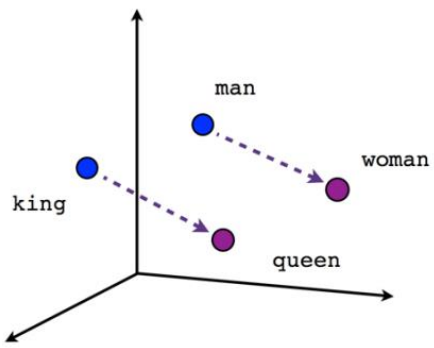
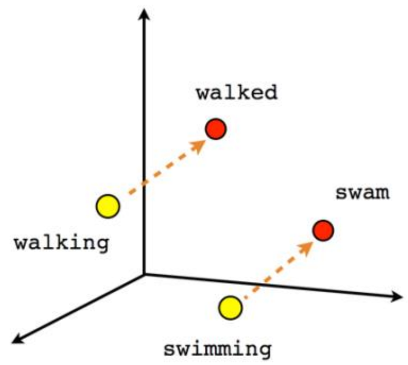
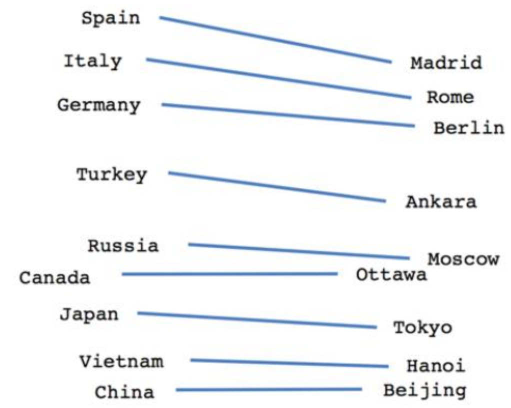


