

대학수학능력시험 및 수능 모의평가 영어 영역 기출문제 학습을 통한 출제문항(어휘, 구문, 지문) 분석 및 예측

2021.12.06

- * 데이터사이언스학과 석사과정 권도윤
- * 데이터사이언스학과 석사과정 김재호
- * 데이터사이언스학과 박사과정 김진홍

주제 선정 배경 (주제명 : 대학수학능력시험 및 수능 모의평가 영어 영역 기출문제 학습을 통한 출제 문항(어휘, 구문, 지문) 분석 및 예측)

- ✓ 월 평균 사교육비는 여전히 지속적인 증가 추세를 나타내며, 그 중 영어 교과와 사교육비 지출비율이 가장 높은 경향을 보이고 있음
- ✓ 수능영어의 절대평가 전환에 따라 변별력 구분을 위한 기형적인 고난도 문항이 없어지고, 논란이 적으며 핵심적이고 중요한 내용 중심으로 출제

사회 > 사회일반

수능영어 절대평가 4년...난이도 널뛰고 사교육비 치솟았다

입력 2021-03-11 06:30:16 수정 2021.03.11 06:30:16 김창영 기자

HOME > 교육일반 > 초·중·고 교육

월 평균 사교육비 30만원 돌파...여전한 '사교육 공화국'

김백두산 기자 | 승인 2021.01.27 16:41 | 댓글 0

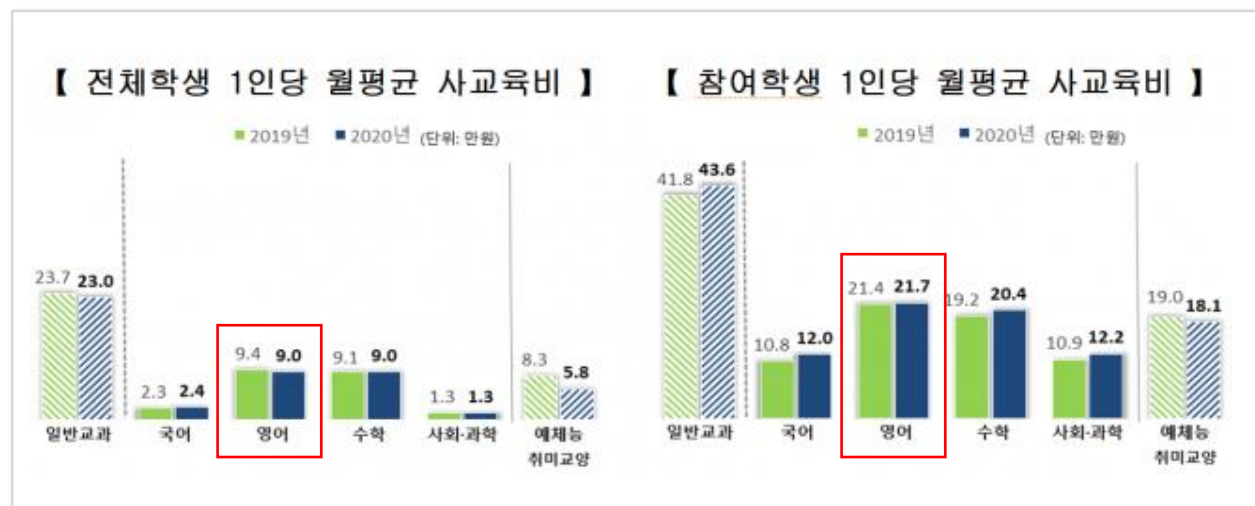
한경연, '우리나라 교육지표 현황과 사교육 영향 분석' 보고서 발표
공교육 확대에도 사교육 참여율 여전히 높아...2019년 초·중고 1인당 월 사교육비 32만 1천원
주당 사교육 참여시간 3.6시간으로 조사 대상 OECD 국가 중 1위...OECD 평균 6배
사교육 받는 학생이 상위권 속달 확률 수확 56.3% ↑, 영어 53.2% ↑

도입 이후 등급별 비중 들쭉날쭉
당초 취지와 달리 학원 의존 심화
월평균 사교육비 21.7만원 '최대'
영어 대신 他영역 반영 비율 높여
특정과목에 당락 갈리는 부작용도

| 등급 | 2018 | 2019 | 2020 | 2021 |
|----|------|------|------|------|
| 1 | 10.0 | 5.3 | 7.4 | 12.7 |
| 2 | 19.6 | 14.3 | 16.2 | 16.5 |
| 3 | 25.4 | 18.5 | 21.9 | 19.7 |
| 4 | 18.0 | 20.9 | 18.5 | 18.6 |
| 5 | 10.5 | 16.5 | 12.3 | 13.5 |
| 6 | 6.7 | 10.7 | 9.2 | 9.0 |
| 7 | 4.7 | 7.4 | 7.4 | 5.6 |
| 8 | 3.5 | 4.6 | 5.2 | 3.4 |
| 9 | 1.5 | 1.7 | 1.9 | 1.0 |

자료: 한국교육과정평가원

공교육 재정이 꾸준히 확대됐지만 사교육의 영향이 여전히 크다는 연구보고서가 발표됐다. 사교육을 받는 학생이 학교성적에서 상위권에 속할 확률이 높다는 분석이다. 초·중고 학생들의 1인당 월평균 사교육비는 2007년 조사 이후 처음으로 30만원을 넘었다. 사진=대학저널 DB



| 연도 | 추진내용 | '18학년도 수험생 |
|------|--|------------|
| '14년 | 2018학년도 수능영어 절대평가 도입 방안 발표(~'14.12월) | 중3 |
| '15년 | 점수체계 및 시험체계 등 세부 도입방안 마련(상반기) ⇒ 2018학년도 대입전형 기본사항 발표시 반영(8월) | 고1 |
| '16년 | 2018학년도 대학별 대입전형시행계획 발표(4월) ⇒ 절대평가 체제의 난이도 안정화 방안 마련·검증(하반기) | 고2 |
| '17년 | 절대평가체제의 모의평가 진행(6월/9월) ⇒ 본 수능 시행(11월) | 고3 |

| 수능 영어 절대평가 전환에 대한 찬반 쟁점 | | |
|---|-------|---|
| 찬성 | 쟁점 | 반대 |
| 쉬운 영어 중심 전환 영어 사교육 심화 해소 | 사교육 | 수학·국어 등 사교육 풍선효과, 현재 영어 사교육은 초·중·고교 중심 |
| 상대평가 과열 경쟁과 등급화 위한 기형적인 고난도 문항 없어져 | 학습 부담 | 쉬운 영어 강화돼 대학들이 별도의 영어전형 도입하면 학생 부담은 옥상옥이 될 가능성 |
| 국가공인 절대평가 체제와 수능 체제 개편해서 영어를 통과(패스) 과목으로 전환 가능 | 평가 방식 | 절대평가 방식인 국가영어능력시험 (NEAT)으로 수능 대체 검토했던 계획은 무산, 수능에서 영어 선택과목 유지 시 타 과목과의 형평성 문제 생겨 |

