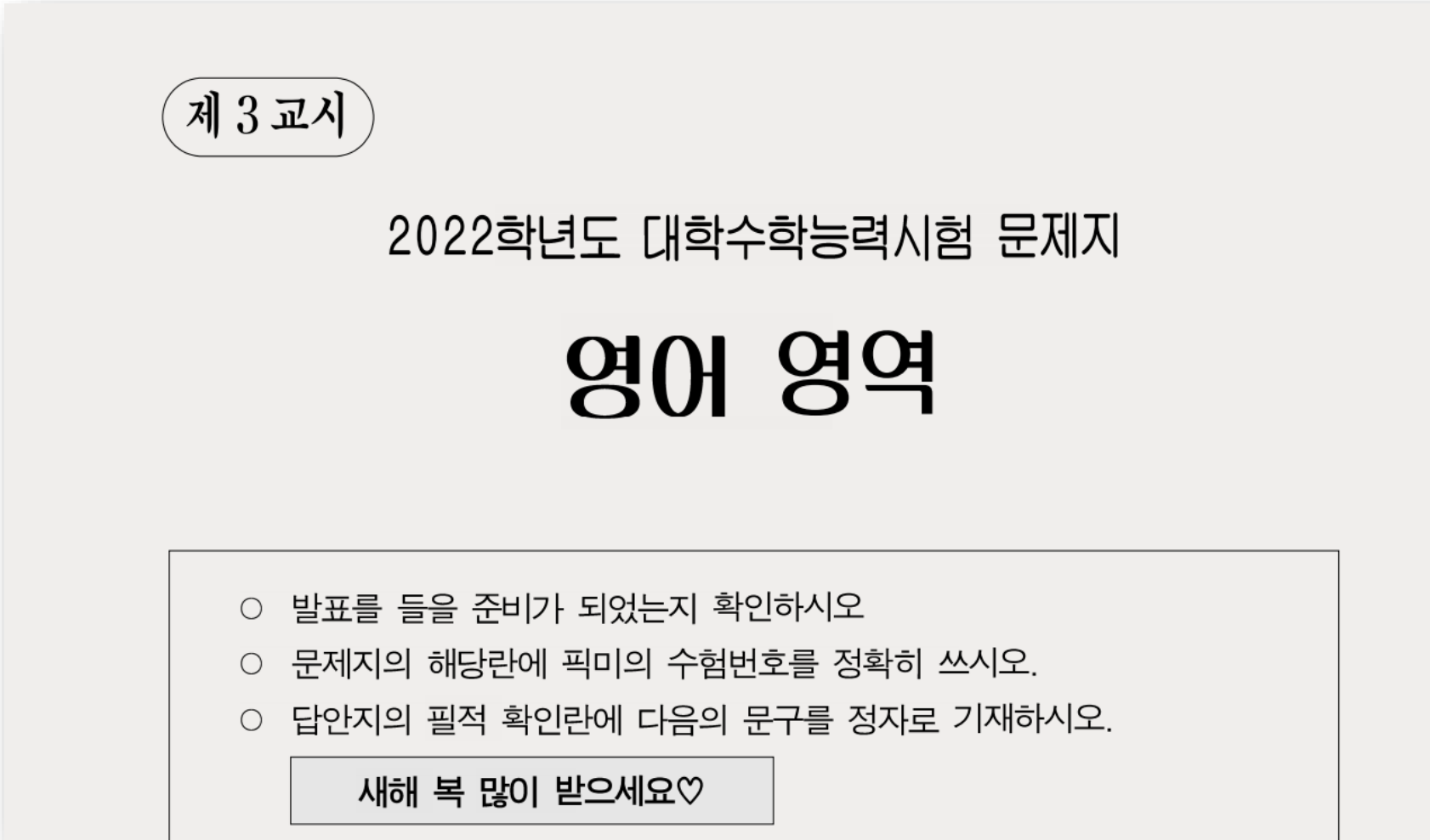
수능영어 풀이봇

픽미(PICK ME)

Mid-Project



1반 3팀
고수진 김수희 오종민
이세준 허정은

LIKELION_AI

어떤 **주제**로 정할까?

