1 0 주 교 육

ToBig's 9기 안상준

Nature Language Process Basic

Contents

Unit 01 | NLP Oveview

Unit 02 | Process

Unit 03 | Tokenizing

Unit 04 | Embedding

Unit 05 | Similarity

Unit 01 | NLP Overview

투 무빅스

<리뷰분석을 통한 영화 손익분기점 예측>

Big Voice

우리 팀 서기는 내 컴퓨터! 화자인식과
음성인식을 이용한 회의 Dialogue

E2A

Extract 요약을 기반으로 한
Abstract 요약 생성

- 이수빈, 이인재, 조양규, 최서현, 최희정 -

투빅스 5회 컨퍼런스

트랜스퍼머

곽대훈 김민정 류호성 박현진 최수정

Unit 01 | NLP Overview

To Movies : TF-IDF, Word2Vec, Counting

Big Voice : MFCC Vector, Cosine Window, Fourier Transform, Mel Filter, log scale

E2A : TextRank, LexRank, SeqToSeq, Tokenize

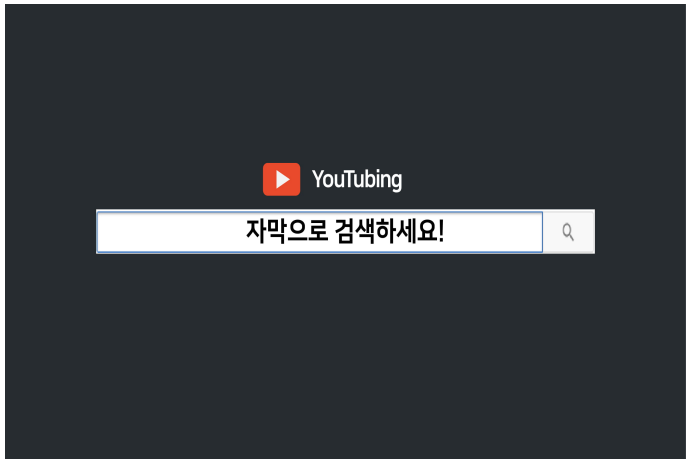
Replace : LDA, LexRank, Wod2Vec, TF-IDF

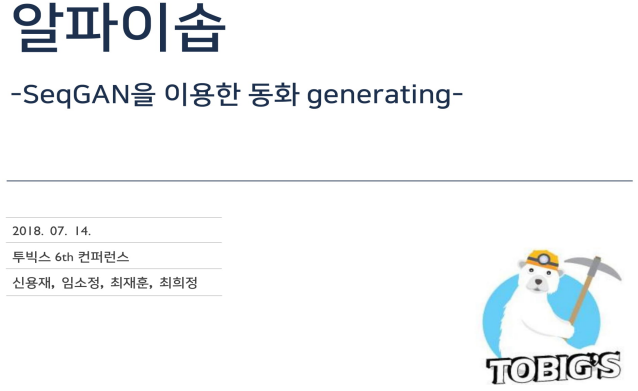
Unit 01 | NLP Overview

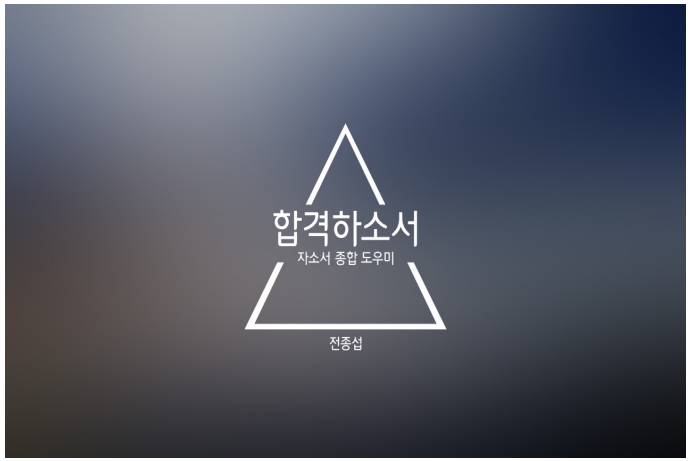
- To Movies : TF-IDF, Word2Vec, Counting
- Big Voice : MFCC Vectors, Cosine Window, Fourier Transform, Mel Filter, log scale
- E2A : TextRank, LexRank, SeqToSeq, Tokenize
- Replace : LDA, LexRank, Wod2Vec, TF-IDF

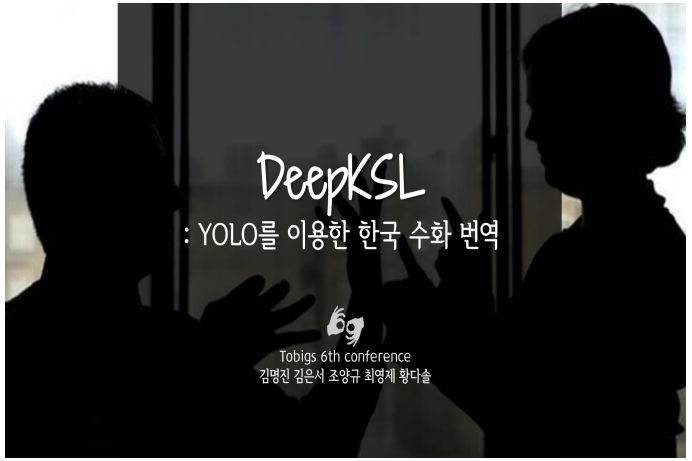
무슨 말이지?

Unit 01 | NLP Overview









Unit 01 | NLP Overview

Youtubing : Tokenizing, Embedding, Similarity

TextRick : Tokenizing, Word2Vec, LDA, LSA, TF-IDF

Alpha Aesop : GAN, Seq, Token, FastText, Bi-Gram

합격하소서 : Pos Tag, Word2Vevc, CBOW, LSTM

Unit 01 | NLP Overview

Youtubing : Tokenizing, Embedding, Similarity

TextRick : Tokenizing, Word2Vec, LDA, LSA, TF-IDF

Alpha Aesop : GAN, Seq, Token, FastText, Bi-Gram

합격하소서 : Pos Tag, Word2Vevc, CBOW, LSTM

무슨 말이지?

같이 배워봅시다!
NLP의 기초를 !

Contents

Unit 01 | NLP Oveview

Unit 02 | Process

Unit 03 | Tokenizing & TDM

Unit 04 | Embedding

Unit 05 | Similarity

Unit 02 | Process

NLP는 대체적으로 어떠한 과정을 거칠까요?