Male-Female



Verb tense



Country-Capital

Lecture 5: Text Representation II
Distributed Representations

Pilsung Kang

School of Industrial Management Engineering

Korea University

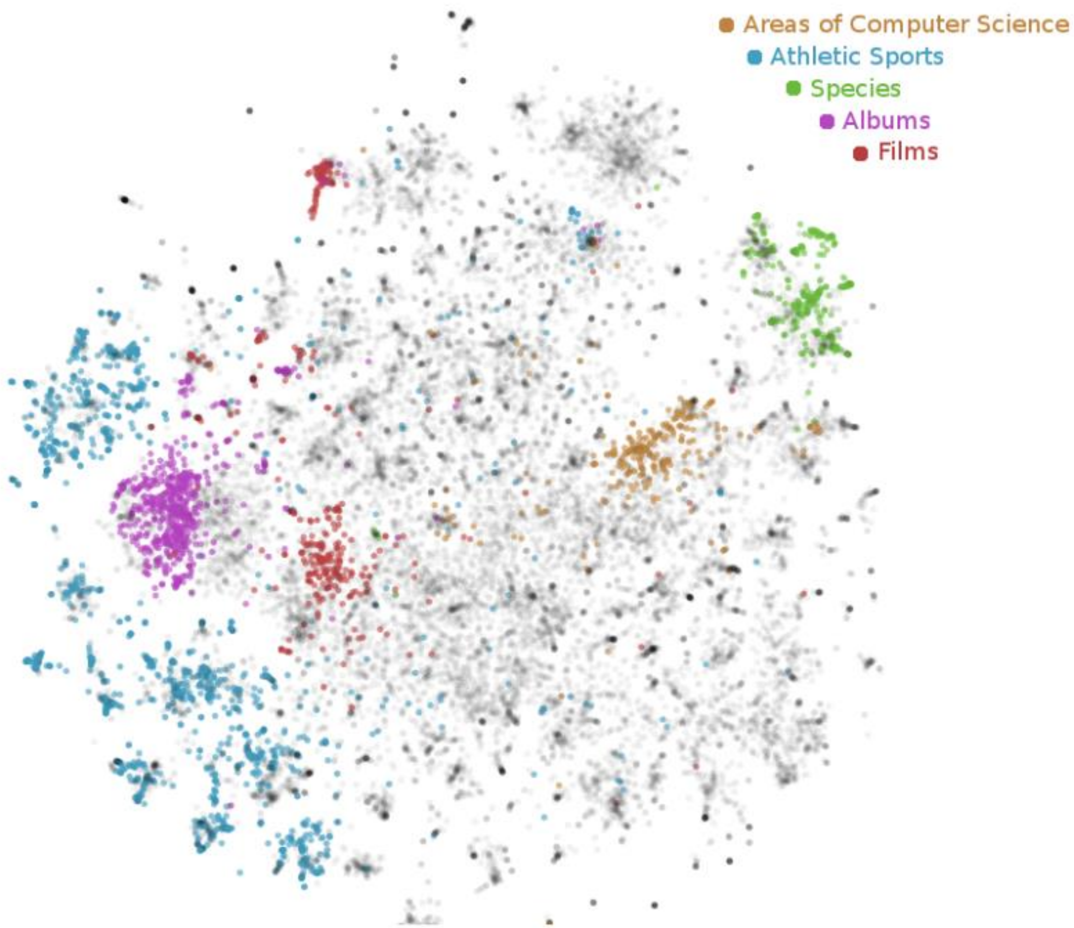
AGENDA

Document Embedding

Dai et al. (2015)

- If we can embed words, why not sentences, phrases, or documents?

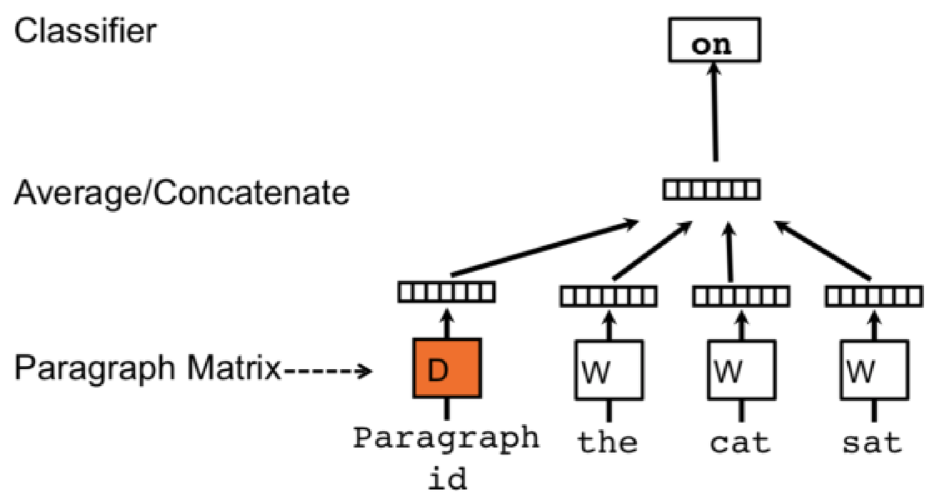
Visualization of Wikipedia paragraph vectors using t-SNE



Document Embedding

Le and Mikolov (2015)

- Paragraph Vector model: Distributed Memory (PV-DM) model



✓ The paragraph vectors are also asked to contribute to the prediction task of the next word given many contexts sampled from the paragraph

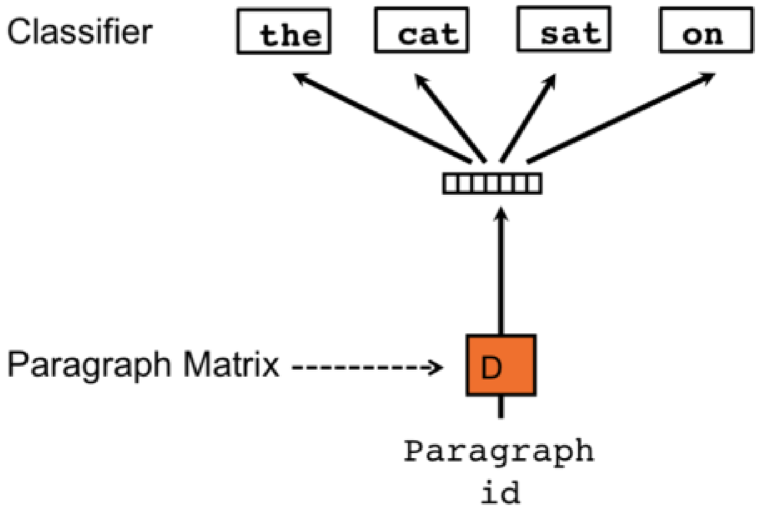
✓ Paragraph vectors are shared for all windows generated from the same paragraph, but not across paragraphs

