TERM PROJECT

Design Pattern: JSoup 기능 확장 및 설계 개선

20134789 김진관

20154547 박철우

20165172 왕천용

목차

1. JSoup 개요

2. Jsoup 설계 및 구현 조사

3. 확장과 설계 개선

4. 테스트 수행 내역

**1. JSoup 개요**

JSoup은 Java로 만들어진 HTML 파서

주요 기능

- URL, 파일, 문자열을 이용해서 HTML 파싱, scrape를 할 수 있다.

- DOM 구조를 순회하거나, CSS selector를 이용해서 원하는 데이터를 찾거나 추출할 수 있다.

- 문서 내의 HTML 요소, 속성, 텍스트를 조작할 수 있다.

- XSS 공격을 방지하기 위해서 사용자 제출 컨텐츠를 안전한 화이트-리스트로 지정된 태그만 남기고 제거할 수 있다.

- 깔끔한 형태의 HTML을 출력할 수 있다.

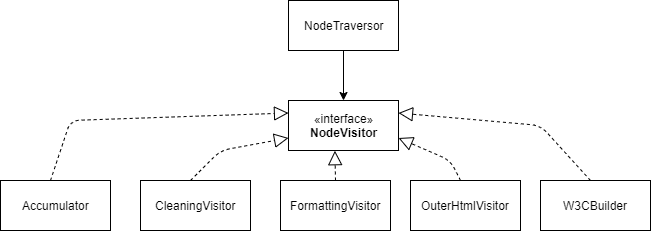
**Class Hierarchy**

java.lang.[**Object**](https://docs.oracle.com/javase/7/docs/api/java/lang/Object.html?is-external=true)