Unit 02 | Process

Crawling

Tokenizing

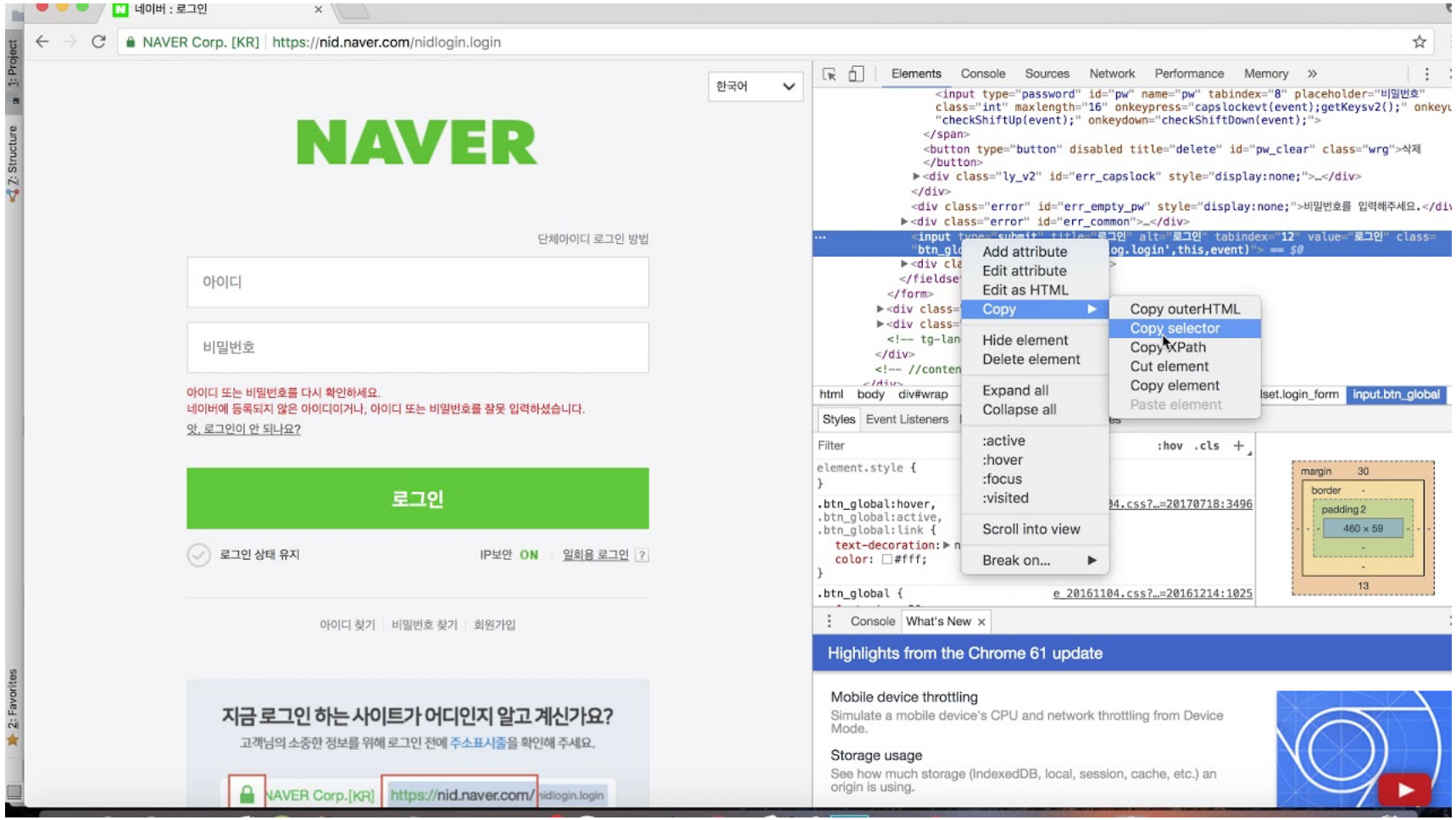
Embedding

Similarity

Network

Unit 02 | Process

Crawling



Tokenizing

투빅스는 과제지옥입니다.

⇩

'투빅스', '는', '과제', '지옥', '입니', '다.'

Embedding

'투빅스', '는', '과제', '지옥', '입니', '다.'

⇩

'투빅스' : 0.1234, '는' : 0.5678, '과제' : 0.9012,
'지옥' : 0.3456, '입니' : 0.7890, '다.' : 0.1234

Similarity

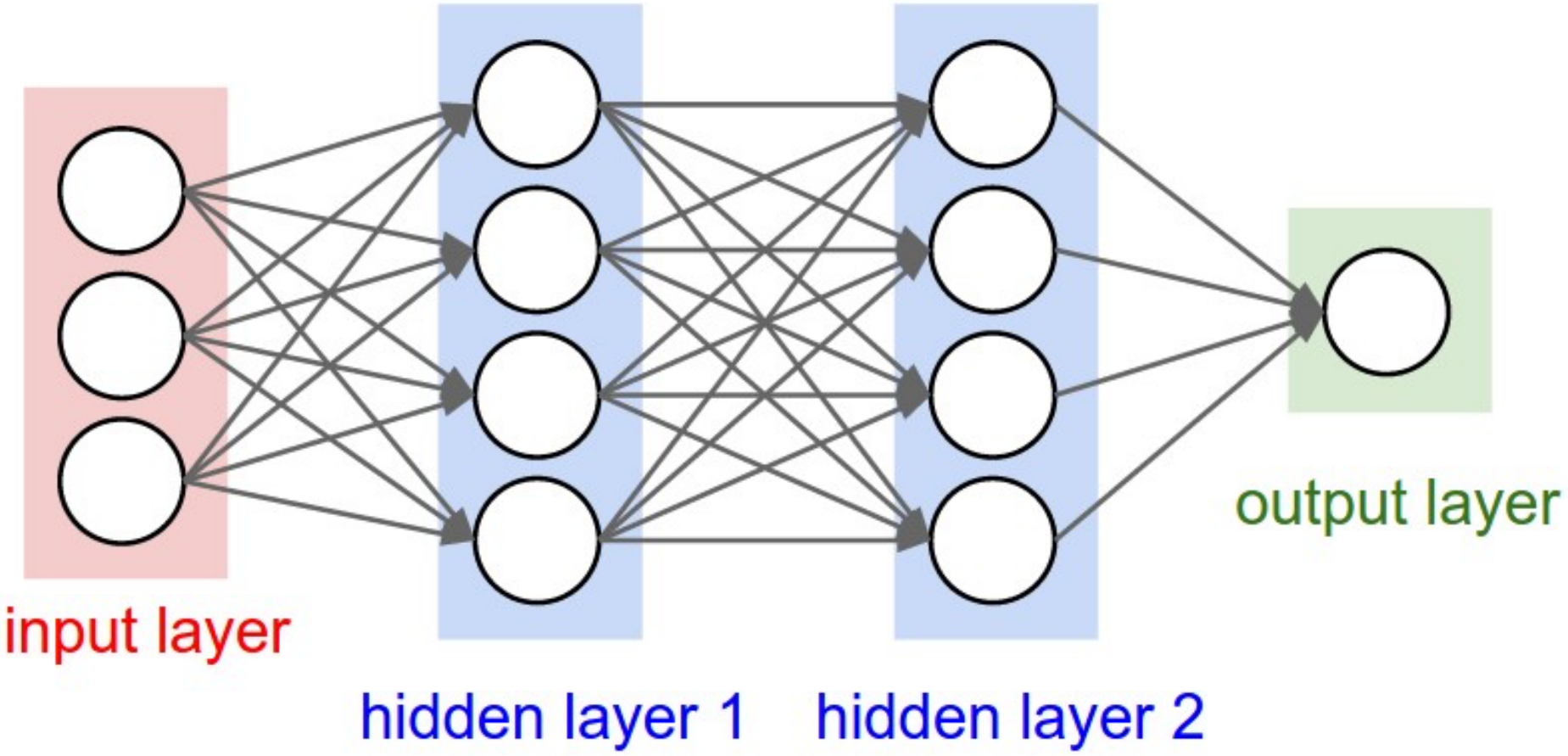
'투빅스' : 0.1234, '는' : 0.5678, '과제' : 0.9012,
'지옥' : 0.3456, '입니' : 0.7890, '다.' : 0.1237

