Detokenization

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Tokenization

1. 영어 원문
   - There's currently over a thousand TED Talks on the TED website.

2. tokenization을 수행하고, 기존 띄어쓰기와 구분을 위해 ▁ (U+2581) 삽입
   - ▁There 's ▁currently ▁over ▁a ▁thousand ▁TED ▁Talks ▁on ▁the ▁TED ▁website .

3. subword segmentation을 수행, 공백 구분 위한 ▁ 삽입
   - ▁▁There ▁'s ▁▁currently ▁▁over ▁▁a ▁▁thous and ▁▁TED ▁▁T al ks ▁▁on ▁▁the ▁▁TED ▁▁we b site ▁.

Detokenization

1. whitespace를 제거
   - ▁▁There▁'s▁▁currently▁▁over▁▁a▁▁thousand▁▁TED▁▁Talks▁▁on▁▁the▁▁TED▁▁website▁.
2. ▁▁을 white space로 치환
   - There▁'s currently over a thousand TED Talks on the TED website▁.
3. ▁를 제거
   - There's currently over a thousand TED Talks on the TED website.

```
sed "s/ //g" | sed "s/__/ /g" | sed "s/_//g"
```