삼성전기 AI전문가 양성과정 - 프로젝트 실습 (비영상)

자연어처리를 위한 Tokenizer & Vocabulary

현청천

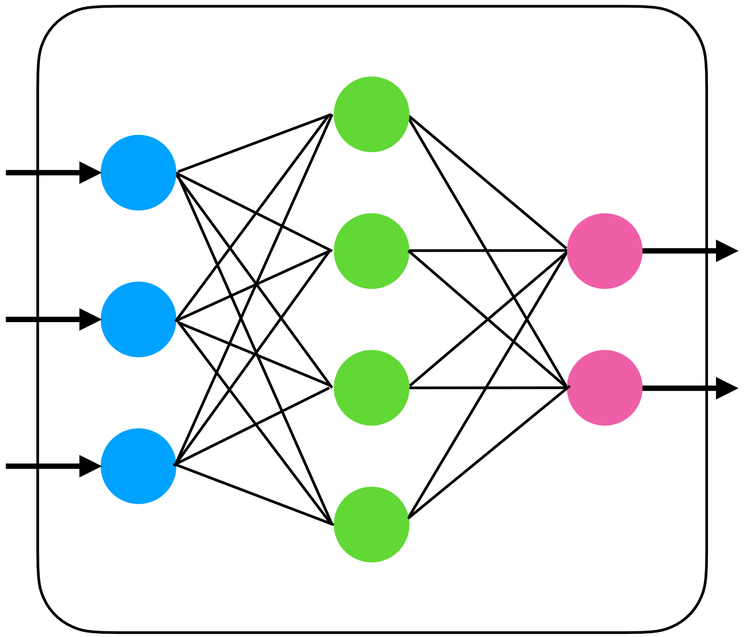
2022.02.28

What is Tokenizer

지미 카터

제임스 얼 "지미" 카터 주니어(, 1924년 10월 1일 ~)는 민주당 출신 미국 39번째 대통령 (1977년 ~ 1981년)이다.
지미 카터는 조지아주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다. 1953년 미국 해군 대위로 예편하였고 이후 땅콩·면화 등을 가꿔 많은 돈을 벌었다. 그의 별명이 "땅콩 농부" (Peanut Farmer)로 알려졌다.
1962년 조지아 주 상원 의원 선거에서 낙선하나 그 선거가 부정선거 였음을 입증하게 되어 당선되고, 1966년 조지아 주 지사 선거에 낙선하지만 1970년 조지아 주 지사를 역임했다. 대통령이 되기 전 조지아주 상원의원을 두번 연임했으며, 1971년부터 1975년까지 조지아 지사로 근무했다. 조지아 주지사로 지내면서, 미국에 사는 흑인 등용법을 내세웠다.
1976년 대통령 선거에 민주당 후보로 출마하여 도덕주의 정책으로 내세워, 포드를 누르고 당선되었다.
카터 대통령은 에너지 개발을 촉구했으나 공화당의 반대로 무산되었다.
카터는 이집트와 이스라엘을 조정하여, 캠프 데이비드에서 안와르 사다트 대통령과 메나헴 베긴 수상과 함께 중동 평화를 위한 캠프데이비드 협정을 체결했다.
그러나 이것은 공화당과 미국의 유대인 단체의 반발을 일으켰다. 1979년 백악관에서 양국 간의 평화조약으로 이끌어졌다. 또한 소련과 제2차 전략 무기 제한 협상에 조인했다.
카터는 1970년대 후반 당시 대한민국 등 인권 후진국의 국민들의 인권을 지키기 위해 노력했으며, 취임 이후 계속해서 도덕정치를 내세웠다.
그러나 주 이란 미국 대사관 인질 사건에서 인질 구출 실패를 이유로 1980년 대통령 선거에서 공화당의 로널드 레이건 후보에게 져 결국 재선에 실패했다. 또한 임기 말기에 터진 소련의 아프가니스탄 침공 사건으로 인해 1980년 하계 올림픽에 반공국가들의 보이콧을 내세웠다.