- 프로젝트명 : 대학수학능력시험 및 수능 모의평가 영어 영역 기출문제 학습을 통한 출제문항 (어휘, 구문, 지문) 분석 및 예측
- 인 원 : 김진홍 (21-2 박사 입학), 권도윤 (21-1 석사 입학), 김재호 (21-2 석사 입학)
- 기간 : 21.10.18 ~ 21.12.06
- 데이터셋 : 2016 ~ 2021년 6월,9월 평가원 기출 12회분, 2016~2021 수능 기출 5회분 (총 17회분)
- 분석환경 구축
 - Google Drive
 - Excel
 - Python 3.7 ~
 - ✓ Pandas, numpy
 - ✓ Matplotlib, seaborn,
 - ✓ sklearn 등
- 커뮤니케이션 :
 - 오프라인 : 주 1회 이상, 프론티어관 EDM-Lab 307-1호
 - 온라인 : KakaoTalk (메신저), Zoom (회의용)

➤ Project R&R

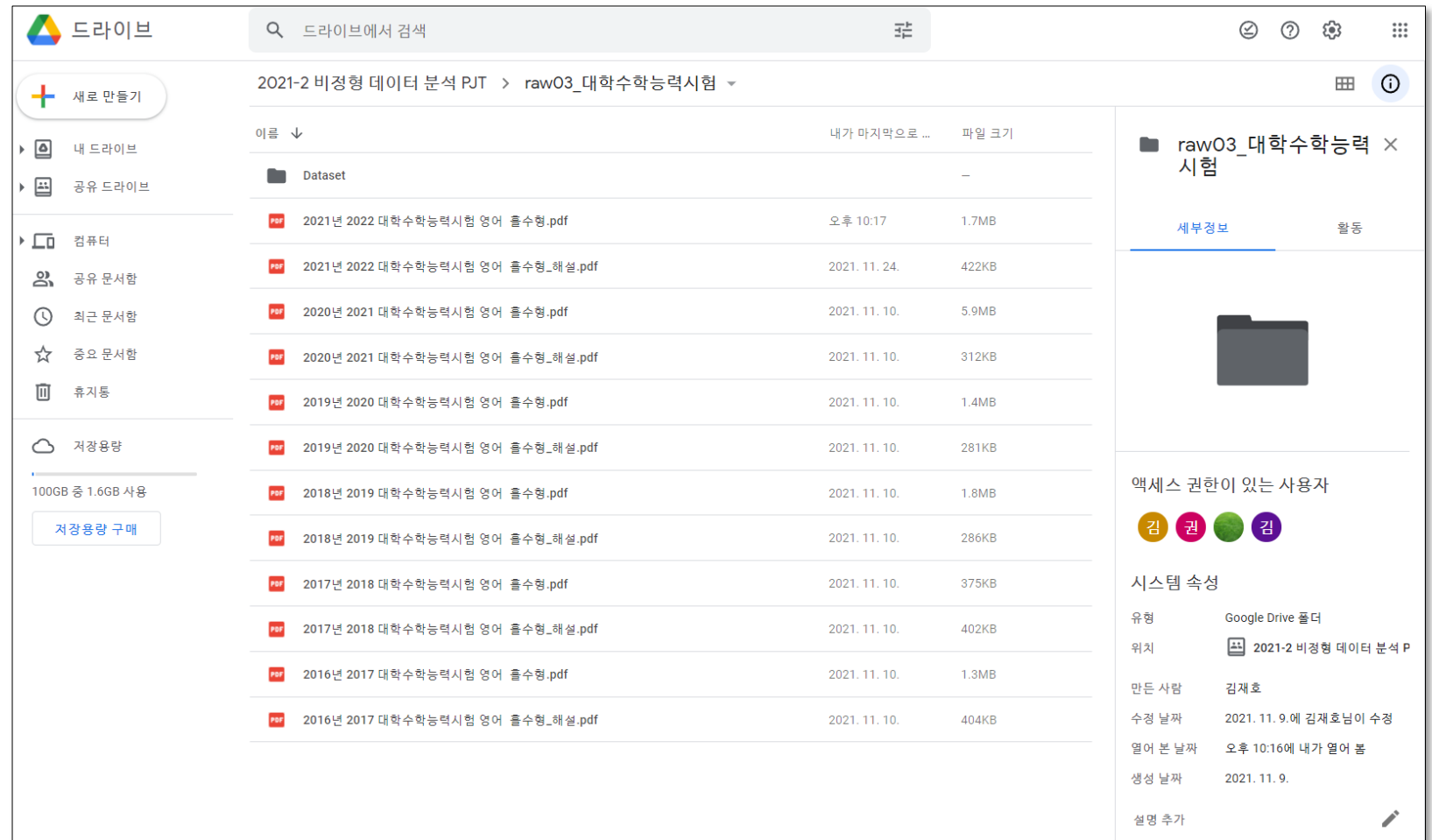
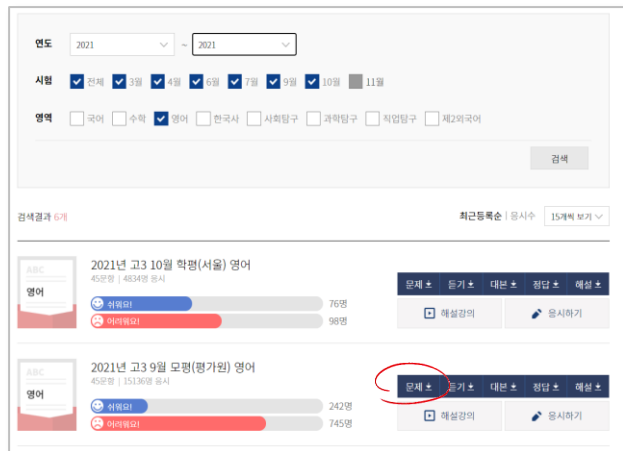
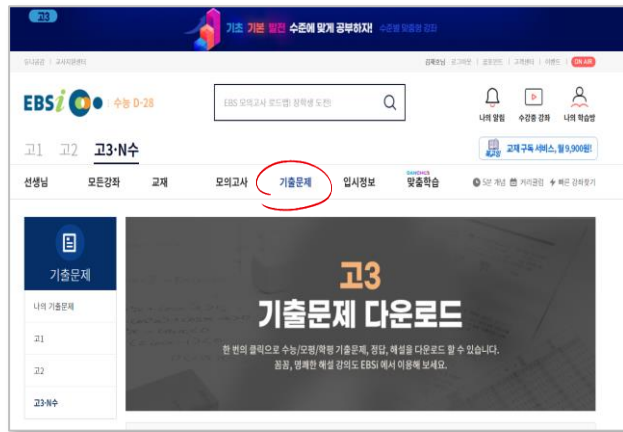
| 이름 | 소속 | Role & Responsibilities |
|-----|--------------------------|--|
| 김진홍 | 서비스혁신연구실 (금영정 교수님 지도) | - 프로젝트 진행관련 Background 조사 및 분석 ★ - 프로젝트 Advanced Part 분석 진행 ★★★ - 데이터 수집 및 전처리 진행 (공통) |
| 권도윤 | 응용확률 Lab 심민규 교수님 지도 | - 데이터 수집을 위한 전처리 기준 수립 및 가이드제작 ★ - 프로젝트 Basic Part 분석 진행 ★★★ - 데이터 수집 및 전처리 진행 (공통) |
| 김재호 | EDM Lab (홍정식 교수님 지도) | - 프로젝트 기획 및 보고서 작성 ★★★ - 프로젝트 진행관리 및 분석결과 검증 ★ - 데이터 수집 및 전처리 진행 (공통) |