수능 영어영역 지문

어떤 **주제**로 정할까?

수능 영어영역 지문

어떤 **주제**로 정할까?

수능영어 지문을 자연어처리하여 **주제/제목/요지**찾기 문제를 풀어보자

PICK ME

1.Data:
수능영어지문
주장/제목/요지문제

2.How:
자연어처리

3.Goal:
5지선다 맞추기

사용한 **Dataset**

23. 다음 글의 주제로 가장 적절한 것은?

Human beings do not enter the world as competent moral agents. Nor does everyone leave the world in that state. But somewhere in between, most people acquire a bit of decency that qualifies them for membership in the community of moral agents. Genes, development, and learning all contribute to the process of becoming a decent human being. The interaction between nature and nurture is, however, highly complex, and developmental biologists are only just beginning to grasp just how complex it is. Without the context provided by cells, organisms, social groups, and culture, DNA is inert. Anyone who says that people are "genetically programmed" to be moral has an oversimplified view of how genes work. Genes and environment interact in ways that make it nonsensical to think that the process of moral development in children, or any other developmental process, can be discussed in terms of nature *versus* nurture. Developmental biologists now know that it is really both, or nature *through* nurture. A complete scientific explanation of moral evolution and development in the human species is a very long way off.

* decency: 예의 ** inert: 비활성의

① evolution of human morality from a cultural perspective
② difficulties in studying the evolutionary process of genes
③ increasing necessity of educating children as moral agents
④ nature versus nurture controversies in developmental biology
⑤ complicated gene-environment interplay in moral development

사용한 Dataset

23. 다음 글의 주제로 가장 적절한 것은?

Human beings do not enter the world as competent moral agents. Nor does everyone leave the world in that state. But somewhere in between, most people acquire a bit of decency that qualifies them for membership in the community of moral agents. Genes, development, and learning all contribute to the process of becoming a decent human being. The interaction between nature and nurture is, however, highly complex, and developmental biologists are only just beginning to grasp just how complex it is. Without the context provided by cells, organisms, social groups, and culture, [illegible] to be moral has an oversimplified view of how genes work. Genes and environment interact in ways that make it nonsensical to think that the process of moral development in children, or any other developmental process, can be discussed in terms of nature *versus* nurture. Developmental biologists now know that it is really both, or nature *through* nurture. A complete scientific explanation of moral evolution and development in the human species is a very long way off.

* decency: 예의 ** inert: 비활성의

① evolution of human morality from a cultural perspective
② difficulties in studying the evolutionary process of genes
③ increasing necessity of educating children as moral agents
④ nature versus nurture controversies in developmental biology
⑤ complicated gene-environment interplay in moral development

94년~2020년 수능 주제/제목/요지 찾기 80문항 지문

사용한 Dataset

text	#1	#2	#3	#4	#5	answer
Twin sirens hide	distinct difference	universal features	historians' efforts	pros and cons of	beliefs that cause	5
Do you have the	reasons for leade	influence of leade	necessity for ana	various ways of s	ways of strengthe	2
Tourism is import	misunderstanding	various ways of c	negative effects c	disappearance of	cultural benefits c	5
Many disciplines	history of science	limitations of lear	importance of lea	effects of intuition	difference betwee	3
The most normal	dangers of playin	beneficial influenc	children's play as	necessity of inter	parental roles in	3
Textiles and cloth	educational funct	ways to diversify	gender difference	different cultural r	nonverbal commu	5
Recently, researc	effects of laughte	benefits of activat	strategies for coc	negative aspects	importance of gro	1
In order to succe	necessity of pare	roles of parental a	consequences of	requirements for	importance of cor	5
Even when scien	problems of maki	needs for an alter	insufficient under	potential benefits	challenge of findi	3
In the nineteenth	the crisis of mode	the various sourc	the media's explo	the importance of	the fragile nature	3
Why is it difficult	reasons for runne	differences betwe	comparison of sp	necessity of build	relationship betw	2
Scientists should	necessary condit	importance of ide	requirements for	guidelines for coll	effective strategie	3
We sometimes e	positive impact of	importance of fan	relationship betw	tests as a means	necessity of inter	4
Hundreds of spec	protective instinct	origin of social or	fish schooling as	necessity of fish	behavioral differer	3
What everyday ru	changes in mater	limitations of disc	parents' concerns	importance of par	effects of thinking	1
Ancient Greek ar	basic characteris	significant transfo	the greatness of	the origin of ancie	difficulties in defir	1
Living things natu	physical balance	inner mechanism	general tendency	major differences	biological proces	3
All of us use the	cultural difference	tragic characteris	the positive funct	human nature an	the process of ac	1

