

# [Paper Review] The Curious Case of Neural Text Degeneration (ICLR 2020)



## Nucleus Sampling ! (Top-p sampling)

### ▼ (0) Abstract

- 인공지능경망 관련한 언어 모델의 성능 급격히 발전함. (이때 당시 GPT-2, GPT-3 )
- 그러나 텍스트 생성을 위한 최적의 디코딩 전략은 여전히 미해결된 문제로 남아있다고 언급
  - ex) 앞서 기존의 최대화 기반 디코딩 방법(Beam Search)이 텍스트 생성에서 일관성 없는 결과를 초래하는 문제
- 이러한 문제를 해결하기 위해 논문에서 "Nucleus Sampling"이라는 간단하면서도 효과적인 디코딩 방법을 제안 !!!
  - 확률 분포의 불안정한 꼬리 부분을 자르고, 확률 질량의 대다수를 포함하는 동적인 핵심(nucleus)에서 샘플링함으로써 이전 디코딩 전략보다 높은 품질의 텍스트 생성 !!
- 최대화 기반과 확률적 디코딩 방법을 인간 텍스트와 비교
  - Nucleus Sampling은 현재로서는 인간 평가를 기준으로 높은 품질을 갖고 동시에 인간이 작성한 텍스트와 유사한 다양성을 제공하는 최고의 디코딩 전략 중 하나로 꼽힘 !

### ▼ (1) Introduction

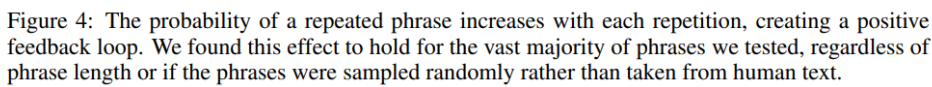
- 2019년에 OpenAI가 발표한 GPT-2.1 기반 논문이 소개 되었는데, 디코딩 방법의 중요성을 강조하면서 무작위성을 활용한 top-k sampling을 언급!
- 그러나 argmax, beam search나 top-k sampling은 텍스트의 품질 및 다양성 측면에서 문제가 있음을 설명
  - 지나치게 generic(일반적인) 문장 생성
  - 반복적이고 긴 문장 생성

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Figure 1 is a line graph showing the probability of Beam Search and Human search over 100 timesteps. The y-axis is labeled 'Probability' and ranges from 0 to 1. The x-axis is labeled 'Timestep' and ranges from 0 to 100. The 'Beam Search' line (blue) is mostly at 1.0, with a few dips. The 'Human' line (orange) fluctuates significantly, often dropping to 0.0 and peaking at 1.0.

확률값 → Human이 생성한 문장은 등장 확률이 낮은  
단어(모델이 선택하지 않은 단어)도 자주 등장



### Pure Sampling:

문맥과 상관 없는 단어들이 자주 등장한다 !! (빨간색 부분)

- 

2

## ▼ (2) Background

### • 2.1 TEXT GENERATION DECODING STRATEGIES

- 최근엔 생성적 적대 신경망(GANs)이 주목받았지만, 품질과 다양성을 함께 고려할 때 GAN이 언어 모델에서 생성된 텍스트보다 성능이 낮다는 연구 결과가 나옴(Caccia et al., 2018; Tevet et al., 2019; Semeniuta et al., 2018)
- unlikelyhood loss 도입 → 토큰에서 train loss를 감소시켜 자주 발생하는 토큰에 대한 gradient를 감소시킨다.
  - negative candidate라고 불리는 특정 토큰들의 확률을 감소시키는 방식

$$\mathcal{L}_{\text{UL-token}}^t(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\alpha \cdot \underbrace{\sum_{c \in \mathcal{C}^t} \log(1 - p_\theta(c|x_{<t}))}_{\text{unlikelyhood}} - \underbrace{\log p_\theta(x_t|x_{<t})}_{\text{likelihood}}.$$

### • 2.2 Directed Generation VS Open-ended Generation

- Directed Generation : [입력문장, 출력 문장] pair 존재
- Open-ended Generation : 입력 문장 존재하지 않거나, context가 주어짐
  - 출력 문장의 길이를 예측할 수 없다. (끝이 정해져 있지 않음) - continuation
  - Greedy, beam과 같은 확률값의 maximization 방식을 사용하면 문제가 발생

## ▼ (3) Language Model Sampling

- Open-ended NLG task → Input, context 주어지면, 이를 바탕으로 자연어를 생성하는 task이다.

$$P(x_{1:m+n}) = \prod_{i=1}^{m+n} P(x_i|x_1 \dots x_{i-1}), \quad (1)$$

m 개의 토큰으로 이뤄진 sequence를 통해 다음 n개의 연속된 토큰을 생성하여 완전한 시퀀스 생성  $x_{m+n} \dots$



특정 decoding 전략을 통해 토큰 생성 !!