⇩

'투빅스' : 0.1234, '다.' : 0.1237

이 두 단어는 유사하다고 판단할 수 있다.

Network



Contents

Unit 01 | NLP Oveview

Unit 02 | Process

Unit 03 | Tokenizing

Unit 04 | Embedding

Unit 05 | Similarity

Tokenizing

투빅스는 과제지옥입니다.

⇩

'투빅스', '는', '과제', '지옥', '입니', '다.'

Tokenizing

투빅스는 과제지옥입니다

특정 기준에 의해서

Sentence -> Token

'투빅스', '는', '과제', '지옥', '입니', '다.'

Unit 03 | Tokenizing

English

NLTK

Korean

KKMA

Twitter

Komoran

Hannanum

Mecab

Unit 03 | Tokenizing

아버지가방에들어가신다.

Hannanum	Kkma	Komoran	Mecab	Twitter
아버지가방에들어가 / N	아버지 / NNG	아버지가방에들어가신다 / NNP	아버지 / NNG	아버지 / Noun
이 / J	가방 / NNG		가 / JKS	가방 / Noun
시ㄴ다 / E	에 / JKM		방 / NNG	에 / Josa
	들어가 / VV		에 / JKB	들어가신 / Verb
	시 / EPH		들어가 / VV	다 / Eomi
	ㄴ다 / EFN		신다 / EP+EC	

Unit 03 | Tokenizing

KKMA

Twitter Komoran

Morphs

Nouns Pos Tagging

Unit 03 | Tokenizing

어항 속에 갇힌 고기들보다 어쩌면 내가 좀 더 멍청할지 몰라

Morphs

KKMA	Twitter	Komoran
'어항', '속', '에', '갇히', 'ㄴ', '고기', '들', '보다', '어쩌', '면', '내가', '좀', '더', '멍청', '하', 'ㄹ자', '몰', 'ㄹ라'	'어항', '속', '에', '갇힌', '고기', '들', '보다', '어쩌면', '내', '가', '좀', '더', '멍청할', '지', '몰라'	'어항', '속', '에', '갇히', 'ㄴ', '고기', '들', '보다', '어쩌면', '내', '가', '좀', '더', '멍청', '하', 'ㄹ지', '모르', '아

Nouns

KKMA	Twitter	Komoran
'어항', '속', '고기', '내가'	'어항', '속', '고기', '어쩌면', '내', '좀', '더'	'어항', '속', '고기'

Pos Tagging

KKMA	Twitter	Komoran
('어항', 'NNG'), ('속', 'NNG'), ('에', 'JKM'), ('갇히', 'VV'), ('ㄴ', 'ETD'), ('고기', 'NNG'), ('들', 'XSN'), ('보다', 'JKM'), ('어쩌', 'VV'), ('면', 'ECE'), ('내가', 'NNG'), ('좀', 'MAG'), ('더', 'MAG'), ('멍청', 'XR'), ('하', 'XSA'), ('ㄹ지', 'ECD'), ('몰', 'VV'), ('ㄹ라', 'ECD')	('어항', 'Noun'), ('속', 'Noun'), ('에', 'Josa'), ('갇힌', 'Verb'), ('고기', 'Noun'), ('들', 'Suffix'), ('보다', 'Josa'), ('어쩌면', 'Noun'), ('내', 'Noun'), ('가', 'Josa'), ('좀', 'Noun'), ('더', 'Noun'), ('멍청할', 'Adjective'), ('지', 'Eomi'), ('몰라', 'Verb')	('어항', 'NNG'), ('속', 'NNG'), ('에', 'JKB'), ('갇히', 'VV'), ('ㄴ', 'ETM'), ('고기', 'NNG'), ('들', 'XSN'), ('보다', 'JKB'), ('어쩌면', 'MAG'), ('내', 'NP'), ('가', 'JKS'), ('좀', 'MAG'), ('더', 'MAG'), ('멍청', 'XR'), ('하', 'XSA'), ('ㄹ지', 'EC'), ('모르', 'VV'), ('아', 'EC')

Contents

Unit 01 | NLP Oveview

Unit 02 | Process

Unit 03 | Tokenizing

Unit 04 | Embedding

Unit 05 | Similarity

왜 Tokenizing을 하셨어요?

각 Token마다 Vector 값을 부여하기 위해서 입니다!

왜 Vector 값을 부여할까요?

기계가 언어의 특성을 이해하기 위해서 입니다!

여기서 언어의 특성이란, 단어의 의미 혹은, 문맥에 대한 해석

Embedding 이란?

Token -> Vector 의 과정

One-Hot Encoding 을 이용하는 건 어때요?

네. One-Hot Encoding 이용하셔도 됩니다.