23. 다음 글의 주제로 가장 적절한 것은?

Human beings do not enter the world as competent moral agents. Nor does everyone leave the world in that state. But development in the human species is a very long way off.

* decency: 예의 ** inert: 비활성의

① evolution of human morality from a cultural perspective
② difficulties in studying the evolutionary process of genes
③ increasing necessity of educating children as moral agents
④ nature versus nurture controversies in developmental biology
⑤ complicated gene-environment interplay in moral development

어떤 **모델**을 사용할까?

Textrank

Word2vec

Doc2vec

Wmd

(Seq2seq + attention)

+ Machine Learning

어떤 **모델**을 사용할까?

Textrank

Word2vec

Doc2vec

Wmd

(Seq2seq + attention)

+ Machine Learning

어떤 **모델**을 사용할까?

Textrank
Word2vec
Doc2vec
Wmd
(Seq2seq + attention)
+ Machine Learning

어떤 **모델**을 사용할까?

Textrank
Word2vec
Doc2vec
Wmd
(Seq2seq + attention)
+ Machine Learning

어떤 **모델**을 사용할까?

Textrank
Word2vec
Doc2vec
Wmd
(Seq2seq + attention)
+ Machine Learning

전략 1 :
지문을 요약한 한 문장과 각 선지를 비교

TextRank로 지문을 한 문장으로 요약하고 그 문장과 각 선지별로 유사도를 비교,
가장 높은 유사도를 가진 지문을 정답으로 고른다.

토큰화한 요약문

['Get', 'start', 'dig', 'benefit', 'multiply', '.',
'heal', 'power', 'flowers—and', 'tree',
'fresh', 'air', 'sweet-smelling', 'soil', '.',
'walk', 'garden', 'matter', 'see', 'one',
'window', 'lower', 'blood', 'pressure',
'reduce', 'stress', 'ease', 'pain', '.']

보기1. ['ways', 'growing', 'flowers']
보기1. distance = 1.2767808544527506

보기2. ['curing', 'high', 'blood', 'pressure']
보기2. distance = 1.1608672050266047

보기3. ['healing', 'effect', 'gardening']
보기3. distance = 1.228558750464308

보기4. ['conditions', 'nursing', 'homes']
보기4. distance = 1.3076918564866575

보기5. ['trends', 'constructing', 'hospitals']
보기5. distance = 1.3256770700500087

전략 2 :
지문전체와 각 선지의 유사도 비교

Wmd와 word2vec으로 영어 지문 전체와 각 선지의 유사도를 비교하여,
가장 높은 유사도를 가진 지문을 정답으로 고른다.

토큰화한 본문

['knowing', 'something', 'happened', 'important.', 'understanding', 'historic', 'events', 'took', 'place', 'also', 'important.', 'this,', 'historians', 'often', 'turn', 'geography.', 'weather', 'patterns,', 'water', 'supply,', 'landscape', 'place', 'affect', 'lives', 'people', 'live', 'there.', 'example,', 'explain', 'ancient', 'egyptians', 'developed', 'successful', 'civilization,', 'must', 'look', 'geography', 'egypt.', 'egyptian', 'civilization', 'built', 'banks', 'nile', 'river,', 'flooded', 'year,', 'depositing', 'soil', 'banks.', 'rich', 'soil', 'could', 'help', 'farmers', 'grow', 'enough', 'crops', 'feed', 'people', 'cities.', 'meant', 'everyone', 'farm,', 'people', 'could', 'perform', 'jobs', 'helped', 'develop', 'civilization.']