✓ Word vectors are shared across all paragraphs

Document Embedding

Le and Mikolov (2015)

- Paragraph Vector model: Distributed Bag of Words (PV-DBOW)



✓ Ignore the context words in the input, and force the model to predict words randomly sampled from the paragraph in the output

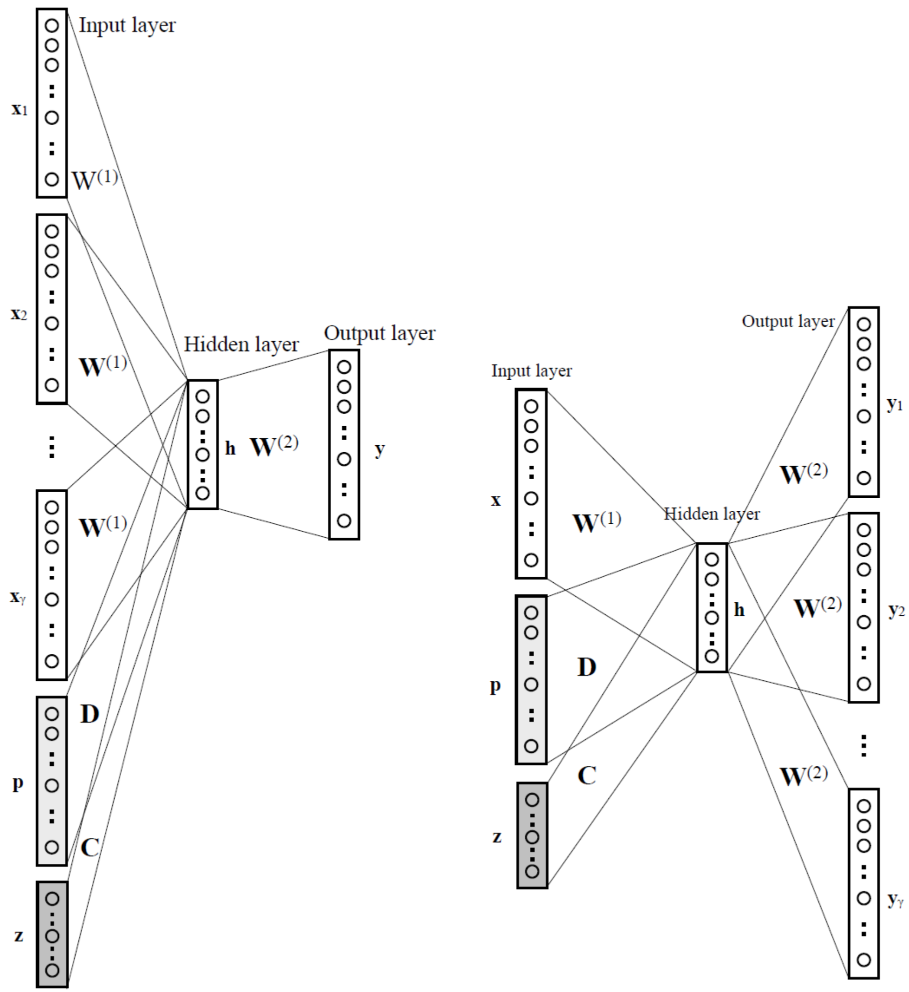
✓ Does not need word vectors

✓ PV-DM alone usually works well for most tasks, but the combination of PV-DM and PV-DBOW are recommended

Let’s Embed Everything!

