Text Analytics

Ch5 : Text Representation

서수원

Business Intelligence Lab.

산업경영공학과, 명지대학교

01

GloVE

Motivation

- Limitations of Word2Vec
  - Word2Vec은 여전히 많이 나오는 단어에 대해 학습을 하는데 많은 시간을 투자한다.
    - ✓ 이는 학습의 불균형을 초래한다.
    - ✓ 학습의 불균형을 해소하기 위한 많은 방법론이 있지만, 여전히 불균형은 있다.
    - ✓ Ex : P(W I the)

Theatre or theater is a collaborative form of fine art that uses live performers to present the experience of a real or imagined event before a live audience in a specific place. The performers may communicate this experience to the audience through combinations of gesture, speech, song, music, and dance. Elements of art and stagecraft are used to enhance the physicality, presence and immediacy of the experience. The specific place of the performance is also named by the word "theatre" as derived from the Ancient Greek (thatron, "a place for viewing"), itself from (theomai, "to see", "to watch", "to observe"). Modern Western theatre comes from large measure from ancient Greek drama, from which it borrows technical terminology, classification into genres, and many of its themes, stock characters, and plot elements. Theatre artist Patrice Pavis denes theatricality, theatrical language, stage writing, and the specificity of theatre as synonymous expressions that differentiate theatre from the other performing arts, literature, and the arts in general. Theatre today, broadly denfied, includes performances of plays and musicals, ballets, operas and various other forms.

Notation

- GloVe
  - X 는 Vocab x Vocab의 행렬이다.
  - X_{ij}는 i랑j가 함께 등장하는 빈도를 의미한다.
  - X_i는 i가 코퍼스에 등장하는 빈도를 의미한다.
  - P_{ij}는 i가 등장할 때 j가 함께 등장할 확률을 의미한다.
  - W는 d차원의 워드 임베딩을 의미한다.
    - $X \in \mathbb{R}^{V \times V}$ word co-occurrence matrix
    - X_{ij} frequency of word i co-occurring with word j
    - $X_i = \sum_k^V X_{ik}$ total number of occurrences of word i in corpus
    - $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ a.k.a. probability of word j occurring within the context of word i
    - $w \in \mathbb{R}^d$ a word embedding of dimension d
    - $\tilde{w} \in \mathbb{R}^d$ a context word embedding of dimension d

Motivation

- GloVe
  - P(k I ice)는 ice가 등장할 때 k가 등장할 확률을 나타낸 것 이다.
  - P(k I steam)은 steam이 등장할 때 k가 등장할 확률을 나타낸 것 이다.
  - P(k I ice) / P(k I steam)은 ice와 steam사이의 비율을 나타낸 것 이다.
    - ✓ 값이 크면 분자와 더 관련이 있다는 것 이다.
      - ❖ Solid
    - ✓ 값이 작으면 분모와 더 관련이 있다는 것 이다.
      - ❖ Steam
    - ✓ 값이 1과 비슷하다면, 둘 다 관련이 있거나, 둘 다 관련이 없다는 것 이다.
      - ❖ Water-> 둘 다 관련이 있다. Fashion ->둘 다 관련이 없다.

Prob. 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}
$\frac{P(k\|ice)}{P(k\|steam)}$	8.9	8.5×10^{-2}	1.36	0.96

Formulation

- GloVe
  - 세 단어의 관계를 F를 통하여 표현한다.

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

  - 세 단어의 관계를 뺄셈을 활용해 표현한다.

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

  - 세 단어의 관계를 내적 스칼라를 활용해 표현한다.
    - ✓ 이는 W_i와 W_j의 차이를 맥락과 연결하기 위함이다.

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

Homomorphism

- GloVe

$$\frac{P(k|ice)}{P(k|steam)} = F\left((v_i - v_j)^T \tilde{v}_k\right) = \frac{P_{ik}}{P_{jk}}$$

$$\frac{P(solid|ice)}{P(solid|steam)} = F\left((ice - steam)^T solid\right)$$

$$\frac{P(solid|steam)}{P(solid|ice)} = F\left((steam - ice)^T solid\right)$$

$$F\left((ice - steam)^T solid\right) = \frac{P(solid|ice)}{P(solid|steam)} = \frac{1}{F\left((steam - ice)^T solid\right)}$$

$$(ice - steam)^T solid = -(steam - ice)^T solid$$

inverse element of addition

$$F\left((ice - steam)^T solid\right) = \frac{1}{F\left((steam - ice)^T solid\right)}$$

inverse element of multiplication

Homomorphism

- GloVe

$$w_i^T \tilde{w}_k = (w_i - w_j)^T \tilde{w}_k + w_j^T \tilde{w}_k$$

$$F(w_i^T \tilde{w}_k) = F\left((w_i - w_j)^T \tilde{w}_k + w_j^T \tilde{w}_k\right)$$

$$= F\left((w_i - w_j)^T \tilde{w}_k\right) \times F(w_j^T \tilde{w}_k)$$

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)} = \frac{P_{ik}}{P_{jk}}$$