선택지1: ['significance', 'geography', 'understanding', 'history']
이 선택지의 wmdistance는 1.2272952741574357 입니다

선택지2: ['effects', 'nile', 'river', 'egyptian', 'farming']
이 선택지의 wmdistance는 1.2180122765246166 입니다

선택지3: ['differences', 'geography', 'geology']
이 선택지의 wmdistance는 1.2825574145683871 입니다

선택지4: ['varieties', 'egyptian', 'civilization']
이 선택지의 wmdistance는 1.2455765729213704 입니다

선택지5: ['development', 'egyptian', 'culture']
이 선택지의 wmdistance는 1.2548030954974887 입니다

텍스트 요약에는 두가지 방법이 있다.

1) 추출적 요약

추출적 요약은 원문에서 중요한 핵심 문장 또는 단어구를 몇 개 뽑아서 이들로 구성된 요약문을 만드는 방법.
그렇기 **때문에 추출적 요약의 결과로 나온 요약문의 문장이나 단어구들은 전부 원문에 있는 문장들**입니다.
추출적 요약의 대표적인 알고리즘으로 머신 러닝 알고리즘인 텍스트랭크(TextRank)가 있다.

이 방법의 단점이라면, 이미 존재하는 문장이나 단어구로만 구성하므로 모델의 언어 표현 능력이 제한된다는 점.

2) 추상적 요약

추상적 요약은 원문에 없던 문장이라도 핵심 문맥을 반영한 새로운 문장을 생성해서 원문을 요약하는 방법.
마치 사람이 요약하는 것 같은 방식인데, 당연히 추출적 요약보다는 난이도가 높다.
이 방법은 주로 인공 신경망을 사용하며 대표적인 모델로 **seq2seq**를 사용한다.

지도학습이기 때문에 추상적 요약을 인공 신경망으로 훈련하기 위해서는 '원문' 뿐만 아니라 '실제 요약문'이라는 레이블 데이터가 있어야 합니다. -> 그렇기 때문에 데이터를 구성하는 것 자체가 하나의 부담.

텍스트 요약에는 두가지 방법이 있다.

1) 추출적 요약

추출적 요약은 원문에서 중요한 핵심 문장 또는 단어구를 몇 개 뽑아서 이들로 구성된 요약문을 만드는 방법.
그렇기 **때문에 추출적 요약의 결과로 나온 요약문의 문장이나 단어구들은 전부 원문에 있는 문장들**입니다.
추출적 요약의 대표적인 알고리즘으로 머신 러닝 알고리즘인 텍스트랭크(TextRank)가 있다.

이 방법의 단점이라면, 이미 존재하는 문장이나 단어구로만 구성하므로 모델의 언어 표현 능력이 제한된다는 점.