개발환경이 정말 좋다면 말이죠.

어항 속에 갇힌 고기들보다 어쩌면 내가 좀 더 멍청할지 몰라

Unit 04 | Embedding

'어항', '속', '에'
'갇힌', '고기', '들', '보다',
'어쩌면', '내', '가', '좀', '더',
'멍청', '할', '지', '몰라'

Unit 04 | Embedding

'어항' : [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'속' : [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'에' : [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'갇힌' : [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'고기' : [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'들' : [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'보다' : [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'어쩌면' : [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
'내' : [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
'가' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
'좀' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
'더' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
'멍청' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
'할' : [0, 0, 10 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
'지' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]
'몰라' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

'어항 속에 갇힌 고기들보다
어쩌면 내가 좀 더 멍청할지 몰라'

↓

**한 문장을 표현하기 위해
16개의 Feature 필요하다.**

Unit 04 | Embedding

어항 속에 갇힌 고기들보다 어쩌면 내가 좀 더 멍청할지 몰라 너가 먹이처럼 던진 문자 몇 통과 너의 부재중 전화는 날 헷갈리게 하지 너의 미모와 옷 입는 스타일로 미루어 보았을 때 너의 어장의 크기는 수족관의 scale 단지 너 하나 때문에 경쟁은 무척 험하고도 아득해 I wanna be your boy friend not one of your 생선 허나 너에겐 늘 누군가가 옆에 있어 이럴 때일수록 내 이미지를 위해선 그저 쏘 쿨한 척 하는 게 최선임을 알기에 I said have a good time 그날 밤 업데이트 된 너의 페북의 담벼락엔 여전히 물고기들이 하악 하악 당연히 나도 그 중 하나 하루 종일 너란 바닷속을 항해하는 나는 아쿠아맨 헤엄 헤엄 헤엄 I'm rolling in the deep inside of you 너의 어장은 너무 캄캄해 헤엄 헤엄 헤엄 손에 꼽을 정도로 아주 가끔씩 엉뚱한 시간에 넌 내가 어딘지 묻지 '어디긴 니 마음이지'라는 본심을 속이며 차분하게 말했지 '지금 집 ' 심상치 않은 징조, 심장은 보다 신속 혹시 모를 급만남이 꿈처럼 이루어 질 수도 있을 것 같은 느낌이 들어 지금 너의 위치가 어디든 let's bring it on girl 통장 잔고는 가까스로 위기모면 입을 옷도 이미 정했어 목걸이는 chrome hearts but you said, 그런 거 아니고 please don't call me no more oh no~~ 내 가슴은 회처럼 조각이 났지 u don't give a shit about my broken heart 넌 딱 잘라 말했지 손톱깎이 같이 I don't give a shit about your stupid heart bitch you gotta get yo mind right 너의 얼굴과 몸이 영원할까 bitch you gotta get yo mind right 너의 얼굴과 몸이 영원할까 nah~~

Unit 04 | Embedding

어항 속에 갇힌 고기들보다 어쩌면 내가 좀 더 멍청할지 몰라 너가 먹이처럼 던진 문자 몇 통과 너의 부재중 전화는 날 헷갈리게 하지 너의 미모와 옷 입는 스타일로 미루어 보았을 때 너의 어장의 크기는 수족관의 scale 단지 너 하나 때문에 경쟁은 무척 험하구도 아득해 I wanna be your boy friend not one of your 생선 하나 되어선 물 무언가가 곁에 있어 어울 때일수록 내 이미지를 위해선 그저 쏘 쿨한 척 하는 [illegible] 그날 밤 업데이트 된 너의 페북의 담벼락엔 여전히 물고기들이 하악 하악 당연히 나도 그 중 하나 하루 종일 너란 바닷속을 항해하는 나는 아쿠아맨 헤엄 헤엄 헤엄 I'm rolling in the deep inside of you 너의 어장은 너무 캄캄해 헤엄 헤엄 헤엄 손에 꼽을 정도로 [illegible] 넌 내가 어딘지 묻지 '어디긴 니 마음이지'라는 본심을 [illegible] 진조, 심장은 보다 신속 혹시 모를 급만남이 꿈처럼 이루어 질 수도 있을 것 같은 느낌이 들어 지금 너의 위치가 어디든 let's bring it on girl 통장 잔고는 가까스로 위기모면 입을 옷도 이미 정했어 목걸이는 chrome hearts but you said, 그런 거 아니고 please don't call me no more oh no~~ 내 가슴은 회처럼 조각이 났지 u don't give a shit about my broken heart 넌 딱 잘라 말했지 손톱깎이 같이 I don't give a shit about your stupid heart bitch you gotta get yo mind right 너의 얼굴과 몸이 영원할까 bitch you gotta get yo mind right 너의 얼굴과 몸이 영원할까 nah~~

One Hot Encoding : N 개의 Feature 생성
(N : 중복되지 않는 Token의 수)

N이 너무 커지기 때문에,
계산하기에 많은 시간이 소모됩니다.

Unit 04 | Embedding

'어항' : [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'속' : [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'에' : [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'갇힌' : [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'고기' : [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'들' : [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'보다' : [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
'어쩌면' : [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
'내' : [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
'가' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
'좀' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
'더' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
'멍청' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
'할' : [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
'지' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]
'몰라' : [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

Sparse Vector

Inner Product = 0
(Independent)

불필요한 계산이 많습니다.

유의어, 반의어 요소를
고려하지 못합니다.

그렇다면 효율적인 Embedding Method는 어떤 종류가 있을까요?

Unit 04 | Embedding

Word2Vec FastText Glove

Word2Vec은 말 그대로 Word to Vector 입니다.

어떠한 방법으로 Vector를 부여할까요?

Word2Vec은 말 그대로 Word to Vector 입니다.