➤ Project Schedule

| Phase | 주요내용 | 10월 | | 11월 | | | | | 12월 |
|-----------|-----------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | | 7주차 (10.18) | 8주차 (10.25) | 9주차 (11.02) | 10주차 (11.08) | 11주차 (11.15) | 12주차 (11.22) | 13주차 (11.29) | 14주차 (12.06) |
| 주제 선정 | 주제선정 | ●.....➡ | | | | | | | |
| | 목표설정 | ●.....➡ | | | | | | | |
| | 제안서 발표 | | ✓ | | | | | | |
| 데이터 수집/분석 | 데이터 수집 | | ●.....➡ | | | | | | |
| | 데이터 전처리 | | | ●.....➡ | | | | | |
| | 데이터 분석/검증 | | | | | ●.....➡ | | | |
| 검증 및 결과보고 | 최종보고 준비 | | | | | | | ●.....➡ | |
| | 최종 발표 | | | | | | | | ✓ |

데이터 수집

- ✓ EBS-i 홈페이지를 통해 2016년 부터 2021년 까지 총 17회 분량의 평가원 및 수능 기출 문제와 해설자료를 다운 받았음
- ✓ 수집한 데이터는 PDF 파일 형태로 분석을 위한 별도의 전처리가 필요하였고, 우선 수집한 데이터는 구글 드라이브 프로젝트 폴더에 보관하여 공유 및 협업 할 수 있도록 준비하였음



- ✓ PDF를 크롤링 하여 문항별로 지문추출 하는 방법을 검토하였으나, 보기, 빈칸 처리 등 변수 사항들이 많은 관계로 조원들과 배분하여 데이터 수집을 수작업으로 진행하였음
- ✓ 데이터 수집을 위한 기준을 마련하였으며, 읽기영역 28문제 / 25개 지문에서 그림 문제 지문 2개를 제외하고 23개 지문에 대해서만 데이터 수집 및 전처리 작업을 진행함

데이터 추출 방법

파일 수정 보기 삽입 서식 도구 부가기능 도움말 권도유님이 10일 전에 마지막으로 수정함

100% | 일반 텍스트 | Arial | 11 | B I U |

2 1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 17 18

- 18~23번: 전체 드래그 해서 엑셀칸에 복사
- 24번: 글만 전체 드래그해서 엑셀칸에 복사, 번호기호 지워도 되고 안지워도됨(전처리할때 지울 수 있음)
- 25번: 전체 드래그 해서 엑셀칸에 복사
- 26~27번: 건너뛰
- 28번: 글만 전체 드래그해서 엑셀칸에 복사, 번호기호 지워도 되고 안지워도됨(전처리할때 지울 수 있음)
- 29번: 글 전체 드래그해서 엑셀에 복사한뒤, 엑셀 칸 그대로 복사해서 정답 찾아 메모장 또는 워드같은 프로그램으로 수정하고 다시 엑셀에 복사 붙여넣기
- 30번: 글만 전체 드래그해서 엑셀칸에 복사, 번호기호 지워도 되고 안지워도됨(전처리할때 지울 수 있음)
- 31~34번: 글 전체 드래그해서 엑셀에 복사한뒤, 엑셀 칸 그대로 복사해서 정답 찾아 메모장 또는 워드같은 프로그램으로 수정하고 다시 엑셀에 복사 붙여넣기
- 35번: 글만 전체 드래그해서 엑셀칸에 복사, 번호기호 지워도 되고 안지워도됨(전처리할때 지울 수 있음)
- 36~37번: 정답 찾아 순서 맞추기, 엑셀 사용 팁! 본문-(B)-(C)-(A) 순서가 답이라면 본문을 B열에 (B)자문을 C열에 (C)자문을 D열에 (A)자문을 E열에 복사 붙여넣기 한 후, BCDE열 4칸을 드래그한 후 복사-붙여넣기를 메모장 or 워드에 하면 수정하기 편함. 메모장이나 워드에 복붙을 한 후 각 지문 사이에 Tap이 들어가있을 수 있으니 잘 확인하고 수정한다. 그 후 다시 복사 붙여넣기를 엑셀에 하면 된다.
- 38~39번: 정답 위치 찾아 넣기, 정답 부분만 괄호와 번호기호를 지워 넣어 넣으면 된다. 나머지부분은 알아서 전처리해 처리 가능
- 40번: 본문 그대로 넣고 이어서 요약문 넣기. 요약문은 정답 찾아 넣는다.
- 41번: 본문에 (A),(b) 정답 찾아 넣기
- 43번: 지문 순서 맞춰 넣기, 본 문제는 가리키는 대상 (a)(b)(c)(d)(e) 모두 지워야함.

집중해서 하면 15분~20분 안에 완료되었음

경리하자면

- 숫자 기호는 지워도 되고 안지워도 됨
- 포스터 or 안내문 지문은 건너뛰기
- 두개 중 정답찾는 문제는 정답 찾아서 넣기
- 빈칸 문제는 답 찾아서 넣기
- 순서 문제는 순서 맞춰서 넣기
- 알맞은 위치에 문장 넣기 문제는 정답 위치 찾아서 넣기
- 숫자기호가 아닌 영어기호라면 순서 전처리하기

32. What story could be harsher than that of the Great Auk, the large black-and-white seabird that in northern oceans took the ecological place of a penguin? Its tale rises and falls like a Greek tragedy, with island populations savagely destroyed by humans until almost all were gone. Then the very last colony found safety on a special island, one protected from the destruction of humankind by vicious and unpredictable ocean currents. These waters presented no problem to perfectly adapted seagoing birds, but they prevented humans from making any kind of safe landing. After enjoying a few years of comparative safety, disaster of a different kind struck the Great Auk. Volcanic activity caused the island refuge to sink completely beneath the waves, and surviving individuals were forced to find shelter elsewhere. The new island home they chose _____ in one terrible way. Humans could access it with comparative ease, and they did! Within just a few years the last of this once-plentiful species was entirely eliminated. [3점]

* savagely: 잔혹하게

① lacked the benefits of the old
② denied of
③ faced un
④ caused c
⑤ had a sim

2017학년도 대학수학능력시험 6월 모의평가
영어영역 정답 및 해설

01. ⑤ 02. ③ 03. ② 04. ① 05. ③ 06. ⑤ 07. ② 08. ② 09. ① 10. ④
11. ④ 12. ③ 13. ⑤ 14. ③ 15. ④ 16. ① 17. ④ 18. ② 19. ① 20. ①
21. ③ 22. ⑤ 23. ⑤ 24. ④ 25. ② 26. ③ 27. ④ 28. ④ 29. ⑤ 30. ⑤
31. ⑤ 32. ① 33. ② 34. ④ 35. ③ 36. ② 37. ① 38. ③ 39. ③ 40. ④
41. ① 42. ① 43. ② 44. ② 45. ④

