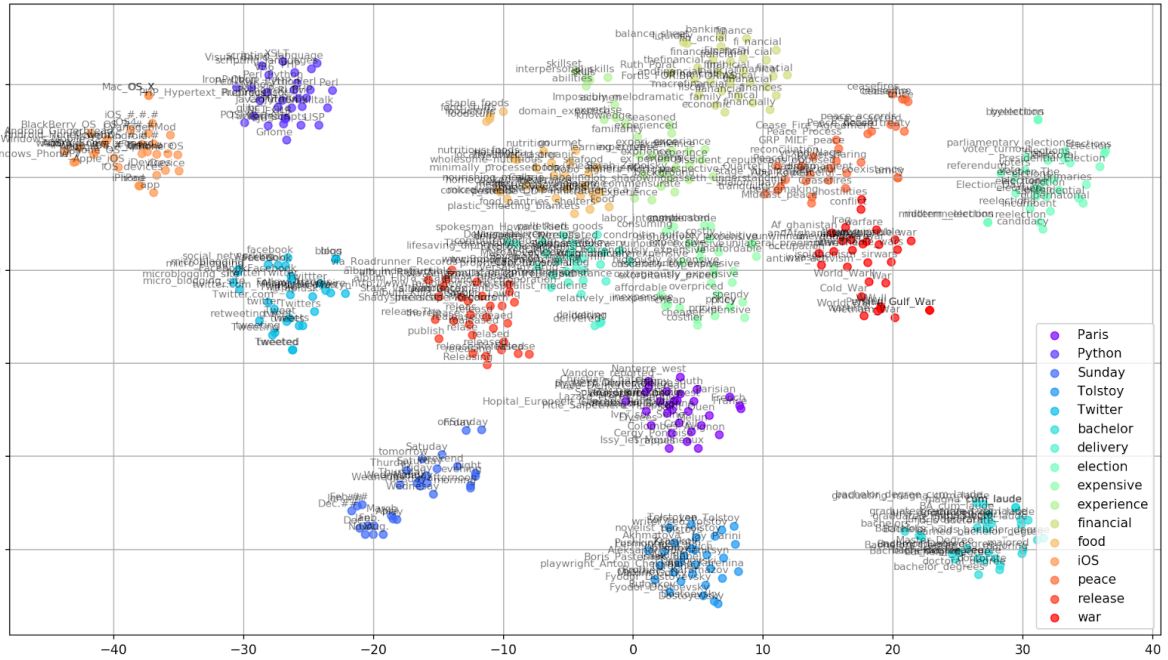
Appendix A

Relationships



Appendix B

Nearest Neighbour



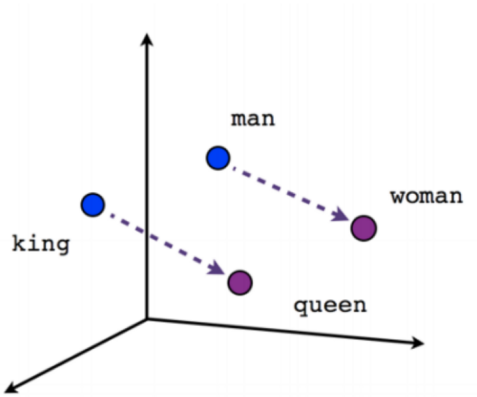
https://miro.medium.commax1400/0*quUwP6EqEhFmNAqX

Vector Embedding of Code

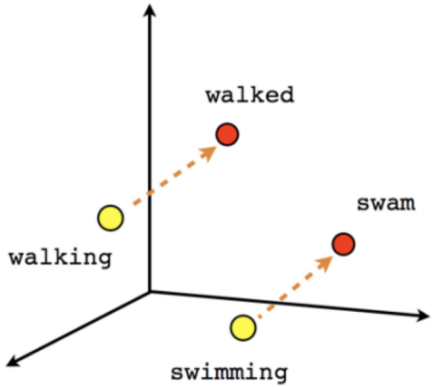
Lando Löper, Code Repository Mining, 02.06.2020