Unit 04 | Embedding

1. Word & Neighbor (CBOW, Skip Gram)

2. Network

Unit 04 | Embedding

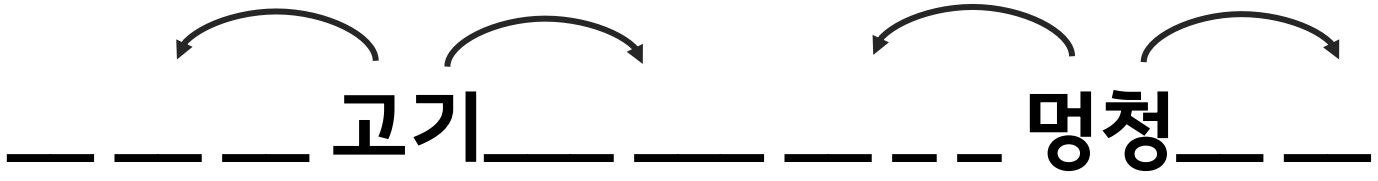
어항 속에 갇힌 고기들보다 어쩌면 내가 좀 더 멍청할지 몰라

어항 속에 갇힌 __들보다 어쩌면 내가 좀 더 멍청할지 몰라

CBOW : 주변에 있는 단어들을 가지고 중심에 있는 단어를 맞추는 방식

어항 속에 갇힌 __들보다 어쩌면 내가 좀 더 __할지 몰라

Skip - Gram : 중심에 있는 단어로 주변의 단어를 맞추는 방식



Window Size : 주변의 단어를 탐색하는 크기

Window Size : 3

어항 속에 갇힌 고기들보다 어쩌면 내가 좀 더 멍청할지 몰라

Unit 04 | Embedding

'어항', '속', '에'
'갇힌', '고기', '들', '보다',
'어쩌면', '내', '가', '좀', '더',
'멍청', '할', '지', '몰라'

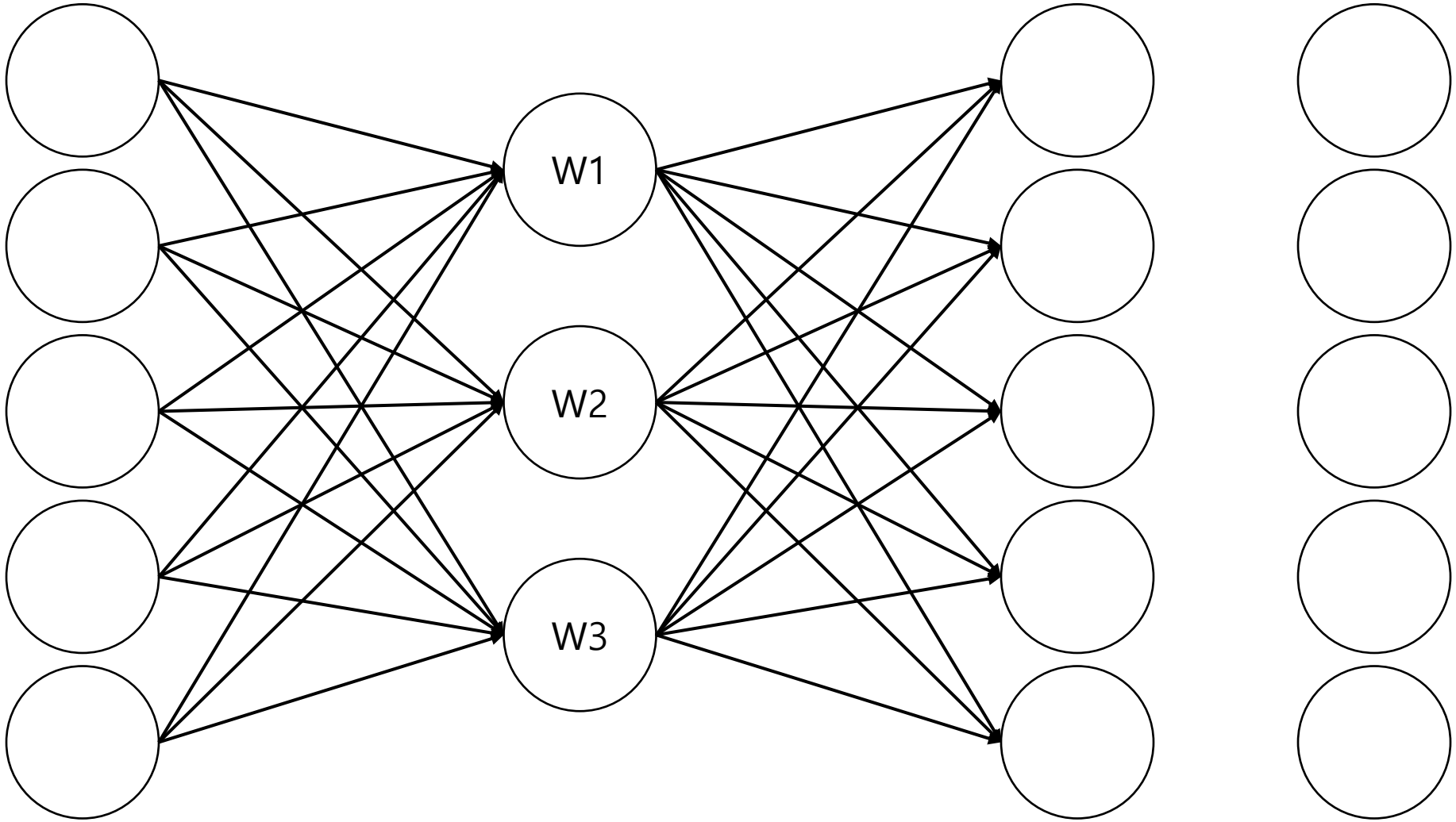
Word	Neighbor (Window Size = 2)
어항	속
속	어항
속	에
에	속
에	갇힌
고기	갇힌
고기	들
들	…