* java.util.[**AbstractCollection**](https://docs.oracle.com/javase/7/docs/api/java/util/AbstractCollection.html?is-external=true)<E> (implements java.util.[Collection](https://docs.oracle.com/javase/7/docs/api/java/util/Collection.html?is-external=true)<E>)
  + java.util.[**AbstractList**](https://docs.oracle.com/javase/7/docs/api/java/util/AbstractList.html?is-external=true)<E> (implements java.util.[List](https://docs.oracle.com/javase/7/docs/api/java/util/List.html?is-external=true)<E>)
    - java.util.[**ArrayList**](https://docs.oracle.com/javase/7/docs/api/java/util/ArrayList.html?is-external=true)<E> (implements java.lang.[Cloneable](https://docs.oracle.com/javase/7/docs/api/java/lang/Cloneable.html?is-external=true), java.util.[List](https://docs.oracle.com/javase/7/docs/api/java/util/List.html?is-external=true)<E>, java.util.[RandomAccess](https://docs.oracle.com/javase/7/docs/api/java/util/RandomAccess.html?is-external=true), java.io.[Serializable](https://docs.oracle.com/javase/7/docs/api/java/io/Serializable.html?is-external=true))
      * org.jsoup.helper.[**ChangeNotifyingArrayList**](https://jsoup.org/apidocs/org/jsoup/helper/ChangeNotifyingArrayList.html)<E>
      * org.jsoup.select.[**Elements**](https://jsoup.org/apidocs/org/jsoup/select/Elements.html)
      * org.jsoup.parser.[**ParseErrorList**](https://jsoup.org/apidocs/org/jsoup/parser/ParseErrorList.html)
* org.jsoup.nodes.[**Attribute**](https://jsoup.org/apidocs/org/jsoup/nodes/Attribute.html) (implements java.lang.[Cloneable](https://docs.oracle.com/javase/7/docs/api/java/lang/Cloneable.html?is-external=true), java.util.[Map.Entry](https://docs.oracle.com/javase/7/docs/api/java/util/Map.Entry.html?is-external=true)<K,V>)
  + org.jsoup.nodes.[**BooleanAttribute**](https://jsoup.org/apidocs/org/jsoup/nodes/BooleanAttribute.html)
* org.jsoup.nodes.[**Attributes**](https://jsoup.org/apidocs/org/jsoup/nodes/Attributes.html) (implements java.lang.[Cloneable](https://docs.oracle.com/javase/7/docs/api/java/lang/Cloneable.html?is-external=true), java.lang.[Iterable](https://docs.oracle.com/javase/7/docs/api/java/lang/Iterable.html?is-external=true)<T>)
* org.jsoup.parser.[**CharacterReader**](https://jsoup.org/apidocs/org/jsoup/parser/CharacterReader.html)
* org.jsoup.safety.[**Cleaner**](https://jsoup.org/apidocs/org/jsoup/safety/Cleaner.html)
* org.jsoup.select.[**Collector**](https://jsoup.org/apidocs/org/jsoup/select/Collector.html)
* org.jsoup.helper.[**DataUtil**](https://jsoup.org/apidocs/org/jsoup/helper/DataUtil.html)
* org.jsoup.nodes.[**Document.OutputSettings**](https://jsoup.org/apidocs/org/jsoup/nodes/Document.OutputSettings.html) (implements java.lang.[Cloneable](https://docs.oracle.com/javase/7/docs/api/java/lang/Cloneable.html?is-external=true))
* org.jsoup.nodes.[**Entities**](https://jsoup.org/apidocs/org/jsoup/nodes/Entities.html)
* org.jsoup.select.[**Evaluator**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.html)
  + org.jsoup.select.[**Evaluator.AllElements**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AllElements.html)
  + org.jsoup.select.[**Evaluator.Attribute**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.Attribute.html)
  + org.jsoup.select.[**Evaluator.AttributeKeyPair**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeKeyPair.html)
    - org.jsoup.select.[**Evaluator.AttributeWithValue**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeWithValue.html)
    - org.jsoup.select.[**Evaluator.AttributeWithValueContaining**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeWithValueContaining.html)
    - org.jsoup.select.[**Evaluator.AttributeWithValueEnding**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeWithValueEnding.html)
    - org.jsoup.select.[**Evaluator.AttributeWithValueNot**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeWithValueNot.html)
    - org.jsoup.select.[**Evaluator.AttributeWithValueStarting**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeWithValueStarting.html)
  + org.jsoup.select.[**Evaluator.AttributeStarting**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeStarting.html)
  + org.jsoup.select.[**Evaluator.AttributeWithValueMatching**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.AttributeWithValueMatching.html)
  + org.jsoup.select.[**Evaluator.Class**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.Class.html)
  + org.jsoup.select.[**Evaluator.ContainsData**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.ContainsData.html)