✓ Finally, we can drive that $F(x) = \exp(x)$

Solution

- GloVe

✓ We know that $F(x) = \exp(x)$ and $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)} = \frac{P_{ik}}{P_{jk}}$

$$\exp\left(w_i^T \tilde{w}_k - w_j^T \tilde{w}_k\right) = \frac{\exp\left(w_i^T \tilde{w}_k\right)}{\exp\left(w_j^T \tilde{w}_k\right)}$$

$$w_i^T \tilde{w}_k = \log P_{ik} = \log X_{ik} - \log X_i$$

$$w_i^T \tilde{w}_k = \log P_{ik} = \log X_{ik} - \log X_i$$

$$w_i^T \tilde{w}_k = \log X_{ik} - b_i - \tilde{b}_k$$

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log X_{ik}$$

Objective Function

- GloVe

✓ A least squared objective function

$$J = \sum_{i,j=1}^{V} \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

$$\Rightarrow J = \sum_{i,j=1}^{V} f\left(X_{ij}\right) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

where f has the following desiderata:

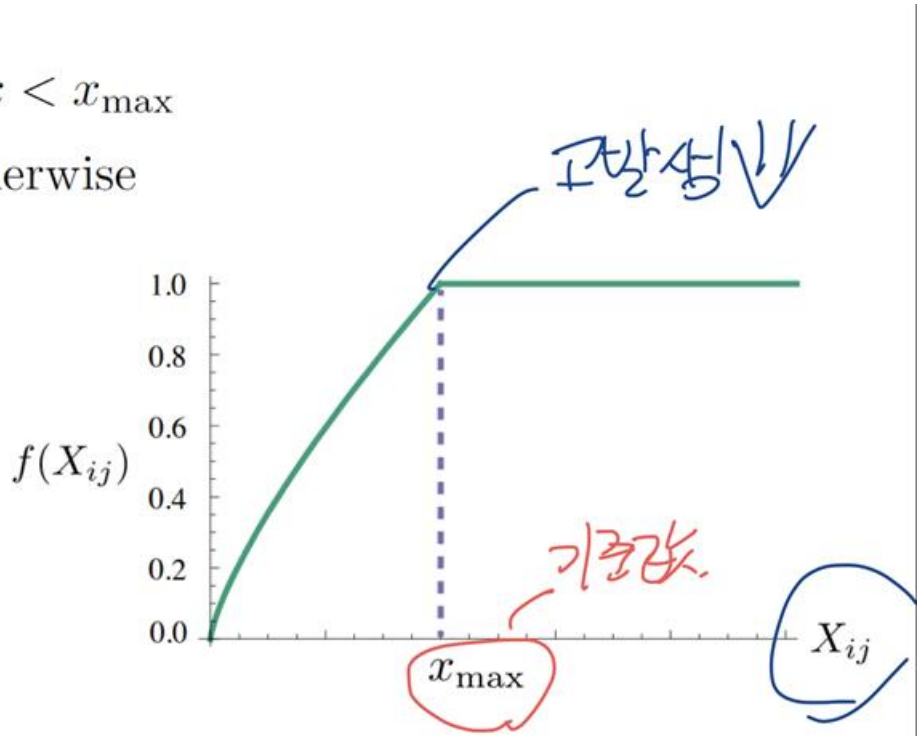
  - F(0) 은 0
  - F(X)는 감소하지 않는다.
  - F(X)는 고 발생빈도 단어에 대해 가중치를 낮춰준다.

Objective Function

- GloVe

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

$$\text{where } f(x) = \begin{cases} \left(\frac{x}{x_{\max}}\right)^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$



Result

- GloVe
  - 우선 코퍼스를 대상으로 상관행렬을 만드는 것으로 학습한다.
    - ✓ 단어의 개수가 1만개라면 요소 개수가 1억이나 되는 행렬을 만들어야 한다.
      - ❖ 계산 복잡성이 크다.

02

FastText

Limitations

- Limitations of NNLM,Word2Vec,and GloVe
  - Morphology의 특성을 무시한다.
  - 형태소 변화가 잦거나, 낮은 빈도의 단어가 많으면 적용이 어렵다.
    - ✓ 터키어

- FastText의 목표
  - 캐릭터 레벨의n-gram을 학습한다.
  - 단어의 분산표상을 n-gram vector의 합으로 표현한다.

Objective Function

- FastText
  - Objective Function

$$J_t(\theta) = \log\sigma(u_o^T v_c) + \sum_{i=1}^{k} E_{i\sim P(w)}\left[\log\sigma(-u_i^T v_c)\right]$$

  - Subword model
    - ✓ N-gram의 세트를 w: $\mathcal{G}_w \subset \{1, ..., G\}$ 게 정의한다.

$$score(w, c) = \sum_{g\in\mathcal{G}_w} \mathbf{z}_g^T \mathbf{v}_c$$

    - ✓ 단어를 N-gram의 벡터 내적 합으로 나타낸다.

FastText

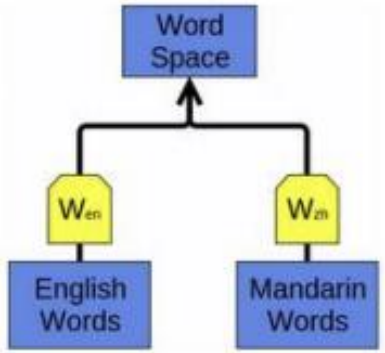
- Subword model
  - N-gram representation
    - ✓ 모든 사이즈의 n-gram을 다 포함한다.
    - ✓ 같은 단어를 공유 하더라도, 서로 다른 벡터가 할당된다.
      - ❖ Apple : Appl,App
        - » 같은 app를 공유 하더라도 둘의 벡터는 서로 다르다.