Unit 04 | Embedding

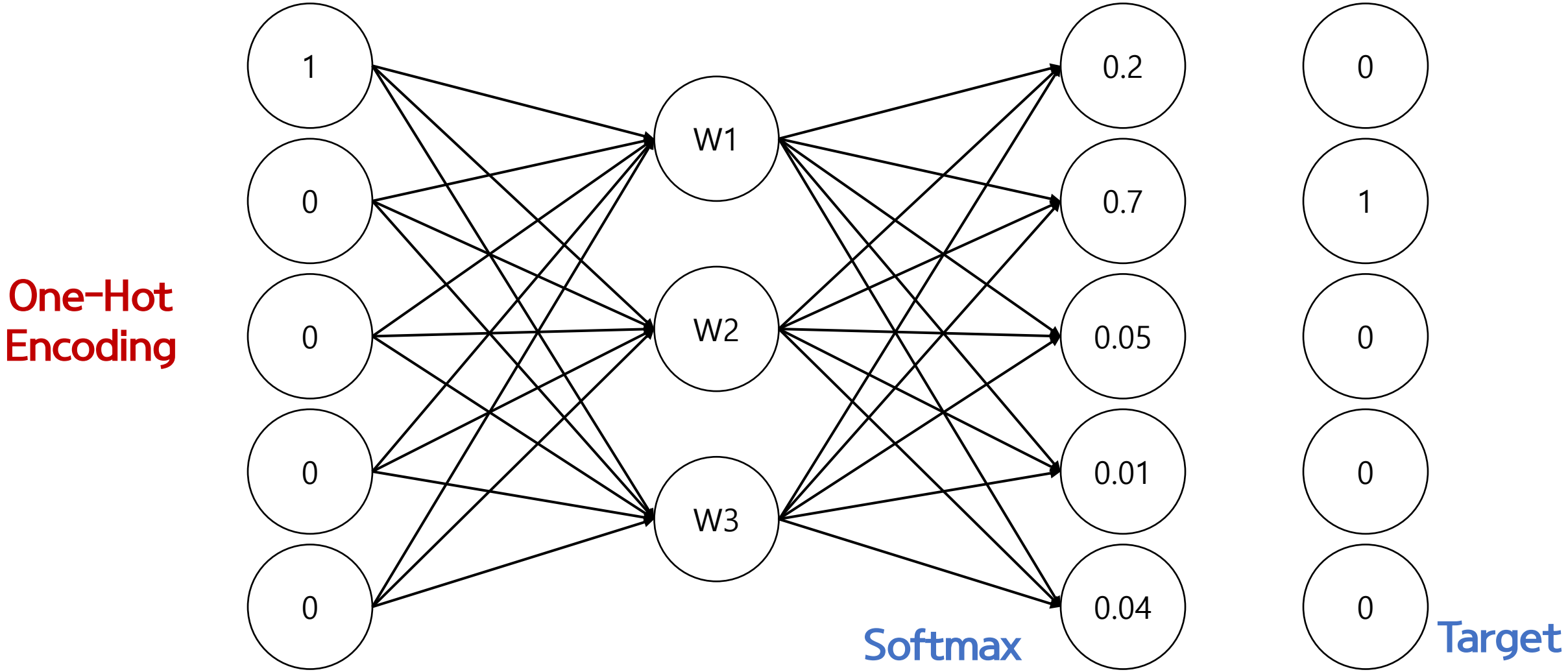
'어항', '속', '에'
'갇힌', '고기', '들', '보다',
'어쩌면', '내', '가', '좀', '더',
'멍청', '할', '지', '몰라'

Input	Target
Word	Neighbor (Window Size = 2)
어항	속
속	어항
속	에
에	속
에	갇힌
고기	갇힌
고기	들
들	…