  + org.jsoup.select.[**Evaluator.ContainsOwnText**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.ContainsOwnText.html)
  + org.jsoup.select.[**Evaluator.ContainsText**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.ContainsText.html)
  + org.jsoup.select.[**Evaluator.CssNthEvaluator**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.CssNthEvaluator.html)
    - org.jsoup.select.[**Evaluator.IsNthChild**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsNthChild.html)
    - org.jsoup.select.[**Evaluator.IsNthLastChild**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsNthLastChild.html)
    - org.jsoup.select.[**Evaluator.IsNthLastOfType**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsNthLastOfType.html)
      * org.jsoup.select.[**Evaluator.IsLastOfType**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsLastOfType.html)
    - org.jsoup.select.[**Evaluator.IsNthOfType**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsNthOfType.html)
      * org.jsoup.select.[**Evaluator.IsFirstOfType**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsFirstOfType.html)
  + org.jsoup.select.[**Evaluator.Id**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.Id.html)
  + org.jsoup.select.[**Evaluator.IndexEvaluator**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IndexEvaluator.html)
    - org.jsoup.select.[**Evaluator.IndexEquals**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IndexEquals.html)
    - org.jsoup.select.[**Evaluator.IndexGreaterThan**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IndexGreaterThan.html)
    - org.jsoup.select.[**Evaluator.IndexLessThan**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IndexLessThan.html)
  + org.jsoup.select.[**Evaluator.IsEmpty**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsEmpty.html)
  + org.jsoup.select.[**Evaluator.IsFirstChild**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsFirstChild.html)
  + org.jsoup.select.[**Evaluator.IsLastChild**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsLastChild.html)
  + org.jsoup.select.[**Evaluator.IsOnlyChild**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsOnlyChild.html)
  + org.jsoup.select.[**Evaluator.IsOnlyOfType**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsOnlyOfType.html)
  + org.jsoup.select.[**Evaluator.IsRoot**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.IsRoot.html)
  + org.jsoup.select.[**Evaluator.Matches**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.Matches.html)
  + org.jsoup.select.[**Evaluator.MatchesOwn**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.MatchesOwn.html)
  + org.jsoup.select.[**Evaluator.MatchText**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.MatchText.html)
  + org.jsoup.select.[**Evaluator.Tag**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.Tag.html)
  + org.jsoup.select.[**Evaluator.TagEndsWith**](https://jsoup.org/apidocs/org/jsoup/select/Evaluator.TagEndsWith.html)
* org.jsoup.examples.[**HtmlToPlainText**](https://jsoup.org/apidocs/org/jsoup/examples/HtmlToPlainText.html)
* org.jsoup.parser.[**HtmlTreeBuilder**](https://jsoup.org/apidocs/org/jsoup/parser/HtmlTreeBuilder.html)
* org.jsoup.helper.[**HttpConnection**](https://jsoup.org/apidocs/org/jsoup/helper/HttpConnection.html) (implements org.jsoup.[Connection](https://jsoup.org/apidocs/org/jsoup/Connection.html))
* org.jsoup.helper.[**HttpConnection.KeyVal**](https://jsoup.org/apidocs/org/jsoup/helper/HttpConnection.KeyVal.html) (implements org.jsoup.[Connection.KeyVal](https://jsoup.org/apidocs/org/jsoup/Connection.KeyVal.html))
* org.jsoup.helper.[**HttpConnection.Request**](https://jsoup.org/apidocs/org/jsoup/helper/HttpConnection.Request.html) (implements org.jsoup.[Connection.Request](https://jsoup.org/apidocs/org/jsoup/Connection.Request.html))
* org.jsoup.helper.[**HttpConnection.Response**](https://jsoup.org/apidocs/org/jsoup/helper/HttpConnection.Response.html) (implements org.jsoup.[Connection.Response](https://jsoup.org/apidocs/org/jsoup/Connection.Response.html))
* java.io.[**InputStream**](https://docs.oracle.com/javase/7/docs/api/java/io/InputStream.html?is-external=true) (implements java.io.[Closeable](https://docs.oracle.com/javase/7/docs/api/java/io/Closeable.html?is-external=true))
  + java.io.[**FilterInputStream**](https://docs.oracle.com/javase/7/docs/api/java/io/FilterInputStream.html?is-external=true)
    - java.io.[**BufferedInputStream**](https://docs.oracle.com/javase/7/docs/api/java/io/BufferedInputStream.html?is-external=true)