지미 카터는 대한민국과의 관계에서도 중요한 영향을 미쳤던 대통령 중 하나다. 인권 문제와 주한미군 철수 문제로 한때 한미 관계가 불편하기도 했다. 1978년 대한민국에 대한 북한의 위협에 대비해 한미연합사를 창설하면서, 1982년까지 3단계에 걸쳐 주한미군을 철수하기로 했다. 그러나 주한미군사령부와 정보기관·의회의 반대에 부딪혀 주한미군은 완전철수 대신 6,000명을 감축하는 데 그쳤다. 또한 박정희 정권의 인권 문제 등과의 논란으로 불협화음을 냈으나, 1979년 6월 하순, 대한민국을 방문하여 관계가 다소 회복되었다.
1979년 ~ 1980년 대한민국의 정치적 격변기 당시의 대통령이었던 그는 이에 대해 애매한 태도를 보였고, 이는 후에 대한민국 내에서 고조되는 반미 운동의 한 원인이 됐다. 10월 26일, 박정희 대통령이 김재규 중앙정보부장에 의해 살해된 것에 대해 그는 이 사건으로 큰 충격을 받았으며, 사이러스 밴스 국무장관을 조문사절로 파견했다. 12·12 군사 반란과 5.17 쿠데타에 대해 초기에는 강하게 비난했으나, 미국 정부가 신군부를 설득하는데, 한계가 있었고 결국 묵인하는 듯한 태도를 보이게 됐다.
퇴임 이후 민간 자원을 적극 활용한 비영리 기구인 카터 재단을 설립한 뒤 민주주의 실현을 위해 제 3세계의 선거 감시 활동 및 기니 벌레에 의한 드라쿤쿠르스 질병 방재를 위해 힘썼다. 미국의 빈곤층 지원 활동, 사랑의 집짓기 운동, 국제 분쟁 중재 등의 활동도 했다.
카터는 카터 행정부 이후 미국이 북핵 위기, 코소보 전쟁, 이라크 전쟁과 같이 미국이 군사적 행동을 최후로 선택하는 전통적 사고를 버리고 군사적 행동을 선행하는 행위에 대해 깊은 유감을 표시 하며 미국의 군사적 활동에 강한 반대 입장을 보이고 있다.
특히 국제 분쟁 조정을 위해 북한의 김일성, 아이티의 세드라스 장군, 팔레인스타인의 하마스, 보스니아의 세르비아계 정권 같이 미국 정부에 대해 협상을 거부하면서 사태의 위기를 초래한 인물 및 단체를 직접 만나 분쟁의 원인을 근본적으로 해결하기 위해 힘썼다. 이 과정에서 미국 행정부와 갈등을 보이기도 했지만, 전직 대통령의 권한과 재야 유명 인사들의 활약으로 해결해 나갔다.
1978년에 채결된 캠프데이비드 협정의 이행이 지지부진 하자 중동 분쟁 분제를 해결하기 위해 1993년 퇴임 후 직접 이스라엘과 팔레인스타인의 오슬로 협정을 이끌어 내는 데도 성공했다.