Park et al. (2016+)

- Supervised Paragraph Vector (SPV) for Class Embedding

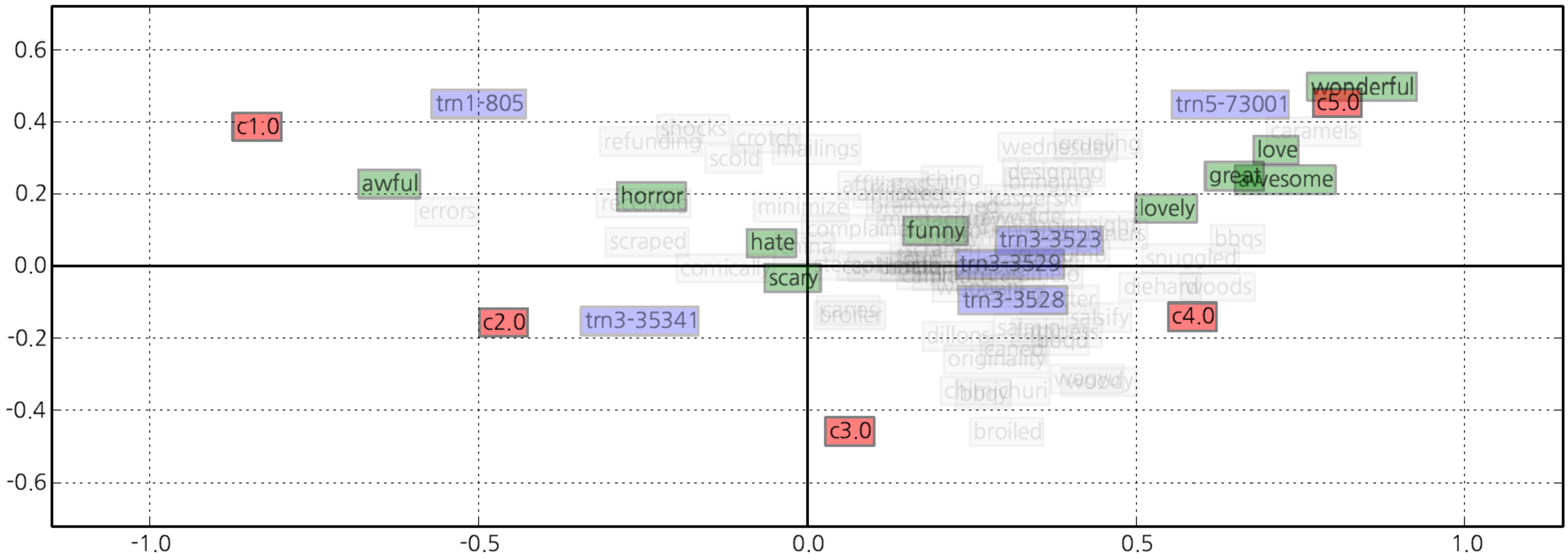


(a) (b)

Let's Embed Everything!

Park et al. (2016+)

- Supervised Paragraph Vector (SPV) for Class Embedding



Let's Embed Everything!

Park et al. (2016+)

- Supervised Paragraph Vector (SPV) for Class Embedding

	imdb							
n(epochs)	1	5	10	30	50	70	100	t
BOW-TF	**85.30**	-	-	-	-	-	-	-
BOW-TFIDF	**85.55**	-	-	-	-	-	-	-
PV-DM	77.06	80.78	80.84	79.68	81.22	81.49	**82.16**	123.14
PV-DBOW	85.89	88.19	**88.47**	88.27	88.22	88.12	88.04	115.22
SPV-DM	82.57	81.68	82.05	**82.66**	82.42	82.53	82.61	121.51
SPV-DBOW	87.58	***88.87**	88.69	88.53	88.51	88.56	88.49	117.33