      * org.jsoup.internal.[**ConstrainableInputStream**](https://jsoup.org/apidocs/org/jsoup/internal/ConstrainableInputStream.html)
* org.jsoup.[**Jsoup**](https://jsoup.org/apidocs/org/jsoup/Jsoup.html)
* org.jsoup.examples.[**ListLinks**](https://jsoup.org/apidocs/org/jsoup/examples/ListLinks.html)
* org.jsoup.nodes.[**Node**](https://jsoup.org/apidocs/org/jsoup/nodes/Node.html) (implements java.lang.[Cloneable](https://docs.oracle.com/javase/7/docs/api/java/lang/Cloneable.html?is-external=true))
  + org.jsoup.nodes.[**Comment**](https://jsoup.org/apidocs/org/jsoup/nodes/Comment.html)
  + org.jsoup.nodes.[**DataNode**](https://jsoup.org/apidocs/org/jsoup/nodes/DataNode.html)
  + org.jsoup.nodes.[**DocumentType**](https://jsoup.org/apidocs/org/jsoup/nodes/DocumentType.html)
  + org.jsoup.nodes.[**Element**](https://jsoup.org/apidocs/org/jsoup/nodes/Element.html)
    - org.jsoup.nodes.[**Document**](https://jsoup.org/apidocs/org/jsoup/nodes/Document.html)
    - org.jsoup.nodes.[**FormElement**](https://jsoup.org/apidocs/org/jsoup/nodes/FormElement.html)
    - org.jsoup.nodes.[**PseudoTextElement**](https://jsoup.org/apidocs/org/jsoup/nodes/PseudoTextElement.html)
  + org.jsoup.nodes.[**TextNode**](https://jsoup.org/apidocs/org/jsoup/nodes/TextNode.html)
    - org.jsoup.nodes.[**CDataNode**](https://jsoup.org/apidocs/org/jsoup/nodes/CDataNode.html)
  + org.jsoup.nodes.[**XmlDeclaration**](https://jsoup.org/apidocs/org/jsoup/nodes/XmlDeclaration.html)
* org.jsoup.select.[**NodeTraversor**](https://jsoup.org/apidocs/org/jsoup/select/NodeTraversor.html)
* org.jsoup.internal.[**Normalizer**](https://jsoup.org/apidocs/org/jsoup/internal/Normalizer.html)
* org.jsoup.parser.[**ParseError**](https://jsoup.org/apidocs/org/jsoup/parser/ParseError.html)
* org.jsoup.parser.[**Parser**](https://jsoup.org/apidocs/org/jsoup/parser/Parser.html)
* org.jsoup.parser.[**ParseSettings**](https://jsoup.org/apidocs/org/jsoup/parser/ParseSettings.html)
* org.jsoup.select.[**QueryParser**](https://jsoup.org/apidocs/org/jsoup/select/QueryParser.html)
* org.jsoup.select.[**Selector**](https://jsoup.org/apidocs/org/jsoup/select/Selector.html)
* org.jsoup.internal.[**StringUtil**](https://jsoup.org/apidocs/org/jsoup/internal/StringUtil.html)
* org.jsoup.parser.[**Tag**](https://jsoup.org/apidocs/org/jsoup/parser/Tag.html)
* java.lang.[**Throwable**](https://docs.oracle.com/javase/7/docs/api/java/lang/Throwable.html?is-external=true) (implements java.io.[Serializable](https://docs.oracle.com/javase/7/docs/api/java/io/Serializable.html?is-external=true))
  + java.lang.[**Exception**](https://docs.oracle.com/javase/7/docs/api/java/lang/Exception.html?is-external=true)
    - java.io.[**IOException**](https://docs.oracle.com/javase/7/docs/api/java/io/IOException.html?is-external=true)
      * org.jsoup.[**HttpStatusException**](https://jsoup.org/apidocs/org/jsoup/HttpStatusException.html)
      * org.jsoup.[**UnsupportedMimeTypeException**](https://jsoup.org/apidocs/org/jsoup/UnsupportedMimeTypeException.html)
    - java.lang.[**RuntimeException**](https://docs.oracle.com/javase/7/docs/api/java/lang/RuntimeException.html?is-external=true)
      * java.lang.[**IllegalStateException**](https://docs.oracle.com/javase/7/docs/api/java/lang/IllegalStateException.html?is-external=true)
        + org.jsoup.select.[**Selector.SelectorParseException**](https://jsoup.org/apidocs/org/jsoup/select/Selector.SelectorParseException.html)
      * org.jsoup.[**SerializationException**](https://jsoup.org/apidocs/org/jsoup/SerializationException.html)
      * org.jsoup.[**UncheckedIOException**](https://jsoup.org/apidocs/org/jsoup/UncheckedIOException.html)
* org.jsoup.parser.[**TokenQueue**](https://jsoup.org/apidocs/org/jsoup/parser/TokenQueue.html)
* org.jsoup.helper.[**Validate**](https://jsoup.org/apidocs/org/jsoup/helper/Validate.html)
* org.jsoup.helper.[**W3CDom**](https://jsoup.org/apidocs/org/jsoup/helper/W3CDom.html)
* org.jsoup.helper.[**W3CDom.W3CBuilder**](https://jsoup.org/apidocs/org/jsoup/helper/W3CDom.W3CBuilder.html) (implements org.jsoup.select.[NodeVisitor](https://jsoup.org/apidocs/org/jsoup/select/NodeVisitor.html))
* org.jsoup.safety.[**Whitelist**](https://jsoup.org/apidocs/org/jsoup/safety/Whitelist.html)
* org.jsoup.examples.[**Wikipedia**](https://jsoup.org/apidocs/org/jsoup/examples/Wikipedia.html)
* org.jsoup.parser.[**XmlTreeBuilder**](https://jsoup.org/apidocs/org/jsoup/parser/XmlTreeBuilder.html)