2) 추상적 요약

** 어텐션을 이용한 텍스트 요약(추상적 요약) : https://wikidocs.net/72820*

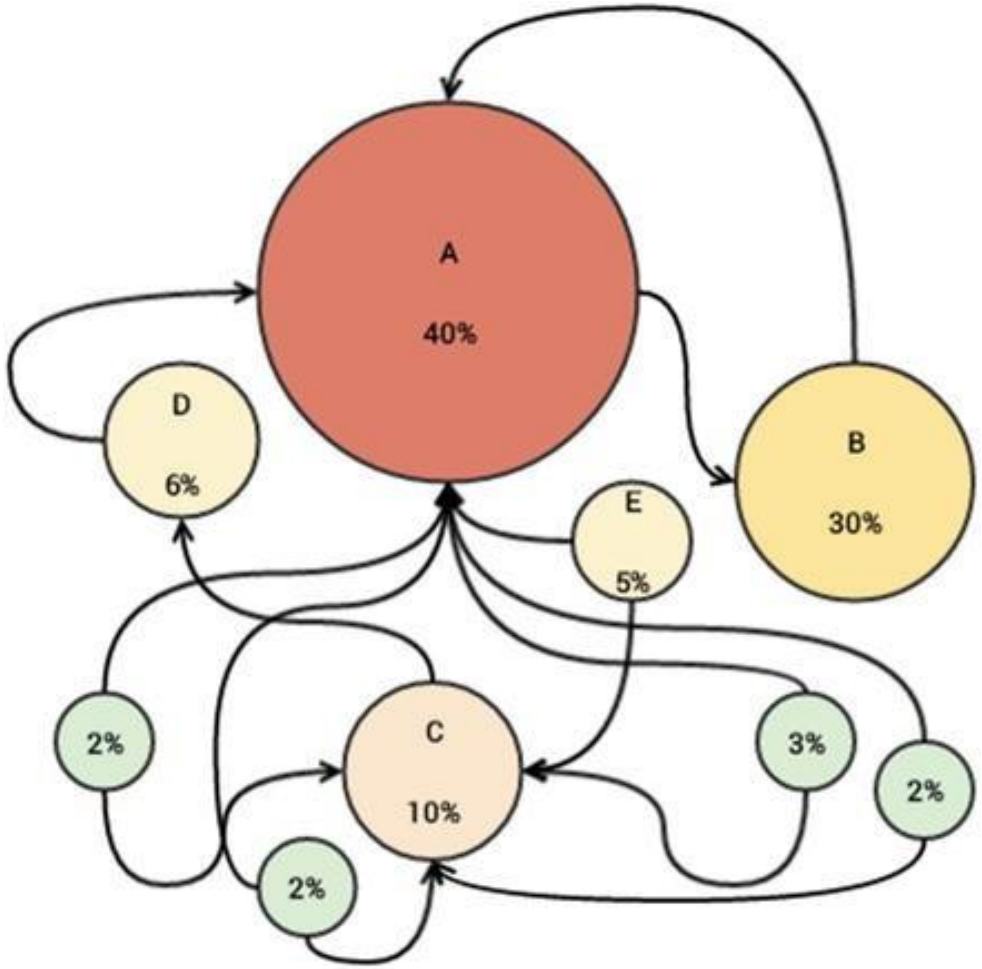
추상적 요약은 원문에 없던 문장이라도 핵심 문맥을 반영한 새로운 문장을 생성해서 원문을 요약하는 방법.
마치 사람이 요약하는 것 같은 방식인데, 당연히 추출적 요약보다는 난이도가 높다.
이 방법은 주로 인공 신경망을 사용하며 대표적인 모델로 **seq2seq**를 사용한다.

지도학습이기 때문에 추상적 요약을 인공 신경망으로 훈련하기 위해서는 '원문' 뿐만 아니라 '실제 요약문'이라는 레이블 데이터가 있어야 한다 -> 그렇기 때문에 데이터를 구성하는 것 자체가 하나의 부담.

Textrank algorithm

하이퍼링크를 가진 웹 문서

Google의 PageRank Algorithm을 이용



유사도 비교를 통해 문장이나 단어에 가중치를 매겨 하나의 문서를 요약하는 알고리즘

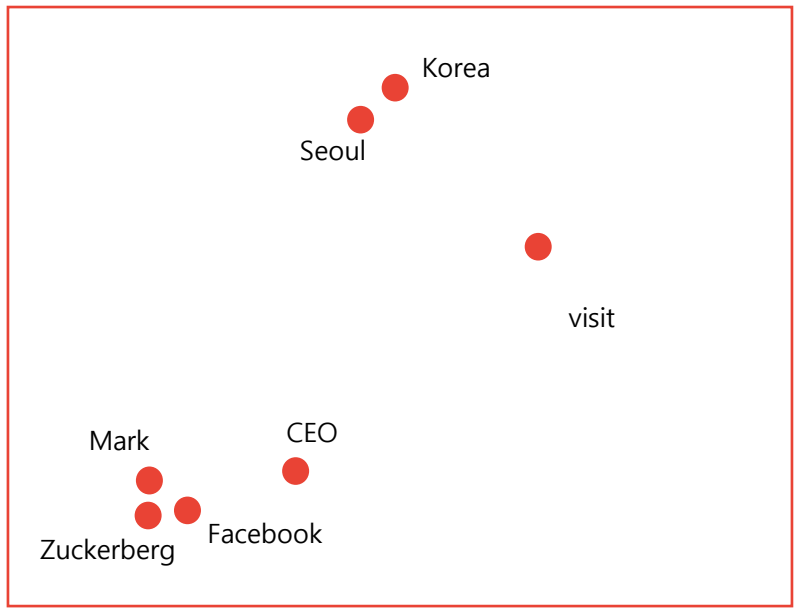
단어는 문장 내 공동 출현 값을 가중치로, 문장은 다른 문장과의 유사도를 가중치로 매긴다.

