Textparsing

Text를 문장단위로 파싱해주는 모듈

적용 Parsing rules (v3기준 20190904)

```
1. 줄바꿈 parsing (GetRefined.cropLine)
2. 공백 제거 (GetRefined.cropLine)
3. BreakIterator
4. [.숫자/알파벳] parsing (GetRefined.confirmSentence)
5. [.] 뒤에 띄어쓰기 없어도 Parsing (GetRefined.confirmSentence)
6. 인용문 내의 ?와 !으로 끝나는 문장 Reattach (GetRefined.Reattach)
7. [숫자목록. \n 문장] 문장들 Reattach (GetRefined.Reattach)
```

Getting Started

Installing

1. 압축파일을 원하는 디렉토리(모듈디렉토리)에 저장 후 압축해제

Prerequisites

1. 파싱하고자 하는 문장들을 Excel 파일로 작성
    - Input Excel file 작성 형식

No	Text
1	Text Ex

    - Input Excel file 예시
2. 작성한 Input Excel file을 모듈디렉토리 내의 input 폴더에 저장
3. Input Excel file 파일명 설정
   (1) 프로그램 실행 시 파일명 입력
   혹은
   (2) 모듈디렉토리/conf 디렉토리 내의 information.xml 파일을 text 편집기(ex notepad)를 이용하여 연 후 filename 태그의 name 속성에 Input Excel file 파일명 입력 후 저장
    - `<filename name=Input Excel file 파일명>`

Running

1. 윈도우 cmd 창에서 압축파일이 해제된 디렉토리로 이동

```
C:\Users\USER>cd C:\Users\USER\Desktop\TextParsing_v3
```

2. jar파일 실행

```
C:\Users\USER\Desktop\TextParsing_v2>java -jar TextParsing_v3.jar
```

Versioning

- Textparsing_v2 (2019-09-02)
- Textparsing_v3 (2019-09-04)
    - parsing rule 추가
    - output 파일의 no를 input 파일의 no로 설정
- Textparsing_v3 수정 (2021-01-19)
    - 첫 번째 Parsing rule(줄바꿈 parsing) \r\n 은 인식 안되는 버그 수정, 현재 \n 과 \r\n 모두 파싱됨.

Authors

- JiyeonLee(woogjlee@gmail.com), 2021