Mining Software Repositories Lab

Week 3
Group 10

Motivation
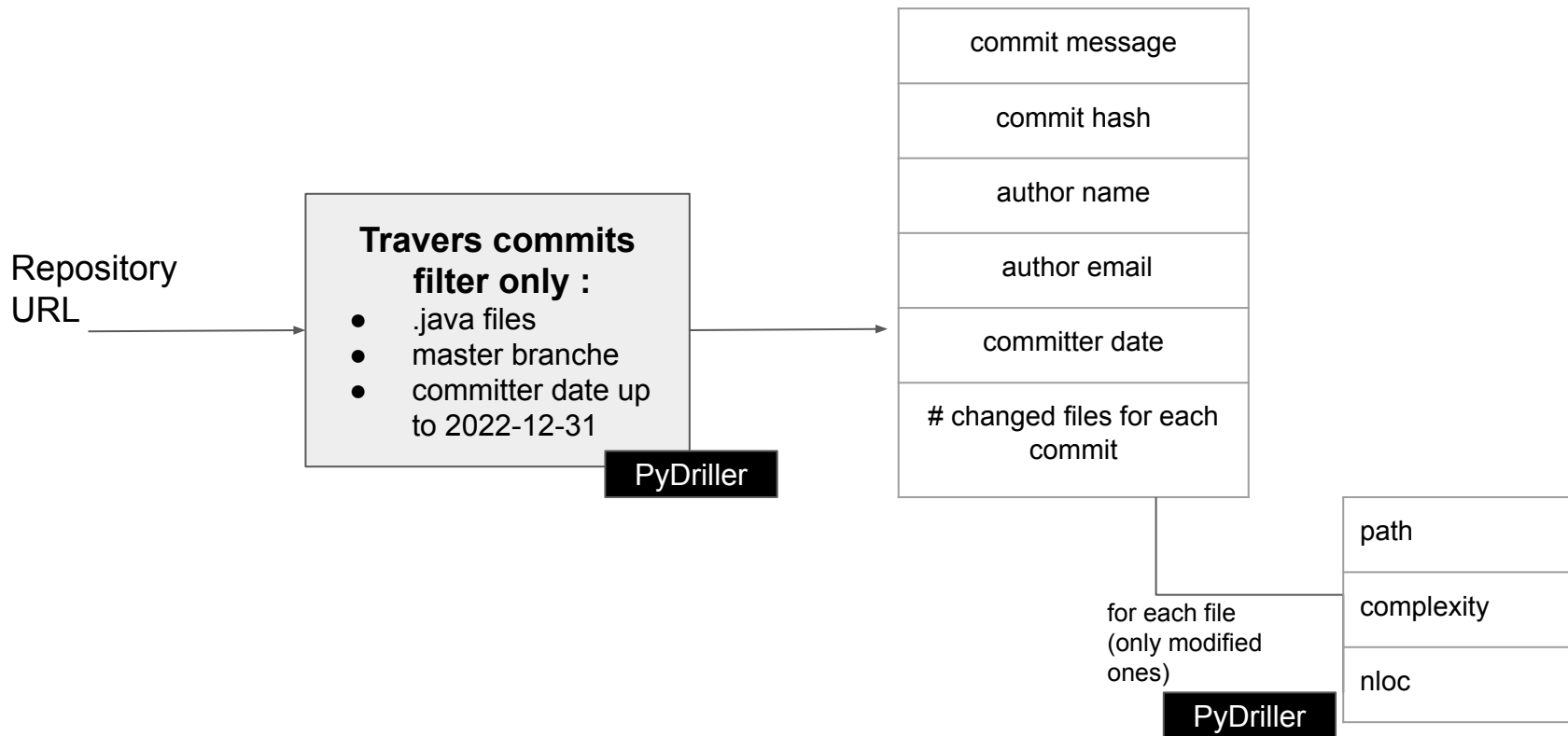
1. Maintainability
2. Collaboration
3. Code quality
4. Time and cost savings
5. User satisfaction
6. Automation and CI/CD

Data extraction :

Repository URL

Travers commits filter only :
- .java files
- master branche
- committer date up to 2022-12-31

PyDriller

commit message

commit hash

author name

author email

committer date

changed files for each commit

for each file (only modified ones)