?



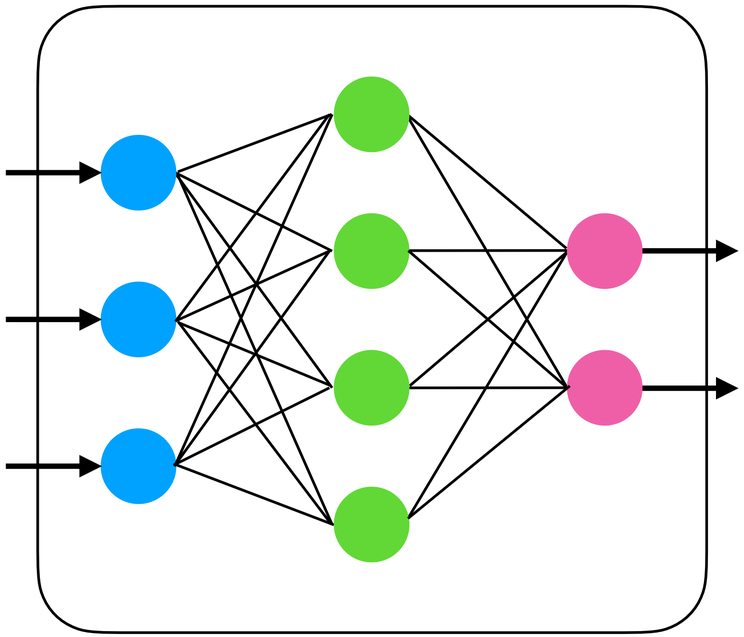
입력문장을 일정한 단위로 분할

Char Tokenizer

지미 카터

제임스 얼 "지미" 카터 주니어(, 1924년 10월 1일 ~)는 민주당 출신 미국 39번째 대통령 (1977년 ~ 1981년)이다.
지미 카터는 조지아주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다. 1953년 미국 해군 대위로 예편하였고 이후 땅콩·면화 등을 가꿔 많은 돈을 벌었다. 그의 별명이 "땅콩 농부" (Peanut Farmer)로 알려졌다.
1962년 조지아 주 상원 의원 선거에서 낙선하나 그 선거가 부정선거 였음을 입증하게 되어 당선되고, 1966년 조지아 주 지사 선거에 낙선하지만 1970년 조지아 주 지사를 역임했다. 대통령이 되기 전 조지아주 상원의원을 두번 연임했으며, 1971년부터 1975년까지 조지아 지사로 근무했다. 조지아 주지사로 지내면서, 미국에 사는 흑인 등용법을 내세웠다.
1976년 대통령 선거에 민주당 후보로 출마하여 도덕주의 정책으로 내세워, 포드를 누르고 당선되었다.
카터 대통령은 에너지 개발을 촉구했으나 공화당의 반대로 무산되었다.
카터는 이집트와 이스라엘을 조정하여, 캠프 데이비드에서 안와르 사다트 대통령과 메나헴 베긴 수상과 함께 중동 평화를 위한 캠프데이비드 협정을 체결했다.
