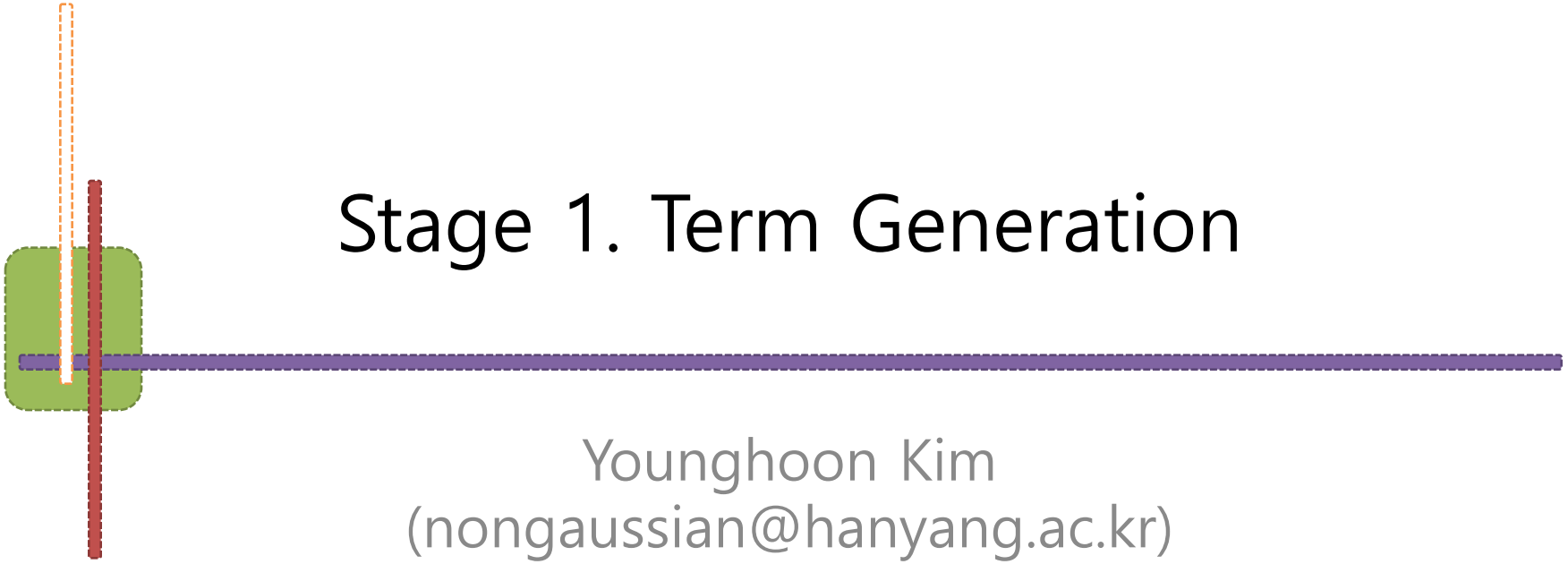


Stage 1. Term Generation

Younghoon Kim
(nongaussian@hanyang.ac.kr)

Note

- The goal of the first stage is
  - To practice how to write and submit your code

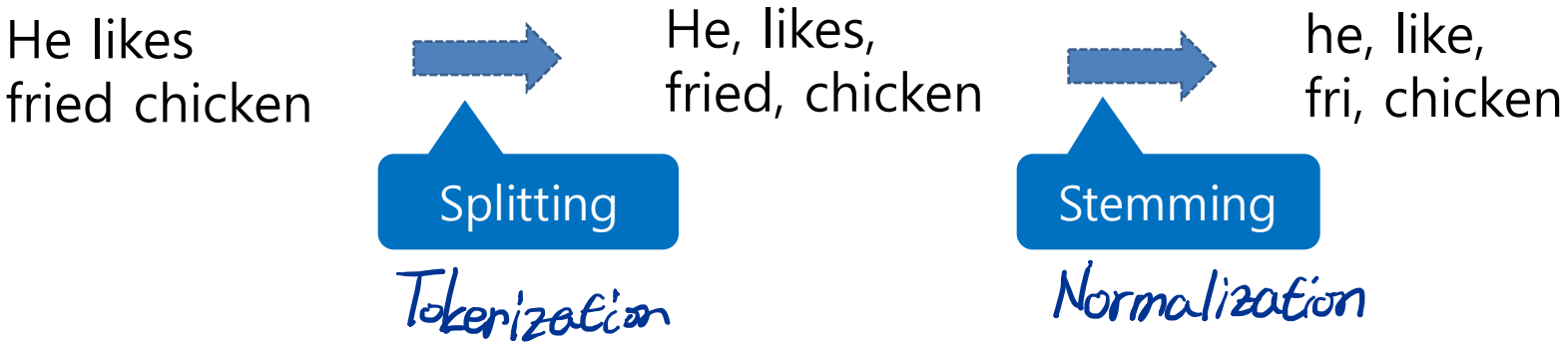
매우 간단.