	yelp							
n(epochs)	1	5	10	30	50	70	100	t
BOW-TF	**58.42**	-	-	-	-	-	-	-
BOW-TFIDF	**58.93**	-	-	-	-	-	-	-
PV-DM	50.59	51.70	52.67	**52.97**	51.73	51.81	52.70	546.10
PV-DBOW	58.53	**59.37**	58.91	59.13	59.10	59.06	59.21	534.03
SPV-DM	51.48	51.42	52.40	53.14	53.77	53.75	**53.84**	552.71
SPV-DBOW	60.13	***60.21**	60.04	59.93	59.85	59.77	59.93	538.95

	amazon							
n(epochs)	1	5	10	30	50	70	100	t
BOW-TF	**85.91**	-	-	-	-	-	-	-
BOW-TFIDF	**85.97**	-	-	-	-	-	-	-
PV-DM	76.47	77.38	78.86	**79.26**	75.83	77.00	78.57	361.52
PV-DBOW	86.87	87.96	88.30	88.34	88.31	**88.42**	88.18	339.14
SPV-DM	79.05	78.35	78.66	80.26	80.36	80.62	**80.70**	371.95
SPV-DBOW	88.58	89.21	89.16	***89.34**	89.08	89.06	89.29	342.40

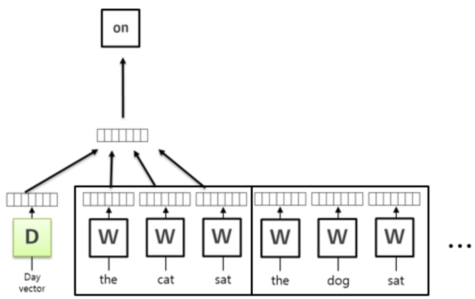
	20news							
n(epochs)	1	5	10	30	50	70	100	t
BOW-TF	**58.58**	-	-	-	-	-	-	-
BOW-TFIDF	**63.43**	-	-	-	-	-	-	-
PV-DM	24.45	41.17	45.36	49.04	52.34	53.85	**54.27**	71.64
PV-DBOW	53.51	69.16	72.51	74.76	75.22	75.16	**75.40**	70.37
SPV-DM	44.76	64.18	67.01	67.84	68.32	**68.34**	67.70	71.88
SPV-DBOW	69.41	76.97	77.95	78.93	78.93	79.47	***79.59**	72.18

AGENDA

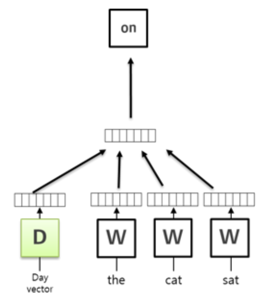
Let’s Embed Everything!