그러나 이것은 공화당과 미국의 유대인 단체의 반발을 일으켰다. 1979년 백악관에서 양국 간의 평화조약으로 이끌어졌다. 또한 소련과 제2차 전략 무기 제한 협상에 조인했다.
카터는 1970년대 후반 당시 대한민국 등 인권 후진국의 국민들의 인권을 지키기 위해 노력했으며, 취임 이후 계속해서 도덕정치를 내세웠다.
그러나 주 이란 미국 대사관 인질 사건에서 인질 구출 실패를 이유로 1980년 대통령 선거에서 공화당의 로널드 레이건 후보에게 져 결국 재선에 실패했다. 또한 임기 말기에 터진 소련의 아프가니스탄 침공 사건으로 인해 1980년 하계 올림픽에 반공국가들의 보이콧을 내세웠다.
…

지	미		카	터										
제	임	스		얼		“	지	미	”	카	터		주	니	.....
지	미		카	터	는		조	지	아	주		섬	터		.....
⋮															



글자 단위로 분할

Char Tokenizer

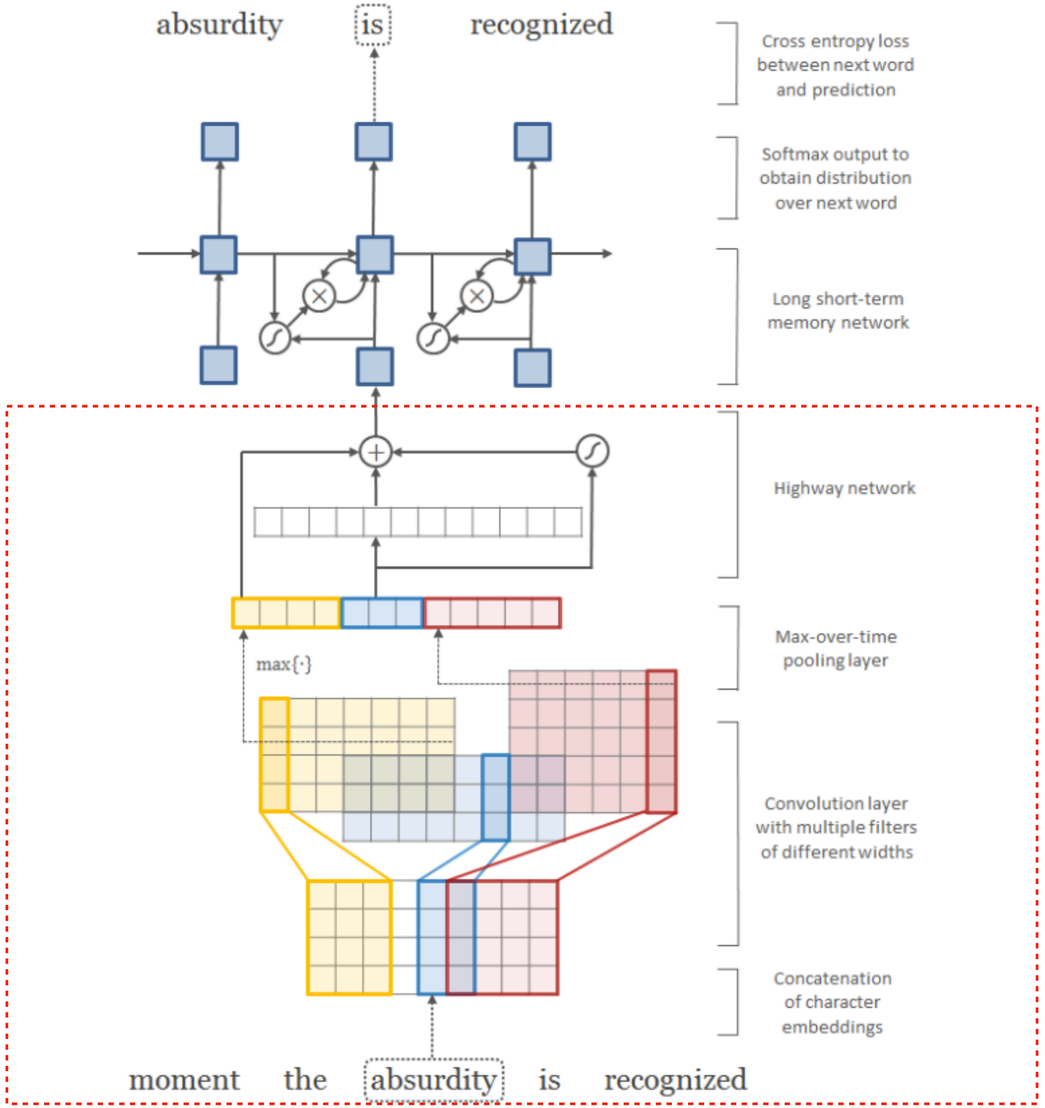
- 장점
  - 모든 문장을 적은 수의 Vocabulary로 표현할 수 있음
  - Vocabulary에 글자가 없어서 '[UNK]'로 표현해야 하는 OOV(Out of Vocabulary) 문제가 발생할 가능성이 낮음

