
Preprocessing Interim Findings



NLP Lab
Report

양민석
2023. 04. 21.

CONTENTS

Summary

Corpus Preprocessing

01 AIHUB_전문분야 말뭉치

AIHUB_professional_corpus.txt

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

03 AIHUB_산업정보 연계 주요국 특허 영-한 데이터

AIHUB_british_korean_data_on_patents_in_major_countries_linked_to_industrial_information.txt

04 AIHUB_논문자료 요약

AIHUB_summary_of_thesis_materials.txt

05 AIHUB_요약문 및 레포트 생성 데이터

AIHUB_summary_and_report_generation_data.txt

06 AIHUB_문서요약 텍스트

AIHUB_document_summary_text.txt

CONTENTS

07 AIHUB_특허 분야 자동분류 데이터

AIHUB_automatic_patent_classification_data.txt

08 AIHUB_뉴스 기사 기계독해 데이터

AIHUB_news_article_machine_reading_data.txt

09 AIHUB_행정 문서 대상 기계독해 데이터

AIHUB_machine_reading_data_for_administrative_documents.txt

10 AIHUB_기계독해

AIHUB_machine_reading.txt

11 AIHUB_도서자료 요약

AIHUB_summary_of_book_materials.txt

12 AIHUB_대규모 구매도서 기반 한국어 말뭉치 데이터

AIHUB_korean_corpus_data_based_on_large_scale_purchase_books.txt

13 AIHUB_도서자료 기계독해

AIHUB_reading_books_by_machine.txt

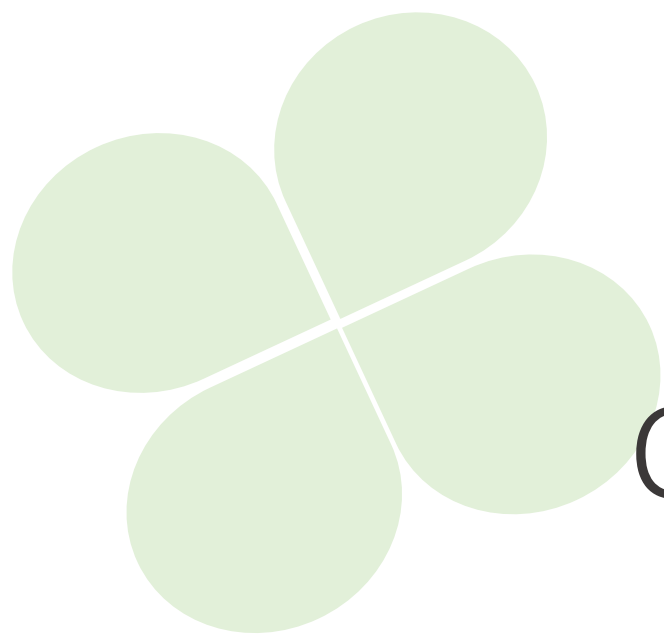
CONTENTS

14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_text_analysis_data.txt

15 일반상식

AIHUB_general_common_sense.txt



Summary

Corpus Preprocessing

Summary

Corpus Preprocessing

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer

- **re.match:** 특수기호로 시작하는 Token 제외

. | , | ◆ | ◇ | △ | ▲ | ▽ | ▼ | ▷ | ► | < | > | (|) | < | 【 | ◆ | 《 | / | ○ | - |

(특수기호 전처리의 경우, 코드를 짜면서 서서히 만들어진 특수기호 모음이기 때문에

이 ‘특수기호 제외 전처리 알고리즘’이 모든 말뭉치 데이터에 적용된 것은 아니다.

예를 들어, 어떤 말뭉치 데이터에서 전처리할 때는 ◆, ○ 제외 알고리즘이 빠져 있는 등 경우가 다르다.

그래서 어떤 데이터에서는 특수기호 ◆, ○로 시작하는 문장이 존재한다.)

조항을 지칭하는 제목 Token 제외

민사법조항 Token. 판례문 Token. 관계법령 Token 제외

- **len:** 공문서상 짧은 Token은 문장이 아닌 경우(이름, 직함, 섹션 등)가 대부분이라 제외



01 AIHUB_전문분야 말뭉치

AIHUB_professional_corpus.txt

01 AIHUB_전문분야 말뭉치

AIHUB_professional_corpus.txt

```
from kss import split_sentences

split_sentences(
    text: Union[str, List[str], Tuple[str]],
    backend: str = "auto",
    num_workers: Union[int, str] = "auto" ,
    strip: bool = True,
    ignores: List[str] = None,
)
```

- **history:** 처음에는 C++로 구현되었고, 파이썬 포팅버전은 Cython으로 연동됨
- **text:** String or List/Tuple of strings
- **backend:** Morpheme analyzer backend MeCab
- **num_workers:** The number of multiprocessing workers
- **strip:** Whether it does strip() for all output sentences or not
- **ignores:** ignore strings to do not split

