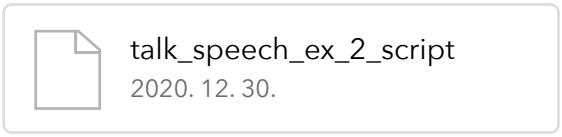
201230

1. 영어, 국사 텍스트 추출

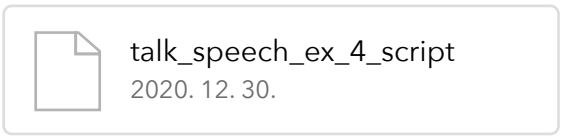
   (1) 영어 - KNS 어휘 특강

   (2) 국사 - http://www.talkkang.com/index.html?mode=read&lecture_idx=947

   (1)은 문제 풀이 강의로 어휘 문제를 푸는 것밖에 없고 강의 내용도 주로 단어 뜻 암기 확인에 그침. 추출한 스크립트를 활용해서 학습카드 만들기에는 단순한 단어 카드 정도밖에 안 될 거 같음. 그래서 부적합한 강의로 판단함.

   (2)를 추출한 결과 +

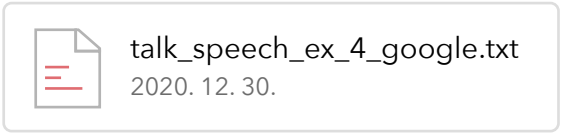
   문맥이나 추출한 키워드 등의 퀄리티가 그냥 그럼

   ➜ 201229에서 고민한 것 중에 원문 스크립트의 질을 높일지, 추출 알고리즘의 정확도를 다방면으로 높여볼지가 있었는데 전자가 아무래도 중요한 것 같음. 요약 모듈들을 보면 잘 만들어진 글(뉴스 기사, 리뷰 문장 등 이미 글로 쓰여진 어쨌든 문장, 지금 음성인식한 스크립트는 문장 형태가 아닌 단어들의 나열에 가까워 보임)에 대해서는 비교적 잘 작동함.

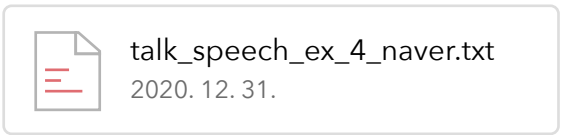
2. 구글 VS 네이버

   국사 1번 문제 설명에 대해

   (1) 구글 API + 1번 문제에서 명사만 따서 힌트로 추가 + 구두점(쉼표, 마침표)을 추가

   (2) 네이버 API

   각각의 키워드, 요약문 결과

   (1-1)

```
('문제/NNG', 2.1290206563052543)('시작/NNG', 1.8521290756880315)('씨족/NNP', 1.4688786490371353)
('단위/NNG', 1.38305099951183)('농경/NNG', 1.2666669174287537)('부분/NNG', 1.2666669174287537)
('기본/NNG', 1.180980276261597)('오늘/NNG', 1.15462009696147)('신석기 시대/NNP', 1.0951137775509545)
('신석기/NNP', 1.0951137775509545)('사람/NNG', 0.9326719509513683)('등장/NNG', 0.8931890724758897)
('주의/NNG', 0.8931890724758897)('단위/NNP', 0.8931890724758897)('채집/NNG', 0.8931890724758897)
('경제생활/NNG', 0.8931890724758897)('사냥/NNG', 0.8931890724758897)('시대/NNG', 0.8072672631897774)
('청동기/NNG', 0.8072672631897774)('부족/NNG', 0.8072672631897774)('조개/NNP', 0.8072672631897774)
('자원/NNP', 0.7209579262598862)('어족/NNP', 0.7209579262598862)('평등/NNP', 0.7209579262598862)
('정답/NNG', 0.6443790673236488)('인식/NNG', 0.6443790673236488)('이야기/NNG', 0.5676257521412484)
```

re_talk_speech_ex_4_goog...
2020. 12. 31.

