




**Hello? NLP~**

11회차 세미나  
발표자 김민주



08) 사전 훈련된 워드 임베딩

---

09) 엘모(Embeddings from Language Model, ELMo)

---

10) 임베딩 벡터의 시각화

---

11) 문서 벡터를 이용한 추천 시스템

---

12) 워드 임베딩의 평균

---

## 08) 사전 훈련된 워드 임베딩



### 1. 케라스 임베딩 층

Embedding() : 인공 신경망 구조 관점에서 임베딩 층(embedding layer)을 구현

1) 임베딩 층은 룩업 테이블이다.

Word → Integer → lookup Table → Embedding vector



```
# 아래의 각 인자는 저자가 임의로 선택한 숫자들이며 의미있는 선정 기준이 아님.  
v = Embedding(20000, 128, input_length=500)  
# vocab_size = 20000  
# output_dim = 128  
# input_length = 500
```

vocab\_size : 텍스트 데이터의 전체 단어 집합의 크기

output\_dim : 워드 임베딩 후의 임베딩 벡터의 차원

input\_length : 입력 시퀀스의 길이

## 08) 사전 훈련된 워드 임베딩

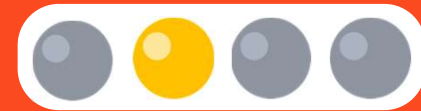


### 2. 사전 훈련된 워드 임베딩 사용하기(실습)

GloVe 다운로드 링크 : <http://nlp.stanford.edu/data/glove.6B.zip>

Word2Vec 다운로드 링크 : <https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM>

## 09) 엘모(Embeddings from Language Model, ELMo)



### ELMo, Embeddings from Language Model

- 사전 훈련된 언어 모델(Pre-trained language model) 사용

Bank	Bank Account(은행 계좌)	[0.2 0.8 -1.2]
[0.2 0.8 -1.2]	River Bank (강둑)	[0.2 0.8 -1.2]

같은 표기의 단어라도 문맥에 따라서 다르게 워드 임베딩을 할 수 있으면 자연어 처리의 성능이 더 올라가지 않을까?

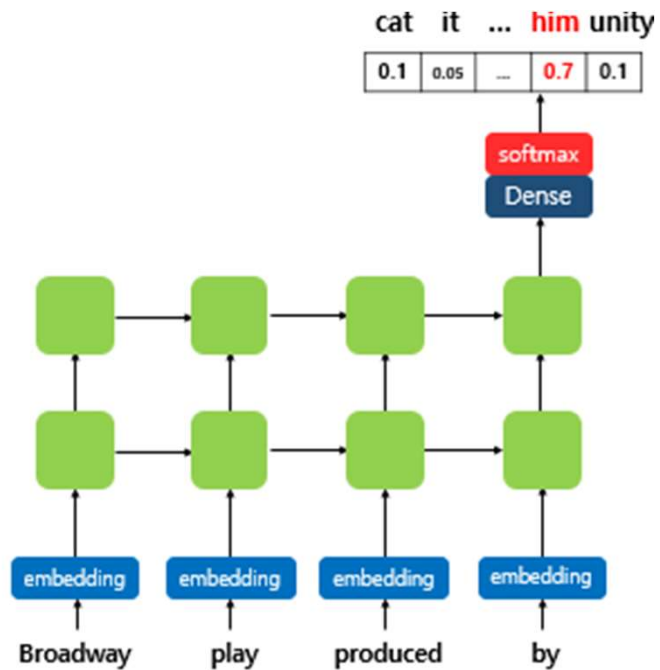
문맥을 반영한 워드 임베딩(Contextualized Word Embedding)