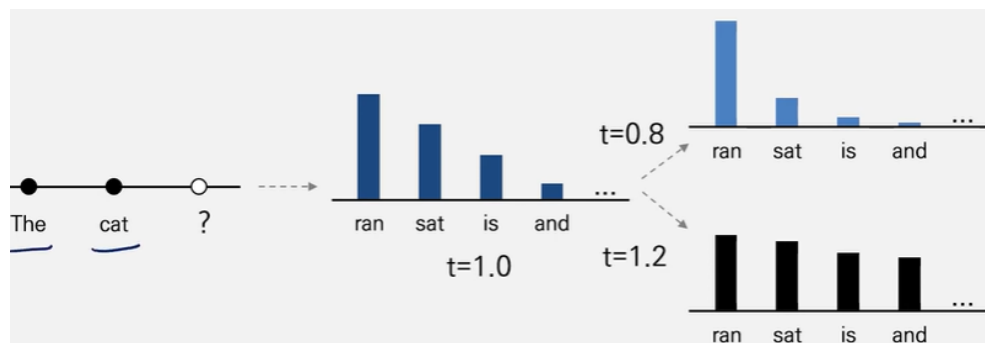
### 1. Sampling with Temperature

$$p(x = V_l | x_{1:i-1}) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u_{l'}/t)}$$

temperature  $t$  파라미터 추가

- Softmax temperature 도입
- Temperature, 0에 가까울수록 one-hot 분포
- 1에 가까울수록 uniform 분포
- 일반적으로 문장 생성시  $t \in [0, 1)$ 을 사용 → 분포 높은 확률 쪽으로 skewed
  - 낮은 확률을 갖는 단어를 덜 샘플링하도록 한다.

▼ 참고 사진



## 2. Top-k sampling

- 단어 분포에서 확률 값 기준 상위  $k$  단어만 선택
- 선택한  $K$ 개를 renormalize해서 새로운 분포를 생성하고 sampling 진행
- GPT2에서 사용 문장 생성 방법으로 우수한 성능을 보이는 방법 (논문  $k = 40$ )
- 적당한  $k$  값 선택 어려움

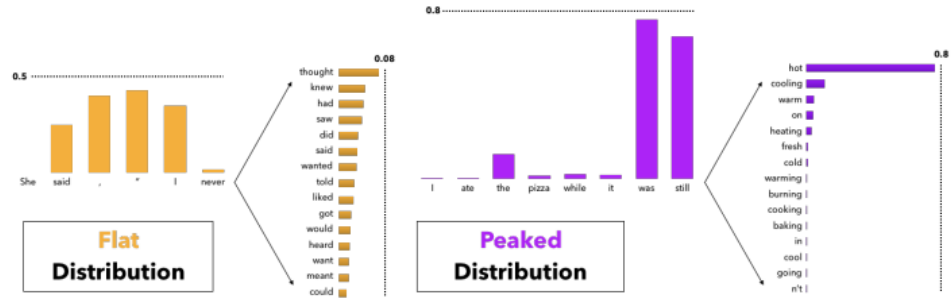


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small  $k$  in top- $k$  sampling problematic, while the presence of peaked distributions makes large  $k$ 's problematic.

확률 차이가 거의 없는 단어들 중에서 포함되지 않을 수 있음 (좌)  
 비교적 확률차이가 많은 단어들이 후보에 포함됨(우)

- $K$  값 작으면, 충분히 뽑힐 가능성이 있는 단어 무시(좌)
- $K$  값 크면, 확률 값이 작은 outlier들이 후보에 포함(우)

### 3. Top-p sampling (Nucleus Sampling)

Top -  $k$  sampling 단점을 보완하고자 이 논문에서 제안 Top-p sampling (Nucleus Sampling)