FastText

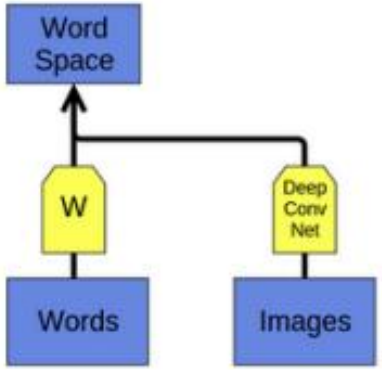
p	...	...	...	...	...
...	...	...	...	...	...
r	...	...	...	...	...
pa	...	...	...	...	...
...	...	...	...	...	...
er	...	...	...	...	...
par	...	...	...	...	...
...	...	...	...	...	...
ameter	...	...	...	...	...
Avg.	...	...	...	...	...

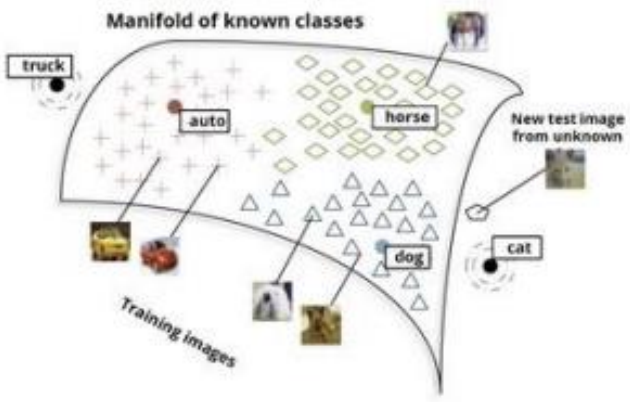
Word Embedding Examples

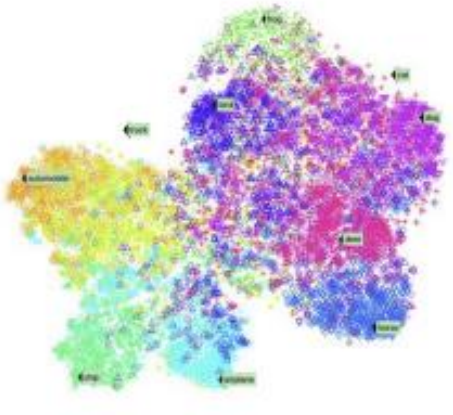
- Word Embedding with two different languages