코드 길이가 20~30 줄.

Problem Definition

- Given
  - A string (e.g., sentence, article) of type String
- Return
  - A list of terms which is split by whitespaces and stemmed
  - Type: List<String>

⇒ Tokenization



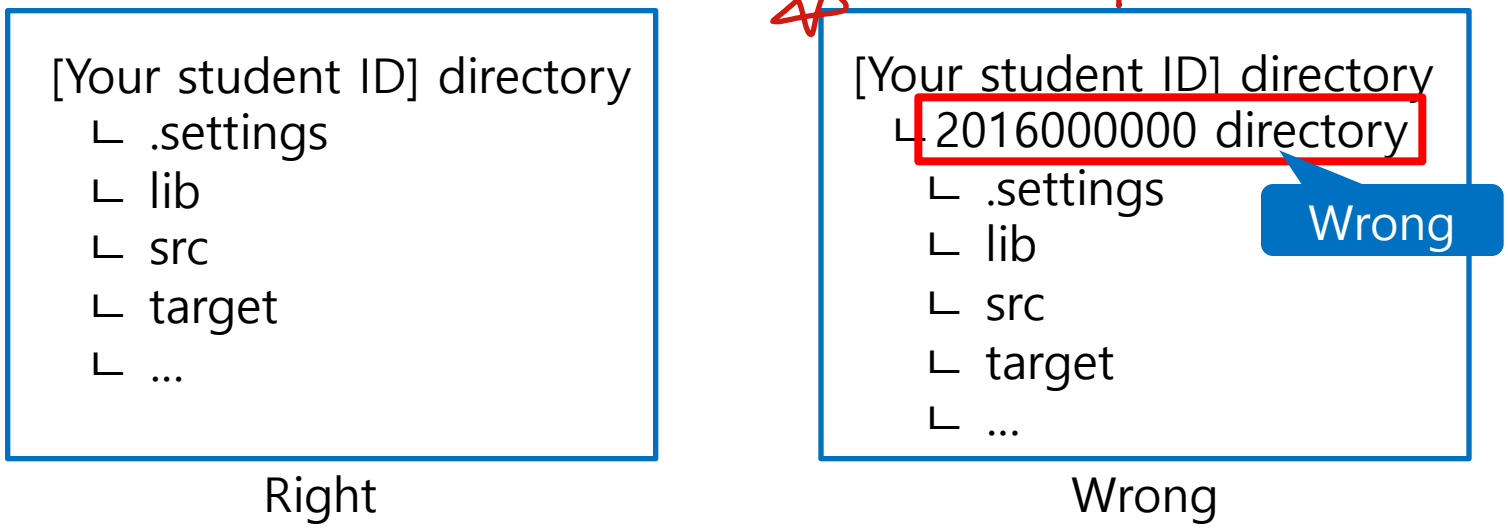
Code Template

- We provide a package of
  - A maven project created in Eclipse
- It contains
  - Template codes (`edu.hanyang.submit.TinySETokenizer.java`)
  - TinySE framework (`lib/tinyse-0.0.1-SNAPSHOT.jar`) ← to be updated on every stage
    - Interface files (e.g., Tokenizer.java)
    - Indexer and query processer codes which will complete a search engine by connecting your submissions

이클립스 쓰기 싫으면 maven만 써도 된다.
- 쓰는게 편할 듯.