01 AIHUB_전문분야 말뭉치

AIHUB_professional_corpus.txt

```
elif '특허_0' in file_name_list[i] or '특허_1' in file_name_list[i] or \
     '법령' in file_name_list[i] or '판례' in file_name_list[i]:

    for j in one_json_sample['data'][:]:

        for sentence in kss.split_sentences(j['sentence'][0]['text']):

            if bool(re.match(r'[.]|[,]|[\u2666]|[\u2667]|[\u25b2]|[\u25b3]|[\u25bc]|[\u25c0]|[\u25ba]|[\u27e8]|[\u27ea]|[\u27e9]|[\u27eb]|[\u27f4]|[\u27f5]|[\u27f6]|[\u27f7]|[\u27f8]|[\u27f9]',
sentence[0])) == False and \
                bool(re.match(r'[0-9]+\.', sentence[:2])) == False and \
                bool(re.match(r'제+[0-9]+항|제+[0-9]+조', sentence[:3])) == False and \
                bool(re.match(r'제+[0-9]+0-9+항|제+[0-9]+0-9+조', sentence[:4])) == False and \
                bool(re.match(r'+[0-9]+', sentence[:3])) == False and \
                len(sentence) > 55:

                sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외. 조항을 지칭하는 제목 Token 제외
- **len:** 공문서상 짧은 Token은 문장이 아닌 경우(이름, 직함, 섹션 등)가 대부분이라 제외

01 AIHUB_전문분야 말뭉치

AIHUB_professional_corpus.txt

[illegible]

「성폭력범죄의 처벌 등에 관한 특례법」 제3조부터 제10조까지 및 제15조(제3조부터 제9조까지의 미수범으로 한정한다)의 죄바.

「아동·청소년의 성보호에 관한 법률」 제7조, 제8조, 제11조부터 제16조까지 및 제17조제1항의 죄4, 「국가공무원법」 제2조 및 「지방공무원법」 제2조에 규정된 공무원과 국가나 지방자치단체에서 일상적으로 공무에 종사하는 「

㉠ 국가보훈처장은 제1항에 따라 이 법의 적용 대상에서 제외된 사람이 다음 각 호의 어느 하나에 해당하게 되면 그 누위친 정도가 현저하다고 인정되는 경우에만 제4조에 따라 등록신청을 받아 이 법의 적용대상자로 결정하여 보상할 수 있다.

제73조(포상금의 지급) ① 국가보훈처장은 다음 각 호의 어느 하나에 해당하는 사람을 신고한 사람에게는 예산의 범위에서 신고포상금을 지급할 수 있다. 1. 제4조를 위반하여 거짓이나 그 밖의 부정한 방법으로 보훈보상대상자, 그 유족 또는 제74조(위임 및 위탁) ① 이 법에 따른 국가보훈처장의 권한은 대통령령으로 정하는 바에 따라 그 일부를 그 소속 기관의 장 또는 제주특별자치도지사에게 위임하거나 다른 행정기관의 장에게 위탁할 수 있다.

제 1 항에 있어서, 상기 바이오매스수지는 폴리에틸렌(polyethylene, PE), 폴리프로필렌(polypropylene, PP), 폴리에틸렌테레프탈레이트(polyethylene terephthalate, PET), 폴리페닐렌에테르(polyp

제 1 항에 있어서, 상기 바이오매스 수지는 전분계 바이오매스 물질을 포함하는 건축 자재 제조 방법.

제 1 항에 있어서, 상기 목재를 제 1 층으로서 준비하는 단계는 로터리 커팅(rotary cutting)을 이용하여 상기 목재를 기 설정된 값 이하의 두께를 갖는 판재 형태로 가공하는 단계를 포함하는 건축 자재 제조 방법.

제 1 항에 있어서, 상기 목재를 제 1 층으로서 준비하는 단계는 기 정해진 제 2 온도 이상의 온도에서 상기 제 1 층의 수분을 제거하는 단계를 포함하는 건축 자재 제조 방법

제 1 항 및 제 3 항 내지 제 13 항 중 어느 한 항에 있어서, 포유동물에서 유연성 및 편안한 움직임을 촉진 또는 향상시키기 위한 약학적 조성물.

제 1 항 및 제 3 항 내지 제 13 항 중 어느 한 항에 있어서, 포유동물에서 뼈 건강, 연골 건강 또는 둘 모두를 지지하기 위한 약학적 조성물.

제 1 항 및 제 3 항 내지 제 13 항 중 어느 한 항에 있어서, 포유동물에서 건강한 뼈 기능, 연골 기능, 뼈의 편안함, 연골의 편안함 또는 이들의 임의의 조합을 촉진하기 위한 약학적 조성물.

조항. 항목



02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

```
source_list = list(pd.DataFrame(one_json_sample['SJML']['text'])['content'])

for source in source_list:

    for sentence in kss.split_sentences(source):

        if bool(re.match(r'[.],[\u25c0][\u25b6][\u25aa][\u25ba][\u25bc][\u25c4][\u25c6][\u25c8][\u25ca][\u25cb][\u25cd][\u25cf][\u25d0][\u25d2][\u25d4][\u25d6][\u25d8][\u25da][\u25dc][\u25de][\u25df][\u25e0][\u25e2][\u25e4][\u25e6][\u25e8][\u25ea][\u25ec][\u25ee][\u25f0][\u25f2][\u25f4][\u25f6][\u25f8][\u25fa][\u25fc][\u25fd][\u25fe][\u25ff]', sentence[0])) == False:

            sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

‘프로젝트a(가칭)’로 진행 중인 해당 게임은 콘솔게임 급의 방대한 시나리오와 애니메이션, 액션연출을 더한 수집형 액션rpg(역할수행게임)다.. .
우수한 게임개발진 확보 및 개발사 인수합병(m&a), 스트리밍게임·멀티플랫폼 등의 서비스기술에 대한 연구개발 투자도 지속해 중장기 성장동력으로 삼는다.. . (이름)
일부 유저는 청와대 청원게시판에 글을 올리면서 불매 운동을 유도하기도 했다. . .

[기술문화①] 시키는 대로만 하면 창조는 없다.

[기술문화②] 전복적 이노베이션 환경 만들어야.

(서울=뉴스1) (이름) 기자 = 깜찍한 sd 캐릭터(2~3등신 캐릭터)가 소형 카트를 타고 경주를 벌이는 '카트라이더 러쉬플러스'가 흥행질주를 이어가고 있다.

게임 과몰입의 직접적 원인은 게임에 빠질 수밖에 없는 환경 때문이라고 생각한다.(이름)

의원이 주최한 정책세미나.

두번째 줄 세번째 인물이 전 매니저. q. <마크>야 널리 알려진 좋은 게임이니까 그렇다 치고, 부모는 아이가 밤새 <리그 오브 레전드>(이하 롤)나 <gta>를 할까 걱정 아닌가?
a.

이 8주년 기사는 누구의 생각인가요?

// 세설 저요.

(당연한가?) 깨

쓰통님은 이제 몇 번째 고참이신가요?

IT/과학 코퍼스: 따옴표. 괄호. 백슬래시. 물음표

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

apex legends hit 50 million players worldwide!

we are humbled by all your support and can't wait to show you what's next. (이메일)

/qtdh57lfvb— apex legends (@playapex) 2019년 3월 4일 해당 내용은 3월 5일, <에이펙스 레전드> 서비스 1개월 기념 트윗을 통해 밝혀졌다.

/ 디스이즈게임 (이름) 기자 # 모바일 환경에 맞춘 구성 먼저, 게임에 들어서면 꽤 간단한 구성으로 되어있음을 볼 수 있다.

■ 정부 혁신성장 성적표 낙제점 수준....

[소박스]☞목재펠릿이란?.

= 삼성전자가 갤럭시s7 언팩행사를 vr(가상현실)로 생중계한다.

kyustar@(이메일)<저작권자© 공감언론 뉴시스통신사. 무단전재-재배포 금지.>

【서울=뉴시스】 (이름) 기자 = 미국 국방부는 '이슬람 국가'(is)가 이라크의 미군 기지에 화학무기 공격을 가했다는 의혹이 사실로 확인됐다고 22일(현지시간) 밝혔다..

copyright © (이메일), 무단 전재 및 재배포 금지

©공감언론 뉴시스 hipyun@(이메일)

'국민 보건에 심대한 해 끼칠 수 있어 책임 엄중'..