- Word Embedding with Images





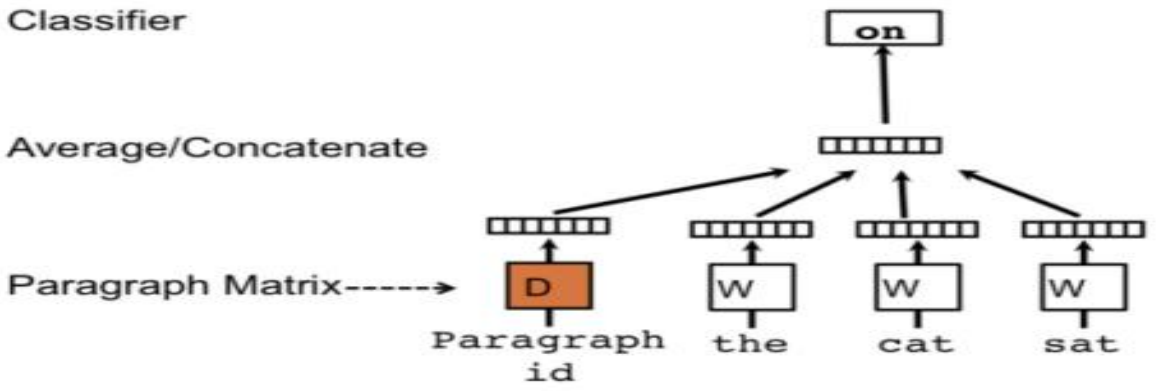


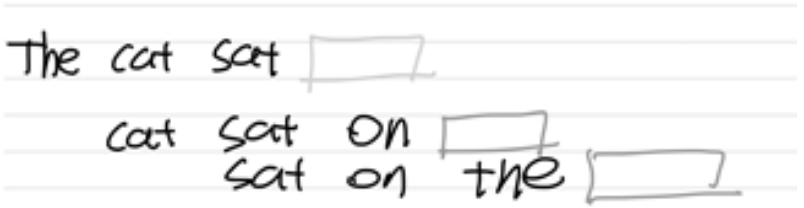
03

High - level

Document Embedding

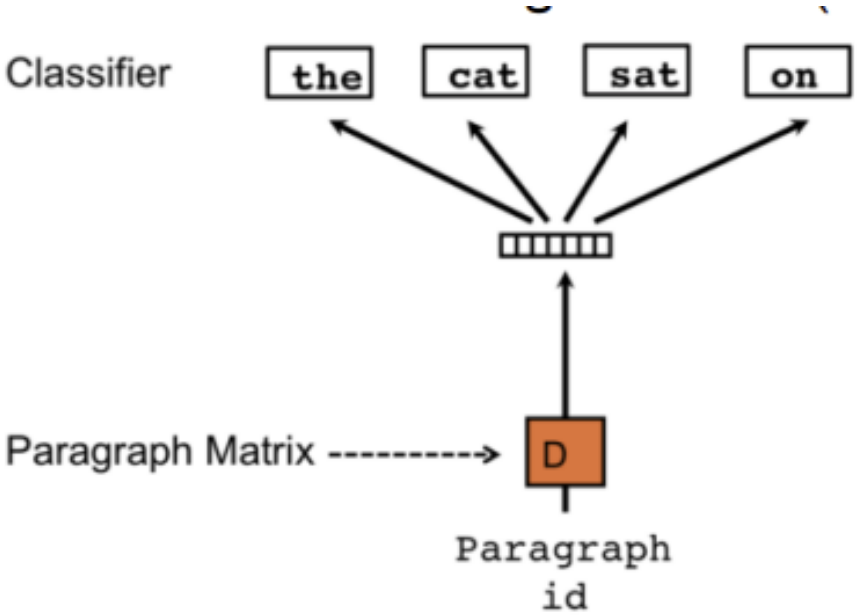
- Document Embedding
  - 단어수준의 임베딩이 가능하다면, 문장, 문단, 문서 단위 임베딩도 가능하지 않을까?
  - 두가지 방법이 존재한다.
    - ✓ Paragraph Vector model : Distributed Memory(PV-DM)model
      - ❖ 문단 마다 id가 존재한다.
      - ❖ Window size는 하이퍼 파라미터 이다.
      - ❖ Word vector는 모든 문단을 공유한다.
        - » Paragraph id : a & cat = Paragraph id : b & cat





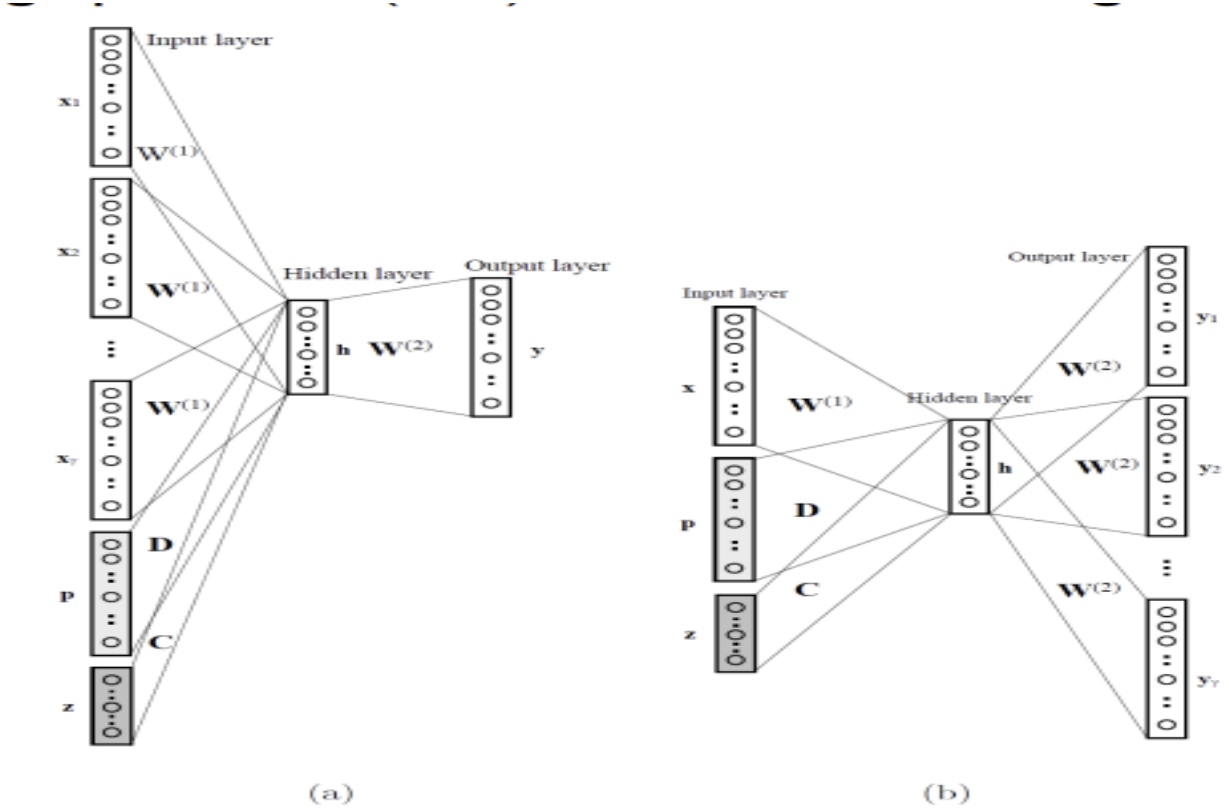
Document Embedding

- Document Embedding
  - 단어수준의 임베딩이 가능하다면, 문장, 문단, 문서 단위 임베딩도 가능하지 않을까?
  - 두가지 방법이 존재한다.
    - ✓ Paragraph Vector model : Distributed Bag of Words(PV-DBOW)
      - ❖ 문단 id만 넣고 일정 개수의 단어를 예측한다.
      - ❖ 단어의 순서 상관없이 예측한다.
        - » The sat on cat
        - » 워드벡터가 필요 없다.
      - ❖ 성능은 PV-DM만으로도 충분 하지만,
        두개를 combine해서 사용하는 것을 추천한다.