# 09) 엘모(Embeddings from Language Model, ELMo)



## 2. biLM(Bidirectional Language Model)의 사전 훈련

은닉층이 2개인 일반적인 단방향 RNN 언어 모델의 언어 모델링



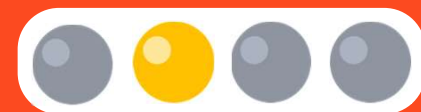
biLM

양쪽 방향의 언어 모델을 둘 다 활용

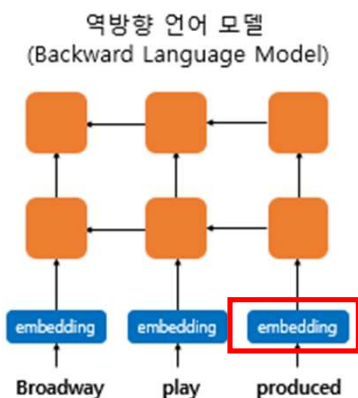
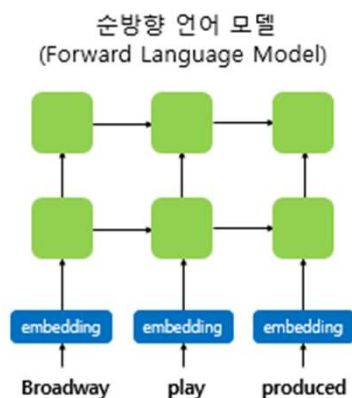
기본적으로 다층 구조(Multi-layer)를 전제

➔ 은닉층이 최소 2개 이상이다!

## 09) 엘모(Embeddings from Language Model, ELMo)



### biLM(Bidirectional Language Model)

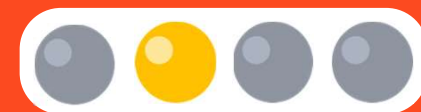


char CNN

글자(character)단위로 계산  
문맥과 상관없이 단어의 연관성을 찾아낼 수 있음  
OOV에도 견고함

\* Out-Of-Vocabulary(단어 집합에 없는 단어)

# 09) 엘모(Embeddings from Language Model, ELMo)



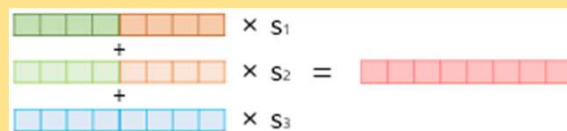
## 2. biLM(Bidirectional Language Model)의 활용



1) 각 층의 출력값을 연결(concatenate)한다.



2) 각 층의 출력값 별로 가중치를 준다.

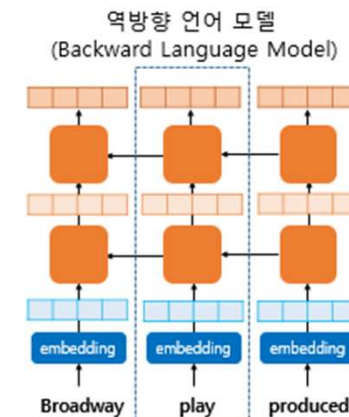
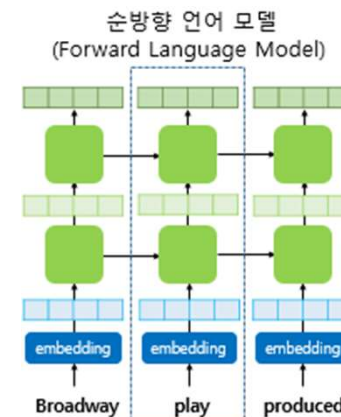


3) 각 층의 출력값을 모두 더한다.



ELMo 표현(representation)

4) 벡터의 크기를 결정하는 스칼라 매개변수를 곱한다.