IT/과학 코퍼스: 영어. 기자. 사각형. 등호. 이메일. 대괄호. Copyright. 제목

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

필요(이름) 생각은 버리고 가을바람을 타고 여행이라도.▶소띠= 참을 인자 세 개면 살인도 면한다.
자기 앞에 주어진 일만 성실히 하면 오해는 시간이 풀어줄 듯.
금전문제로 약간 힘에 겹지만 쥐, 말띠에게 도움 청하면 해결될 듯.
애정에는 방해자가 나타나니 감정을 억제.▶호랑이띠= 진실한 마음과 마음으로 이어져야 사랑에 갈등이 없다.
책임있는 말과 행동이 요구되는 날이다. ㅂ
추진 중인 일은 서서히 성사될 수.▶토끼띠= 지금까지 지연되었던 일들이 순조롭게 풀리니 차분한 마음으로 힘들어도 하나씩 추진하라.
처음보다 나중에 큰 이익이 따라주니 투자하면 대성할 수.
5, 7, 9월생 여자는 시집식구와 의견충돌로 마음고생이 크겠다.
자신감을 되찾고 냉정한 판단이 요구됨.
동, 북쪽에서 도움 줄 자 나타나니 새롭게 시작한다는 각오가 필요. ㄹ
마음먹은 일 밀고 나갈 때 귀인의 도움이 있겠다. ㅅ
의견대립으로 거래처와 다툼이 예상된다.
서, 남쪽이 길.▶말띠= 착실함을 제일로 하라.
성급함이나 경솔한 태도를 보이지 않는다면 최고의 날이 될 듯.
사적인 일보다는 공적인 일에서 성과 있겠다.
7, 9, 12월생 아무리 상대에게 잘 해준다 해도 멀어져 가는 것이 애정이다.
잡으려고만 말고 자유를 주어라.
매사 느긋한 자세로 참고 노력하라.
의외로 좋은 결과가 기다린다.
힘을 가져라.
하면 된다는 용기가 절대 필요한 오늘이다.

건강 코퍼스: 운세

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터

AIHUB_web_data_based_korean_corpus_data_source.txt

[오피니언].
양상문 감독과lg 그리고 맨유의 교훈.
kbs는 공정할 수 없다.
ceo보다 더 중요한게 기업문화. .
[(이름)의 증권 반세기].
파동과 휴장의 연속, '바람 잘 날 없는' 명동 증권가. .
[wealth].
똑소리나는 모바일 장바구니테크 고객님~ '엄지도장'만 찍으세요. .
[book].
슬픔에 젖은 대(이름) 지금 필요한 리더는.
다른듯 닮은 예술 일상을 위로하다. .
-유양디앤유는 22일 운영자금 목적의 100억원 규모 전환사채권 발행을 결정했다고 공시했다.. .
※ cap스탁론 상담센터 : 1644-1728. 바로가기 /.
굽네
치킨만의 특제 파우더를 입히고 오븐에 구워내 겉은 바사삭하고 속은 촉촉하다.

경제 코퍼스: 제목. 하이픈.

02 AIHUB_웹데이터 기반 한국어 말뭉치 데이터 AIHUB_web_data_based_korean_corpus_data_source.txt

트렌치코트가 멋스럽게 잘 어울린다.
트렌치코트, 터틀넥, 트윌 팬츠, 벨트, 키 체인, 스니커즈 모두 가격미정 디올 옴므.
파리의 오래된 건물 안에 주저앉아 먼 곳을 바라보는 그.
터틀넥, 팬츠 모두 가격미정 디올 옴므.
부츠 21만원 닥터마틴.
양말 스타일리스트 소장품.드라마 <하백의 신부> 촬영이 끝나자마자 파리로 온 거죠?
어떻게 이 드라마를 선택하게 됐나요?
일단 대본이 정말 재미있었어요.
'하백'이라는 캐릭터가 '신'이라는 점이 (이름)면서 매력적으로 다가오기도 했고요.
학창 시절엔 운동선수였고, 이후 모델로 시작해 배우로 커리어가 안정적으로 자리 잡게 됐어요.
지금 생각해보면 시간이 또 다르게 느껴질 것 같아요.
어렸을 때는 농구 선수가 되는 게 꿈이었기 때문에 배우나 모델은 생각해본 적이 없어요.
부상당하고 진로를 다시 고민하면서 이쪽 일을 새롭게 꿈꾸게 됐죠.
지금 생각해보면 꽤 신기한 일이에요.
어린 시절 남주혁은 어떤 아이였나요?
지금보다도 훨씬 개구쟁이였어요.
지금도 장난기가 많긴 하지만 소심한 면이 있고 차분한 편이기도 하거든요.
보고 있으면 빠져들 것만 같은 그의 눈빛.
터틀넥 가격미정 디올 옴므.
시계, 팔찌 모두 가격미정 까르띠에.
best exfoliator 스킨푸드 블랙슈가 마스크 워시오프 7천7백원 '싼 게 비지떡'이라는 건 이제 편견에 지나지 않는다.
구조 요청!

라이프스타일/문화 코퍼스: 잡지 인터뷰. 영어



03 AIHUB_산업정보 연계 주요국 특허 영-한 데이터

AIHUB_british_korean_data_on_patents_
in_major_countries_linked_to_industrial_information.txt