PyDriller

path

complexity

nloc

Data extraction :

before version → Extract the readability for each changed file → Readability before and after changes

after version →
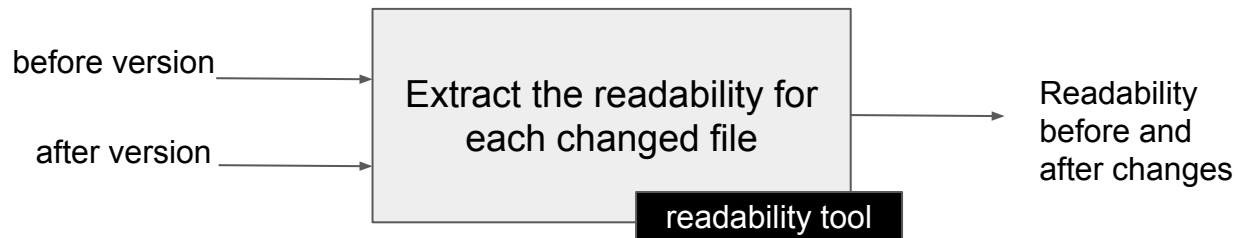
readability tool

Repository used	real size	# unique commits
commons-vfs	11931	1911
commons-bcel	15138	1332
commons-codec	4081	1294
javalin	1386	558

RQ1:Do developers increase readability deliberately?

Collected Data → Calculate **readability variation** → **Aggregated** over commit messages and **summing** the readability variation

Readability increased
readability variation >0

2535

Readability decreased
readability variation <0

2712
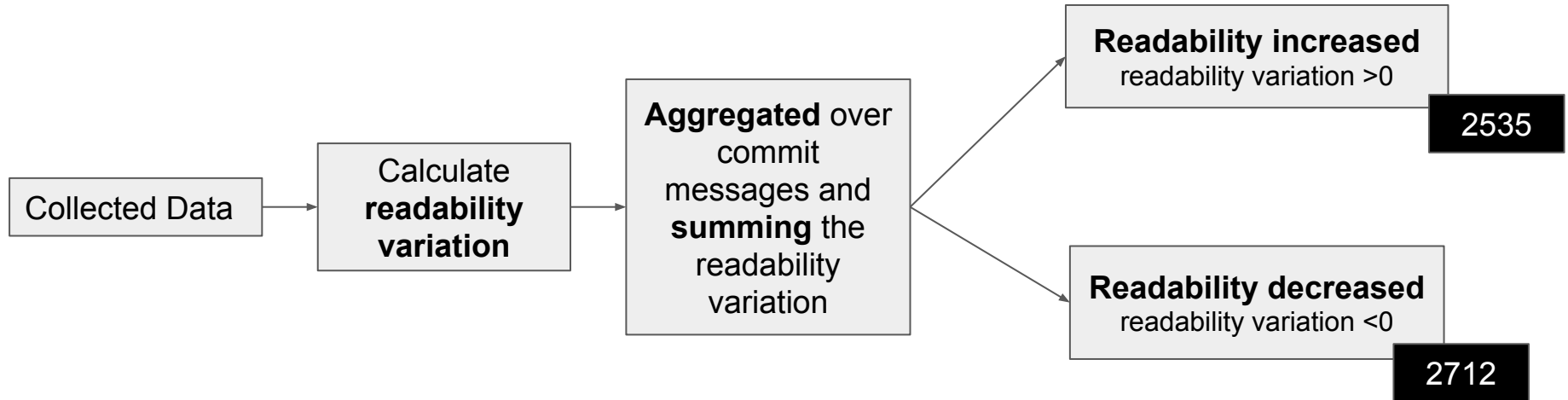
Text Processing pipeline

list of commit messages →

Text Cleaning:
- remove unicodes
- digits
- punctuation
- whitespaces
- hashes
- URLs