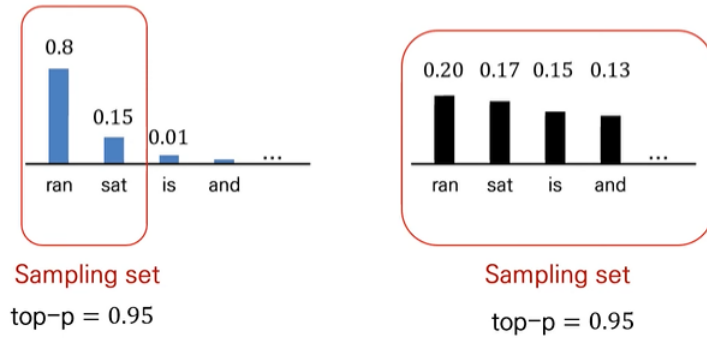
- 단어 분포에서 누적확률 분포 기준 상위  $p$ 까지의 단어 선택
- Nucleus(핵) 명명 이유
  - $p$  높게 설정할 경우(ex 0.95) 선택된 일부 단어에 대부분의 확률값 (probability mass) 집중됨 !!




**확률분포의 모양에 따라 sampling set의 크기가 가변적으로 변한다 !!!**

ex) skewed 된 경우, sampling set 적다! // uniform 형태일 경우, sampling set 비교적 많다 !


▼ 참고 사진




### ▼ Example generations continuing an initial sentence




WebText




Beam Search,  $b=16$




Pure Sampling




Sampling,  $t=0.9$




Top-k,  $k=640$



Top-k,  $k=40$ ,  $t=0.7$



Nucleus,  $p=0.95$



WebText

**An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.**

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Pumping Station #3 shut down due to construction damage Find more at: [www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html](http://www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html)  
 "In the top 10 killer whale catastrophes in history:  
 1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Figure 3: Example generations continuing an initial sentence. Maximization and top- $k$  truncation methods lead to copious repetition (highlighted in blue), while sampling with and without temperature tends to lead to incoherence (highlighted in red). Nucleus Sampling largely avoids both issues.

파란색 : repetition 빨간색 : incoherence(관계 없는 문장) 의미

## ▼ (4) Evaluation

### ▼ Likelihood, distribution statistical Evaluation

GPT2-Large 모델을 사용해서 평가  
Context가 주어진 채로 생성된 5000개 문서(continuation)를  
비교함

#### 1. Perplexity(PPL)

- 모델이 단어를 생성할 때 얼마나 헛갈리는지 나타내는 척도 (낮을수록 좋음 !!)

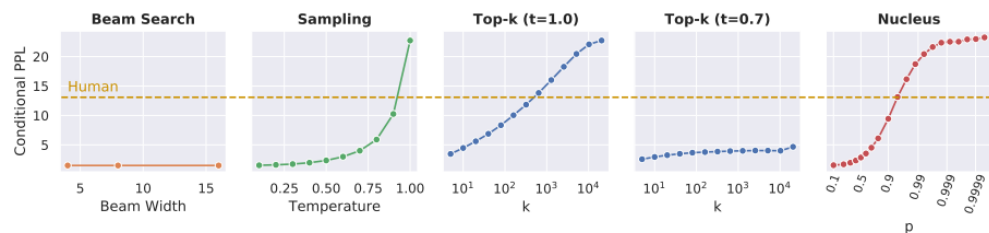


Figure 6: Perplexities of generations from various decoding methods. Note that beam search has unnaturally low perplexities. A similar effect is seen using a temperature of 0.7 with top- $k$  as in both Radford et al. (2019) and Fan et al. (2018). Sampling, Top- $k$ , and Nucleus can all be calibrated to human perplexities, but the first two face coherency issues when their parameters are set this high.

Beam, Top-k ( $t=0.7$ ) x  
나머지 파라미터 잘 조절하면 Human 이상의 성능

#### 2. Zipf Distribution Analysis

지프의 법칙: 텍스트 중 어구별 빈출 순위와 빈도의 관계에서  $k$ 번째로 많은 어구의 빈도가 1번째로 많은 어구의 빈도의  $1/k$ 의 값이 되는 법칙

즉, 모든 단어를 빈도순으로 정렬했을 때, 단어의 빈도는 순위에 반비례한다!