03 AIHUB_산업정보 연계 주요국 특허 영-한 데이터

AIHUB_british_korean_data_on_patents_in_major_countries_linked_to_industrial_information.txt

```
source_list = list(pd.DataFrame(one_json_sample['labeled_data'])['astrt_cont_kor'])

for source in source_list:

    for sentence in kss.split_sentences(source):

        if bool(re.match(r'[.],[\u25c0][\u25b6][\u25aa][\u25ba][\u25d4][\u25d6][\u25e0][\u25e2][\u25f4][\u25f6][\u2600][\u2602][\u2610][\u2612][\u2620][\u2622][\u2630][\u2632][\u2640][\u2642][\u2650][\u2652][\u2660][\u2662][\u2670][\u2672][\u2680][\u2682][\u2690][\u2692][\u26a0][\u26a2][\u26b0][\u26b2][\u26c0][\u26c2][\u26d0][\u26d2][\u26e0][\u26e2][\u26f0][\u26f2][\u2700][\u2702][\u2710][\u2712][\u2720][\u2722][\u2730][\u2732][\u2740][\u2742][\u2750][\u2752][\u2760][\u2762][\u2770][\u2772][\u2780][\u2782][\u2790][\u2792][\u27a0][\u27a2][\u27b0][\u27b2][\u27c0][\u27c2][\u27d0][\u27d2][\u27e0][\u27e2][\u27f0][\u27f2][\u2800][\u2802][\u2810][\u2812][\u2820][\u2822][\u2830][\u2832][\u2840][\u2842][\u2850][\u2852][\u2860][\u2862][\u2870][\u2872][\u2880][\u2882][\u2890][\u2892][\u28a0][\u28a2][\u28b0][\u28b2][\u28c0][\u28c2][\u28d0][\u28d2][\u28e0][\u28e2][\u28f0][\u28f2][\u2900][\u2902][\u2910][\u2912][\u2920][\u2922][\u2930][\u2932][\u2940][\u2942][\u2950][\u2952][\u2960][\u2962][\u2970][\u2972][\u2980][\u2982][\u2990][\u2992][\u29a0][\u29a2][\u29b0][\u29b2][\u29c0][\u29c2][\u29d0][\u29d2][\u29e0][\u29e2][\u29f0][\u29f2][\u2a00][\u2a02][\u2a10][\u2a12][\u2a20][\u2a22][\u2a30][\u2a32][\u2a40][\u2a42][\u2a50][\u2a52][\u2a60][\u2a62][\u2a70][\u2a72][\u2a80][\u2a82][\u2a90][\u2a92][\u2aa0][\u2aa2][\u2ab0][\u2ab2][\u2ac0][\u2ac2][\u2ad0][\u2ad2][\u2ae0][\u2ae2][\u2af0][\u2af2][\u2b00][\u2b02][\u2b10][\u2b12][\u2b20][\u2b22][\u2b30][\u2b32][\u2b40][\u2b42][\u2b50][\u2b52][\u2b60][\u2b62][\u2b70][\u2b72][\u2b80][\u2b82][\u2b90][\u2b92][\u2ba0][\u2ba2][\u2bb0][\u2bb2][\u2bc0][\u2bc2][\u2bd0][\u2bd2][\u2be0][\u2be2][\u2bf0][\u2bf2][\u2c00][\u2c02][\u2c10][\u2c12][\u2c20][\u2c22][\u2c30][\u2c32][\u2c40][\u2c42][\u2c50][\u2c52][\u2c60][\u2c62][\u2c70][\u2c72][\u2c80][\u2c82][\u2c90][\u2c92][\u2ca0][\u2ca2][\u2cb0][\u2cb2][\u2cc0][\u2cc2][\u2cd0][\u2cd2][\u2ce0][\u2ce2][\u2cf0][\u2cf2][\u2d00][\u2d02][\u2d10][\u2d12][\u2d20][\u2d22][\u2d30][\u2d32][\u2d40][\u2d42][\u2d50][\u2d52][\u2d60][\u2d62][\u2d70][\u2d72][\u2d80][\u2d82][\u2d90][\u2d92][\u2da0][\u2da2][\u2db0][\u2db2][\u2dc0][\u2dc2][\u2dd0][\u2dd2][\u2de0][\u2de2][\u2df0][\u2df2][\u2e00][\u2e02][\u2e10][\u2e12][\u2e20][\u2e22][\u2e30][\u2e32][\u2e40][\u2e42][\u2e50][\u2e52][\u2e60][\u2e62][\u2e70][\u2e72][\u2e80][\u2e82][\u2e90][\u2e92][\u2ea0][\u2ea2][\u2eb0][\u2eb2][\u2ec0][\u2ec2][\u2ed0][\u2ed2][\u2ee0][\u2ee2][\u2ef0][\u2ef2][\u2f00][\u2f02][\u2f10][\u2f12][\u2f20][\u2f22][\u2f30][\u2f32][\u2f40][\u2f42][\u2f50][\u2f52][\u2f60][\u2f62][\u2f70][\u2f72][\u2f80][\u2f82][\u2f90][\u2f92][\u2fa0][\u2fa2][\u2fb0][\u2fb2][\u2fc0][\u2fc2][\u2fd0][\u2fd2][\u2fe0][\u2fe2][\u2ff0][\u2ff2]', sentence[0])) == False:

            sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외

03 AIHUB_산업정보 연계 주요국 특허 영-한 데이터

AIHUB_british_korean_data_on_patents_in_major_countries_linked_to_industrial_information.txt

A: 셀룰로오스 수지, B: 화학식 1로 표시되는 유기산과 다가 알코올을 축합하여 얻어지는 에스테르 화합물, C: 지방족 폴리에스테르 또는 지방족-방향족 공중합체.

[과제] 피압연재의 사행 및 캠버의 발생을 억제한다.

[해결 수단] 본 발명의 감광성 수지 조성물은, (A) 하기 일반식 (1) 또는 (2)로 나타나는 에스테르기 함유 폴리이미드 전구체와, (B) 감광제를 포함하는 것을 특징으로 한다.

너비가 w_1 인 $cn \dots$

신호(s_1, s_2, p)의 쌍을 이루는 순간(t_i', t_i').

(상부(2)와 하부(3)(그림 20).

또한 식용 색소를 미리 식용유지


<화학식 I><화학식 II><화학식 III>

$\Delta B50x \leq 0.3 \dots$ 식 (1)

무기 화합물 입자를 포함한 잉크 비흡수성 피기록 매체에 있어서 기록물의 내찰성을 향상시키는 처리 방법, 기록 방법 및 잉크 세트를 제공하는 것.

s_i , 0.20 ~ 0.80 wt. %

코일 스프링 (7) 의 일방의 돌출부 (72) 를, 내측 링크 (4) 의 일방의 측판부 (41) 에 캠 부재 (91) 를 개재하여 누름과 함께, 비틀림



04 AIHUB_논문자료 요약

AIHUB_summary_of_thesis_materials.txt

04 AIHUB_논문자료 요약

AIHUB_summary_of_thesis_materials.txt

(한국교총, 2015a).

(계영희, 2005; 백선수·김원경, 2007; 박현미·강신포·김성준, 2007; 계영희·김종민, 2008; 임해경·박은영, 2002; 김남균, 2004).

핀은 빠짐

이러한 HAB는 대부분 *Microcystis* sp., *Anabaena* sp., *Aphanizomenon* sp.

A: Chem., 295권, 52쪽 (2008), F. Guozhi, F.S. Ichiro, Z.



05 AIHUB_요약문 및 레포트 생성 데이터

AIHUB_summary_and_report_generation_data.txt

05 AIHUB_요약문 및 레포트 생성 데이터

AIHUB_summary_and_report_generation_data.txt

```
for sentence in kss.split_sentences(one_json_sample['Meta(Refine)']['passage']):  
    if bool(re.match(r'[.],|[\u25c0]|\u25b6]|\u25aa]|\u25a0]|\u25b2]|\u25bc]|\u25c2]|\u25c4]|\u25ba]|\u25c6]', sentence[0])) ==  
False:  
    sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외

05 AIHUB_요약문 및 레포트 생성 데이터

AIHUB_summary_and_report_generation_data.txt

황 부회장은 ”

오히려 어른들이 만드는 KBS ‘TV 유치원’ 시청률은 고작 0.2%다.

54.9%.

이어 용인서울고속도로(79.4%), 서수원평택고속도로(79.0%), 수원광명고속도로(77.4%), 구리포천(75.2%) 순이었다.

다음은 홍교수와의 일문일답.

‘죽산군’은 1895년까지 인근 군·현(郡縣)들을 관할하던 죽산부(竹山府)가 있었기 때문에 붙은 이름이라고 한다. 일제강점기 시절 일본은 이 지명을 탐탁지 않게 여겼다고 한다. 이에 조선총독부령 제111호에 따라 전국의 군·면이 통폐합하던 1914년 ‘죽산군’을 없앴다. 대신 이들 지역을 안성군으로 편입시킨 뒤 ‘죽일면(竹一面)’이라는 이름이 붙었다. 하지만 곧 민원이 제기됐다.

“자리를 유지하게 해 달라”고 부탁했다.

05 AIHUB_요약문 및 레포트 생성 데이터

AIHUB_summary_and_report_generation_data.txt

㉠ 2020년 미국 대선은 중국 때리기 경연장이 될 것이다.

(www.
google.
com/covid19/mobility)

- 유럽과 북미에서는 공원 방문이 증가하고 있다.

- 오세아니아에서는 식품점·약국 방문이 상대적으로 늘고 있다.

㉡ 한국 일터 이동, 코로나 전 92% 회복 • 한국은 일터로의 복귀가 거의 이뤄졌다.

