PyCon Korea 2019

딥러닝 NLP 손쉽게 따라해보기
- GluonNLP-

Embedding

Contents

Word Embedding

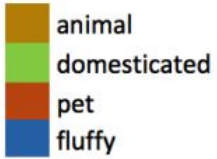
- Skip-Gram
- fastText
- GloVe

Word Embedding

- Word Embedding
  - 비정형 Text를 숫자로 바꾸어 줌으로써 사람의 언어를 컴퓨터로 번역하는 행위

Dimensions

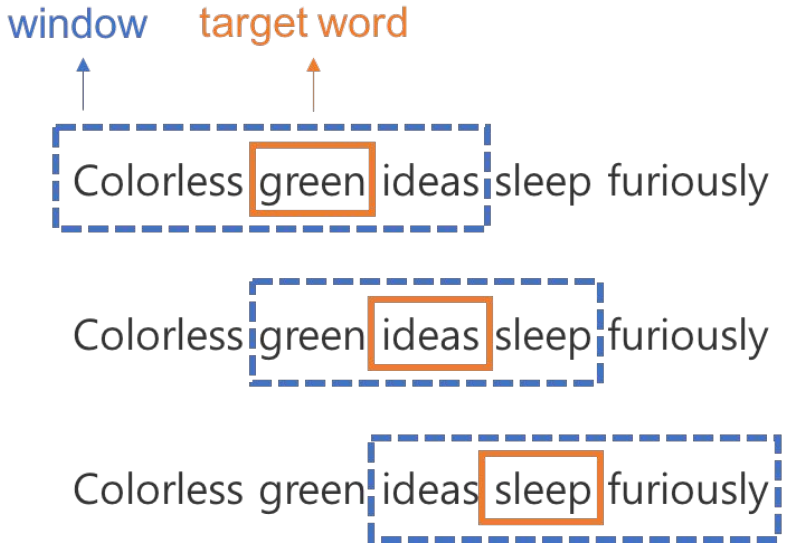
Word vectors				
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37



https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf

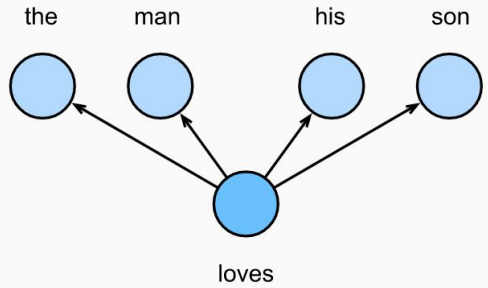
Word2Vec

- Approach
  - 주변 단어를 통해서 해당 단어를 유추



$$\mathbb{P}(\text{"the"}, \text{"man"}, \text{"his"}, \text{"son"} \mid \text{"loves"}).$$

$$\mathbb{P}(\text{"the"} \mid \text{"loves"}) \cdot \mathbb{P}(\text{"man"} \mid \text{"loves"}) \cdot \mathbb{P}(\text{"his"} \mid \text{"loves"}) \cdot \mathbb{P}(\text{"son"} \mid \text{"loves"}$$



https://dreamgonfly.github.io/machine/learning./natural/language/processing/2017/08/16/word2vec_explained.html
https://www.d2l.ai/chapter_natural-language-processing/word2vec.html#the-skip-gram-model

fastText

Out of Vector 를 해결하기 위한 방법으로 고안된 방법

Approach

- 학습 단위를 문자 단위의 n-gram으로 학습하고 word의 embedding을 sub-word 의 조합으로 생성함

"<where>" ⇨ "<wh", "whe", "her", "ere", "re>",

n=3

$$\mathbf{u}_w = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g.$$

https://www.d2l.ai/chapter_natural-language-processing/fasttext.html

GloVe

- 각 단어의 동시 등장 확률(**Co-occurrence Probability**) 을 기반으로 한 embedding 모형을 구현함

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

loss function

$$\sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} h(x_{ij}) \left(\mathbf{u}_j^\top \mathbf{v}_i + b_i + c_j - \log\, x_{ij} \right)^2 .$$

https://www.d2l.ai/chapter_natural-language-processing/glove.html

fastText 기준 실습

- Pre-Trained fastText 활용
  - 전세계 294개 언어로 된 pre-trained 모델을 제공함
  - 한국어의 경우 wiki.ko 기준으로 활용이 가능함

github에 PPT 자료 및 실습 코드 활용

https://github.com/seujung/gluonnlp_tutorial.git

END OF DOCUMENT