To Use Code Template

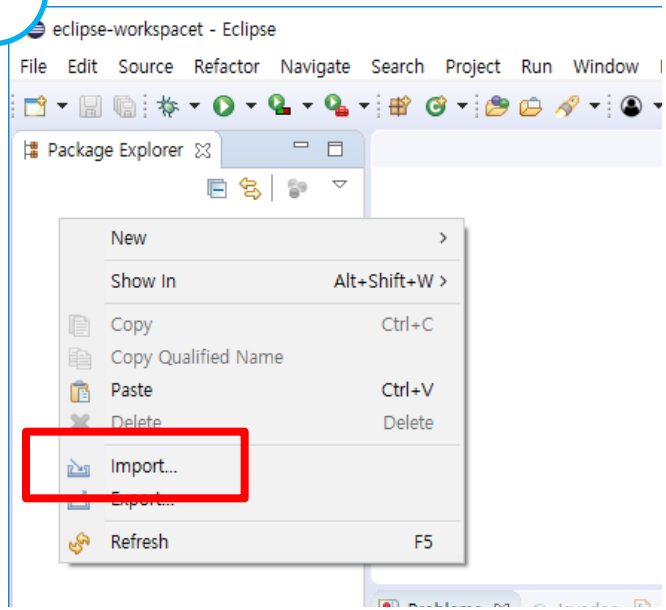
- Download the template Eclipse project
- Rename the project directory
  - 2016000000 ➔ [your student ID]
  - **i.e.** 2016000000 ➔ 2018123456