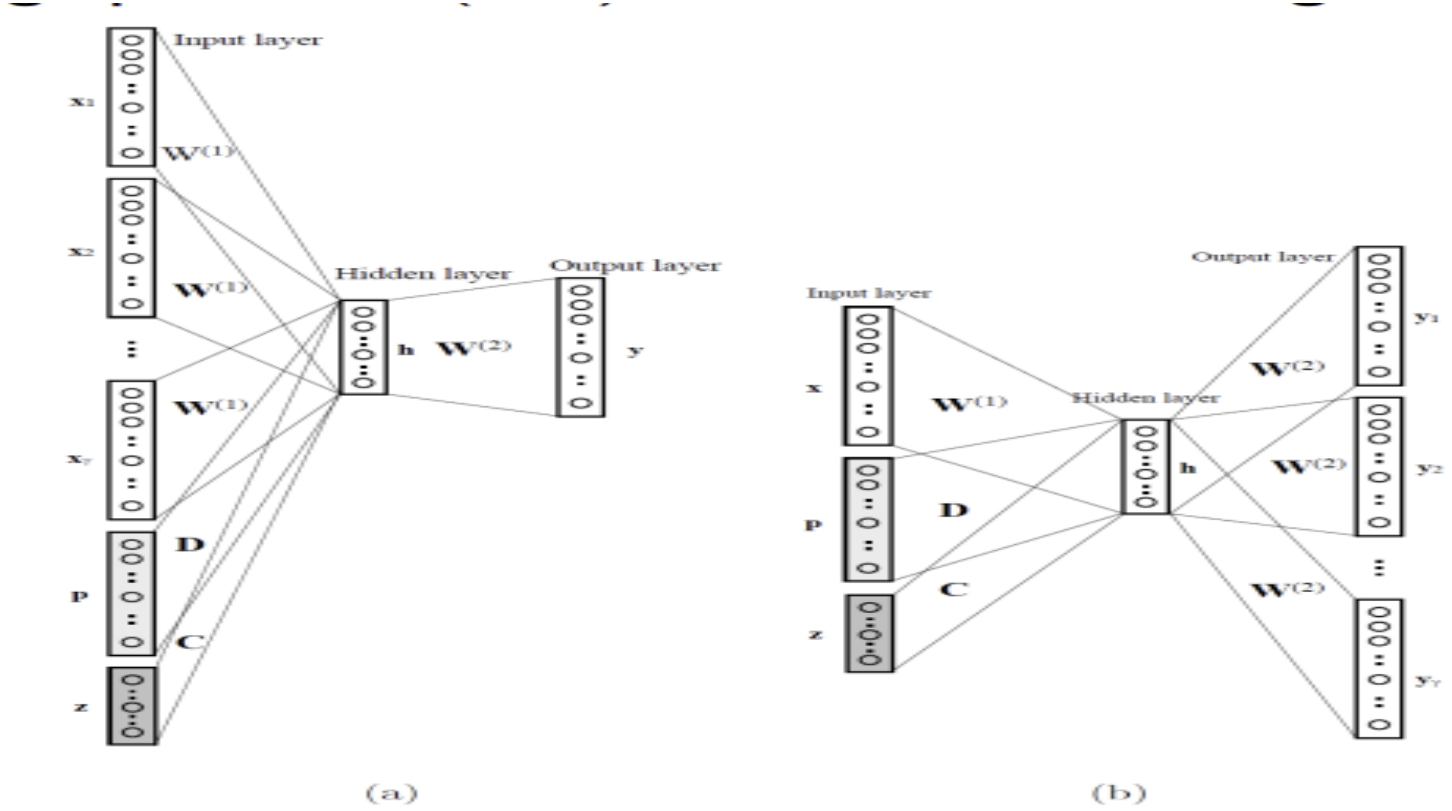
Document Embedding

- Document Embedding
  - 점수에 따른 데이터의 벡터 위치를 알고싶다.
    - ✓ Z는 평점의 벡터, p는 문장의 벡터, x는 리뷰데이터 이다.