Stop words removal → Tokenization → Stemming → **unuseful list words*** filtering → Filtering keeping only english words →
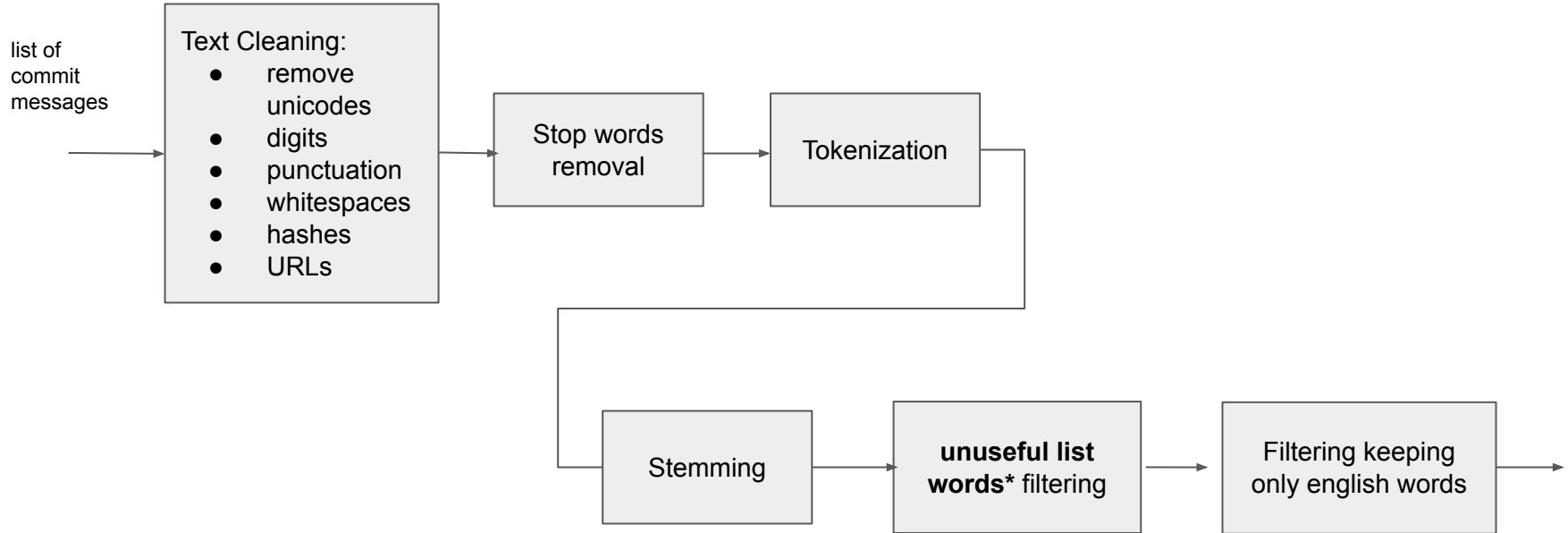
unuseful list words = ['git', 'apache', 'svn', 'id', 'method', 'class', 'don', 'private', 'commons', 'test', 'use', 'fix', 'add', 'code', 'file', 'provider', 'string', 'static', 'upgrade', 'patch','static', 'package', 'instead', 'case','new' , 'final']

RQ1: Preliminary experiment

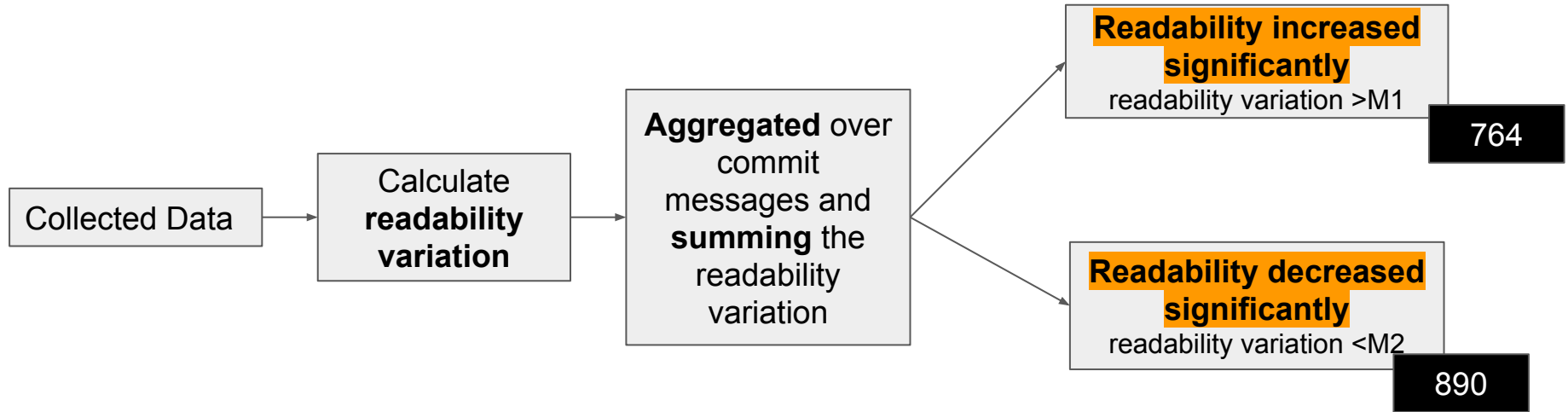We used 0 to split data into
increased and decreased