- 단점
  - 글자 단위로 분할하기 때문에 token 수가 많아짐

    token 수가 많으면 연산이 많아지고 학습속도가 늦어짐
  - 각 글자 하나하나를 벡터로 표현할 경우 단어의 의미를 표현한다고 할 수 없음

    예) $['b','o','o','k'] \neq 'book'$
  - 이런 문제를 해결하기 위해서 char 기반의 Neural Network은 많은 layer를 필요로 함

Char Tokenizer



char에서 단어 의미 추출

Image source: Character-Aware Neural Language Models(https://arxi

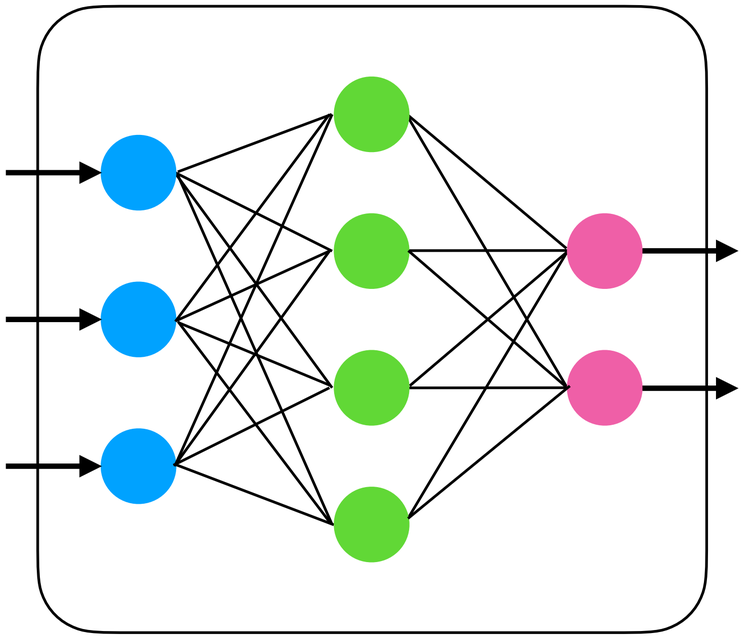
Word Tokenizer

지미 카터
제임스 얼 "지미" 카터 주니어(, 1924년 10월 1일 ~)는 민주당 출신 미국 39번째 대통령 (1977년 ~ 1981년)이다.
지미 카터는 조지아주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다. 1953년 미국 해
1962년 조지아 주 상원 의원 선거에서 낙선하나 그 선거가 부정선거 였음을 입증하게 되어 당선되고, 1966년 조지아 주 지사 선거에 낙선하지만 1970년 조지아 주 지사를 역임했
1976년 대통령 선거에 민주당 후보로 출마하여 도덕주의 정책으로 내세워, 포드를 누르고 당선되었다.
카터 대통령은 에너지 개발을 촉구했으나 공화당의 반대로 무산되었다.
카터는 이집트와 이스라엘을 조정하여, 캠프 데이비드에서 안와르 사다트 대통령과 메나헴 베긴 수상과 함께 중동 평화를 위한 캠프데이비드 협정을 체결했다.
그러나 이것은 공화당과 미국의 유대인 단체의 반발을 일으켰다. 1979년 백악관에서 양국 간의 평화조약으로 이끌어졌다. 또한 소련과 제2차 전략 무기 제한 협상에 조인했다.
카터는 1970년대 후반 당시 대한민국 등 인권 후진국의 국민들의 인권을 지키기 위해 노력했으며, 취임 이후 계속해서 도덕정치를 내세웠다.
그러나 주 이란 미국 대사관 인질 사건에서 인질 구출 실패를 이유로 1980년 대통령 선거에서 공화당의 로널드 레이건 후보에게 저 결국 재선에 실패했다. 또한 임기 말기에 터진
지미 카터는 대한민국과의 관계에서도 중요한 영향을 미쳤던 대통령 중 하나다. 인권 문제와 주한미군 철수 문제로 한때 한미 관계가 불편하기도 했다. 1978년 대한민국에 대한 북한
1979년 ~ 1980년 대한민국의 정치적 격변기 당시의 대통령이었던 그는 이에 대해 애매한 태도를 보였고, 이는 후에 대한민국 내에서 고조되는 반미 운동의 한 원인이 됐다. 10월
퇴임 이후 민간 자원을 적극 활용한 비영리 기구인 카터 재단을 설립한 뒤 민주주의 실현을 위해 제 3세계의 선거 감시 활동 및 기니 벌레에 의한 드라쿤쿠르스 질병 방재를 위해 힘
카터는 카터 행정부 이후 미국이 북핵 위기, 코소보 전쟁, 이라크 전쟁과 같이 미국이 군사적 행동을 최후로 선택하는 전통적 사고를 버리고 군사적 행동을 선행하는 행위에 대해 깊
특히 국제 분쟁 조정을 위해 북한의 김일성, 아이티의 세드라스 장군, 팔레인스타인의 하마스, 보스니아의 세르비아계 정권 같이 미국 정부에 대해 협상을 거부하면서 사태의 위기를
1978년에 체결된 캠프데이비드 협정의 이행이 지지부진 하자 중동 분쟁 분제를 해결하기 위해 1993년 퇴임 후 직접 이스라엘과 팔레인스타인의 오슬로 협정을 이끌어 내는 데도 성
1993년 1차 북핵 위기 당시 북한에 대한 미국의 군사적 행동이 임박했으나, 미국 전직 대통령으로는 처음으로 북한을 방문하고 미국과 북 양국의 중재에 큰 기여를 해 위기를 해결
미국의 관타나모 수용소 문제, 세계의 인권문제에서도 관심이 깊어 유엔에 유엔인권고등판무관의 제도를 시행하도록 노력하여 독재자들의 인권 유린에 대해 제약을 하고, 국제형사재판

지미 | 카터

제임스 | 월 | “ | 지미 | ” | 카터 | 주니어 | (| , | 1924년 | 1월 | 1일 | ~ |) | 는 .....

지미 | 카터는 | 조지아주 | 섬터 | 카운티 | 플레인스 | 마을에서 | 태어났다 .....