사람이 생성하는 문서 → 지프의 법칙을 따른다고 알려짐

- 각 방식을 사용해서 문서를 생성하고 zipf 분포에 fitting 시,

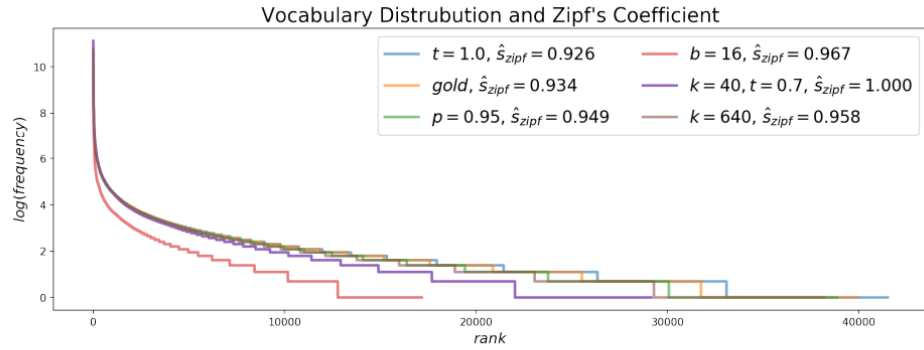


Figure 7: A rank-frequency plot of the distributional differences between  $n$ -gram frequencies of human and machine text. Sampling and Nucleus Sampling are by far the closest to the human distribution, while Beam Search clearly follows a very different distribution than natural language.

pure sampling과 top-p sampling이 실제 문서로 피팅한 zipf 분포 파라미터와 가장 유사함 !!

### 3. Self-BLEU Score

- BLEU score - 일반적으로 translation에서 많이 사용하는 metric
  - 모델이 생성한 문장과 실제 정답 문장간 유사성을 비교하는 척도
  - 서로 겹치는 단어 or n-gram을 이용해 측정
- Self-BLEU score
  - 모델이 unconditional하게 문장 생성 → 1개 문장 측정 대상 선정 후 나머지 4999개를 reference로 삼아서 BLEU score 계산
  - **Self-BLEU score 높으면 모델이 생성한 문장들이 서로 유사한 것 !!**
  - **Self-BLEU score 낮으면 유사성이 적어 diversity가 높은 것 !!**



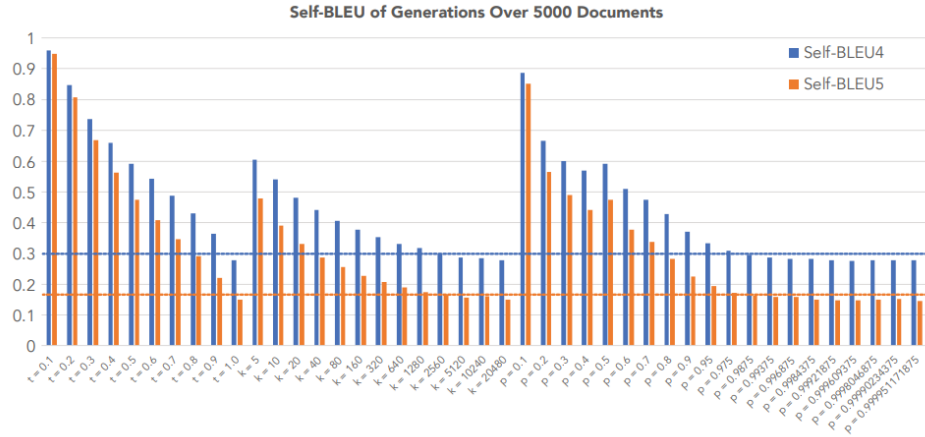


Figure 8: Self-BLEU calculated on the unconditional generations produced by stochastic decoding methods; lower Self-BLEU scores imply higher diversity. Horizontal blue and orange lines represent human self-BLEU scores. Note how common values of  $t \in [0.5, 1]$  and  $k \in [1, 100]$  result in high self-similarity, whereas “normal” values of  $p \in [0.9, 1]$  closely match the human distribution of text.

temperature, top-k, top-p 나와있음

- 중간의 파란색, 주황색 점선 → 사람이 생성한 문장들의 self-BLEU score 의미
- normal value 파라미터 값을 사용하면, 사람이 생성한 문장과 유사한 성능을 낸다 !!