Increased readability : ('remove', 152), ('proper', 109), ('better', 105),
('support', 74), ('exception', 58), ('local', 53), ('type', 53), ('binary', 53),
('bug', 52), ('set', 52), ('unused', 51), ('fixed', 51), ('line', 50), ('throw', 45)

Decreased readability :
('remove', 133), ('proper', 113), ('better', 69), ('exception', 65), ('binary',
59), ('type', 51), ('missing', 51), ('unused', 48), ('support', 46), ('avoid',
46), ('need', 45), ('version', 44), ('fixed', 42), ('unnecessary', 40),
('check', 39), ('simplify', 39), ('system', 39), ('local', 39), ('bug', 38),

RQ1:Do developers increase readability deliberately?

```
Collected Data  →  Calculate
                   readability
                   variation
```
→
Aggregated over commit messages and **summing** the readability variation

Readability increased significantly
readability variation >M1

764

Readability decreased significantly
readability variation <M2

890

M1:mean of readability increased group :0.01107
M2:mean of readability decreased group :-0.01203

RQ1: Word Clouds

Decreased readability

('set', 18), ('allow', 16), ('better', 14), ('configur', 14),
('support', 13), ('resourc', 13), ('base', 11), ('type', 11),
('local', 10),('format', 10), ('handl', 9), ('port', 9),
('coverag', 9), ('replac', 9), ('miss', 8)
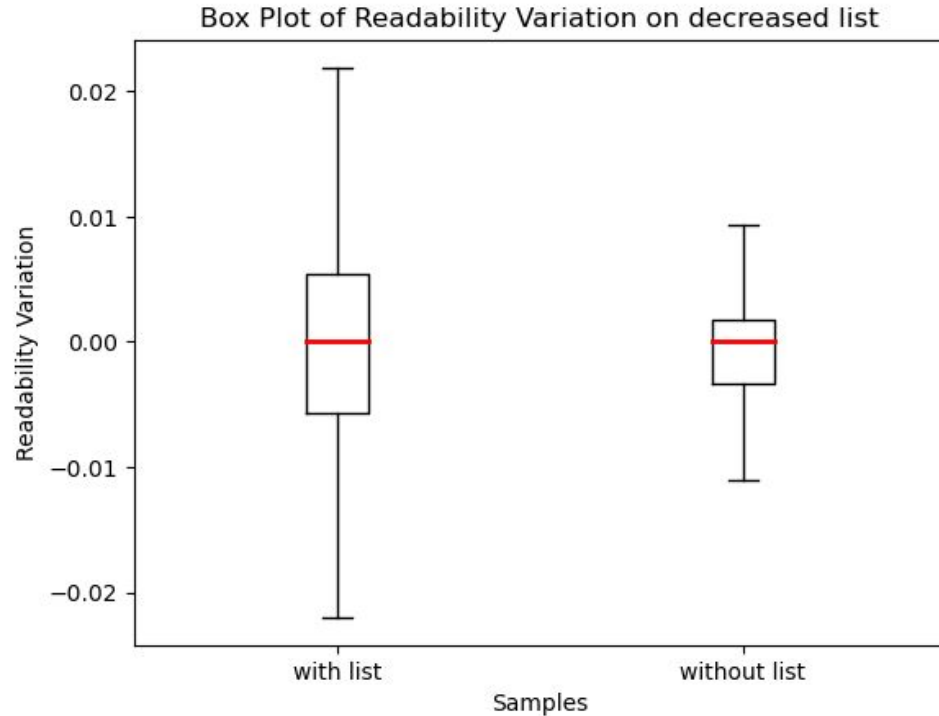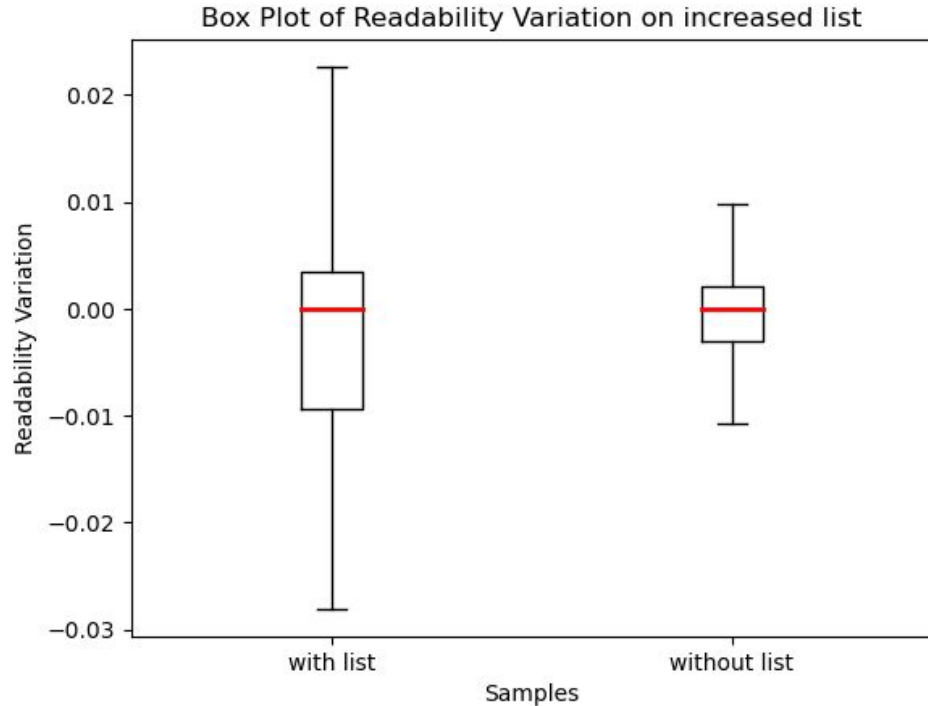
Increased readability