01_2016.6월 .XLSX ☆ 100% | \$ % .0 .00 123 | 기본값 (Cal... | 5일 전에

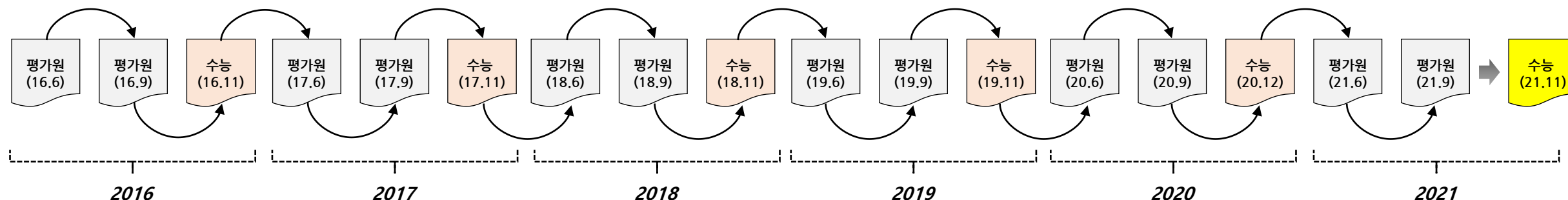
파일 수정 보기 삽입 서식 데이터 도구 도움말

100% | \$ % .0 .00 123 | 기본값 (Cal... | 5일 전에

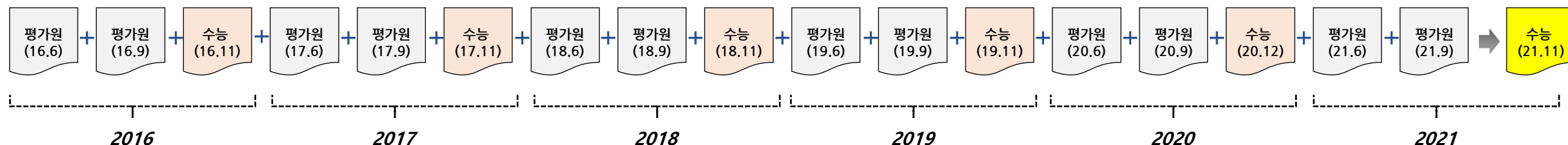
| | A | B |
|-----|---|---|
| Z13 | | |
| 12 | | Once a hand or gripper has been directed to an object by reaching, it can be grasped. Grasping requires that fingers hold an object securely. A secure grip is one in which the object won't slip or move, especially when displaced by an external force. Your grasp on a hammer, for example, would not be secure if knocking against something caused you to drop it. |
| 13 | | 31 One precondition of a firm grasp is that the forces applied by the fingers balance each other so as not to disturb the object's position. The characteristics of an object such as its geometric configuration and mass distribution may demand that some fingers apply greater force than others to maintain stability. The grasp and support forces must also match overall object mass and fragility. An egg requires a more delicate touch than a rock. 32 What story could be harsher than that of the Great Auk, the large black-and-white seabird that in northern oceans took the ecological place of a penguin? Its tale rises and falls like a Greek tragedy, with island populations savagely destroyed by humans until almost all were gone. Then the very last colony found safety on a special island, one protected from the destruction of humankind by vicious and unpredictable ocean currents. These waters presented no problem to perfectly adapted seagoing birds, but they prevented humans from making any kind of safe landing. After enjoying a few years of comparative safety, disaster of a different kind struck the Great Auk. Volcanic activity caused the island refuge to sink completely beneath the waves, and surviving individuals were forced to find shelter elsewhere. The new island home they chose lacked the benefits of the old in one terrible way. Humans could access it with comparative ease, and they did! Within just a few years the last of this once-plentiful species was entirely eliminated. |

- ✓ 제안발표 당시 수능 출제 문항은 기출문제들의 seq to seq 모델링을 따를 것이라는 개념으로 생각하고 학습을 통한 토큰 예측을 진행하고자 하였으나, 학습지식 부족 및 프로젝트 진행 스케줄 등의 이유로 각 기출문제 데이터 전체를 일괄적으로 병합하여 분석을 진행하였음

■ As-is (Proposal Presentation, seq to seq modeling concept)



■ To-Be



Basic part



- 2021년 수능 기출문제 지문분석 (토큰수, 사용빈도수 문장수 등)
- 2016~2021 모의평가 및 수능 기출 분석 (토큰수, 누적토큰수, 일치율 등)

Advanced Part

- 품사 태깅 및 표제어 추출(Lemmatization)을 통한 토큰 일치율 비교
- 코사인 유사도 비교를 통한 지문/장르 일치율(%) 비교

Basic Part I (2021년 수능 기출문제 분석)



- ✓ 2021년 수능 영어 기출(21,11,18일 시행) 문제의 읽기영역 지문 23개를 분석한 결과 23,519개 문자로 총 문장수는 183개, 불용어 제외시 2,236개 토큰이 사용되었음
- ✓ 수능 지문은 평균적으로 1,030여개 문자로 8개의 문장, 불용어 제외시 97개의 토큰으로 구성되었음

1

| | A | B | C | D |
|----|--------|---|---|---|
| 1 | number | contents | | |
| 2 | | 18 Dear Ms. Green, My name is Donna Williams, a science teacher at Rogan Hi | | |
| 3 | | 19 It was Evelyn's first time to explore the Badlands of Alberta, famous across C | | |
| 4 | | 20 One of the most common mistakes made by organizations when they first co | | |
| 5 | | 21 Scientists have no special purchase on moral or ethical decisions; a climate s | | |
| 6 | | 22 Environmental hazards include biological, physical, and chemical ones, along | | |
| 7 | | 23 Scientists use paradigms rather than believing them. The use of a paradigm in | | |
| 8 | | 24 Mending and restoring | | |
| 9 | | 25 The above graphs show | | |
| 10 | | 28 Donato Bramante, born | | |
| 11 | | 29 Like whole individuals, | | |
| 12 | | 30 It has been suggested | | |
| 13 | | 31 Humour involves not jus | | |
| 14 | | 32 News, especially in its | | |
| 15 | | 33 Elinor Ostrom found th | | |
| 16 | | 34 Precision and determin | | |
| 17 | | 35 Since their introduction | | |
| 18 | | 36 According to the marke | | |
| 19 | | 37 In spite of the likenes | | |
| 20 | | 38 Introduction of robots in | | |
| 21 | | 39 Cinema is valuable not | | |
| 22 | | 40 Philip Kitcher and Wes | | |
| 23 | | 41 Classifying things toge | | |
| 24 | | 43 In the gym, members o | | |
| 25 | | | | |