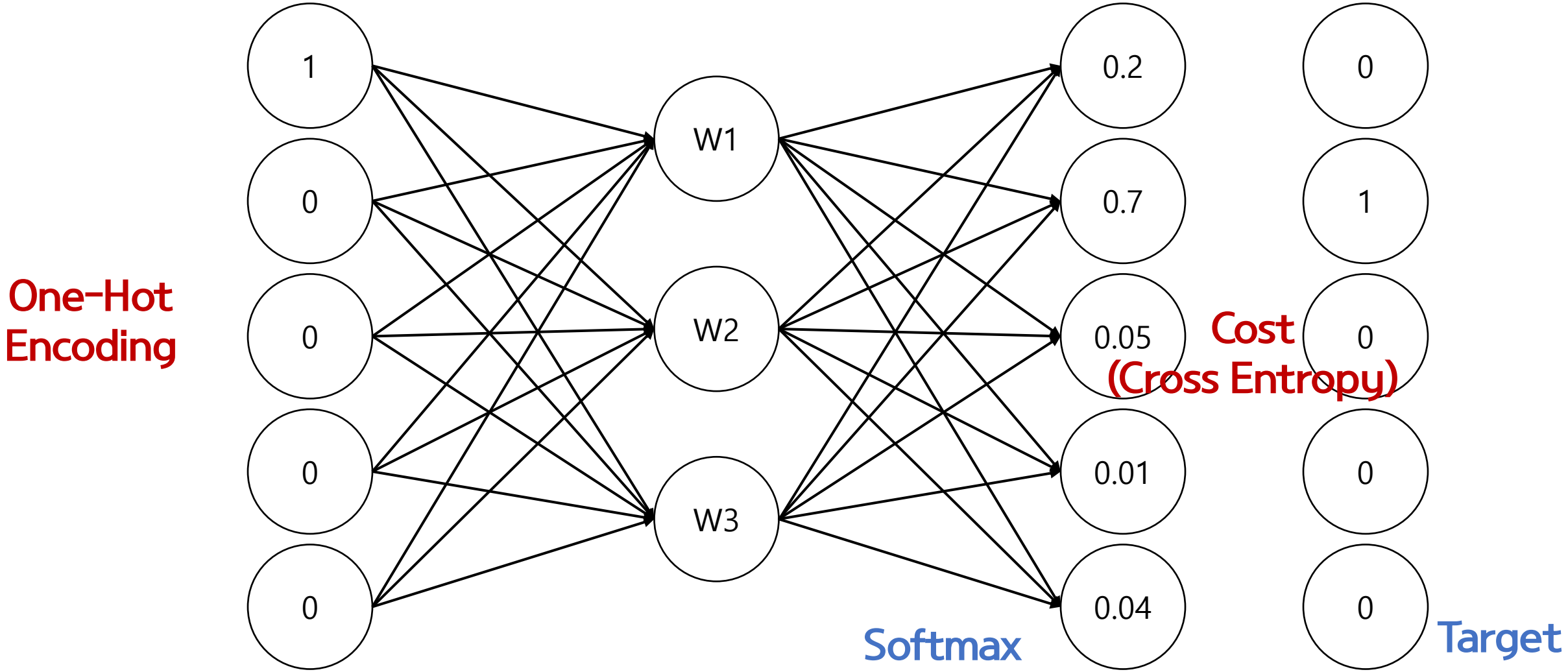
Unit 04 | Embedding



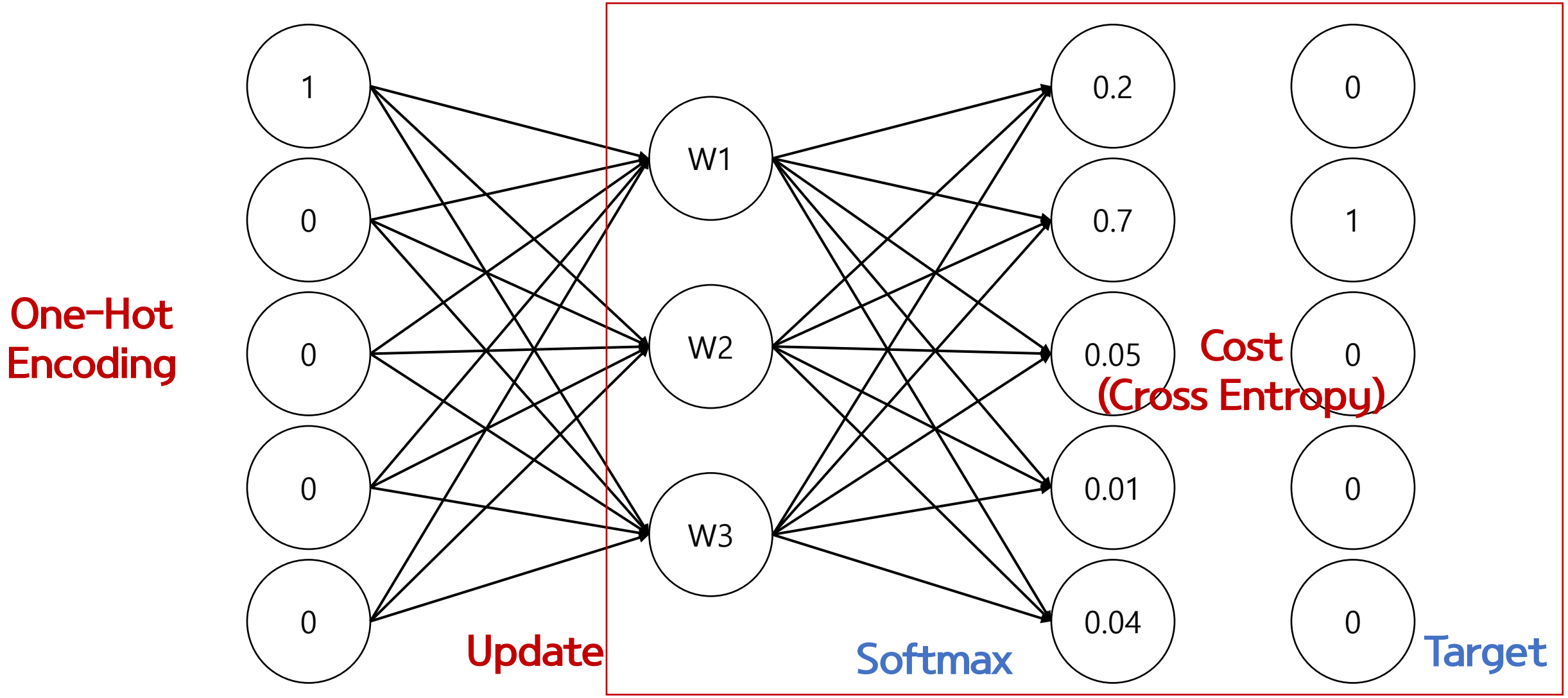
Unit 04 | Embedding



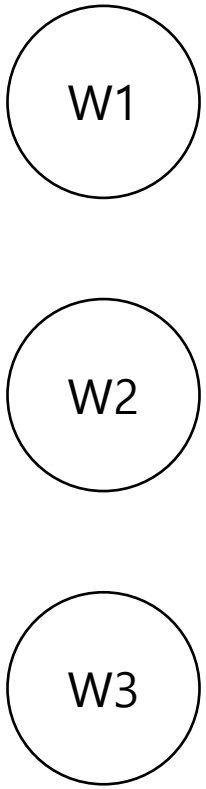
Unit 04 | Embedding



Unit 04 | Embedding



Unit 04 | Embedding



Word2Vec

FastText

단어가 아닌 n-gram을 최소단위로 !

"apple"

(0,1,0,0, ·····0)

min = 3, max = 6

"<ap", "app", "appl", "apple", "apple>",
"ppl", "pple", "pple>", "ple", "ple>", "le>"

(1,0,0,0, ·····0), (0,1,0,0, ·····0),
(0,0,1,0, ·····0), (0,0,0,1, ·····0), ·····

Glove

[This is as solid as ice.]

[Ice is solid.]

[I have some ice in my fridge.]

[Steam is usually hot]

.
.
.
.

두 단어가 한 말뭉치(한 문장)에
동시 등장할 확률 !

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

특정 단어 k가 주어졌을 때 ice와 steam의 내적값이 이 값과 같아지도록 학습 !