('need', 11), ('system', 11), ('log', 11), ('remov', 10),
('depend', 9), ('miss', 9), ('interfac', 9), ('assert', 9),
('gener', 8), ('better', 8), ('close', 8),('comment', 8),
('type', 7), ('style', 7),

9

RQ1: Suggested keywords lists

Increased readability : 'support', 'set', base,allow, 'assert', 'implement', 'local', 'line', 'configur', 'miss','refactor', 'enhanc' , 'clean' ,'simplifi','duplicat'

Decreased readability : 'implement', 'lock', 'need', 'miss', 'base', 'system', 'better', 'close', 'check', 'remov', 'log', 'simplifi','featur' ,' conflict' , 'configur' ,'paramet'

RQ1: Distribution of commit message readability variation based on keywords extracted
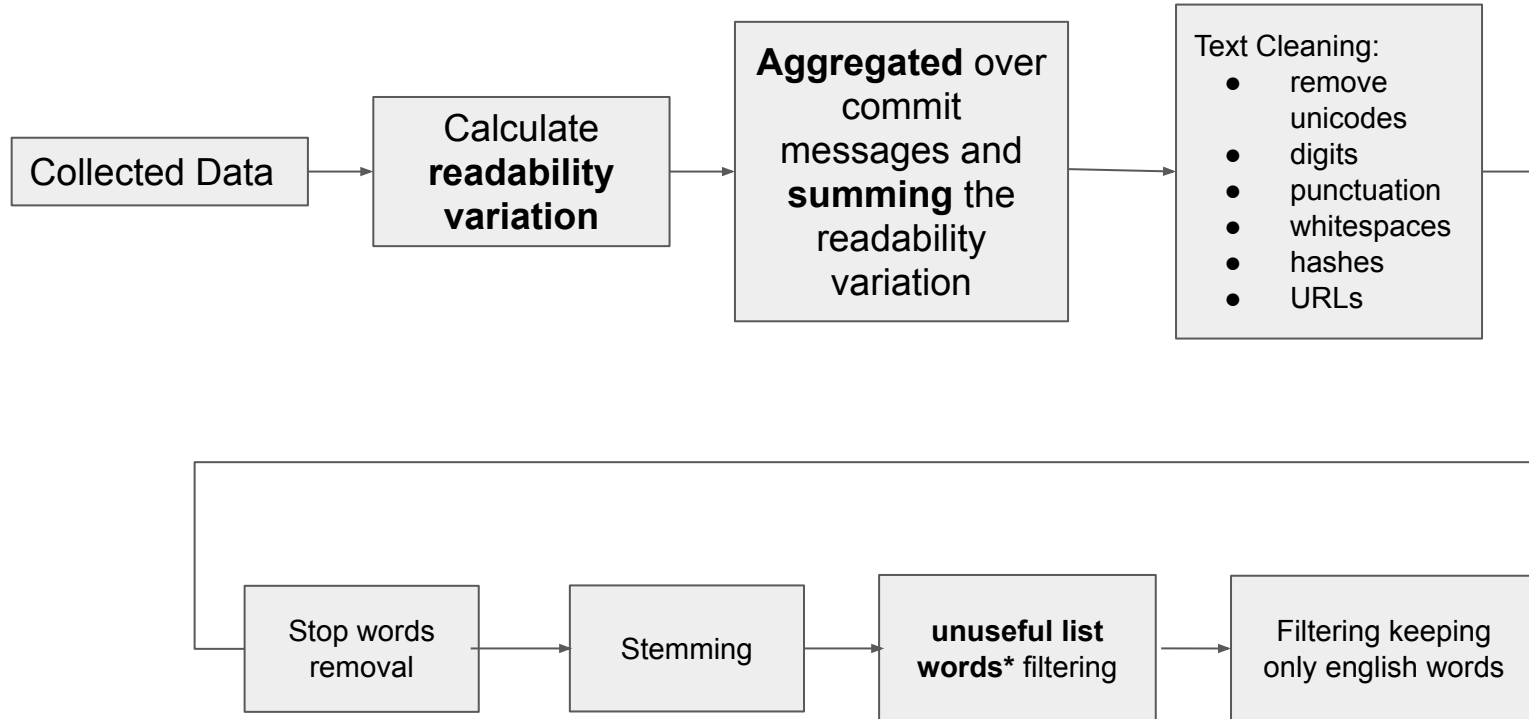
RQ1 :Do developers increase readability deliberately?

Answer

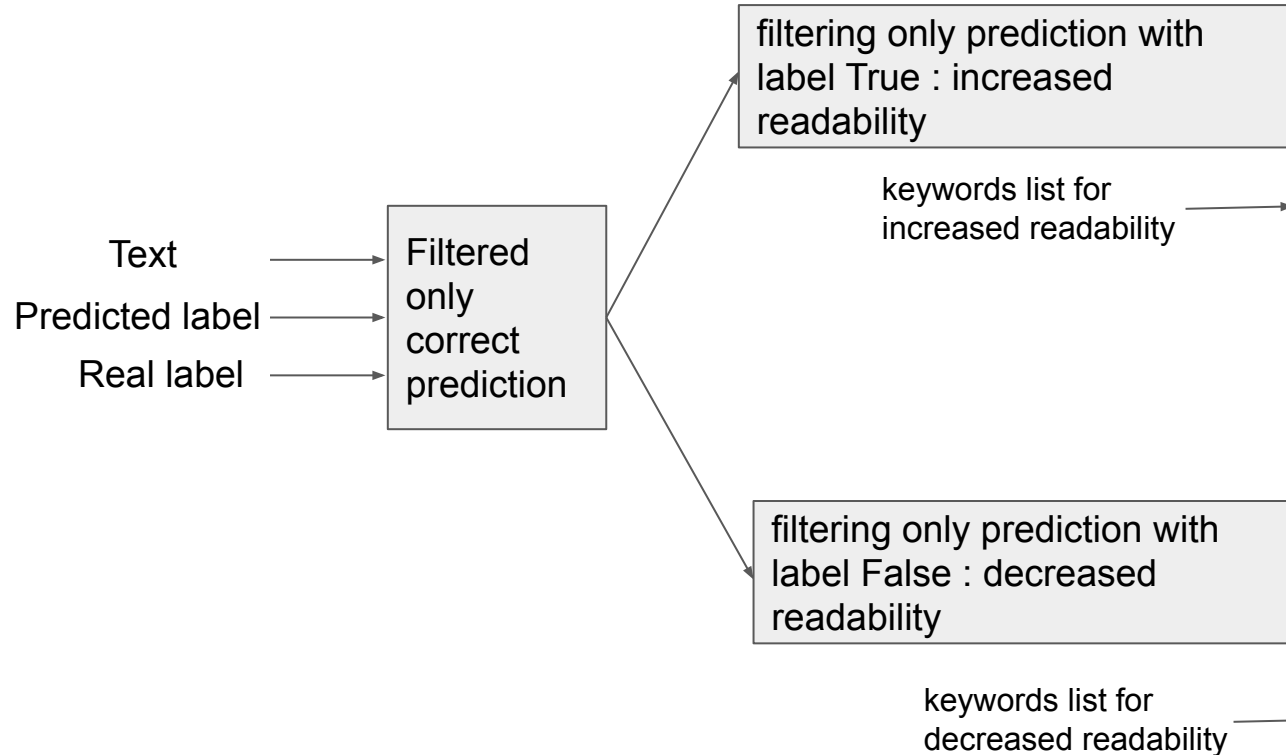RQ2: Can we predict whether the readability of a file will be improved?