2

<코드1>

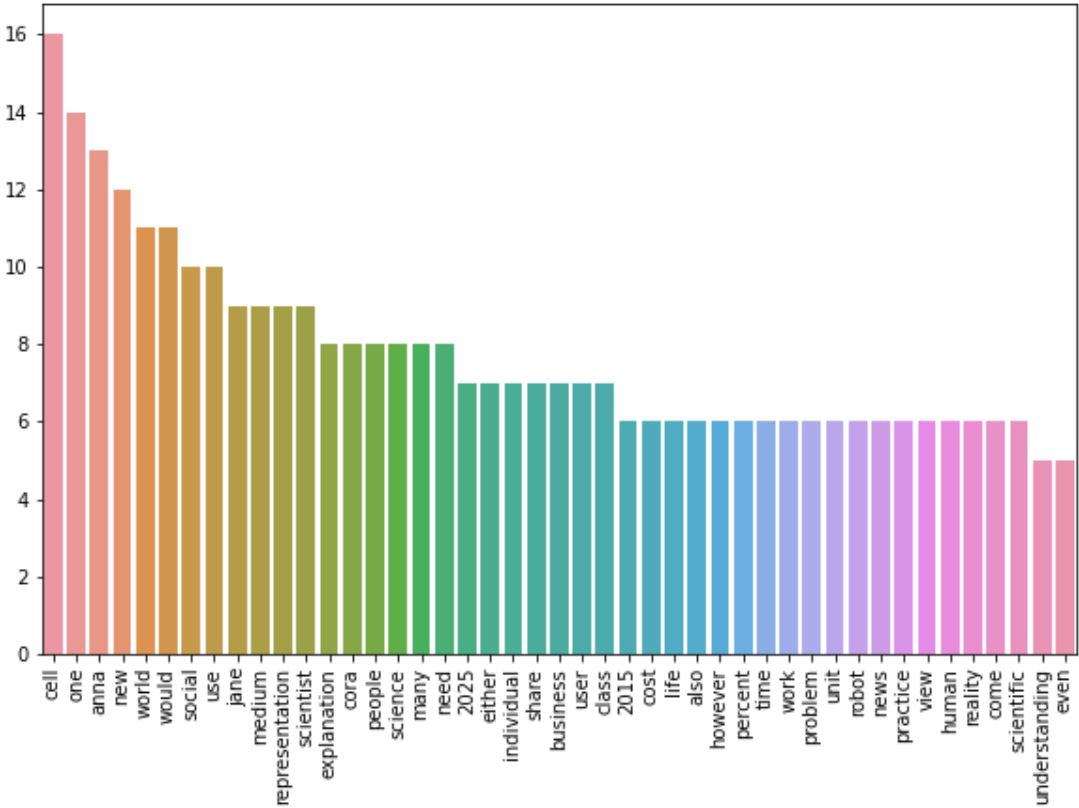
```
num_len=0
sent_num=0
max_sent=0
min_sent=0
nstop_token=0
stop_token=0
stop_words = set(stopwords.words('english'))
da=np.array([])
for i in range(len(df)):
    num_len=(len(df["contents"][i]))
    sent_tokens = sent_tokenize(df["contents"][i])
    sent_num=(len(sent_tokens))
    max_sent=(len(max(sent_tokens, key=len)))
    min_sent=(len(min(sent_tokens, key=len)))
    mean_sent=0
    for k in (sent_tokens):
        mean_sent+=len(k)
    mean_sent=mean_sent/len(sent_tokens)
    df["contents"][i]=re.sub("[^a-zA-Z0-9]", " ",df["contents"][i])
    nstop_token=(len(word_tokenize(df["contents"][i])))
    result = []
    for w in word_tokenize(df["contents"][i]):
        if w not in stop_words:
            result.append(w)
    stop_token=(len(result))
    da=np.append(da,np.array([[df["number"].values[i],num_len,sent_num,max_sent,min_sent,mea
```

▶ 2021년 수능영어 지문 분석

| 문항 번호 | 지문길이 (len) | 문장수 | 문장길이(len) | | | 토큰수 | |
|----------|---------------|-----|-----------|----------|------|---------|---------|
| | | | Max | Min | Mean | (불용어포함) | (불용어제외) |
| 18 | 637 | 8 | 139 | 25 | 79 | 110 | 69 |
| 19 | 697 | 9 | 123 | 22 | 77 | 122 | 73 |
| 20 | 968 | 6 | 249 | 38 | 161 | 165 | 90 |
| 21 | 1,021 | 8 | 269 | 31 | 127 | 171 | 99 |
| 22 | 1,002 | 6 | 239 | 83 | 166 | 149 | 93 |
| 23 | 1,005 | 7 | 224 | 52 | 143 | 157 | 85 |
| 24 | 994 | 8 | 178 | 40 | 123 | 161 | 95 |
| 25 | 823 | 6 | 196 | 110 | 136 | 158 | 87 |
| 28 | 753 | 10 | 123 | 44 | 74 | 131 | 80 |
| 29 | 1,028 | 9 | 153 | 47 | 113 | 170 | 105 |
| 30 | 1,083 | 6 | 272 | 105 | 180 | 168 | 105 |
| 31 | 825 | 6 | 235 | 72 | 137 | 142 | 74 |
| 32 | 1,084 | 6 | 285 | 114 | 180 | 171 | 102 |
| 33 | 1,022 | 6 | 240 | 139 | 170 | 166 | 91 |
| 34 | 1,165 | 5 | 306 | 194 | 232 | 181 | 94 |
| 35 | 874 | 6 | 217 | 103 | 145 | 133 | 84 |
| 36 | 1,035 | 7 | 267 | 42 | 147 | 163 | 94 |
| 37 | 1,023 | 7 | 210 | 75 | 145 | 168 | 89 |
| 38 | 980 | 6 | 321 | 84 | 163 | 155 | 89 |
| 39 | 1,071 | 7 | 210 | 92 | 152 | 174 | 97 |
| 40 | 1,244 | 9 | 182 | 49 | 137 | 180 | 100 |
| 41 | 1,427 | 11 | 283 | 81 | 129 | 257 | 137 |
| 43 | 1,918 | 24 | 187 | 14 | 79 | 349 | 204 |
| 합계 | 23,519 | 183 | 321 (max) | 14 (min) | - | 3,901 | 2,236 |
| 평균 | 1,030 | 8 | - | - | 139 | 170 | 97 |