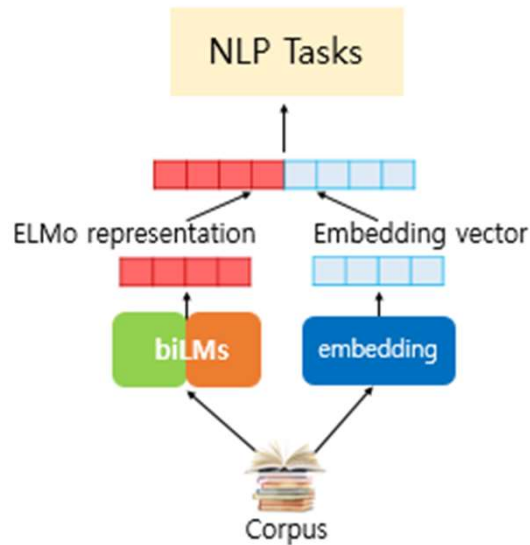




## 09) 엘모(Embeddings from Language Model, ELMo)

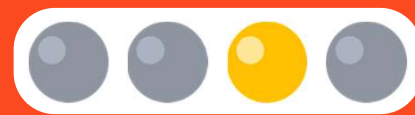


### 2. biLM(Bidirectional Language Model)의 활용



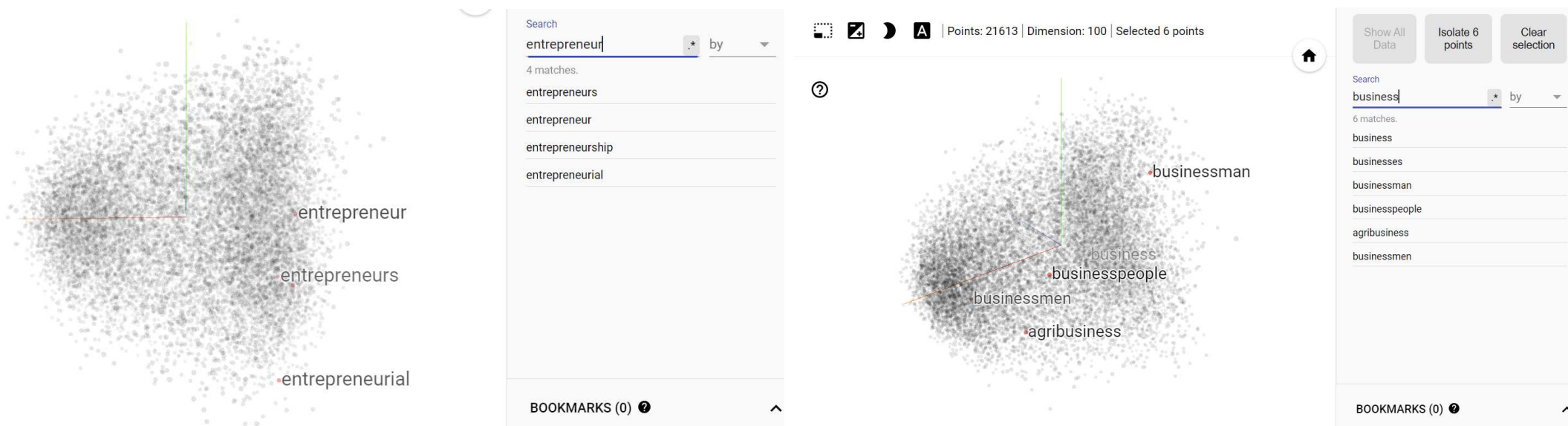
ELMo 표현을 사용해서 스팸 메일 분류하기 (실습)

# 10) 임베딩 벡터의 시각화

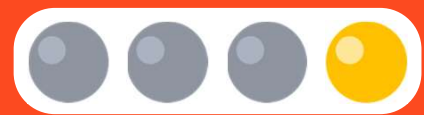


구글이 지원하는 시각화 툴, 임베딩 프로젝트

더 알고 싶다면? → <https://arxiv.org/pdf/1611.05469v1.pdf>



# 11) 문서 벡터를 이용한 추천 시스템



## 유사도 비교

1. Doc2Vec / Sent2Vec : 문서 벡터로 변환

### Today's Topic

2. 문서에 존재하는 단어 벡터들의 평균을 구하는 것

## 12) 워드 임베딩의 평균(실습)



깃헙에서 실습결과를 확인하세요^^



*Thank you for listening*