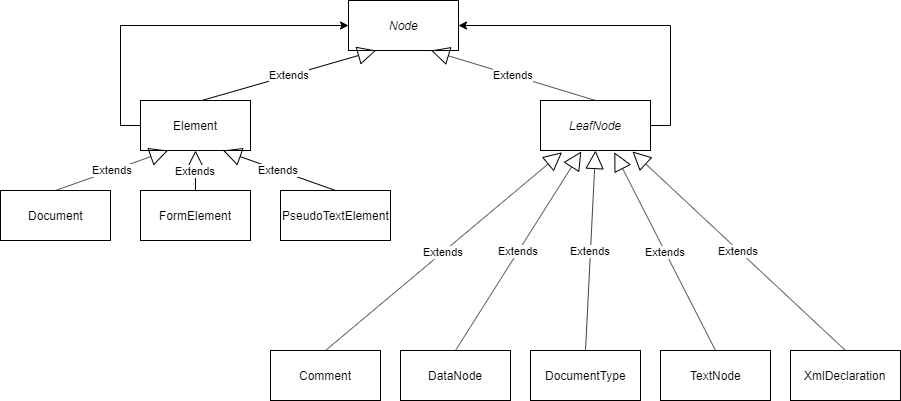
**2. JSoup 설계 및 구현 조사**

**Visitor Pattern**



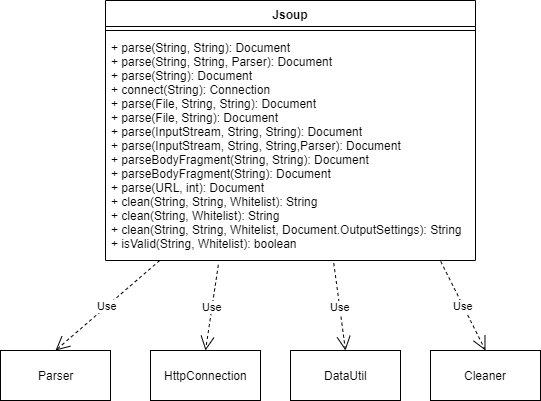
NodeTraversor에는 Node를 depthwise traverse를 하는 traverse 메소드가 존재한다. 이 알고리즘은 Node 클래스와 나눠서 해당 클래스를 오염시키지 않도록 했다. traverse 메소드에서 Node를 순회하기 위해서 추상화된 Visitor가 필요한데 이에 해당되는 것이 NodeVisitor Interface이다. 각 concrete한 visitor는 NodeVisitor를 implement해서 구현되어서 NodeTraversor의 traverse 메소드로 depthwise traverse를 할 때 각각의 concrete한 visitor 클래스를 사용한다.