"올해는 진짜 대형 센터백이라는 얘기 듣고 싶어요."
"

(지원예산) 92.12억 원

(지원대상) 국내외 창업전문기관, 실험실창업혁신단(실험실창업탐색팀)

1. 개정이유

소하천정비법일부개정법률안이 공포('19.12.10.)됨에 따라 소하천 관련 기술을 연구를 목적으로 하는 법인 등에게 소하천 설계기준에 관한 도서(圖書) 등을 작성하

2. 주요내용

가.

소하천 설계기준 신설(안 제2조의2)

행정안전부장관은 소하천등 정비의 품질과 안전을 확보하고 관련 기술의 향상을 도모하기 위하여 설계기준에 관한 도서(圖書) 등을 작성하여 보급할 수 있도록 함.

05 AIHUB_요약문 및 레포트 생성 데이터

AIHUB_summary_and_report_generation_data.txt

그 밖의 참고 사항 등

제출의견 보내실 곳

일반우편 : 세종특별자치시 정부2청사로 13 행정안전부 재난경감과

전자우편 : wide@korea.kr

팩 스 : 0442055149

김진태 위원] "이번에는 그게 들어가 있나요?"

법무부장관 황교안] "지금 안 들어가 있는 것으로 알고 있습니다. 이번에 안 들어가 있는 것으로 알고 있습니다."

(※주공3단지 49.

9㎡의 올해 공시가는 8870만원으로 1억원이 안 된다.

)



06 AIHUB_문서요약 텍스트

AIHUB_document_summary_text.txt

06 AIHUB_문서요약 텍스트

AIHUB_document_summary_text.txt

㉔ 타인과의 관계에 있어 선출원이어야 한다.

"COMPUTER"는 "컴퓨터"로 "VISION"은 "보는 것,

乙이 제3자로부터 정자를 제공받아 시험관시술로 丙을 출산하였고 이후 혼외 관계로 丁을 출산하였으며,

甲이 丙과 丁을 甲과 乙의 자녀로 출생신고를 하였는데,

乙이 甲의 동의를 얻어 제3자로부터 정자를 제공받아 시험관시술을 통한 인공수정의 방법으로 丙을 임신·출산하였으므로,

丙은 민법 제844조 제1항에 따라 甲의 친생자로 추정되어 丙에 대한 친생자관계부존재확인의 소는 부적법하고,

甲과 丁 사이에는 친생자관계가 존재하지 않으나,

甲은 늦어도 丁이 초등학교 5학년 무렵 교통사고를 당했을 당시에 병원 검사를 통하여 丁이 甲의 친자가 아니라는 사실을 이미 알고 있었는데도 상당히 오랜 기간 동안

甲이 丁의 입양을 추진하고

甲과 丁 사이에 파양에 의하여 양친자관계를 해소할 필요가 있는 등의 특별한 사정이 없으므로,

丁에 대한 친생자관계부존재확인의 소는 확인의 이익이 없어 부적법하다고 한 사례.

乙은 수험생들에게 수능시험 시행기본계획에 따라 시험장 반입이 금지되는 물품과 휴대가 가능한 물품에 대하여 명확한 안내를 할 주의의무가 있는데,

'시각표시, 교시별 잔여시간 표시,

甲의 문의에 대해서도 막연히 어떠한 기능이 있는 시계라면

乙과 국가는 공동하여 甲이 휴대 가능한 시계를 소지하지 못한 채 시험을 치름으로써 입은 정신적 손해를 배상할 책임이 있다고 한 사례.

"충격흡수용 차량진입 방호방지대"를 대상물품으로 하는 등록디자인 " ",

" "이 비교대상디자인 1 " ",

" "과 유사하다는 등의 이유로 甲이 디자인등록무효심판청구를 한 사안에서,

[보령] 보령시는 지난 25일 오후 시장실에서 김동일 시장과 박주필 행복하우스 대표, 최석길 새뜰마을 추진위원회 대표 등 관계자가 참석한 가운데 명천6통 취락지구 집

TV조선 '연애의 맛' 캡처

[박영채 기자 ycpark@imaeil.com]

06 AIHUB_문서요약 텍스트

AIHUB_document_summary_text.txt

■ 대구FC, 창단 최초 스플릿A
‘패션 디자이너 대구관’ 개소
亞 투수 최초...평균자책점 2.32

= 일본 수입맥주의 점유율이 하락하고 있는 가운데 29일 서울 시내 한 대형마트에 일본맥주가 진열돼 있다.

?더불어민주당 이인영 원내대표가 12일 오전 서울 여의도 국회에서 열린 최고위원회의에서 발언하고 있다.

【 앵커멘트 】 검찰이 가슴기 살균제 사건의 2차 수사 결과를 발표했습니다.

#1. 임수정의 충격 엔딩, 로맨스 적신호 커졌나?

BNK금융경영연 동남권연구센터, ‘동남권의 일본 수출입보고서 발표
iusm

- 2단계 1차분(효자동·평화동) 올해 상반기 공사완료 추진 김선희 기자1 ksh9887@hanmail.net



07 AIHUB_특허 분야 자동분류 데이터

AIHUB_automatic_patent_classification_data.txt

07 AIHUB_특허 분야 자동분류 데이터

AIHUB_automatic_patent_classification_data.txt

메시지를 수신하고, 상기 푸시 알림

그 결과 본 발명에 따른 하수처리장 배관막힘



08 AIHUB_뉴스 기사 기계독해 데이터

AIHUB_news_article_machine_reading_data.txt

08 AIHUB_뉴스 기사 기계독해 데이터

AIHUB_news_article_machine_reading_data.txt

(문의 ☎ 043-201-4058)

이곳에는 카라반 3대, 5t트럭 1대, 3.5t트럭 2대.
대형버스 2대.
미니버스 4대 등 총 12대의 대형차가 주차 중이다.

김인식 미라클 감독은 “코로나19로 인해 상황이 어렵지만 열심히 훈련하는 선수들의 프로 진출을 위해 최선을 다하겠다”라며 “올해 경기도리그 첫 우승을 통해 연천군
“고 밝혔다.

과천=김형표기자

그는 “우승을 자축하는 시간보다 남들과 기쁨을 나누는데 (시간을) 더 할애했다”고 했다. 이어 “많은 지인과 계속 연락하고 얘기했더니 시간이 금방 갔다. (좋아서)

김세영은 2015년 LPGA 데뷔 후 매 시즌 1승 이상 거뒀다. 어느새 두 자릿수 우승(11승). 아무리 그래도 메이저 대회는 긴장될 법했다. 최종 라운드 전날 알람 시계

대회 첫날, 김세영은 우승권이 아니었다. 1오버파였다. 2라운드에서 5타를 줄여 선두로 뛰어오르면서 우승 기회를 잡았다. 언젠가 해야 할 메이저 우승. 모처럼 온

코로나19 탓에 많은 갤러리 앞에서 우승하는 건 불가능했다. 김세영은 그래도 우승 기분을 한껏 즐겨보려 했다. 그는 “대회 당일 펜실베이니아주 당국에서 대회가 열

김세영은 특히 우승 경쟁을 펼친 박인비에 대한 감사 인사를 빼놓지 않았다. 그에게 5타 뒤진 기록으로 준우승한 박인비는 경기 후 인터뷰에서 “김세영은 언터처블이

김세영은 별명이 많다. 역전의 명수, 연장의 여왕, 승부사, 빨간 바지의 마법. 여기에 메이저 쿼를 추가했다. 이번 우승에 칭찬이 쏟아지자 그는 “과찬이다. 그런데

김세영은 이번 우승 덕분에 여자골프 세계 2위까지 올라섰다. 개인 최고 랭킹이다. “세계 2위에 오르게 됐다”고 귀땀하자 그는 “아 진짜”라고 되묻더니 “오 나이스

08 AIHUB_뉴스 기사 기계독해 데이터

AIHUB_news_article_machine_reading_data.txt

다음은 아이오케이컴퍼니 입장문 전문
안녕하세요.