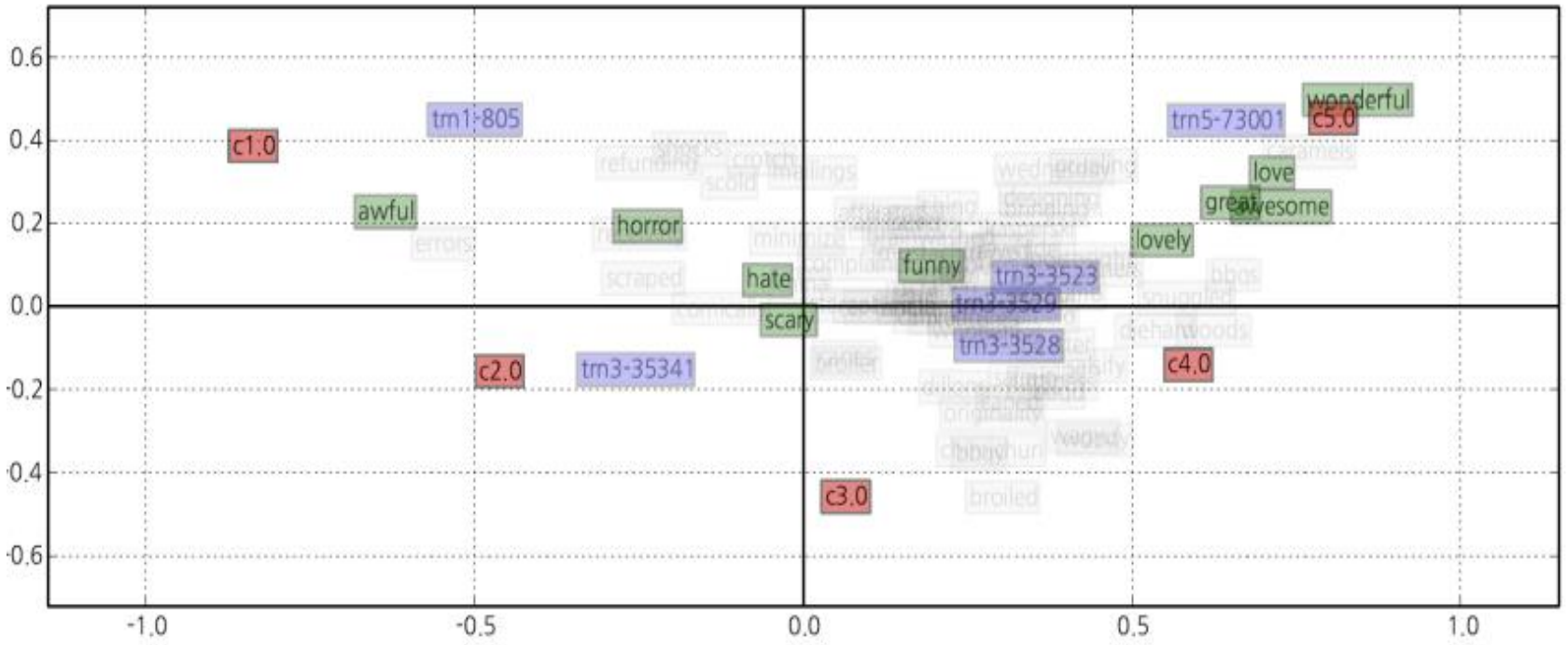
Document Embedding

- Supervised Paragraph Vector
  - 점수에 따른 데이터의 벡터 위치를 알고싶다.
    - ✓ Z는 평점의 벡터, p는 문장의 벡터, x는 리뷰데이터 이다.



Document Embedding

- Supervised Paragraph Vector
  - 점수에 따른 데이터의 벡터 위치를 알고싶다.
    - ✓ 빨간색은 평점, 초록색은 단어, 파란색은 리뷰 문장이다.