**Composite Pattern**



Node들은 트리 형태의 구조를 갖는다. 트리 형태를 구현하기 위한 패턴으로 Composite 패턴이 주로 사용되는데 여기서도 Composite 패턴이 사용되었다. Dom은 Node에 모든 계층 정보를 담고 있다. 모든 요소를 Node로 접근하기 위해서 Node 클래스를 abstract class로 만들고, Dom을 구성하는 여러 요소들을 Node를 상속받아 구현하게 만들었다. 이에 해당되는 클래스가 Element 클래스와 LeafNode 클래스이다. Element 클래스는 그 자체로 인스턴스를 만들 수 있는 concrete한 클래스이고 이를 상속받아 Document, FormElement, PseudoTextElement 클래스를 생성할 수 있다. 말단 노드에 해당하는 LeafNode 클래스는 추상 클래스를 이를 상속해서 Comment, DataNode, DocumentType, TextNode, XmlDeclaration이라는 concrete한 객체를 만들 수 있는 클래스를 만들었다. 실제 코드를 보게 되면 Element와 LeafNode가 **private** **static** **final** List<Node> ***EMPTY\_NODES*** = Collections.*emptyList*(); 코드에서 볼 수 있듯이 여러 개의 Node를 리스트로 갖고 있는 것으로 보아, Element와 LeafNode가 Composition Relationship이 있음을 알 수 있다. 이런 관계 때문에 Decorator Pattern이 아닌 Composite Pattern이라고 생각했다.

**Adapter Pattern**



Jsoup 클래스가 Parser, HttpConnection, DataUtil 클래스 등에서 만든 여러 메소드들의 Adatper 역할을 한다. 여러 메소드들을 직접 사용하지 않고 Jsoup 클래스의 메소드가 간접적으로 각 클래스의 메소드들을 호출한다.

호출되는 메소드

- Parser.parse

- Parser.parseInput

- HttpConnection.connect

- DataUtil.load

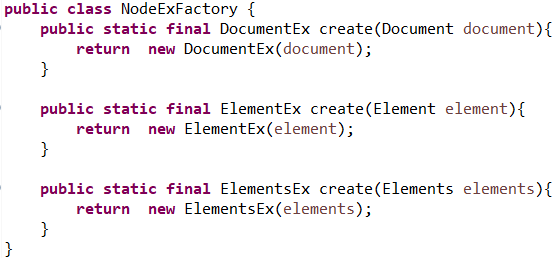
- Cleaner.clean

- Cleaner.isValidBodyHtml

**3. 확장과 설계 개선**

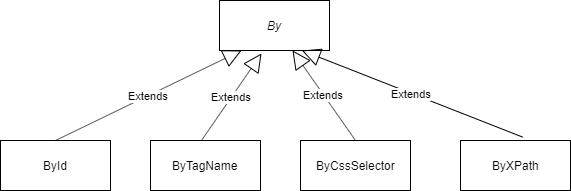
Jsoup 기능 중에서 XML과 XPATH 관련된 기능이 상대적으로 부실한 것을 알 수 있어서 XML과 XPATH 관련 기능 위주로 개선을 했다.

**Factory Method Pattern**



기존 Jsoup의 클래스인 Document, Element, Elements 클래스를 받아서 확장시킨 DocumentEx, ElementEx, ElementsEx 클래스를 만들 수 있도록 하였다. 인스턴스를 만드는 역할은 NodeExFactory에서 create 메소드를 통해 Runtime에서 필요한 인스턴스를 만들 수 있도록 Factory Method Pattern을 적용했다.