Contents

Unit 01 | NLP Oveview

Unit 02 | Process

Unit 03 | Tokenizing & TDM

Unit 04 | Embedding

Unit 05 | Similarity

Distance

Cosine Similarity

Unit 05 | Similarity

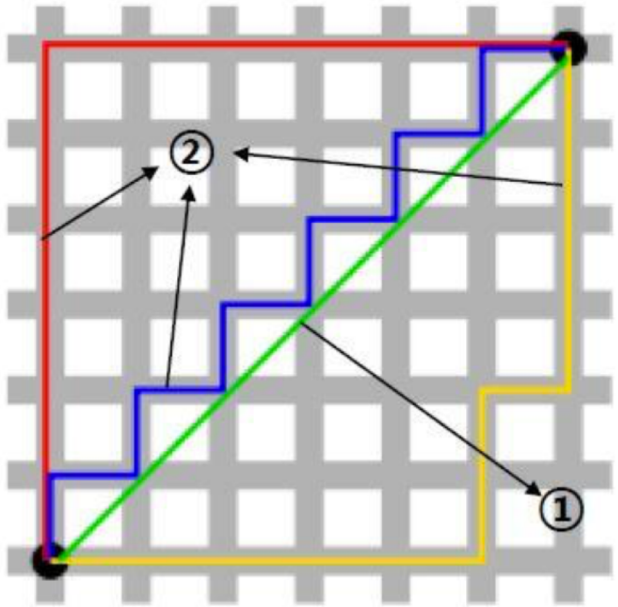
Distance

$$d(x,y) = \left(\sum_{i=1}^{p}(x_i - y_i)^2\right)^{1/2}$$

$$d(x,y) = \sum_{i=1}^{p}|x_i - y_i|$$

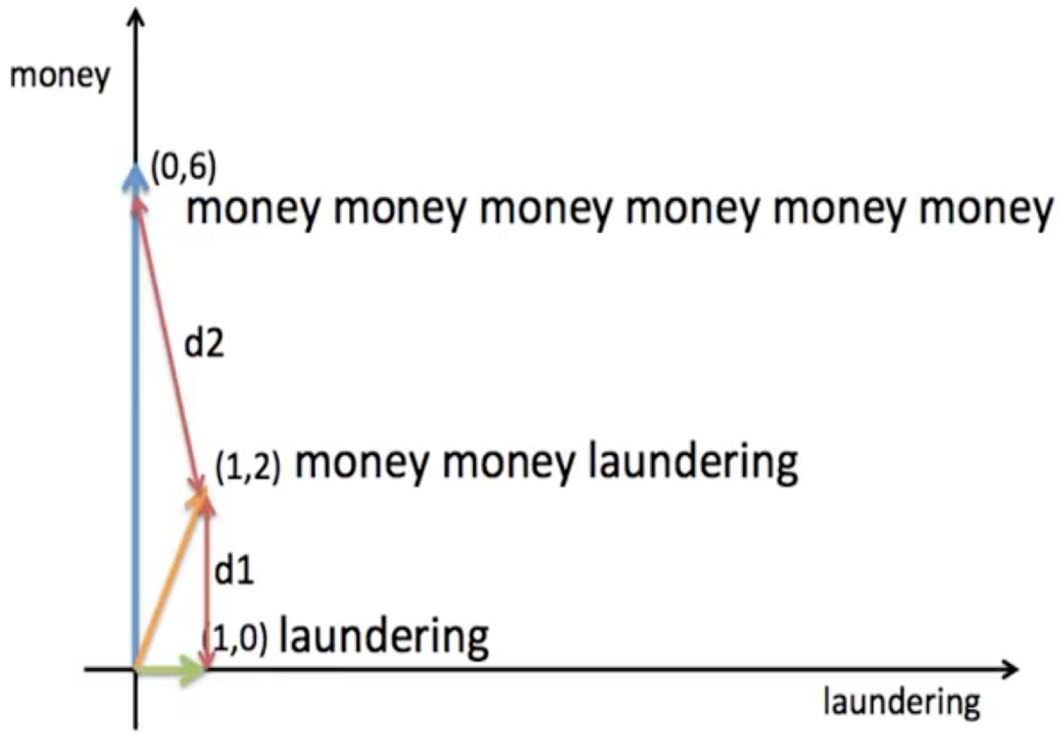
$$d(x,y) = \left(\sum_{i=1}^{p}(x_i - y_i)^2/s_i^2\right)^{1/2}$$

$$d(x,y) = \left(\sum_{i=1}^{p}(x_i - y_i)^m\right)^{1/m}$$



Unit 05 | Similarity

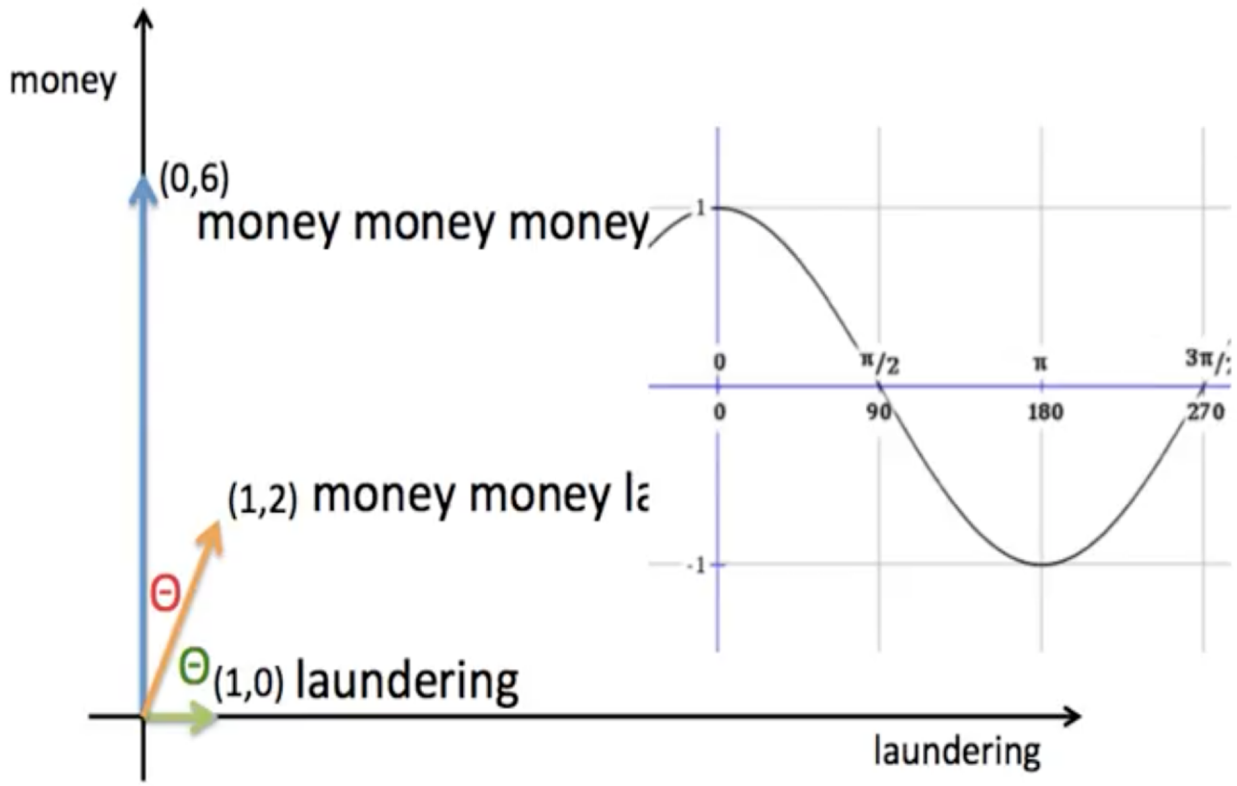
Cosine Similarity



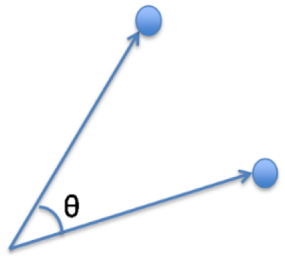
Money Money Money Money Money가 더 유사하지 않을까?

Unit 05 | Similarity

Cosine Similarity



$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



벡터의 크기를 무시하고 방향성만 고려하여 계산
두 벡터가 이루고 있는 각도를 활용하여,
작을수록 유사하다고 판단

Unit 01 | NLP Overview

Youtubing : Tokenizing, Embedding, Similarity

TextRick : Tokenizing, Word2Vec, LDA, LSA, TF-IDF

Alpha Aesop : GAN, Seq, Token, FastText, Bi-Gram

합격하소서 : Pos Tag, Word2Vevc, CBOW, LSTM

Q & A

들어주셔서 감사합니다.