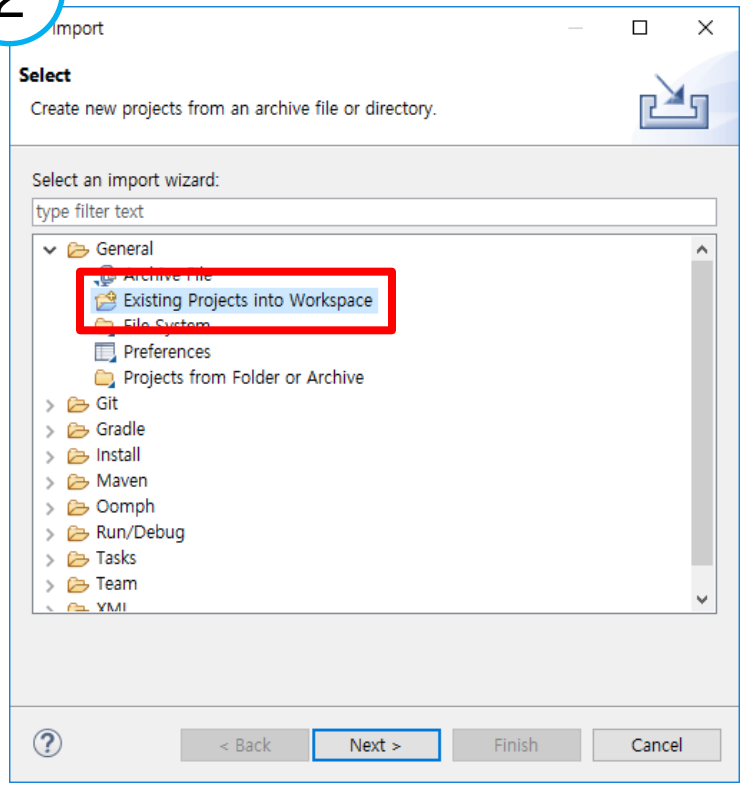
To Use Code Template

- Import the project in eclipse IDE

1



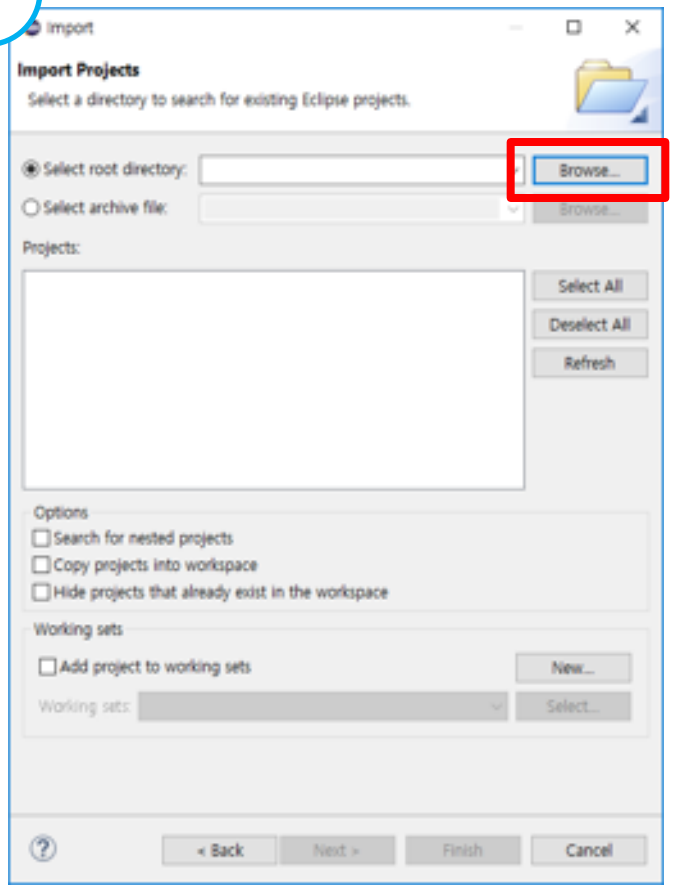
2



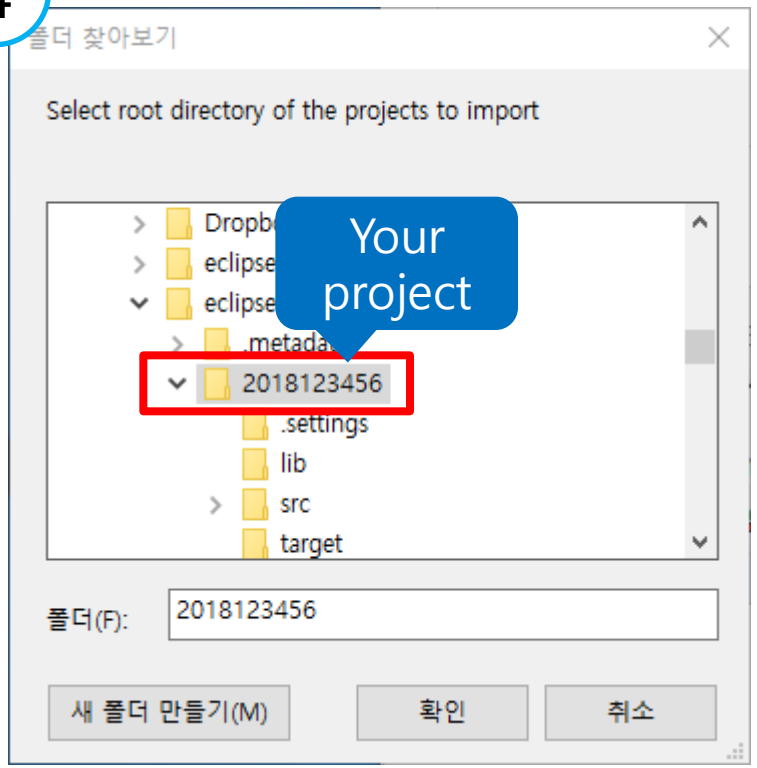
To Use Code Template

- Import the project in eclipse IDE

3



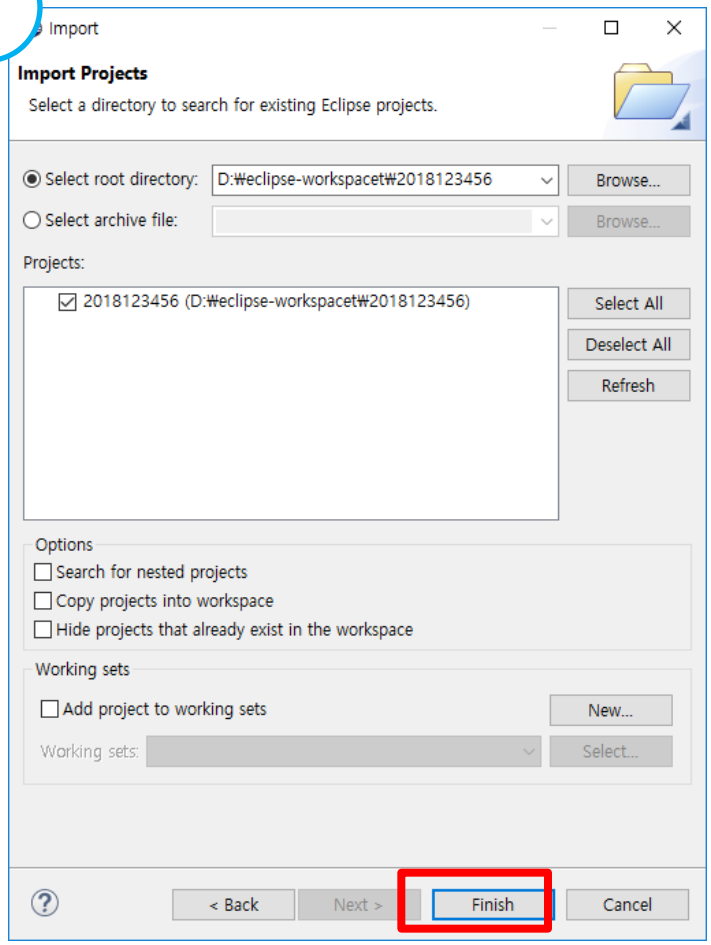
4



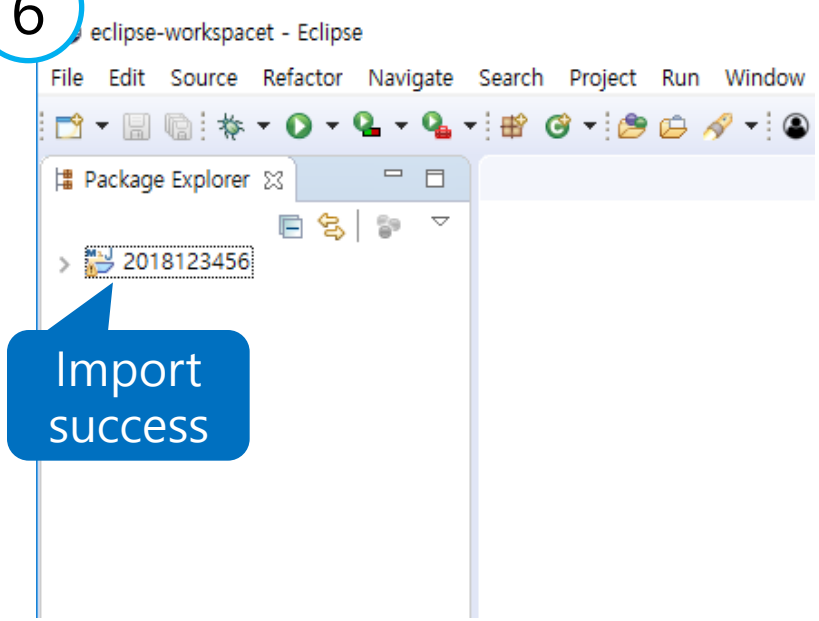
To Use Code Template

- Import the project in eclipse IDE

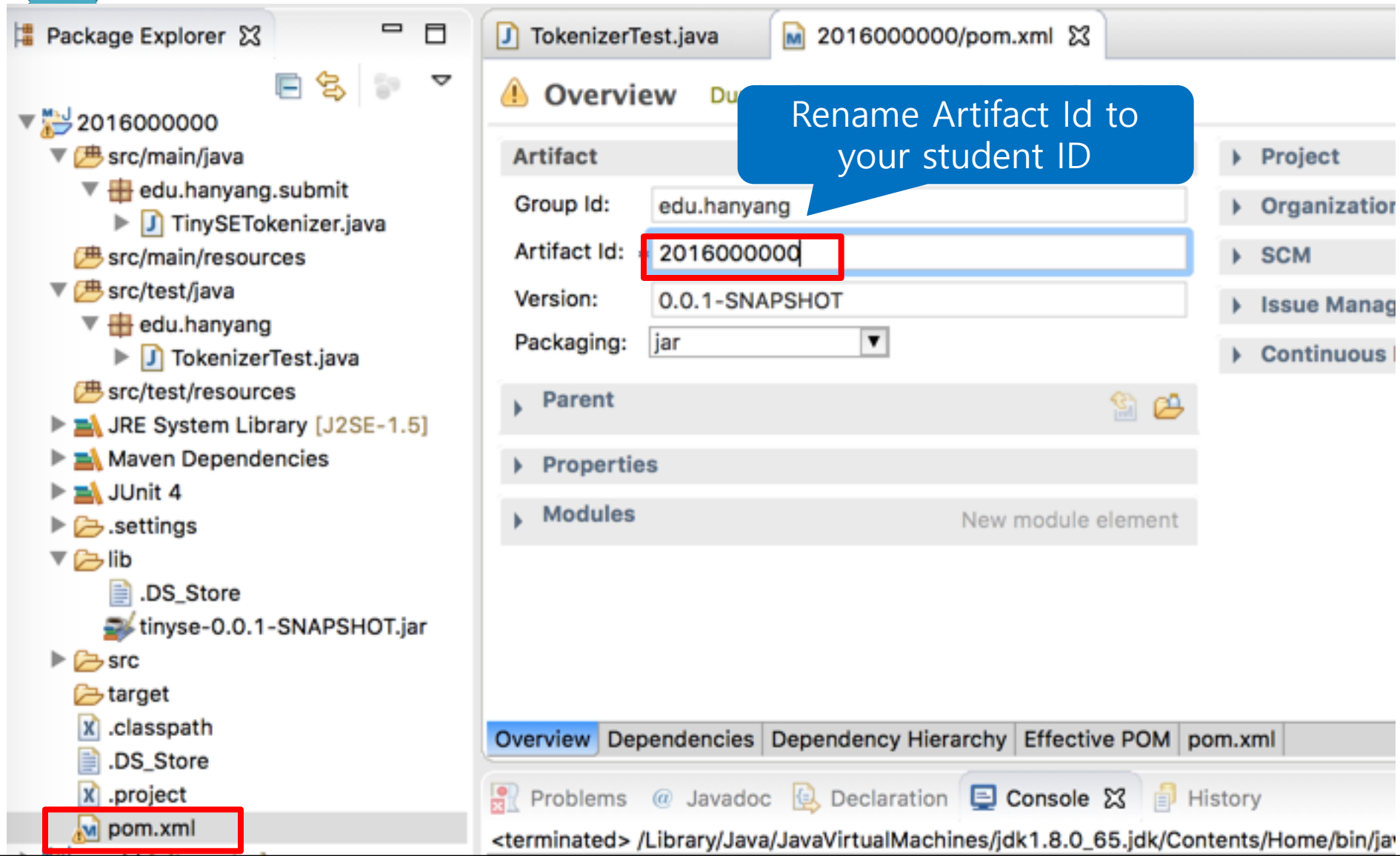