출현 빈도가 높은 단어나 문장을 추출하는 성격의 요약 알고리즘

Word2vec algorithm

고양이 + 애교 = 강아지
한국 - 서울 + 도쿄 = 일본
박찬호 - 야구 + 축구 = 호나우두

저차원에 단어의 의미를 여러 차원에 분산하여 표현



<One Hot Vector>
Mark = [1 0 0 0 0 0 0]
Zuckerberg = [0 1 0 0 0 0 0]
Visit = [0 0 1 0 0 0 0]
Korea = [0 0 0 1 0 0 0]
CEO = [0 0 0 0 1 0 0]
Facebook = [0 0 0 0 0 1 0]
Seoul = [0 0 0 0 0 0 1]

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = 0$$

Dimension = # of words = 7

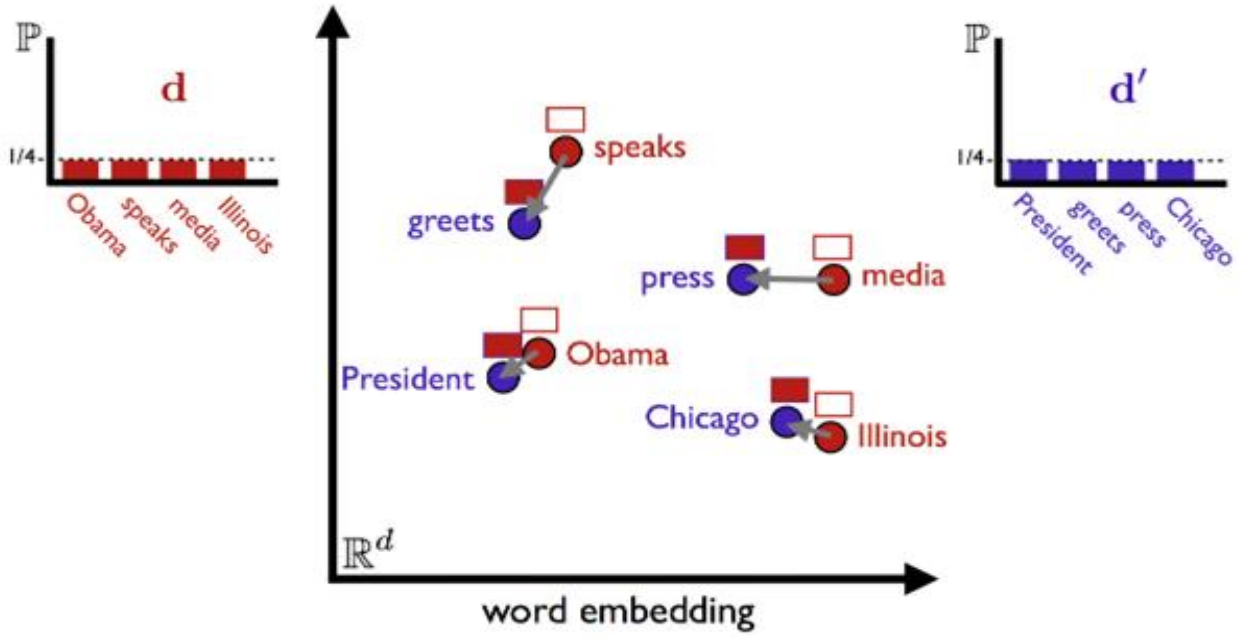
Mark Zuckerberg visited Korea.
The CEO of Facebook visited Seoul.