- Day Embedding in News corpus

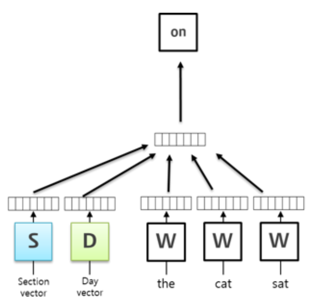
Approach 01: 하루 동안의 기사제목들을 병합 후 Day 태그



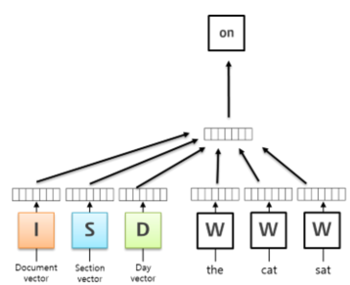
Approach 02: 각 뉴스기사 제목에 Day 태그



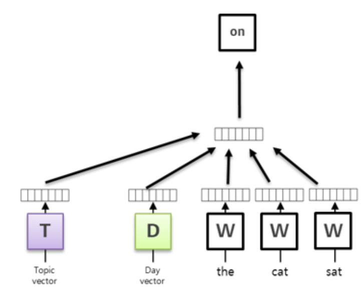
Approach 03: 각 뉴스기사제목에 Day, Section 태그



Approach 04: 각 뉴스기사제목에 Day, Section, IDX 태그

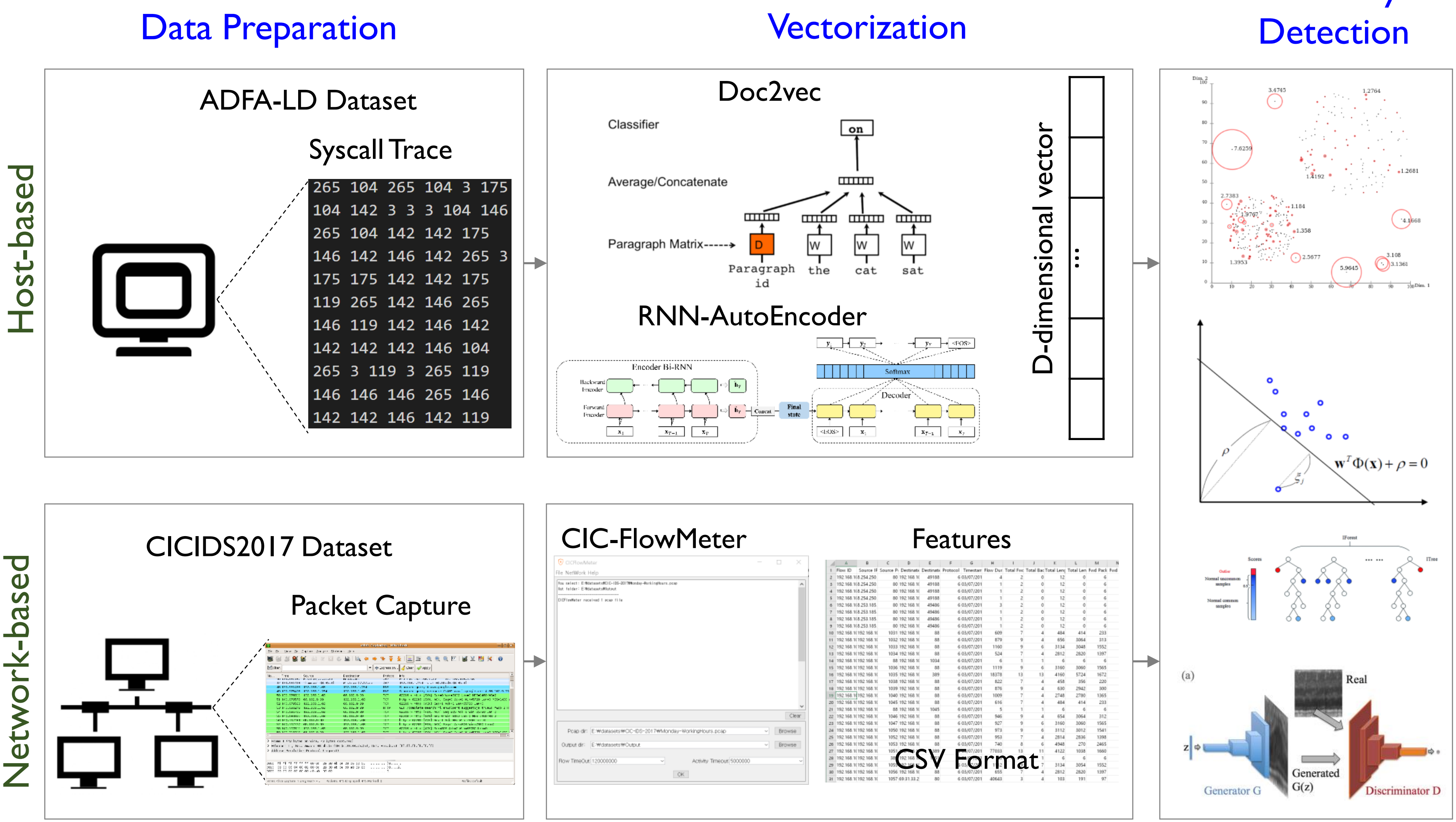


Approach 05: 각 뉴스기사제목에 Day, Topic 태그



Let's Embed Everything!