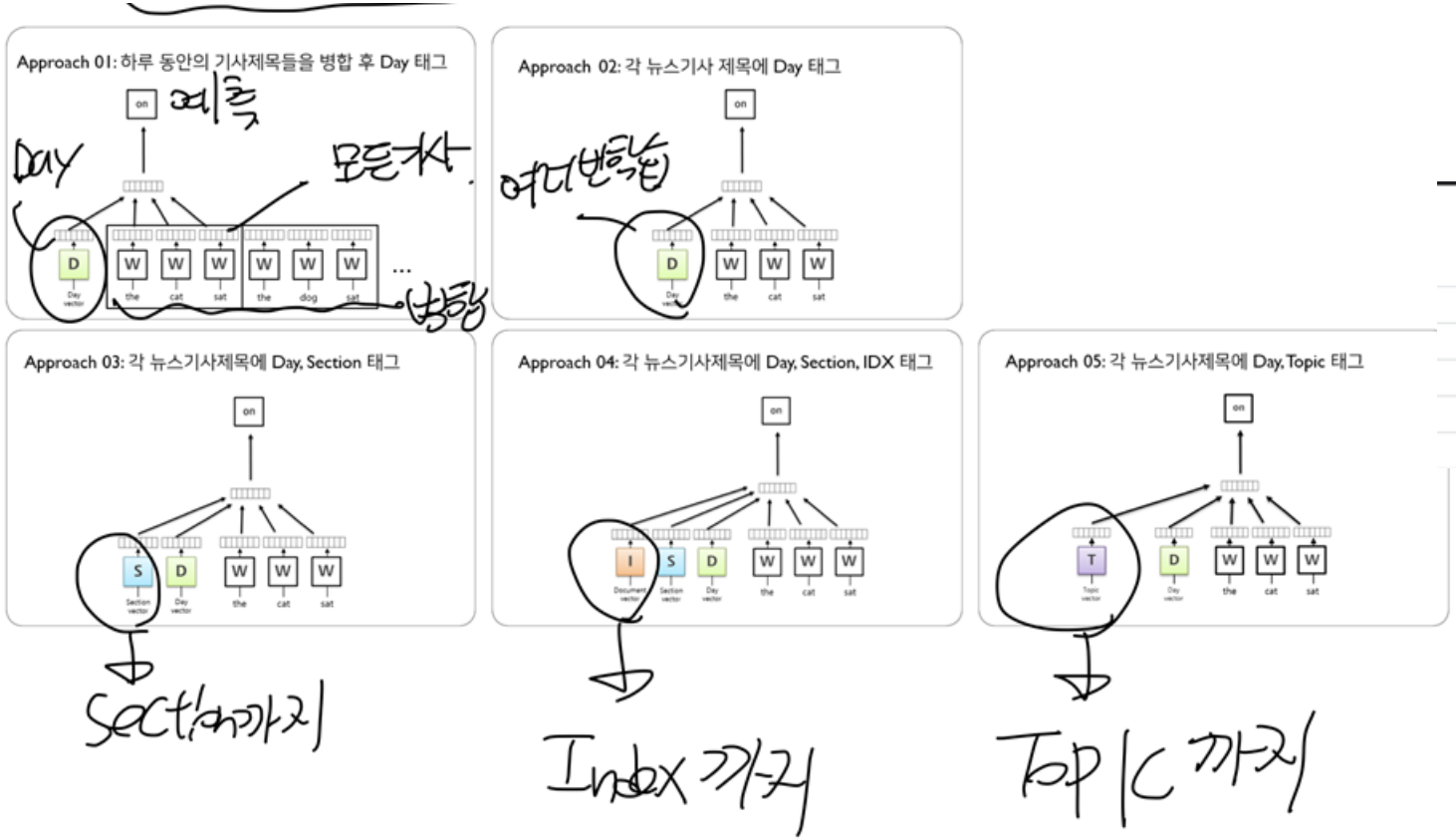


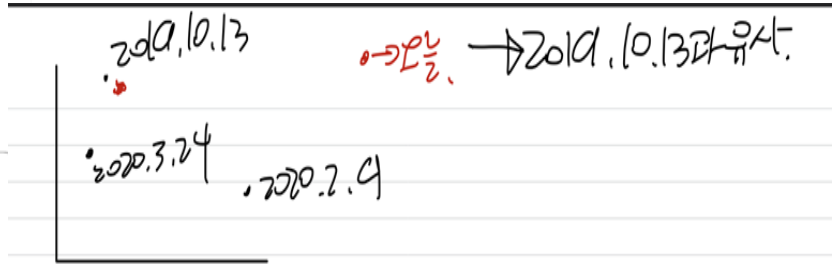
04

More Things to Embed?

Day Embedding

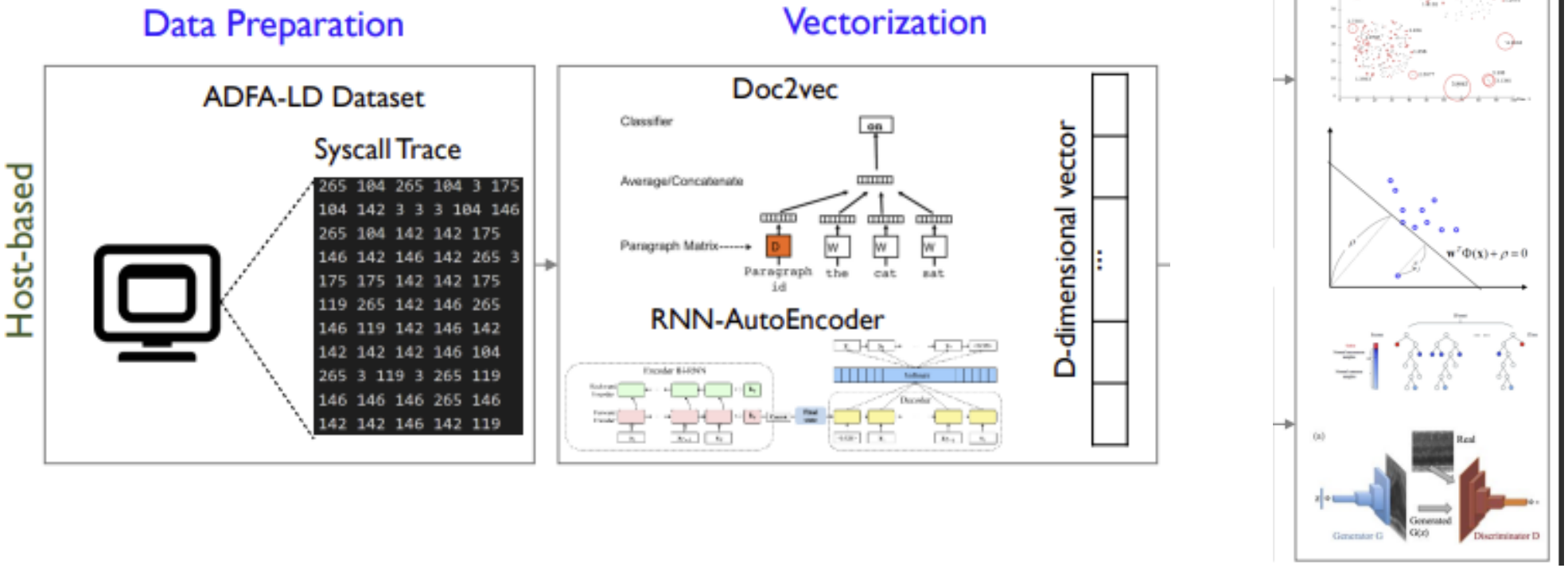
- Day Embedding
  - 뉴스단어, 기사를 임베딩 한다.
    - ✓ 그날 대표 기사는 어디에 임베딩 되는지 알기 위함이다.