5



6

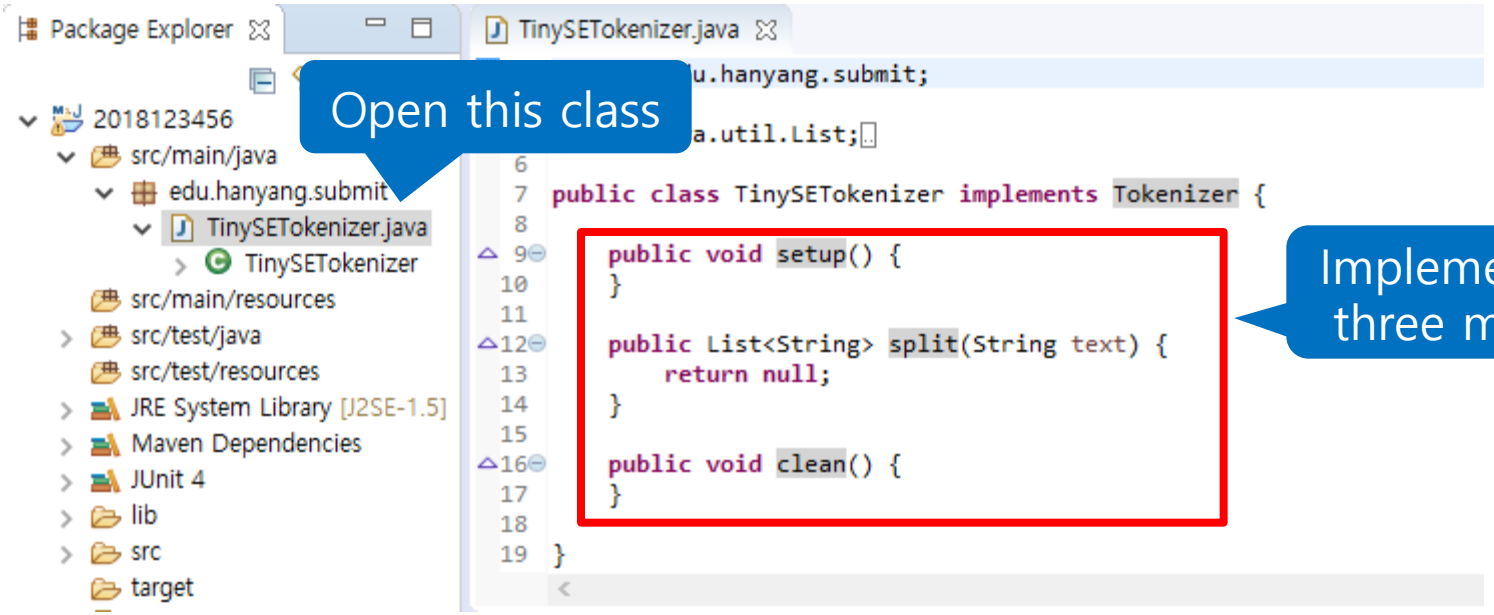


To Use Code Template



To Use Code Template

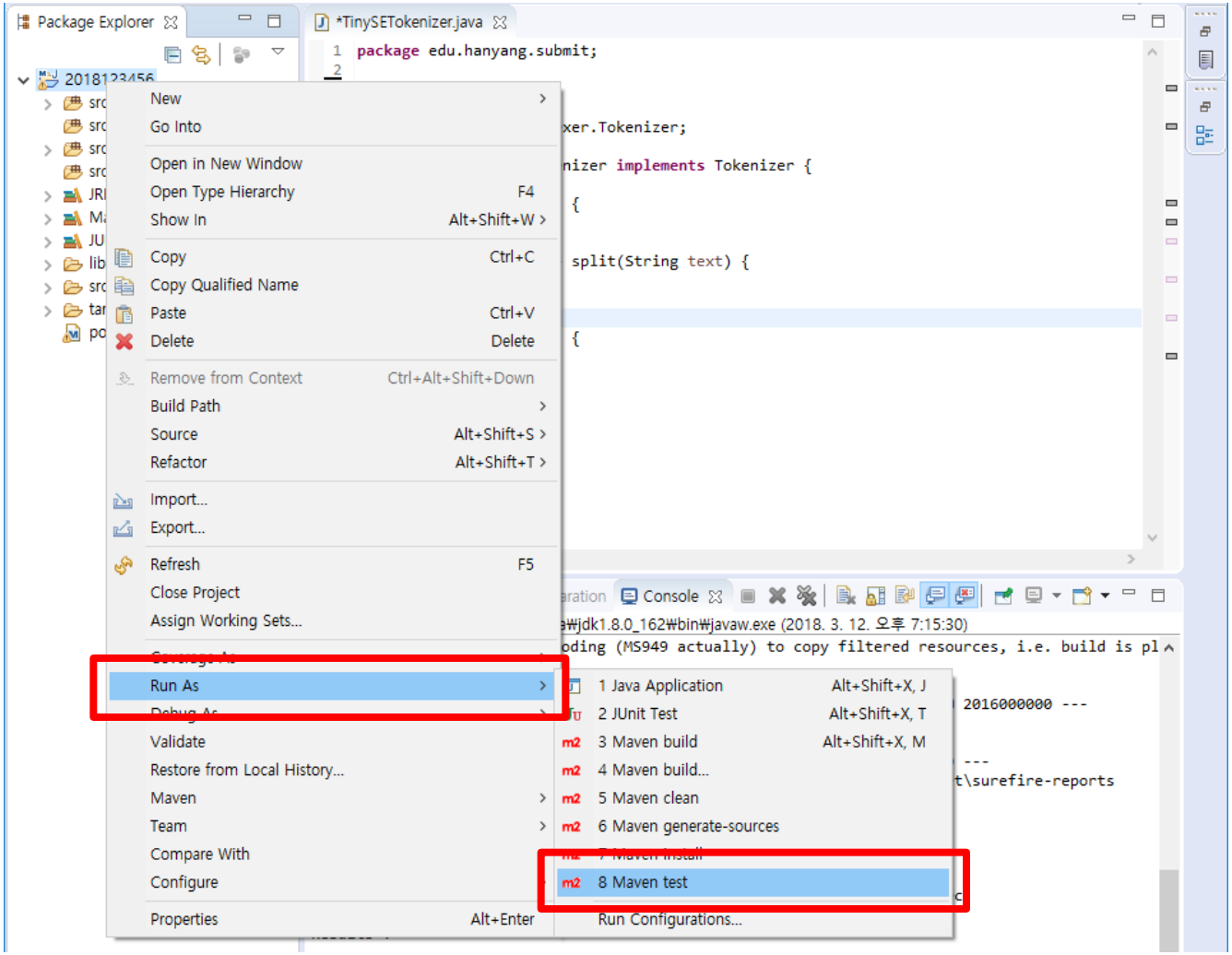
- Complete *edu.hanyang.submit.TinySETokenizer*



You use *SimpleAnalyzer* and *PorterStemmer* class.
If you use other class, you may get wrong result.

To Use Code Template

- 6. Test your code



To Use Code Template

- 6. Test your code

```
-------------------------------------------------------