※ 그림영역 2개 문항의 지문은 제외하고 계산 하였음

✓ 2021년 수능 영어 기출문제에 출제된 토큰과 빈도수를 분석하여 시각화를 진행함

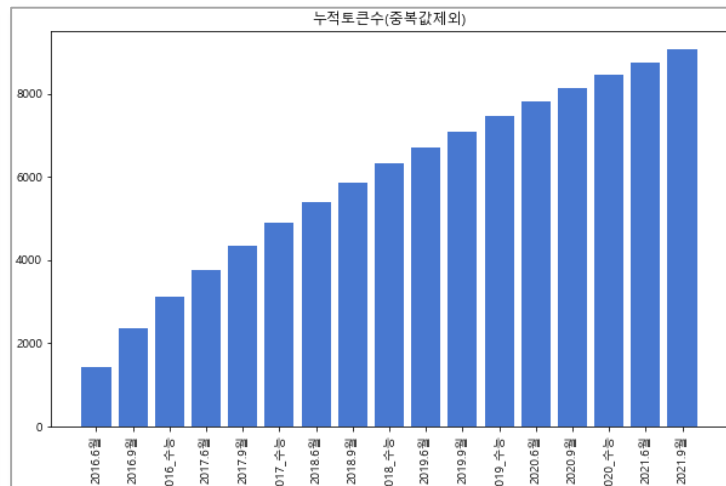
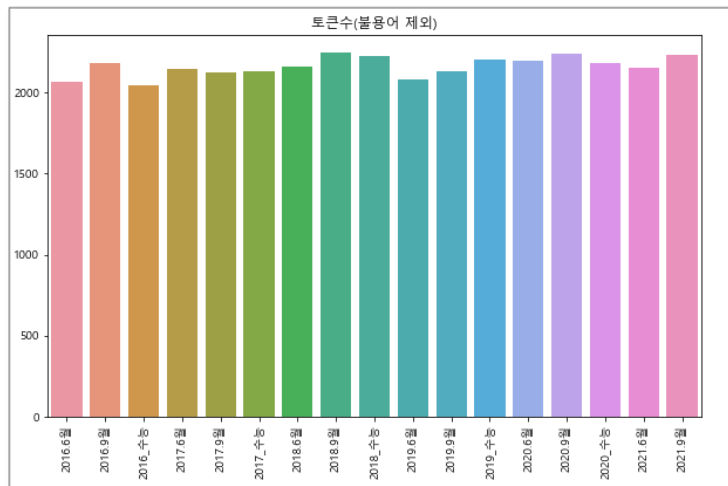


▶ 2021년 수능영어 시험 토큰 리스트 (불용어 제거)

| 순 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | ... | 1208 | 1209 | 1210 |
|-----|------|-----|------|-----|-------|-------|--------|-----|------|--------|----------------|-----------|-------------|------|--------|---------|------|------|-----|----------|--------|------|
| 단어 | cell | one | anna | new | world | would | social | use | jane | medium | representation | scientist | explanation | cora | people | science | many | need | ... | regional | latest | went |
| 빈도수 | 16 | 14 | 13 | 12 | 11 | 11 | 10 | 10 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 8 | ... | 1 | 1 | 1 |

- ✓ 각 연도별 모의평가 기출 단어와 수능시험 기출 단어를 분석하였으며 불용어 제외시 평균 2,162개 토큰으로 지문이 구성되어 있음을 확인함
- ✓ 각 회차가 누적 될 수록 신규 토큰수 들이 점차 감소되고 있음을 확인할 수 있었음

▼ 2016~2021.9 모의평가 및 수능기출출제 토큰 워드클라우드



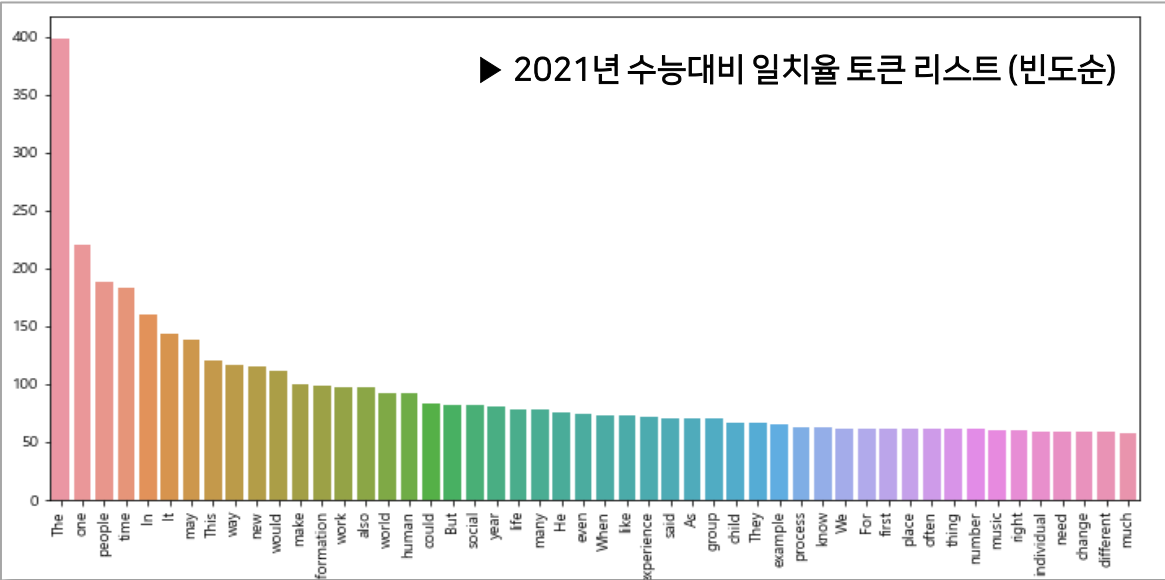
| 구분 | | 2016.6월 | 2016.9월 | 2016.수능 | 2017.6월 | 2017.9월 | 2017.수능 | 2018.6월 | 2018.9월 | 2018.수능 | 2019.6월 | 2019.9월 | 2019.수능 | 2020.6월 | 2020.9월 | 2020.수능 | 2021.6월 | 2021.9월 |
|-----------------|-----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 토큰수 | | 3,651 | 3,916 | 3,544 | 3,727 | 3,809 | 3,749 | 3,745 | 3,907 | 3,798 | 3,688 | 3,868 | 3,824 | 3,773 | 3,987 | 3,914 | 3,800 | 3,804 |
| 누적 토큰수 | | 3,651 | 7,567 | 11,111 | 14,838 | 18,647 | 22,396 | 26,141 | 30,048 | 33,846 | 37,534 | 41,402 | 45,226 | 48,999 | 52,986 | 56,900 | 60,700 | 64,504 |
| 누적 토큰수 (중복값 제거) | | 1,415 | 2,354 | 3,123 | 3,769 | 4,338 | 4,890 | 5,399 | 5,865 | 6,318 | 6,711 | 7,096 | 7,479 | 7,828 | 8,149 | 8,459 | 8,763 | 9,062 |
| 불용어 제외 | 토큰수 | 2,068 | 2,183 | 2,044 | 2,143 | 2,126 | 2,135 | 2,162 | 2,245 | 2,229 | 2,079 | 2,130 | 2,201 | 2,198 | 2,237 | 2,180 | 2,155 | 2,234 |
| | 누적토큰수 | 2,068 | 4,251 | 6,295 | 8,438 | 10,564 | 12,699 | 14,861 | 17,106 | 19,335 | 21,414 | 23,544 | 25,745 | 27,943 | 30,180 | 32,360 | 34,515 | 36,749 |
| | 누적 토큰수 (중복값 제거) | 1,302 | 2,227 | 2,992 | 3,636 | 4,204 | 4,753 | 5,262 | 5,728 | 6,180 | 6,573 | 6,958 | 7,340 | 7,689 | 8,009 | 8,319 | 8,623 | 8,921 |
| | 신규 토큰수 | - | ▲ 925 | ▲ 765 | ▲ 644 | ▲ 568 | ▲ 549 | ▲ 509 | ▲ 466 | ▲ 452 | ▲ 393 | ▲ 385 | ▲ 382 | ▲ 349 | ▲ 320 | ▲ 310 | ▲ 304 | ▲ 298 |

Basic Part III (2021년 수능 영어시험 대비 기출문제 일치율 분석)