- System Call Trace Embedding for System Anomaly Detection

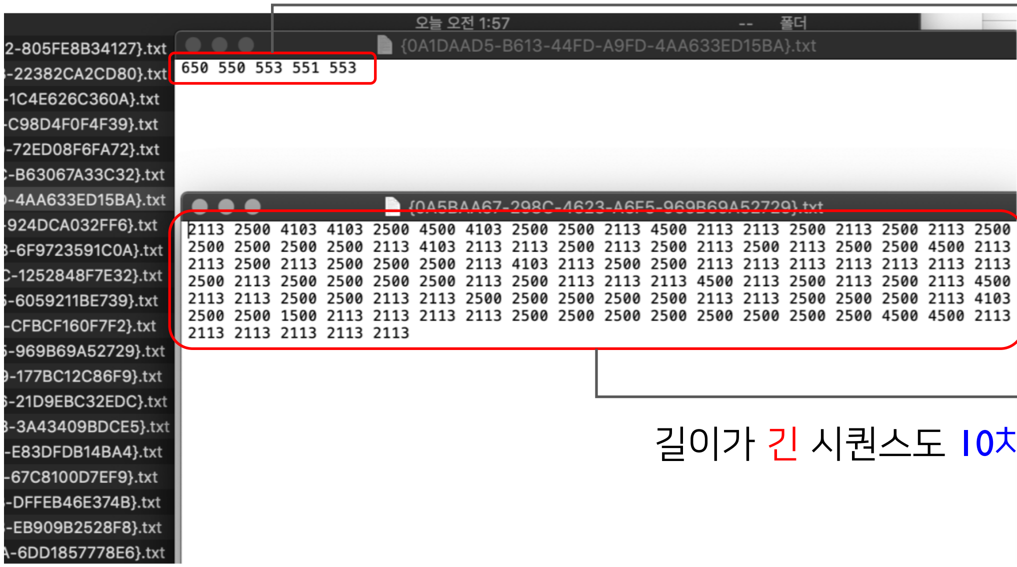


Let's Embed Everything!

- Question
  - ✓ 어떻게 하면 가변 길이의 Syscall Trace를 고정 길이의 벡터로 변환할 수 있을까?

EXAMPLE

길이가 짧은 시퀀스도 10차원 벡터로



0.5	0.2	1.3	2.5	0.7	0.6	1.8	0.4	0.8	0.7

0.7	1.0	4.3	1.8	0.2	1.3	1.1	2.4	1.8	0.5

길이가 긴 시퀀스도 10차원 벡터로

Let's Embed Everything!

- Sequence Embedding based on Doc2Vec
  - ✓ Syscall2Vec: 하나의 System Call Trace를 Document로 취급하고, 개별 syscall을 word로 취급하여 임베딩 수행