RQ2: Data Preprocessing

```
Collected Data → Calculate readability variation → Aggregated over commit messages and summing the readability variation → Text Cleaning:
  ● remove unicodes
  ● digits
  ● punctuation
  ● whitespaces
  ● hashes
  ● URLs

Stop words removal → Stemming → unuseful list words* filtering → Filtering keeping only english words
```

RQ2: Can we predict whether the readability of a file will be improved?

Dataset	All Repositories
Train set	80%
Test set	20%
Model	Random Forest Classifier
Vectorizer	Tf idf Vectorizer

Accuracy	60%
Precision	50%
recall	38%
f1	43%

RQ2: Feature importance

occurrence of the words from the list

filtering only prediction with label True : increased readability

keywords list for increased readability

filtering only prediction with label False : decreased readability

keywords list for decreased readability

Text

Predicted label

Real label

Filtered only correct prediction
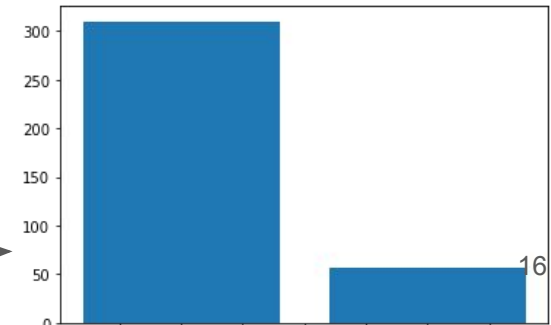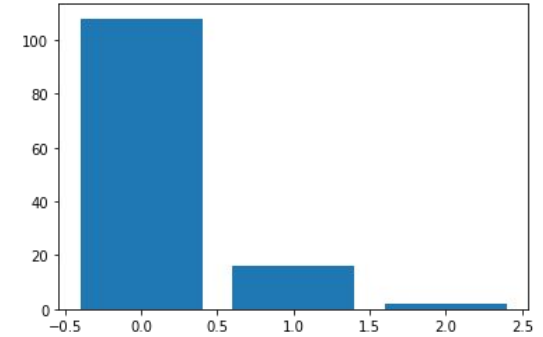
RQ2: Evaluation Cross-projects

Train data repository	Test data repository	Performance
commons-codec commons-bcel (2285)	commons-vfs (1708)	Accuracy: 58% precision: 46% recall: 39% f1: 42%
commons-vfs commons-bcel (2866)	commons-codec (1122)	Accuracy: 58% precision: 47% recall: 32% f1: 38%
commons-vfs commons-codec (2811)	commons-bcel(1181)	Accuracy: 60% precision: 51% recall: 29% f1: 37%

RQ2: Evaluation within-project

Repository	Performance
commons-vfs (1708)	Accuracy: 59% precision: 61% recall: 22% f1: 33%
commons-codec (1122)	Accuracy: 59% precision: 52% recall: 31% f1: 39%
commons-bcel(1181)	Accuracy: 56% precision: 54% recall: 39% f1: 45%
javalin(469)	Accuracy: 56% precision: 45% recall: 28% f1: 34%

Train set	80%
Test set	20%

RQ2: Can we predict whether the readability of a file will be improved?

Answer

Thank you for your attention