(1-2)

```
'습니다 신이라고 와서 볼게요 오늘 먹으면 경제생활은 농경과 더불어 채집과 사냥을 좀업했다. 요게 그런데 조금 주의하시고 부분인데 신석기 혁명이라고 해서 농경과 목적은 시작된겁니다. 시작되었다고 해서 전체적으로 농경 과목 죽이는게 아니라 이제 시작되었기 때문에 서서히 그런 게 나타나기 시작했다. 이런 얘기예요. 일반적으로 주된 경제생활은 여전히 채집과 사냥이 주된 경제 생활이었다. 요건 여러분이 주의를 기울여야 되는 부분인데 어쨌든 신석기 시대에는 부분이죠. 딱 보니까 산본 고기 스스로 하늘의 아들이라고 우리는 선민사상 등장이라고 나왔지요. 바로 이선민 사장은 청동기시대의 등장을 한 거니까 혼돈을 이렇게 없습니다. 다 4번 씨족을 기본단위로 씨족을 직원도 아니고 씨가 같은 접속입니다. 얘네를 기본단위로 부족 단위로 넘어가게 되는데 바로 이걸 뭘 가지고 갈 수 있냐면 조개 오네.\n'

이라고 하는 걸 알 수 있습니다. 조개 올레시장 조계훈 다른 씨족과 이혼 일을 통해서 씨족 단위가 부족 단위로 넘어가게 되는데, 그때 시종이 기본단위이며 적 중요한 것은 공동체적인 평등사회를 여전히 유지하고 있다. 구석기 신석기는 여전히 평등사회 없습니다. 청동기시대에 비로소 계급이 발생 되거든요. 가자 영어로 바로 신석기 시대 단계는 문제였죠. 정답은 그러니까 3번이 되는데 이제 산에 올라가서 인식이 사람들은 주로 살았다. 구석기인들은 주생활 살이고 신석기 사람들은 후방기에 해수면이 상승. 그러면서 어족자원이 풍부해집니다. 이러다 보니까 많은 어족자원을 얻기 위해서 바로 해안가로 삶의 무게가 옮겨서 갔다. 오는 거 아시면 딱 신석기 시대의 나 그리고.\n'


'자 오늘 준비가 된 문제는 그렇게 먹어 여러분들에게 난도가 이렇게 센 그런 문제는 아닙니다. 내가 죽어 앞으로 이제 나 오늘 문제를 보시면 알겠지만 그게 문제가 아니니까 좀 편안한 마음으로 아침 좀 출발 하시길 바라고요 오늘 뭐 52% 정답 나왔습니다. 1번부터 자 그럼 일본 문제를 한번 보시는데 다음 중 인식이라는 설명으로 옳지 않아 이렇게 물었지요. 자 그러면이 시기에 사람들이라고 이야기했는데 사실 이런 문제를 풀 때 요령 같은 경우를 하나만 더 말씀해 드릴게요. 물론 여기서 만약에 옳은 것을 하라고 이야기를 하면 약간 다를 수도 있습니다. 그러면 시기가 여러 개 다 나오니까 하지만 옳지 않은 것이라고 먹게 되면 선사시대 문제를 모를 때 옳지 않은 걸 찾으라 그러면 세계 보기는 어떻게 나옵니까? 팬시 드려야겠죠? 사료를 보지 않고 바로 옳지 않은 건 문제에서는 바로 보기로 보겠습니다. 가락바퀴 뼈바늘 이용했다. 원시적 수공업이 시작되었다. 바로 어느 쪽입니까?\n'
```

re_sen_talk_speech_ex_4_...
2020. 12. 31.

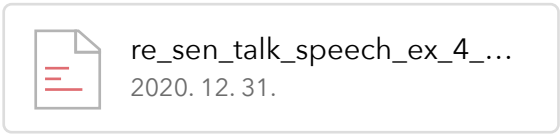
(2-1)

```
('단위/NNG', 2.794453160663375)('시작/NNG', 1.9130027519811175)('기본/NNG', 1.7665732046552938)
('생활/NNG', 1.6639963746547717)('경제/NNG', 1.3507989492495995)('부분/NNG', 1.3507989492495995)
('문제/NNG', 1.2002862886407466)('농경/NNG', 1.1856083085700262)('신석기/NNP', 1.0555642293997574)
('이번/NNG', 1.0334647254504425)('시중/NNG', 1.0305664672540913)('주의/NNG', 0.9687878913518054)
('채집/NNG', 0.9687878913518054)('사냥/NNG', 0.9687878913518054)('등장/NNG', 0.9668535865228525)
('사장/NNG', 0.9668535865228524)('선민/NNP', 0.9668535865228524)('시대/NNG', 0.9415032056504422)
('사람/NNG', 0.8950002889271456)('개원/NNP', 0.8755410641722697)('부족/NNG', 0.8755410641722697)
('사실/NNG', 0.858150886739723)('목축/NNP', 0.7358043498472889)('어족/NNP', 0.7158983647252475)
('무대/NNG', 0.7158983647252475)('자원/NNP', 0.7158983647252475)('청동기 시대/NNP', 0.6986607295513441)
('시기/NNG', 0.6443550314682088)('해안/NNG', 0.5937432661732231)('이야기/NNG', 0.5883760478171389)
```

re_talk_speech_ex_4_naver...
2020. 12. 31.