['Mark', 'Zuckerberg', 'visit', 'Korea']
['CEO', 'Facebook', 'visit', 'Seoul']

WMD algorithm (Word Mover's Distance)

Find the minimum traveling distance

doc1의 분포를 doc2의 분포로 '이동' 시키는 최적의 값을 구하는 것



word2vec vector embedding

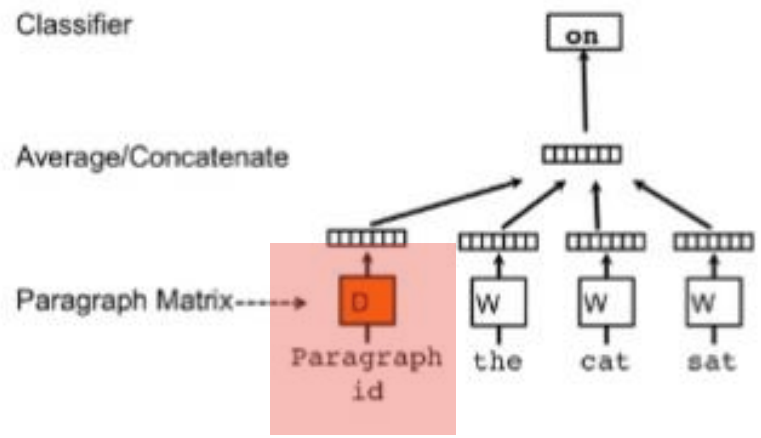
보통 직접 만들기에 데이터 양이 적고,
시간적인 한계가 존재

1st try : Using pretrained word2vec model by Google -> GoogleNews-Vectors-negative300-SLIM(불필요한 데이터가 없고 용량이 작음)

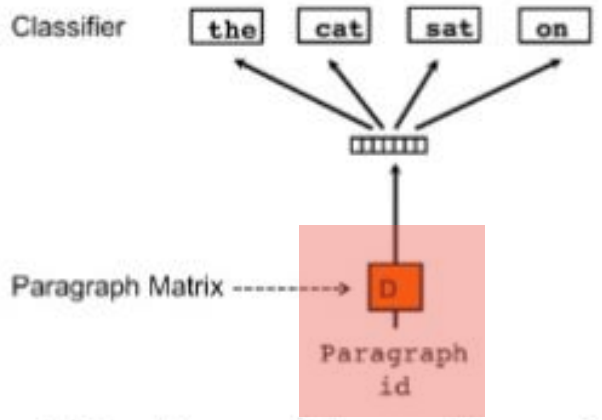
2nd try : 영어지문 데이터로 직접 word2vec 모델을 학습시켜서 문제풀이에 적용

Doc2vec algorithm

Word2vec이 확장된 임베딩



Paragraph vector with distributed memory (PV-DM)



Distributed bag of words version (PVDBOW)

PV-DM

이런 paragraph vector 와 앞의 단어들을 사용해서 다음에 나오는 단어를 유추.

window라는 정해진 사이즈의 단어들을 문맥정보(context)로 사용하며 맨 앞에서부터 한 단어씩 옆으로 이동하면서 훈련 데이터로 사용합니다.

PV-DBOW

이전 방식에서 나오는 context 단어들을 사용하지 않고 paragraph id 만 가지고 이 패러그래프에서 나오는 단어를 랜덤하게 예측하는 방식을 사용합니다.

input은 패러그래프 벡터이고 output은 패러그래프에서 random하게 뽑인 단어들입니다.

PICK ME

각 모델의 자세한 사용방법은 코드리뷰에서!

픽미는 **몇 등급일까?**

(우리 픽미는요..)

	(최고정답률기준) PICK ME 등급컷	
정답률	59%	4등급
*너그러운 정답률	73%	3등급..
*더 너그러운 정답률	84%	2등급…^^

*유사도가 제일 높은 2개/3개의 선지 중에 정답이 있을 확률

+ **Code** Review

잘한 점, 아쉬운 점

잘한 점, 아쉬운 점

잘한 점, 아쉬운 점