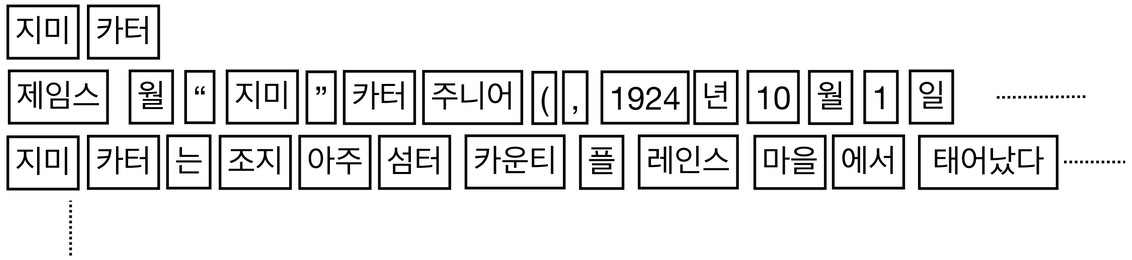
띄어쓰기 단위로 분할하는 방법

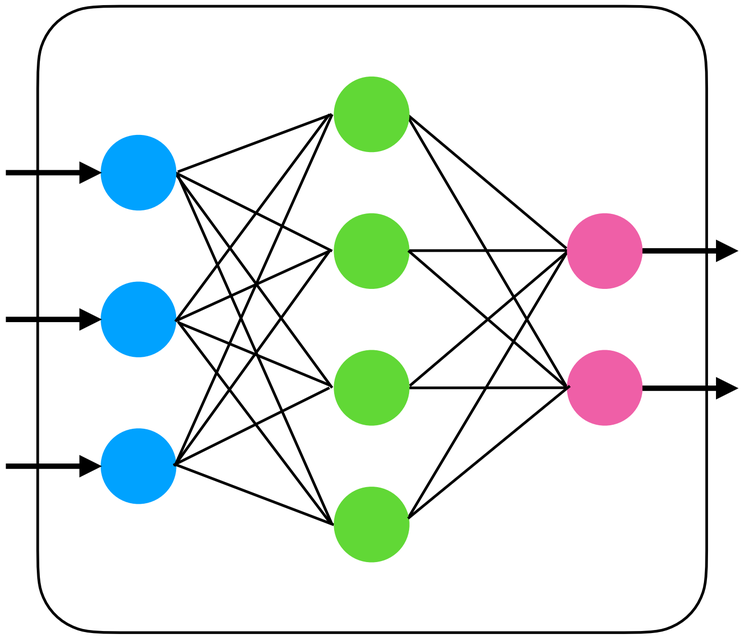
Word Tokenizer

- 장점
  - 띄어쓰기 단위로 분할하기 때문에 token 수가 적음
- 단점
  - 어미변화로 인해 유사 단어들이 많아짐
    - ‘책’, ‘책을’, ‘책에다’, ‘책에서’, ‘책이’ 등
    - ‘play’, ‘plays’, ‘playing’, ‘played’ 등
  - 어미변화로 인한 단어를 표현하는 벡터들이 다른 의미를 가질 수 있음
  - Vocabulary 개수가 매우 많아짐
    - 메모리 사용량 증가
    - 연산량 또한 증가

Morph Tokenizer

지미 카터
제임스 얼 "지미" 카터 주니어(, 1924년 10월 1일 ~)는 민주당 출신 미국 39번째 대통령 (1977년 ~ 1981년)이다.
지미 카터는 조지아주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다. 1953년 미국 해
1962년 조지아 주 상원 의원 선거에서 낙선하나 그 선거가 부정선거 였음을 입증하게 되어 당선되고, 1966년 조지아 주 지사 선거에 낙선하지만 1970년 조지아 주 지사를 역임했
1976년 대통령 선거에 민주당 후보로 출마하여 도덕주의 정책으로 내세워, 포드를 누르고 당선되었다.
카터 대통령은 에너지 개발을 촉구했으나 공화당의 반대로 무산되었다.
카터는 이집트와 이스라엘을 조정하여, 캠프 데이비드에서 안와르 사다트 대통령과 메나헴 베긴 수상과 함께 중동 평화를 위한 캠프데이비드 협정을 체결했다.
그러나 이것은 공화당과 미국의 유대인 단체의 반발을 일으켰다. 1979년 백악관에서 양국 간의 평화조약으로 이끌어졌다. 또한 소련과 제2차 전략 무기 제한 협상에 조인했다.
카터는 1970년대 후반 당시 대한민국 등 인권 후진국의 국민들의 인권을 지키기 위해 노력했으며, 취임 이후 계속해서 도덕정치를 내세웠다.
그러나 주 이란 미국 대사관 인질 사건에서 인질 구출 실패를 이유로 1980년 대통령 선거에서 공화당의 로널드 레이건 후보에게 저 결국 재선에 실패했다. 또한 임기 말기에 터진
지미 카터는 대한민국과의 관계에서도 중요한 영향을 미쳤던 대통령 중 하나다. 인권 문제와 주한미군 철수 문제로 한때 한미 관계가 불편하기도 했다. 1978년 대한민국에 대한 북한
1979년 ~ 1980년 대한민국의 정치적 격변기 당시의 대통령이었던 그는 이에 대해 애매한 태도를 보였고, 이는 후에 대한민국 내에서 고조되는 반미 운동의 한 원인이 됐다. 10월
퇴임 이후 민간 자원을 적극 활용한 비영리 기구인 카터 재단을 설립한 뒤 민주주의 실현을 위해 제 3세계의 선거 감시 활동 및 기니 벌레에 의한 드라쿤쿠르스 질병 방재를 위해 힘
카터는 카터 행정부 이후 미국이 북핵 위기, 코소보 전쟁, 이라크 전쟁과 같이 미국이 군사적 행동을 최후로 선택하는 전통적 사고를 버리고 군사적 행동을 선행하는 행위에 대해 깊
특히 국제 분쟁 조정을 위해 북한의 김일성, 아이티의 세드라스 장군, 팔레인스타인의 하마스, 보스니아의 세르비아계 정권 같이 미국 정부에 대해 협상을 거부하면서 사태의 위기를
1978년에 체결된 캠프데이비드 협정의 이행이 지지부진 하자 중동 분쟁 분제를 해결하기 위해 1993년 퇴임 후 직접 이스라엘과 팔레인스타인의 오슬로 협정을 이끌어 내는 데도 성
1993년 1차 북핵 위기 당시 북한에 대한 미국의 군사적 행동이 임박했으나, 미국 전직 대통령으로는 처음으로 북한을 방문하고 미국과 북 양국의 중재에 큰 기여를 해 위기를 해결
미국의 관타나모 수용소 문제, 세계의 인권문제에서도 관심이 깊어 유엔에 유엔인권고등판무관의 제도를 시행하도록 노력하여 독재자들의 인권 유린에 대해 제약을 하고, 국제형사재판