(2-2)

```
[(12, 1.8868598376483456, '사실 이번 보기에 혼돈을 일으킬 우려도 없습니다. 4번 시중을 기본 단위로. 시중을 기본 단위로. 시가 같은 족속입니다. 예내를 기본 단위로 부족 단위로 넘어가게 되는데 바로 이걸 뭘 가지고 갈 수 있냐면 조개원의 시작이라고 하는 걸 알 수 있습니다. \n'),

(6, 1.4083178289183573, '선사시대 문제를 물을 때 옳지 않은 걸 찾으라 그러면 세기의 보기는 어떻게 나옵니까. 같은 시대를 이야기하겠죠. 사료를 보지 않고 바로 옳지 않은 것. \n'),

(8, 1.197476135849838, '이번을 보면 경제 생활은 농경과 더불어 채집과 사냥을 주로 했다. 여기에 여러분들 조금 주의하시는 부분인데 신석기 혁명이라고 해서 농경과 목축은 시작된 겁니다. \n'),

(3, 1.1134448826993784, '1번부터 1번 문제를 한 번 보시는데 다음 중 이 시기에 해당하는 설명으로 옳지 않은 이렇게 물었죠. \n'),

(5, 1.0838053168806634, '물론 여기서 만약에 오른 것을 찾아라고 이야기를 하면 약간 다를 수도 있습니다. 왜냐하면 시기가 여러 개가 나오니까. 하지만 옳지 않은 것이라고 묻게 되면.\n'),

(18, 1.0071998605294852, '이러다 보니까 많은 어족자원을 얻기 위해서 바로 감가와 해안가로 삶의 무대가 옮겨져 갔다. 이렇게 여러분들이 아시면 딱 신속시대라 그리고 원형'),

(16, 1.0, '정답은 그러니까 3번이 되는데 이제 사례에 올라가서 인식이 사람들은 주로 강가와 해안가에 살았다. \n')]
```

(0) 문제 풀이에 사용된 문제 원문
(보기) '이 시기'의 사람들은 주로 강가와 해안가에 살았다.
(선지)
집 자리는 보통 4~5명 정도의 가족이 살기에 알맞은 크기로 바닥은 원형 또는 모가 둥근 사각형이다.
가락바퀴나 뼈바늘을 이용하여 원시적 수공업을 하였다.
경제생활은 농경과 더불어 채집과 사냥을 주로 하였다.
스스로 하늘의 아들이라고 믿는 선민사상이 등장하였다.
씨족을 기본 단위로 하는 원시공동체적 평등사회였다.

키워드를 일단 비교해 보면 **둘이 비슷**한 듯. 네이버 쪽에는 문제 풀이에 사용된 문제(0)의 텍스트 정보를 안 넣어서 씨족이 시중으로 나오고 이런 정도만 감안하면 비슷. 문제 풀이 설명을 전체를 듣지는 않아서 확신할 수는 없지만 학생들이 실제로 보고 푸는 선지의 주요 단어들(문제를 실제로 보면 집 크기, 가족원 수, 가락바퀴, 뼈바늘, 농경, 채집, 사냥, 선민사상, 씨족, 원시공동체, 평등사회 정도)이 실제 추출된 키워드에는 절반 넘게 포함된 듯? 물론 이거는 강사의 설명에서 언급된 비중에 따라 다르겠지만. **문제를 직접 듣고 요약해본 다음에 정확도를 비교해봐야 할 듯** 음성인식 결과인 스크립트와 키워드에서 좀 짚고 넘어가야 할 부분은 조계원, 조개 등등으로 된 부분인데 이게 내가 알기론 '족외혼'인데 이렇게 인식된 듯. 족외혼은 (0)에 없고 실제로 api 측 데이터베이스에 있을리가 없는 단어이므로 **인식되기가 힘듦. 그렇지만 강사의 설명에서 족외혼은 나름 중요한 키워드였음.** 이런 경우에 대해서는 어떻게 대처해야되는지. 수동으로 고치는 수밖에 없나?

요약문 추출에서는 **네이버 사용한 자료가 훨씬 나음.** 문장 길이는 내가 네이버 스크립트 저장을 한 문장 내지 두 문장으로 끊어서 저장해서 그럼. 구글은 음성인식한 그대로 가져온 거라 \n 있는 간격이 좀 큼. 확실히 음성인식 결과 스크립트 자체만 보면 네이버가 문장 형태를 더 갖추고 있다고 느꼈음. 그래서 요약문 추출에서 장점이 두드러진 결과가 나온 것 같음. 구글 쪽은 사실 뭐라하는지 잘 모르겠지만 네이버 쪽은 꽤 알아 들을 수 있음. 요약문을 얻는 쪽 만큼은 네이버로 가는게 좋을 듯 싶음.

3. 내일은 뭘 해야 하나... 일단 과학 쪽 데이터 준비해 놓은 것 음성인식 돌려보고 다양한 과목, 그리고 강의 데

이터를 얻는 것을 병행하면서
요약문을 얻는 것에 네이버가 훨 나았으니 네이버 api 사용 방법을 익혀야 할 듯
그리고 키워드 추출에서 선민 사상, 평등 사회 등 두 단어씩 붙어있는게 훨씬 나은 키워드들이 있는데 이게 쪼개져서 나오니까, 이거를 보완할 방법을 찾아보고
과목 강의 데이터 자체가 좀 많아지면 그때 요약 모듈도 좀 다방면으로 사용하며 실험해봐야 할 것 같음
키워드 추출을 더 깔끔하게 하려면 불용어 리스트를 좀 만들어야 할 듯 데이터 좀 더 얻어서 돌리고 얻은 공통적인 것이나, 아니면 한 종류의 강의 안에서 통용되는 리스트가 더 활용도 높을 가능성도 무시못할 듯