Appendix B

Relationships



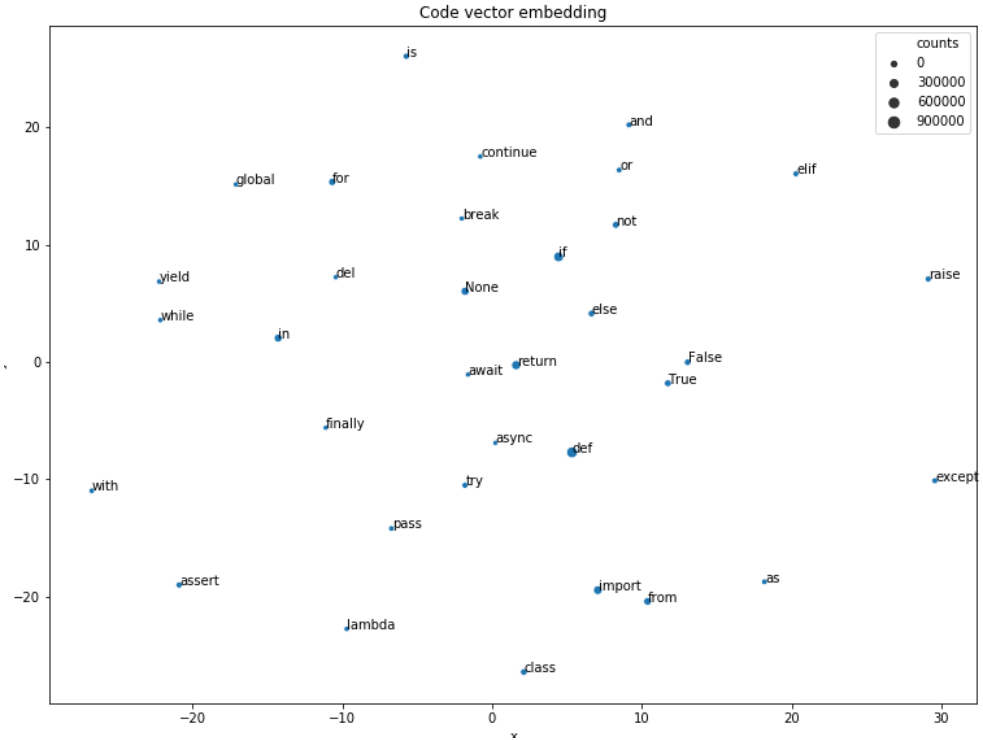
Male-Female



Verb tense

https://miro.medium.commax1400/1*jpnKO5X0Ii8PVdQYFO2z1Q.png

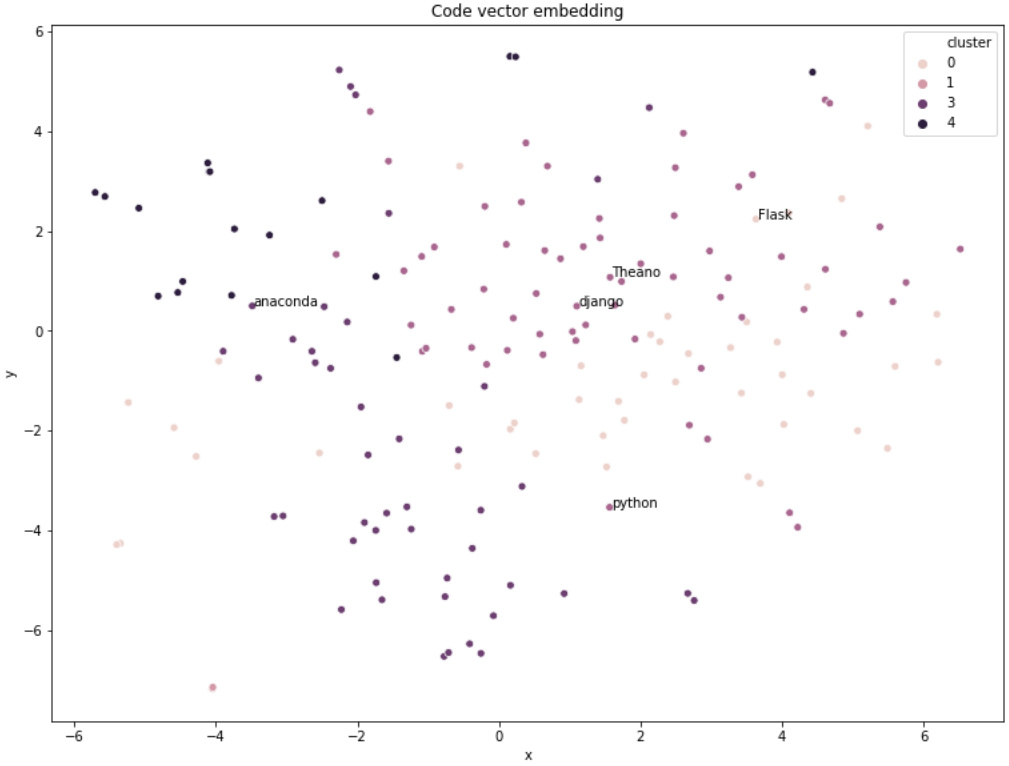
Appendix C



Vector Embedding of Code

Lando Löper, Code Repository Mining, 02.06.2020

Appendix C



Vector Embedding of Code

Lando Löper, Code Repository Mining, 02.06.2020