인공지능 학습활동 5

1829008 김민영

0. Set Up

```
In [125]: import nltk
          from nltk.corpus import *
          corpus_root = "C:/Users/KimMinyoung/nltk_data/corpora/Genomics-Informatics-Corpus-master/Genomics-Informatics-C
```

```
In [126]: GNI = nltk.corpus.PlaintextCorpusReader(corpus_root,'gni-14-1.txt',encoding='utf-8')
          giRaw = GNI.raw('gni-14-1.txt')
          GNISents = nltk.sent_tokenize(giRaw)

          problem = []
```

1) Gni-14-1.txt

1. "."만을 기준으로 문장을 분리하기 때문에, 특수문자와 제목이 한 문장에 들어갔다.

```
In [127]: problem.append(GNISents[0])
          GNISents[0]
```

```
Out[127]: '₩n============Title=========₩nEditor' s Introduction to This Issue.'
```

"." 만을 기준으로 문장을 분리하기 때문에, 특수문자랑, 제목이랑 같이 들어갔다.

개선 방안 -> '---' 나 '==='가 반복적으로 나타나면 문장분리기호로 인식하도록 개선한다. 혹은 \n을 추가적으로 문장분리기호로 인식하도록 개선한다.

-> **개선방안** : '---' 나 '==='가 반복적으로 나타나면 문장분리기호로 인식하도록 개선한다. 혹은 ₩n을 추가적으로 문장분리기호로 인식하도록 개선한다.

2,3. "."을 기준으로 문장을 분리하기 때문에 Vol. No. 에서 문장으로 잘렸다

```
In [128]: problem.append(GNISents[1])
          GNISents[1]
```

```
Out[128]: '============Cor Author==========₩nwww.genominfo.orgGenomics & Informatics Vol.'
```

```
In [129]: problem.append(GNISents[2])
          GNISents[2]
```

```
Out[129]: '14, No.'
```

-> **개선방안** : Vol. No. 등 "."로 끝나는 약어사전을 만들어 문장을 분리할 때에 제거한다.

4. "."을 기준으로 문장을 분리하기 때문에, 이메일 전화번호 등등 여러가지가 하나의 문장으로 들어가게 되었다.

```
In [130]: problem.append(GNISents[3])
          GNISents[3]

Out[130]: '1, 2016eISSN2234-0742Genomics Inform 2016;14(1):1http://dx.doi.org/10.5808/GI.2016.14.1.1*Corresponding autho
          r: Tel: +82-2-2258-7343, Fax: +82-2-537-0572, E-mail: yejun@catholic.ac.krWn==========Author=========WnYeun-
          Jun Chung* IRCGP, College of Medicine, The Catholic University of Korea, Seoul 06591, KoreaIn the post-genome
          era, understanding protein biomarkers is becoming more important.'
```

-> 개선방안 : 마찬가지로 '---' 나 '==='가 반복적으로 나타나면 문장분리기호로 인식하도록 개선한다. 혹은 ₩n을 추가적으로 문장분리기호로 인식하도록 개선한다.

5. "."을 기준으로 문장을 분리하기 때문에 Dr. 에서 문장으로 잘렸다.

```
In [131]: problem.append(GNISents[11])
          GNISents[11]

Out[131]: 'Regarding neurobiology, Dr.'
```

-> 개선방안 : Dr. 등 "."로 끝나는 약어사전을 만들어 문장을 분리할 때에 제거한다.

6. "."을 기준으로 문장을 분리하기 때문에, 이메일 전화번호 등등 여러가지가 하나의 문장으로 분리되어버렸다.

```
In [132]: problem.append(GNISents[-1])
          GNISents[-1]

Out[132]: 'Dr. Seon-Young Kim's group (KRIBB, Korea) suggests that public datasets should not be expected to be error-f
          ree and, whenever possible, that we should check the consistency of the data.For further details, please visit
          the G&I homepage (http://www.kogo.or.kr/webapp/kogo_publish/genomics_and_informatics/)Wn==========Keywords===
          ======WnWn==========Abstract==========WnWn==========Main Text==========Wn.'
```

개선방안 -> 마찬가지로 '---' 나 '==='가 반복적으로 나타나면 문장분리기호로 인식하도록 개선한다. 혹은 ₩n을 추가적으로 문장분리기호로 인식하도록 개선한다.

2) Gni-14.2.txt

```
In [133]: giRaw2 = GNI.raw('gni-14-2.txt')
          GNISents2 = nltk.sent_tokenize(giRaw2)
```

7. Result, After~ 등등이 한꺼번에 다 같은 문장에 들어가게 되었고 마지막에 Fig.에서 짤렸다.

```
In [134]: problem.append(GNISents2[20])
          GNISents2[20]

Out[134]: 'Results₩n₩nPeptide and protein identification₩n₩nAfter obtaining proteomic data from tissue or body fluid sam
          ples using liquid chromatographytandem mass spectrometry (LC-MS/MS) analysis, the tandem mass spectrometry (M
          S/MS) spectra are first searched against a protein sequence database (e. g., SWISS-Prot or UniProt) to identi
          fy the peptide sequences for individual MS/MS spectra (peptide/protein identification) (Fig.'
```

개선방안 -> Fig. 등 "."로 끝나는 약어사전을 만들어 문장을 분리할 때에 제거한다. 또한 ₩n₩n 을 추가적으로 문장분리기호로 인식하도록 개선한다.

8. "."을 기준으로 문장을 분리하기 때문에 홈페이지주소가 . 단위로 잘렸다.

```
In [136]: problem.append(GNISents2[64])
          GNISents2[64]

Out[136]: 'For example, the enrichment analysis of gene ontology biological processes (GOBPs) or Kyoto Encyclopedia of G
          enes and Genomes (KEGG) pathways can be applied to the DEPs using DAVID [38] and PANTHER [39], and commercial
          tools, such as MetaCore [40] and Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, http://www.'

In [137]: problem.append(GNISents2[65])
          GNISents2[65]

Out[137]: 'qiagen.'
```

개선방안 -> "."이후에 ₩s 가 오는 경우에만 문장을 분리하도록 개선한다.

9. "."을 기준으로 문장을 분리하기 때문에 Fig.에서 잘렸다.

```
In [138]: problem.append(GNISents2[90])
          GNISents2[90]

Out[138]: 'Next, the network model are analyzed to identify network modules or clusters each of which includes a set of
          the nodes densely connected in the network (Fig.'
```

개선방안 -> Fig. 등 "."로 끝나는 약어사전을 만들어 문장을 분리할 때에 제거한다. 또한 ₩n₩n 을 추가적으로 문장분리기호로 인식하도록 개선한다.

10. "."을 기준으로 문장을 분리하기 때문에 홈페이지주소가 . 단위로 잘렸다.

```
In [141]: problem.append(GNISents2[127])
          GNISents2[127]

Out[141]: 'A number of the tools have been developed to understand the subnetworks (network clusters) of the multi-layer
          ed networks whose perturbations are collectively indicated by different types of global datasets, including IP
          A (QIAGEN Redwood City, http://www.'
```

개선방안 -> -> "."이후에 ₩s 가 오는 경우에만 문장을 분리하도록 개선한다.