#### 4. Repetition

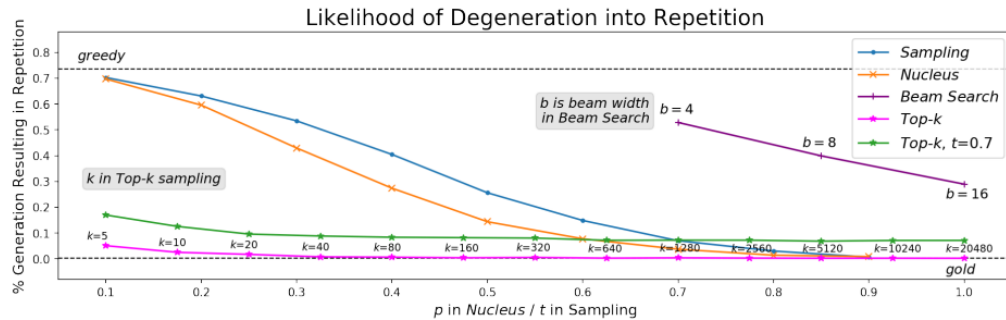


Figure 9: We visualize how often different decoding methods get “stuck” in loops within the first 200 tokens. A phrase (minimum length 2) is considered a repetition when it repeats at least **three** times at the *end* of the generation. We label points with their parameter values except for  $t$  and  $p$  which follow the x-axis. Values of  $k$  greater than 100 are rarely used in practice and values of  $p$  are usually in  $[0.9, 1]$ ; therefore Nucleus Sampling is far closer to the human distribution in its usual parameter range. Sampling with temperatures lower than 0.9 severely increase repetition. Finally, although beam search becomes less repetitive according to this metric as beam width increases, this is largely because average length gets shorter as  $b$  increases (see Appendix A).

모델이 문서 생성하며 첫 200token 동안 반복적인 구(3회 이상)를 생성하는 비율 측정

- top-k : k 100 이하로 할 경우(normal value), repetition이 등장하는 구간 존재
- Sampling temperature : t 0.9 아래로 설정 시, repetition 구간 존재

- top-p : p 0.9이상의 값(normal value) 설정해도, repetition 구간 등장하지 않는다 !!
- Beam search : beam size를 그래프 보다 늘리면 repetition 줄긴 하지만, 생성되는 평균 길이 짧아지는 단점 존재 !!

## ▼ Human Evaluation

### 5. Human Unified with Statistical Evaluation (HUSE)

- 앞선 likelihood, distribution statistical 평가 방식들은 생성된 문장의 일관성(coherence) 평가 x
- 생성된 문장들의 퀄리티 비교를 위해 human evaluation 필수
- 하지만 단점 존재
  - 사람의 평가 → 대부분 생성된 문장의 다양성(diversity) 고려 x
- HUSE score
  - knn 분류기를 학습하고 성능 평가 (모델이 생성, 사람이 생성 구분)



즉 분류기가 구분을 못할 수록(error 값이 높을수록) 모델이 생성한 문장이 사람과 유사한 것으로 해석할 수 있다.

Method	Perplexity	Self-BLEU4	Zipf Coefficient	Repetition %	HUSE
Human	12.38	0.31	0.93	0.28	-
Greedy	1.50	0.50	1.00	73.66	-
Beam, b=16	1.48	0.44	0.94	28.94	-
Stochastic Beam, b=16	19.20	0.28	0.91	0.32	-
Pure Sampling	22.73	0.28	<b>0.93</b>	0.22	0.67
Sampling, t=0.9	10.25	0.35	0.96	0.66	0.79
Top-k=40	6.88	0.39	0.96	0.78	0.19
Top-k=640	13.82	<b>0.32</b>	0.96	<b>0.28</b>	0.94
Top-k=40, t=0.7	3.48	0.44	1.00	8.86	0.08
Nucleus p=0.95	<b>13.13</b>	<b>0.32</b>	0.95	0.36	<b>0.97</b>

Table 1: Main results for comparing all decoding methods with selected parameters of each method. The numbers closest to human scores are in **bold** except for HUSE (Hashimoto et al., 2019), a combined human and statistical evaluation, where the highest (best) value is **bolded**. For Top-k and Nucleus Sampling, HUSE is computed with interpolation rather than truncation (see §6.1).

### 성능 결과값

- HUSE score → top-p가 가장 높고, top-k sampling(k=640) 두번째로 높음 (크게 별 차이 없다 !!)
- 대부분 Nucleus sampling이 성능이 좋다. (Top-k=640도 성능이 꽤나 높게 나온다 !!)

## ▼ (5) Conclusion

결론 : Nucleus Sampling 성능 좋다 !!

top-p sampling으로 하는 것도 성능이 좋게 나오지만, top-k sampling으로도 파라미터만 잘 조절하면 top-p sampling 못지 않게 성능이 잘 나오는 것 같아서 상황에 맞게 쓰는 게 가장 좋을 것 같다 !

## Referance

<https://arxiv.org/pdf/1904.09751.pdf>

<http://dsba.korea.ac.kr/seminar/?mod=document&uid=1345>

<https://sleekdev.tistory.com/19>