 T E S T S
-------------------------------------------------------
Running edu.hanyang.TokenizerTest
Tests run: 1, Failures: 0, Errors: 0, Skipped: 0, Time elapsed: 0.071 sec

Results :

Tests run: 1, Failures: 0, Errors: 0, Skipped: 0

[INFO] ------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------
[INFO] Total time: 1.020 s
[INFO] Finished at: 2018-03-12T19:15:32+09:00
[INFO] Final Memory: 10M/243M
[INFO] ------------------------------------------------
```

If your code passes our unit test, you can see no failures, no errors and no skipped on console

External Libraries

- Use SimpleAnalyzer and PotterStemmer in Lucene 7.2.1
  - SimpleAnalyzer is a tokenizer that splits a sentence with whitespaces
  - PotterStemmer is a well-known and simple stemmer for English
- JavaDoc
  - SimpleAnalyzer: https://lucene.apache.org/core/7_2_1/analyzers-common/index.html?org/apache/lucene/analysis/core/SimpleAnalyzer.html
  - PorterStemmer: https://lucene.apache.org/core/7_2_1/analyzers-common/org/tartarus/snowball/ext/PorterStemmer.html

Constrain & Submit

- How to submit
  - Run “maven package”
  - Submit <your student ID>-0.0.1-SNAPSHOT.jar file (you can find it from <project dir>/target/)
- If any question, contact TA
  - Keonwoo Kim (김건우)
  - kdbml314@gmail.com
  - Room: 4공학관 314호

어디다가 제출? HY-IN 과제란

기한: 3/20 까지.