형태소 단위로 분할하는 방법

Morph Tokenizer

- 장점
  - 형태소 단위로 분할하기 때문에 token 들이 적당한 의미를 가짐
  - Word tokenizer와 char tokenizer의 중간쯤의 token 개수를 가짐
- 단점
  - 형태소 분석기의 발전 속도가 언어의 발전 속도에 비해서 느림
  - 형태소 분석기들이 어느 정도의 오류를 가지고 있음
  - Word에 비해서는 vocabulary 개수가 많이 줄어들지만 여전히 많음

BPE(Byte Pair Encoding)

- 압축 알고리즘을 이용하여 띄어쓰기 단위의 단어를 subword 단위로 분할
  - 빈도수가 가장 많은 subword pair를 새로운 subword로 변환

Bi-gram 발생빈도 기준으로 분할하는 방법

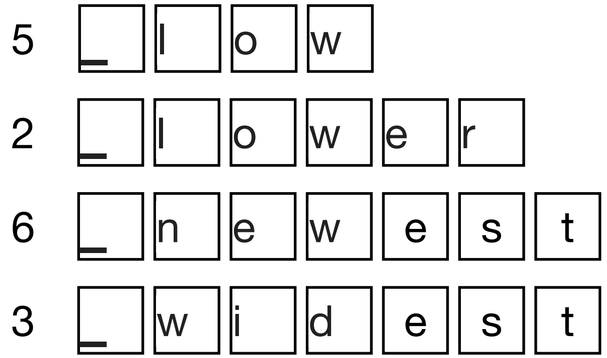
BPE(Byte Pair Encoding)

low	lower	newest	widest
low	lower	newest	widest
low		newest	widest
low		newest	
low		newest	
		newest	

Count words

low (5)　　lower (2)　　newest (6)　　widest (3)

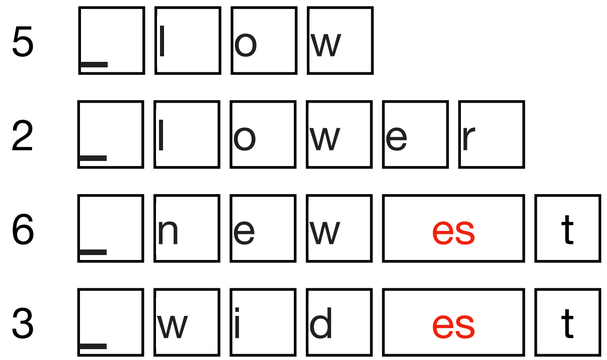
BPE(Byte Pair Encoding)



Vocabulary (11)

_, l, o, w, e, r, n, s, t, i, d

BPE(Byte Pair Encoding)