Anomaly Detection

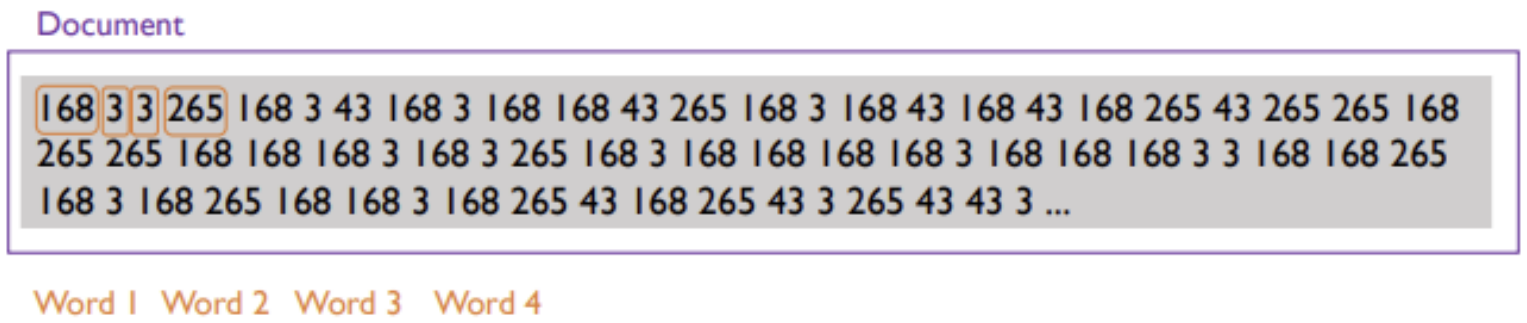
- Anomaly Detection
  - Syscall Trace
    - ✓ 각각의 숫자가 접속했을 때 남는 로그이다.
      - ❖ 265 : 무슨 행위 했니?
      - ❖ 행위들의 집합이 한 사람이 한 행동이다.



Anomaly Detection

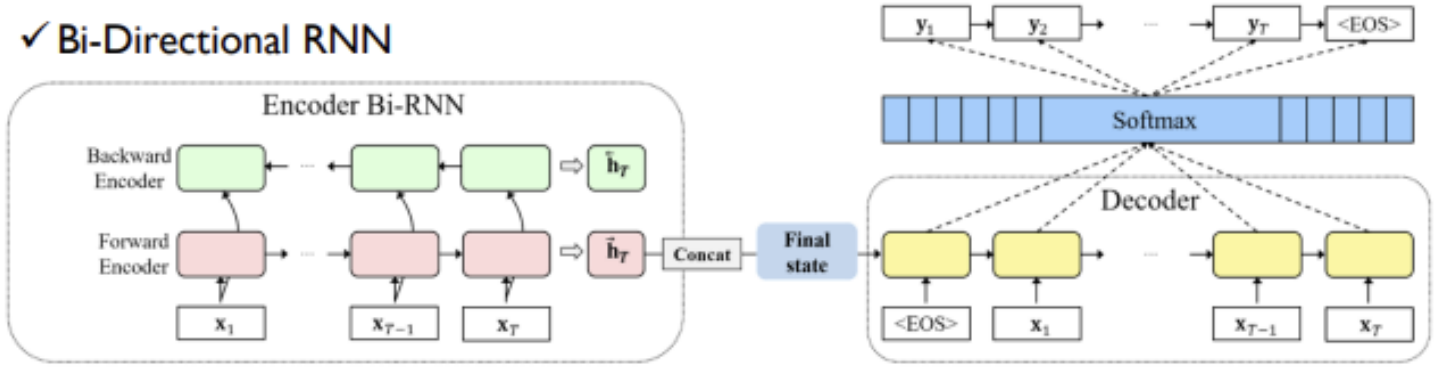
- Syscall Trace
  - 각각의 숫자가 접속했을 때 남는 로그이다.
    - ✓ 사용자 마다 길이가 다름으로 가변길이의 벡터이다.
    - ✓ 고정길이의 벡터로 변환하는 과정이 필요하다.

  - Syscall2Vec
    - ✓ Doc2Vec 기반이다.
    - ✓ 하나의 System Call Trace를 Document로 취급하고, 개별 syscall을 word로 취급한다.

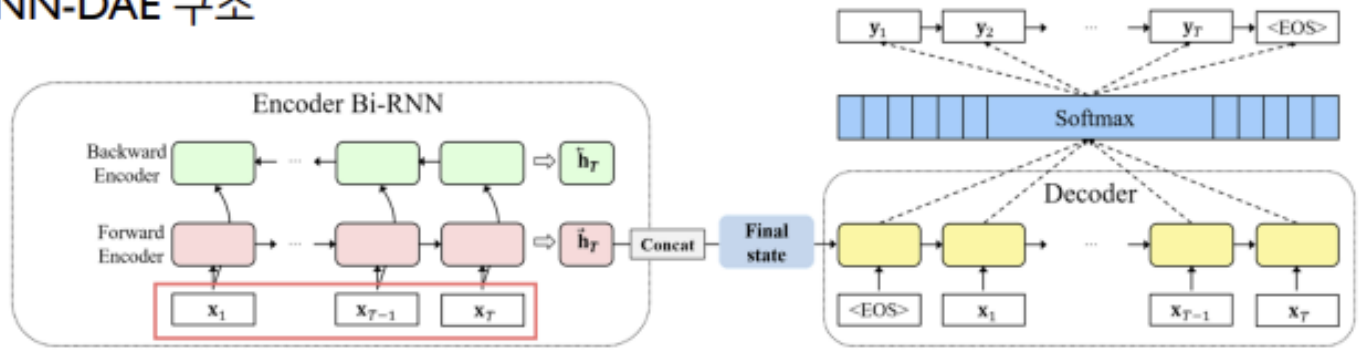


Anomaly Detection

- RNN-AE



- RNN-DAE
  - RNN-DAE 구조



Anomaly Detection

- Live2Vec in afreecaTV

[그림11] 유사도 측정 비교(Nearest neighbor vs Live2Vec)