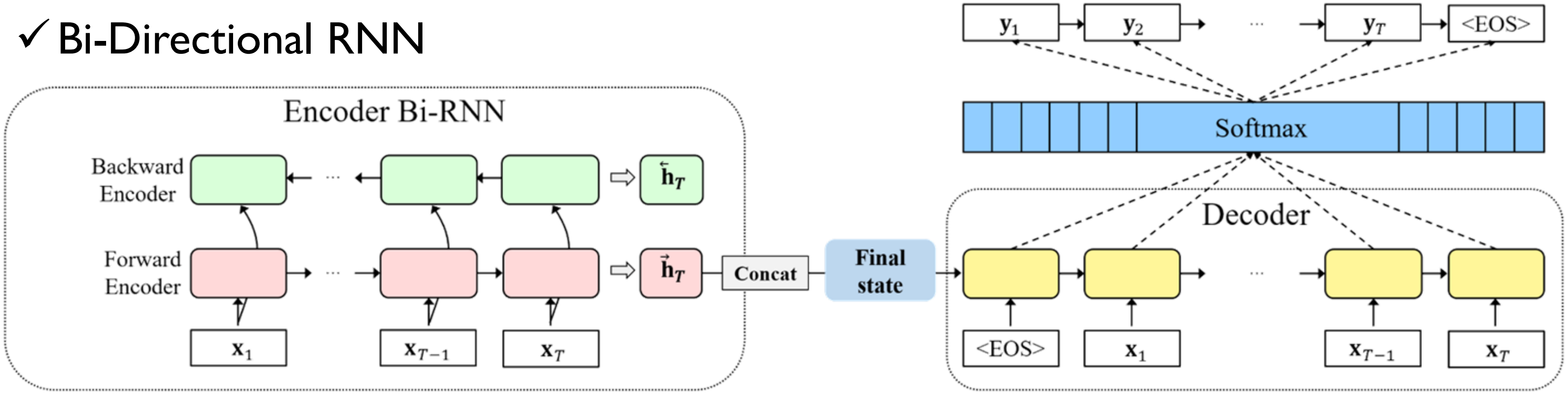
Document

168 3 3 265 168 3 43 168 3 168 168 43 265 168 3 168 43 168 43 168 265 43 265 265 168 265 265 168 168 168 3 168 3 265 168 3 168 168 168 168 3 168 168 168 3 3 168 168 265 168 3 168 265 168 168 3 168 265 43 168 265 43 3 265 43 43 3 ...

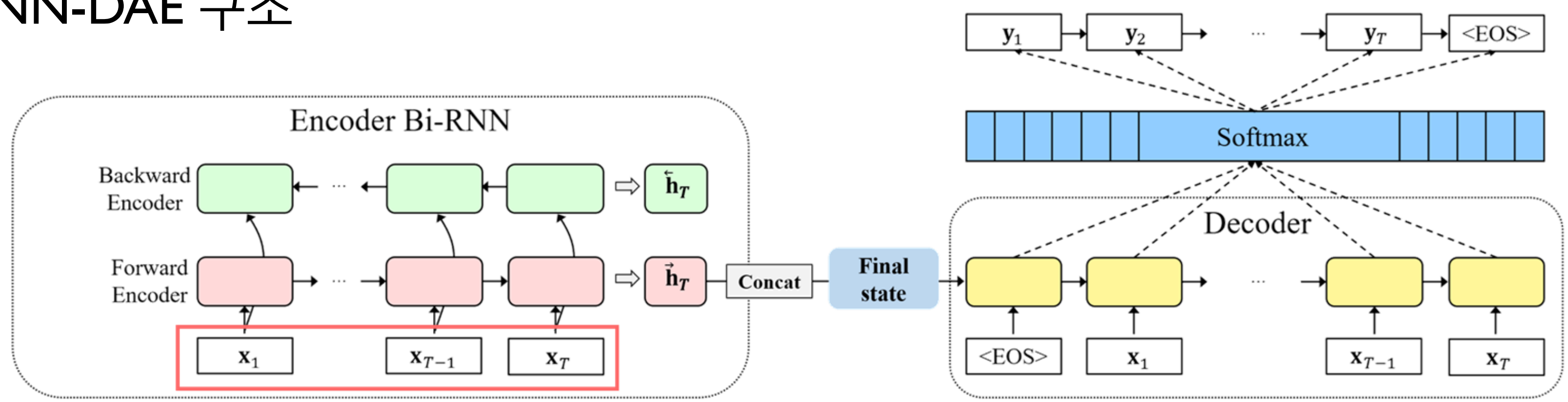
Word 1 Word 2 Word 3 Word 4

Let's Embed Everything!

- RNN-AE 구조
  - ✓ Bi-Directional RNN



- RNN-DAE 구조



Corruption Model

- 임의의 syscall을 p의 확률로 drop
- 임의의 syscall sequence를 permutation

Let’s Embed Everything!

- Live2Vec in afreecaTV

[그림10] 문장과 Live 방송의 추론의 예시

Let’s Embed Everything!

- Live2Vec in afreecaTV

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


ANY
questions?