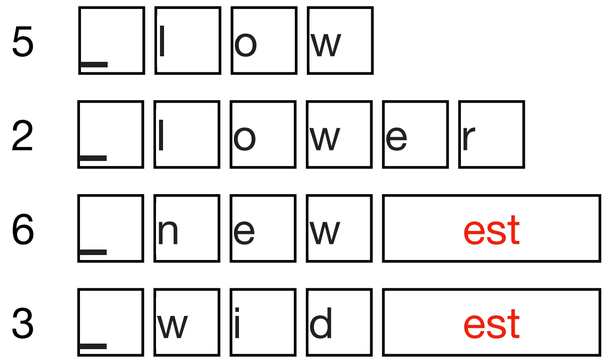


Vocabulary (12)

_, l, o, w, e, r, n, s, t, i, d, es

(e, s) most frequent (9) bi-gram

BPE(Byte Pair Encoding)

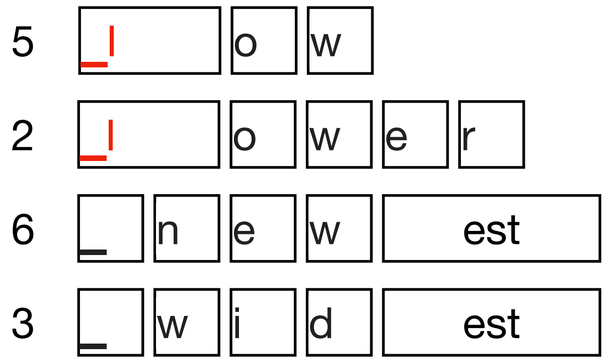


Vocabulary (13)

_, l, o, w, e, r, n, s, t, i, d, es, est

(es, t) most frequent (9) bi-gram

BPE(Byte Pair Encoding)

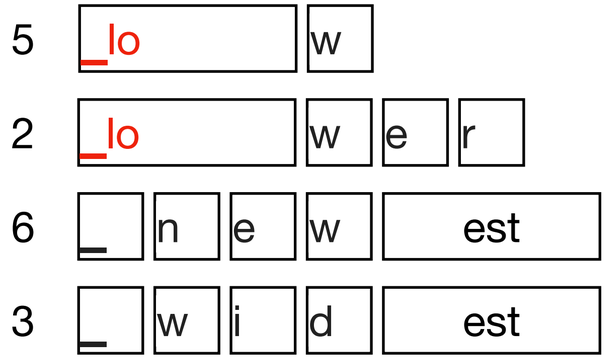


Vocabulary (14)

_, l, o, w, e, r, n, s, t, i, d, es, est, _l

(_, l) most frequent (7) bi-gram

BPE(Byte Pair Encoding)



Vocabulary (15)

_, l, o, w, e, r, n, s, t, i, d, es, est, _l, _lo

(_l, o) most frequent (7) bi-gram

BPE(Byte Pair Encoding)

- 장점
  - 말뭉치가 있다면 비교적 간단하게 만들 수 있음
  - 적은 수의 vocabulary를 가지고 OOV(Out of Vocabulary)를 최소화
- 단점
  - Subword의 분할이 의미 기준이 아닐 수 있음
    - ‘수원에’라는 문장을 분할할 때 [‘▁수원’, ‘에’]가 아닌 [‘▁수’, ‘원에’]로 분할 될 수 있음
    - ‘대한민국을’, ‘대한민국은’, ‘대한민국으로’ 등의 빈도수가 많은 단어들은
      [‘대한민국’, ‘을’, ‘은’, ‘으로’] 형태가 아닌
      [‘대한민국을’, ‘대한민국은’, ‘대한민국으로’] 그대로 분류되기도 함

Sentencepiece

- Google에서 제공하는 Tokenizer tool
  - https://arxiv.org/abs/1808.06226
  - https://github.com/google/sentencepiece
  - Char, word, BPE, Unigram 등 다양한 Tokenizer 제공
- 예제
  - https://github.com/google/sentencepiece/blob/master/python/sentencepiece_python_module_example.ipynb
- 설치방법
  - pip install sentencepiece

Sentencepiece

함수	설명
encode_as_pieces	문자열을 token으로 분할하는 함수
decode_pieces	token을 문자열로 복원하는 함수
encode_as_ids	문자열을 숫자로 분할하는 함수
decode_ids	숫자를 문자열로 복원하는 함수
piece_to_id	token을 숫자로 변경하는 함수
id_to_piece	숫자를 token으로 변경하는 함수

Sentencepiece with Morph

- 형태소 분석기와 sentencepiece를 동시에 사용
  - 형태소 분석기로 문장을 우선 tokenize
  - sentencepiece를 이용해 다시한번 tokenize
- 성능
  - sentencepiece를 그냥 사용하는 것보다 성능이 좋다고 알려져 있음

감사합니다.