**Strategy Pattern + Adapter Pattern**



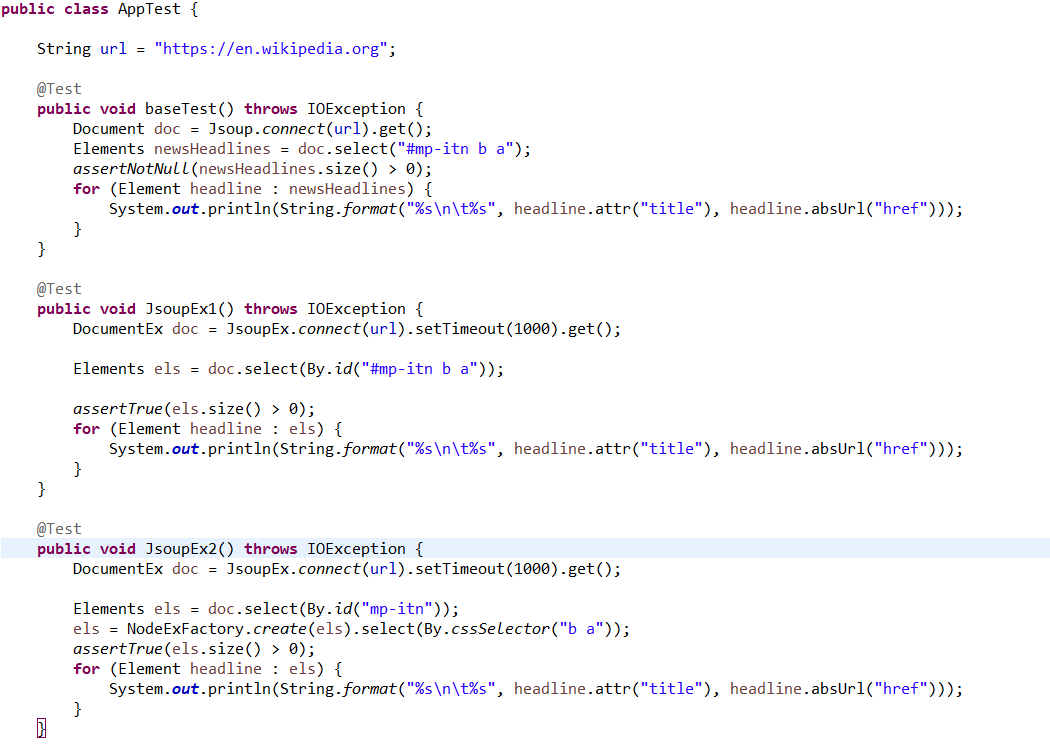
By라는 Abstract 클래스를 만들고 이 클래스를 상속받아 FindElements 메소드를 각 클래스에 맞게 구현했다. ById는 Element를 Id로 찾고, ByTagName은 Element를 Tag로 찾고, ByCssSelector는 Element를 CssSelector로 찾는다. 여기까지는 Jsoup을 재사용했다고 할 수 있다. ByXPath는 JsoupXPath 라이브러리에서 XPath를 이용해서 Element를 찾는 메소드를 우리가 확장해서 사용할 수 있도록 ByXPath라는 Adapter를 통해서 동일한 인터페이스로 이용할 수 있도록 했다. 그리고 DocumentEx, ElementEx, ElementsEx에서 Element를 찾는 여러 알고리즘 중에 하나를 선택해서 사용할 수 있도록 Strategy Pattern을 적용했다. 따라서 클라이언트에 해당하는 DocumentEx, ElementEx, ElementsEx 클래스는 select 메소드를 사용할 때 필요한 알고리즘에 해당되는 클래스를 파라미터로 사용해서 적절한 탐색을 할 수 있다.

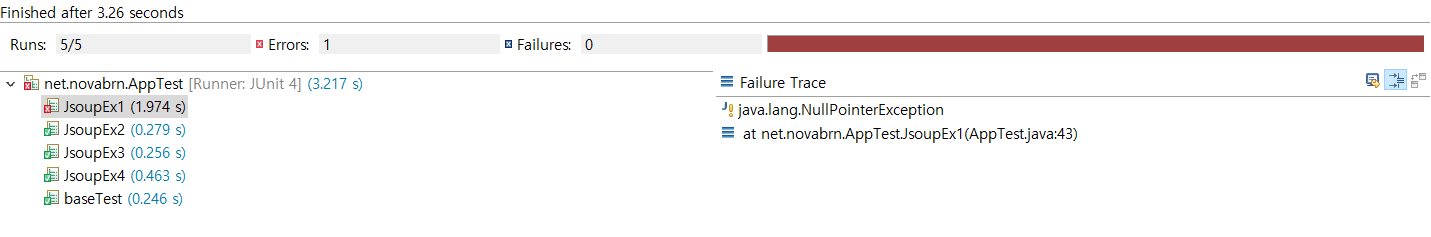
각 클래스의 select 메소드



**4. Test**

select 메소드 동작 Test



콘솔 일부 캡쳐