※알려드립니다.

고 해명했다.



09 AIHUB_행정 문서 대상 기계독해 데이터

AIHUB_machine_reading_data_for_administrative_documents.txt

09 AIHUB_행정 문서 대상 기계독해 데이터

AIHUB_machine_reading_data_for_administrative_documents.txt

```
for sentence in kss.split_sentences(j['paragraphs'][0]['context']):  
    if bool(re.match(r'[.][,][◆][◇][△][▲][▽][▼][▷][▶][<][>][0-9][《][/][○][-  
][ ]|[ ]|[○][I][II]', sentence[0])) == False:  
        sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외.

09 AIHUB_행정 문서 대상 기계독해 데이터

AIHUB_machine_reading_data_for_administrative_documents.txt

제목 2021년 제1차 강북구 치매안심센터 자문위원회 서면 개최

■ 추진개요

○ 건 명: 2021년 제1차 강북구 치매안심센터 자문위원회 서면개최

○ 일시: 2021년 6월 28일(월)

○ 회의방법: 서면회의(코로나19 확산 예방을 위하여 서면으로 안건검토 및 회신) ○ 참석자: 자문위원 9명(외부 및 내부 위원)

○ 안건

- 치매예방 인식개선을 위한 교육 확대 방안

- 치매안심센터 인식도 증대 홍보방안

제목 건축위원회(건축자산전문위원회) 위원 위촉(연임)

■ 관련근거

■ 한옥 등 건축자산의 진흥에 관한 법률

■ 서울특별시 한옥등건축자산의진흥에관한 조례 제27조(건축자산전문위원회 설치·운영)

■ 건축법 제4조(건축위원회)

■ 서울특별시시 건축 조례 제5조(구성)

■ 서울특별시 각종 위원회의 설치·운영에 관한 조례 제8조(위원회의 구성)

○ 건축자산전문위원회 위원 변경 및 운영 개편 방침(한옥건축자산과-8373, 20.8.20.)

■ 연임 개요

○ 연임 사유 : 임기 동안 위원회 업무를 성실히 수행 한 위원을 대상으로 연임신청을 받아 연임 위촉하고자 함.

○

위촉(연임)기간 : 2021. 7. 5.~2023. 7. 4.(2년)

■ 행정사항

○ 위촉장 제작

- 소요예산 : 60,000원(위촉장 6부 × 10,000원)

- 예산과목 : 도시재생실 한옥건축자산과, 한옥건축자산보전진흥, 전통문화계승발전(일반), 건축자산전문위원회 운영, 사무관리비

○ 위원별 연임 위촉 통보

일자리플러스센터 직업훈련교육 운영 계획(ver.1)

09 AIHUB_행정 문서 대상 기계독해 데이터

AIHUB_machine_reading_data_for_administrative_documents.txt

2 요양보호사 양성교육

교육기간 : 2021. 9~10월 [2개월 이내]

※ 2021. 11. 6.(토) 요양보호사 자격시험 대비

교육인원 : 구민 15명 (신규자 대상)

교육내용

이론 및 실기(160시간) : 기본요양보호기술, 가사 및 일상생활 지원 등

현장실습(80시간) : 노인요양시설 실습, 재가요양서비스 실습

목표 : 수료율 90%, 취업률 70% 이상

소요예산 : 10,000천원(구비 100%)

추진방법 : 요양보호사교육기관 위탁추진(공개모집)

협상순서는 합산점수의 고득점순에 의하여 결정하되 합산점수가 동일한 제안자가 2인 이상일 경우 기술능력 평가점수가 높은 제안자

○ 기술평가점수도 동일한 경우에는 평가위원회에서 추첨에 의하여 결정

제목 2021년 6월 신규공사장 안전기술지도 시행계획 변경

1. 2021년 본부장 신년 업무보고(2021.1.27.)와 관련하여 「신규공사장 안전기술지도」2021년 6월 시행계획을 다음과 같이 보고드립니다.

지도일정 : 2021년 6월 10일(목), 29일(화) 오전

나.

지도강사 : 시설국장, 안전관리과장, 안전관리과 정진혁, 이태행

다.

교육시간 : 오전 09:30 ~ 12:30(3.0H)

라.

교육장소 : 도시기반시설본부 10층 회의실

마.

교육대상 : 현장소장, 감리단장, 안전관리자 등

바.

교육내용

1) 건설공사장 안전관리 방향

2) 건설기술진흥법, 산업안전보건법 상의 안전관리 내용과 구비 및 관리 서류(1.5H)

3) 도시기반시설본부 건설안전 관련 각종 방침 및 지침

추진기간 : 착공일 ~ 2021. 7. 31.

주요내용 : 층별 목재데크 및 등벤치 교체 (철거공사 및 기타 보수공사 포함)

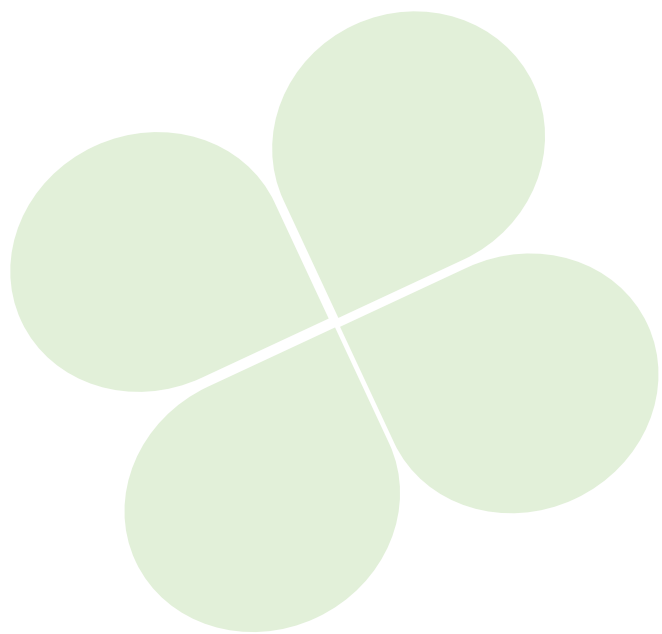
교체범위 : 지하 1층 ~ 지상 3층 데크(아래 붉은색) 및 등벤치(13개)

09 AIHUB_행정 문서 대상 기계독해 데이터

AIHUB_machine_reading_data_for_administrative_documents.txt

□ 감염병 증상(설사, 구토, 발열, 기침 등)이 발생했을 때에는?
④ 설사, 구토, 발열, 기침 등 감염병 증상이 나타나면, 즉시 소속 단체 관리자나 지역보건소로 신고합니다.
* 신고시 보건소가 검체 채취 및 역학조사 실시
④ 감염병의 전파를 막기 위해, 업무를 중단하고 숙소에 마련된 격리장소에서 증상이 없어질 때까지 머물러야 합니다.
* 손씻기, 기침예절 지키기, 다른 사람과의 신체접촉 제한이 중요
□ 감염병 예방수칙 지키기
노로바이러스 식중독은 오염된 물이나 식품 섭취, 또는 오염된 환경이나 감염된 환자와의 접촉을 통해 전염됩니다.
동남아시아 5개국, 한국의 선진 기록관리 경험 공유한다!
□ ‘국제 기록문화유산 관리과정’은 국가기록원이 2008년부터 개발도상국을 대상으로 국제교류협력차원에서 추진해 온 교육과정이다.
질병관리본부 정례브리핑 참고자료 - 감염병별 발생 동향 및 주요 대책 -

* (주관) KAIST, (참여) 포항공대, 한국청년기업가정신재단
- 5월 19일부터 21일까지 개최한 부트 캠프(Boot Camp)에서 비즈니스모델 설계, 엘리베이터 피칭, 기업가정신 등에 대한 집중 실습



10 AIHUB_기계독해
AIHUB_machine_reading.txt



11 AIHUB_도서자료 요약

AIHUB_summary_of_book_materials.txt

11 AIHUB_도서자료 요약

AIHUB_summary_of_book_materials.txt