- ✓ 각 회차별 출제된 토큰과 2021년 수능 출제 토큰과 일치여부를 비교시 평균 856개의 토큰이 일치하고, 1,380개의 토큰이 불일치 하였으며 평균 38.3%의 일치율을 보여주었음
- ✓ 각 회차별 토큰이 누적될 수록 토큰수 일치율이 지속적으로 증가하는 경향을 보여주었으며 최종 누적 일치율은 84.5%까지 증가하는 것을 확인함 (2021수능 토큰 : 2,237개)

| 2021 수능대비 일치여부 | | | 2016.6월 | 2016.9월 | 2016.수능 | 2017.6월 | 2017.9월 | 2017.수능 | 2018.6월 | 2018.9월 | 2018.수능 | 2019.6월 | 2019.9월 | 2019.수능 | 2020.6월 | 2020.9월 | 2020.수능 | 2021.6월 | 2021.9월 |
|----------------|---------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 일치 토큰수 | | | 771 | 839 | 845 | 858 | 882 | 799 | 860 | 877 | 887 | 810 | 858 | 913 | 862 | 883 | 891 | 806 | 921 |
| 불일치 토큰수 | | | 1,466 | 1,398 | 1,392 | 1,379 | 1,355 | 1,438 | 1,377 | 1,360 | 1,350 | 1,427 | 1,379 | 1,324 | 1,375 | 1,354 | 1,346 | 1,431 | 1,316 |
| 토큰수 일치율 | | | 34.5% | 37.5% | 37.8% | 38.4% | 39.4% | 35.7% | 38.4% | 39.2% | 39.7% | 36.2% | 38.4% | 40.8% | 38.5% | 39.5% | 39.8% | 36.0% | 41.2% |
| 회차별 누적 | 중복 감제 외 | 일치 토큰수 | 771 | 1,081 | 1,251 | 1,405 | 1,517 | 1,583 | 1,643 | 1,691 | 1,722 | 1,746 | 1,769 | 1,789 | 1,810 | 1,828 | 1,849 | 1,874 | 1,891 |
| | | 불일치 토큰수 | 1,466 | 1,156 | 986 | 832 | 720 | 654 | 594 | 546 | 515 | 491 | 468 | 448 | 427 | 409 | 388 | 363 | 346 |
| | | 토큰수 일치율 | 34.5% | 48.3% | 55.9% | 62.8% | 67.8% | 70.8% | 73.4% | 75.6% | 77.0% | 78.1% | 79.1% | 80.0% | 80.9% | 81.7% | 82.7% | 83.8% | 84.5% |



```
result = np.array([1])
all_corpus = []
word_tokens = []
stacked_token = 0
all_stacked_token = []
all_stacked_len = 0
stacked_stop_token = 0
word_token_len = []
all_stop_stacked_token = []
stop_words = set(stopwords.words('english'))

for file in file_list:
    if file[0][0:1] == "6월":
        df = pd.read_csv(file, header=None)
    else:
        df = pd.read_csv(file)

    df = df.dropna()
    df.columns = ["number", "contents"]
    df_word = df

    for i in range(len(df)):
        df_word["contents"][i] = re.sub("[0-9]", "", df_word["contents"][i])
        df_word["contents"][i] = re.sub("[a-zA-Z0-9]", "", df_word["contents"][i])

    df_2021 = pd.read_csv("18_2021수능.csv")
    for i in range(len(df_2021)):
        df_2021["contents"][i] = re.sub("[0-9]", "", df_2021["contents"][i])
        df_2021["contents"][i] = re.sub("[a-zA-Z0-9]", "", df_2021["contents"][i])

    sent_corpus = [sent_tokenize(s) for s in df_word["contents"]]
    corpus = []
    for i in sent_corpus:
        corpus += sent_corpus[i]

    sent_corpus_21 = [sent_tokenize(s) for s in df_2021["contents"]]
    corpus_21 = []
    for i in sent_corpus_21:
        corpus_21 += sent_corpus_21[i]
```

<코드2>

```
#토큰수 불일치 제외
stop_result_21 = []
for i in range(len(corpus_21)):
    for w in word_tokenize(corpus_21[i][0]):
        if w not in stop_words:
            stop_result_21.append(w)

stop_result = []
for i in range(len(corpus)):
    for w in word_tokenize(corpus[i][0]):
        if w not in stop_words:
            stop_result.append(w)

same_num = 0
for i in stop_result_21:
    if i in stop_result:
        same_num = same_num + 1

for i in sent_corpus:
    all_corpus = all_corpus + i

#누적 토큰수 불일치 제외
stop_result = []
for i in range(len(all_corpus)):
    for w in word_tokenize(all_corpus[i]):
        if w not in stop_words:
            stop_result.append(w)

tt = pd.Series(np.array(stop_result)).unique()

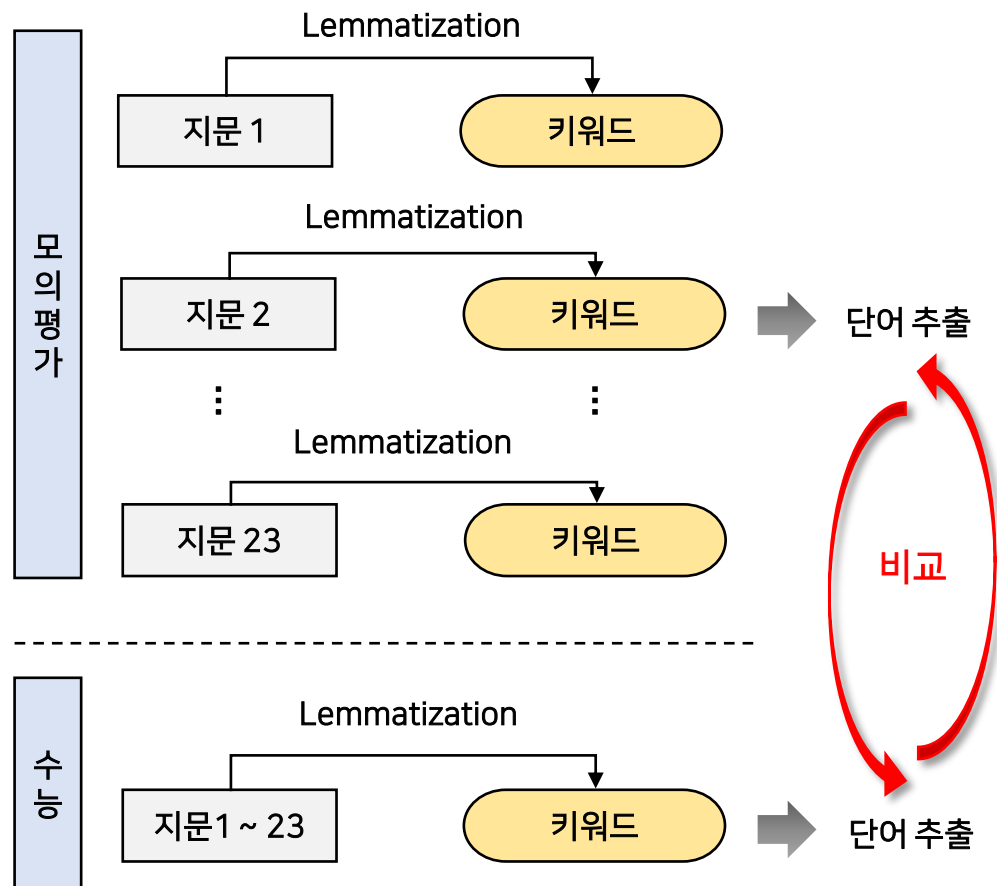
all_same_num = 0
for i in stop_result_21:
    if i in tt:
        all_same_num = all_same_num + 1

result = np.append(result, np.array([same_num, len(stop_result_21) - same_num, same_num / len(stop_result_21)]))
```

Advanced Part I (단어 일치율 비교)

- ✓ 각 연도별 모의평가 기출 단어와 수능시험 기출 단어의 품사 태깅 및 표제어 추출(Lemmatization) 처리를 진행 후 일치율의 비교 분석을 진행함
- ✓ 6개년도 데이터에 대해 비정형 분석을 수행한 결과, 평균적으로 수능단어의 47.7%가 6월 혹은 9월 모의고사 빈출단어와 일치하는 것으로 산출됨
- ✓ 일치하는 단어의 난이도는 평이한 수준이며, 고등학생이라면 반드시 알아야만 하는 단어로 구성됨

▶ 단어비교 모델



<코드3>

```
def tagwn(tag):
    return {
        'N': wn.NOUN,
        'V': wn.VERB,
        'R': wn.ADV,
        'J': wn.ADJ
    }.get(tag[0], wn.ADJ)

def normalize(text):
    for token, tag in nltk.pos_tag(nltk.wordpunct_tokenize(text)):
        token = token.lower()
        if token in stopwords or token in punctuation:
            continue
        token = lemmatizer.lemmatize(token, tagwn(tag))
    yield token
```

⇒ Pos tagging

⇒ Lemmatization

▶ 단어일치율 (%)

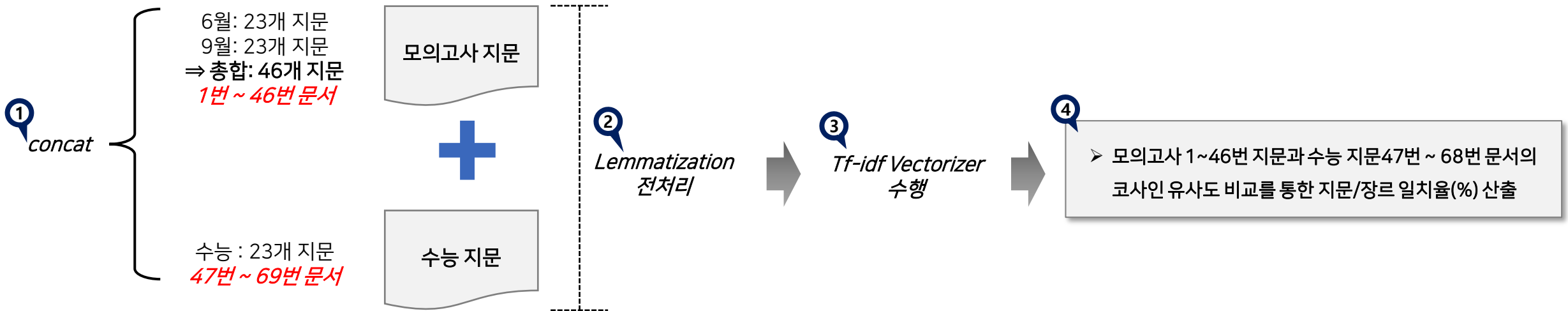
| 수능년도 | 6월 모평 | 9월 모평 |
|------|--------|--------|
| 2016 | 44.60% | 46.10% |
| 2017 | 47.20% | 45.50% |
| 2018 | 48.50% | 47.70% |
| 2019 | 48.90% | 48.00% |
| 2020 | 48.90% | 49.20% |
| 2021 | 46.50% | 51.40% |

▶ 단어 리스트 (일치 빈도순)

Counter({'food': 14, 'one': 13, 'time': 11, 'experience': 10, 'would': 10, 'also': 10, 'right': 10, 'people': 10, 'sale': 10, 'give': 10, 'new': 9, 'make': 9, 'increase': 9, 'year': 9, 'less': 8, 'many': 8, 'work': 7, 'high': 7, 'school': 7, 'eat': 7, 'produce': 7, 'say': 7, 'human': 7, 'great': 6, 'world': 6, 'among': 6, 'life': 6, 'need': 6...})

- ✓ 6개년도 데이터에 대해 비정형 분석을 수행한 결과, 평균적으로 수능지문의 79.7%가 6월 혹은 9월 모의고사 지문과 유사한 것으로 산출됨
- ✓ 또한 검출된 유사지문의 79%가 지문의 장르가 일치하는 것으로 분석됨 (안내문, 설명문, 소설, 도표 해석 등)
- ✓ 이는 글의 전개방식이나 성격에 따라 사용되는 어휘의 차이로 인한 결과라고 추측되며, 따라서 수능 영어에서 요구하는 능력은 글의 종류에 따른 전개방식에 능숙해지는 것임을 추측함

▶ 지문유사도 비교모델



```
import numpy as np
import pandas as pd
import re
from nltk import sent_tokenize
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

read6 = pd.read_csv('2016.6월.csv')
read9 = pd.read_csv('2016.9월.csv')
exam = pd.read_csv('2016_수능.csv')
```

<코드4>

```
read6 = read6['contents'].dropna(axis=0)
read9 = read9['contents'].dropna(axis=0)
exam = exam['contents'].dropna(axis=0)
doc = pd.concat([read6, read9, exam], ignore_index=True) ⇒ Concat

def find_most_similar_news_idx(index, tfidf, corpus):
    idx = (-cosine_similarity(tfidf[index], tfidf[0]).argsort()[1])
    return {'index': idx, 'contents': corpus[idx]}
```

⇒ 코사인 유사도 산출 함수 정의

▶ 지문유사율 / 장르일치율 (%)

| 구분 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|------------|--------|--------|--------|--------|--------|--------|
| 모의고사 지문유사율 | 78.30% | 73.90% | 78.30% | 91.30% | 69.60% | 87.00% |
| 장르 일치율 | 95.70% | 69.60% | 73.90% | 78.30% | 73.90% | 82.60% |

Conclusion

- ✓ 데이터를 수집 및 전처리 하는 과정에서 많은 시간이 소요되었으며, 향후 더 많은 학습데이터를 수집하기 위해서는 문제 유형에 따른 PDF를 자동으로 크롤링 해주는 것이 필요
- ✓ 수능 및 평가원 기출문제 외 EBS 지문과 학평 (3,4,7,10월) 시험문제 등이 반영되면 더 좋은 결과를 예측할 수 있을 것이라 판단됨
- ✓ 평가원 기출문제 출제 위원이 실제로도 수능 출제위원으로 문제를 제출하기 때문에 기존에 고안하였던 seq to seq 모델링을 통해 분석한다면 의미있는 결과가 있지 않을까 예상함



- 감사합니다. -