```
for sentence in kss.split_sentences(passage):
    summary

    if bool(re.match(r'[.],[\u25c0\u25b6][\u25a0\u25a1][\u25b2\u25bc][\u25c4\u25c6][\u25d0\u25d2][\u25d4\u25d6][\u25d8\u25da][\u25dc\u25de][\u25e0\u25e2][\u25e4\u25e6][\u25e8\u25ea][\u25ec\u25ed]', sentence[0])) ==
False and \
        len(sentence) > 16:
        sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외.
- **len:** 공문서상 짧은 Token은 문장이 아닌 경우(이름, 직함, 섹션 등)가 대부분이라 제외.



12 AIHUB_대규모 구매도서 기반 한국어 말뭉치 데이터

AIHUB_korean_corpus_data_based_on_large_scale_purchase_books.txt

12 AIHUB_대규모 구매도서 기반 한국어 말뭉치 데이터

AIHUB_korean_corpus_data_based_on_large_scale_purchase_books.txt

```
for sentence in kss.split_sentences(passage):
    summary

    if bool(re.match(r'[.],|[\u25c6]|\u25c7|\u25b2]|\u25b3]|\u25bc]|\u25bd]|\u25c0]|\u25b6]|\u25c2]|\u25c4]|\u25d4]', sentence[0])) ==
False and \
        len(sentence) > 16:
            sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외
- **len:** 공문서상 짧은 Token은 문장이 아닌 경우(이름, 직함, 섹션 등)가 대부분이라 제외

12 AIHUB_대규모 구매도서 기반 한국어 말뭉치 데이터

AIHUB_korean_corpus_data_based_on_large_scale_purchase_books.txt

```
[Object]-[Path]-[Simplify]를 실행합니다.  
[Simplify] 대화상자에서는 원본에서 많이 벗어나지 않도록 수치를 조절하고 [OK]합니다.  
휴!
```

A. 요청
- 영화 <매트릭스> 속 등장인물의 대사

```
http://localhost:8080/pds/register (POST)B.
```

```
th> t2.new = "new string"  
Rs.CursorType = 1'Keyset Only Cursor  
형식
```

- 변수 x 에 리스트 test 의 값을 하나씩 추출하여, 출력한다.
- 프로그래밍을 하는 과정에 대한 기록을 남기는 것이 좋다.

처음
입사했을 때는 회사 업무 중 제대로 아는 분야가 하나도 없었다.

: 해밀토니언 경로 찾기

이 흥 보 히 김 태 천 표 상 선 윤 박 지 덩 철 황 차
■ 자연시간

«헨리 6세(Henry VI)» 1, 2, 3부: 장미 전쟁 시기.

12 AIHUB_대규모 구매도서 기반 한국어 말뭉치 데이터

AIHUB_korean_corpus_data_based_on_large_scale_purchase_books.txt

+ 부호 연산의 연산 결과는 피 연산자 값이어서 표시하지 않은 것과 결과가 같아서 거의 사용하지 않아요.

그리고 - 부호 연산자는 피 연산자의 값에 -1을 곱한 값이 연산 결과예요.

◆ 부호 연산자#include <stdio.h>int main(){printf("%d %d \n", +5, -5);return 0;}◆ 실행 결과5 -52. 사칙 연산자+, -, *, /사칙 연산은 두 개의 피 연산자 만약 두 개의 피 연산자가 모두 정수형이면 연산 결과는 정수예요.

✚700메가바이트=700×220(메가)×8(바이트)=5,872,025,600비트

| <레퀴엠>을 둘러싼 도작 의혹 |

* 수학을 좋아하는 독자를 위해 설명하면, 활성화 함수로 두 가지 선택이 인기인데, 사이노이드 함수($x \mapsto 1/(1 + e^{-x})$)와 램프 함수($x \mapsto \max\{0, x\}$)가 그것이다.

@timestamp필드는 데이터 수집 과정에서 로그스테시가 추가한 시간이고,timestamp필드는 실제로 아파치 웹 서버에 로그가 작성된 시간이다.

(...) 이러한 엄청난 수요를 설명해주는 것은 바로 변종(variety)이다.

- 함수 bigger() 는 arguments 값으로 first 와 second 를 가진다.

설명

고전주의 대표 음악가(1756.1.27~1791.12.5).

너무 빨리 그리고 너무 잘 하려고 하지 마라

.웹서버에 독립적으로 패킷을 실시간으로 DB에 저장

★원자 중심에는 핵이 있고, 핵을 중심으로 돌고 있는 것이 전자인데 핵과 전자 사이에 빈 공간이 아주 많다.

그것으로 이뤄진 원자 속은 거의 텅 비어 있는 것이나 마찬가지임을 의미하는 문장이다.


[그림4-4] 계층형 트리 그래프의 예

이신입이 작성한 라우트\$app->post('/task/update',TaskController::class.':update');# URI를 자원의 위치 관점에서 보고 작성한 라우트\$app->post('/task/



13 AIHUB_도서자료 기계독해

AIHUB_reading_books_by_machine.txt



14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_text_analysis_data.txt

14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_t
ext_analysis_data.txt

```
if '약관' in file_name_list[i]:  
    num = 0  
  
    for sentence in kss.split_sentences(' '.join(re.sub(r"\n", "", str(root_text)).split())):  
        if bool(re.match(r'[.]|[,]|[\u2666]|[\u2667]|[\u25b3]|[\u25b2]|[\u25bc]|[\u25c0]|[\u25ba]|[\u27e8]|[\u27e9]|[\u2013]|[\u201c]|[/]|[\u2460]|\[|-]|', sentence[0])) == False:
```

- **KSS(Korean Sentence Segmentation):** MeCab 기반으로 돌아가는 Tokenizer
- **re.match:** 특수기호로 시작하는 Token 제외

14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_text_analysis_data.txt

```
elif '01.민사' in file_name_list[i]:  
    num = 0  
  
    for sentence in kss.split_sentences(' '.join(re.sub(r"\n", "", str(root_text)).split())):  
  
        if bool(re.match(r'[.]|[,]|[\◆]|\[◇\]|[\△]|\[▲]|\[▽]|\[▼]|\[▷]|\[▶]|\[<]|\[>]|[0-9]|[\《]|\[/]|\[○]|[-]  
]|   ', sentence[0])) == False:  
  
            if '2. 원고의 청구에 대한 판단' in sentence:  
                num += 1  
                if num == 1 and '2. 원고의 청구에 대한 판단' not in sentence:  
                    sentence_list.append(sentence)
```

- **re.match:** 특수기호로 시작하는 Token 제외. 민사법조항 Token 제외

14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_t
ext_analysis_data.txt

```
elif '02.형사' in file_name_list[i]:  
    num = 0  
  
    for sentence in kss.split_sentences(' '.join(re.sub(r"\n", "", str(root_text)).split())):  
  
        if bool(re.match(r'[.]|[,]|[\u2666]|\u2667|[\u25b3]|\u25b2]|\u25bc]|\u25c0]|\u25ba]|\u27e8]|\u27ea]|\u2794]|\u2796]|\u2797]|\u2798]|\u2799]', sentence[0])) == False:  
  
            if '판례 검색' in sentence:  
                num += 1  
                if num == 1 and '판례 검색' not in sentence:  
                    sentence_list.append(sentence)
```

- **re.match:** 특수기호로 시작하는 Token 제외. 판례문 Token 제외

14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_t
ext_analysis_data.txt

```
elif '03.행정' in file_name_list[i]:  
    num = 0  
  
    for sentence in kss.split_sentences(' '.join(re.sub(r"\n", "", str(root_text)).split())):  
  
        if bool(re.match(r'[.]|[,]|[\◆]|\[◇\]|[\△]|\[▲\]|[\▽]|\[▼]|\[▷]|\[▶]|\[<]|\[>]|[\0-9]|[\《]|\[/]|\[○]|[-]|'|']', sentence[0])) == False:  
  
            if '관계 법령' in sentence:  
                num += 1  
                if num == 1 and '관계 법령' not in sentence:  
                    sentence_list.append(sentence)
```

- **re.match:** 특수기호로 시작하는 Token 제외. 관계법령 Token 제외

14 AIHUB_법률 규정 (판결서 약관 등) 텍스트 분석 데이터

AIHUB_legal_regulations_(such_as_terms_and_conditions_of_judgment)_text_analysis_data.txt

파기절차 및 방법은 다음과 같다. ○

별도 DB로 옮겨진 개인정보는 법률에 의한 경우에만 보관되고 이외의 다른 목적으로 사용되지 않는다. ○

파기방법 □ 전자적 파일형태로 저장된 개인정보는 기록을 재생할 수 없는 기술적 방법을 사용하여 삭제한다. □

그러나 다음 경우에는 예외로 한다. □

㉠ 상조서비스의 제공지역은 다음과 같습니다. ㄱ

(단. 외부업체 출입이 제한된 특정 장소는 역무 행사를 제공할 수 없으며, 천재지변 및 기타 교통 체증 등 불가항력으로 행사에 신속한 대응을 못할 사정이 있을 경우에도

제1조(목적)본 계약은 예식홀을 운영하는 사업자(이하 사업자)와 예식홀을 이용하는 예식 당사자등(이하 이용자) 간의 예식홀 이용에 관한 제반 계약사항을 규정함

제4조(이용자의 의무)1) 이용자는 사업자의 시설관리 및 질서유지에 관한 운영규정을 준수하고, 예식의 원활한 진행을 위하여 협력하여야 합니다.2)

제 3 장 책 임

제 6 장 정보의 제공 제1조 정보의 제공 (1) 000는 이용자가 서비스 이용 중 필요가 있다고 인정되는 다양한 정보에 대해서는 전자우편이나 서신우편 등의 방법으로

(i) 000 사업의 관리와 경영 및 000 서비스의 홍보와 제공에 관한 모든 문제 (ii) 대한민국 안팎으로 이러한 정보를 이전 (iii) 모든 법령, 규칙, 법원명령 또는 규제
다만 적용되는 법령에 따라, 대한민국 내외에 존재하는지를 불문하고, 다음의 법인 또는 개인에게 공개 또는 그에 의하여 관련된 정보가 이용될 수 있다.

(i) 000, 000의 계열회사 및 그룹 (ii) 000 및 계열회사 업무를 수행하는 경우에 한하여 000 및 계열회사의 모든 관리자, 직원 또는 피용인 (iii) 모든 대리인, 계약
(i) 개인 정보의 입수 청구 (ii) 개인 정보에 대한 정정 청구 (iii) 정보 입수 또는 정정이 거부되거나 40일 이내에 제공 정정되지 않은 경우 그 이유의 제공.

제 1 관 목적 및 용어의 정의 제 1 조 (목적) 이 보험계약(제 2 조의 2(용어의 정의 2) 제 1 항에서 정한 보장계약 과 적립계약 을 말하며, 이하 계약 이라 합니다
(이하 보장계약 의 보험료를 보장보험료 , 적립계약의 보험료를 적립보험료 라 하며 보장보험료 와 적립보험료 를 합하여 보험 료 라 합니다.)

제 2 조의 3 (다발성 소아암 의 정의 및 진단 확정) ㉠ 이 계약에 있어서 다발성 소아암 이라 함은 한국표준질병,사인분류 중 별표3(다발성 소아암 관련 악성 신생물

(1) 산후조리원의 귀책사유로 인한 경우 (총 이용금액에서 이용기간에 해당하는 금액을 공제한 잔액을 환급하며, 총 이용금액의 10퍼센트를 보상한다.)



15 AIHUB_일반상식

AIHUB_general_common_sense.txt

15 AIHUB_일반상식

AIHUB_general_common_sense.txt

```
if 'ko_wiki_v1_squad' in file_name_list[i]:
    for j in one_json_sample['data']:
        for sentence in kss.split_sentences(j['paragraphs'][0]['context']):
            if sentence[-1] == ".":
                if sentence != ".":
                    sentence_list.append(sentence)
else:
    for j in one_json_sample['sentence']:
        for sentence in kss.split_sentences(j['text']):
            if sentence[-1] == ".":
                if sentence != ".":
                    sentence_list.append(sentence)
```

- **KSS(Korean Sentence Segmentation)**: MeCab 기반으로 돌아가는 Tokenizer
- **re.match**: 특수기호로 시작하는 Token 제외
- **Slicing**: 마침표로 끝나지 않는 Token 제외