EWHA WOMANS UNIVERSITY, Department of Statistics

Computational Statistics Final Project

— Using MCMC to break classical ciphers

2조 | 222STG24 김민지 | 222STG25 이다은 | 222STG26 이은효

EWHA WOMANS UNIVERSITY, Department of Statistics

INDEX

01 **Introduction**

주제 선정 동기

MCMC Background

Cipher text Background

02 **본론1 – Decryption using MCMC**

Metropolis Algorithm in Decryption

Decipher Algorithm

03 **본론2 – Optimize and Apply**

Tuning parameter

Apply to Exercise

04 **Conclusion**

결론 및 보완점

참고 문헌

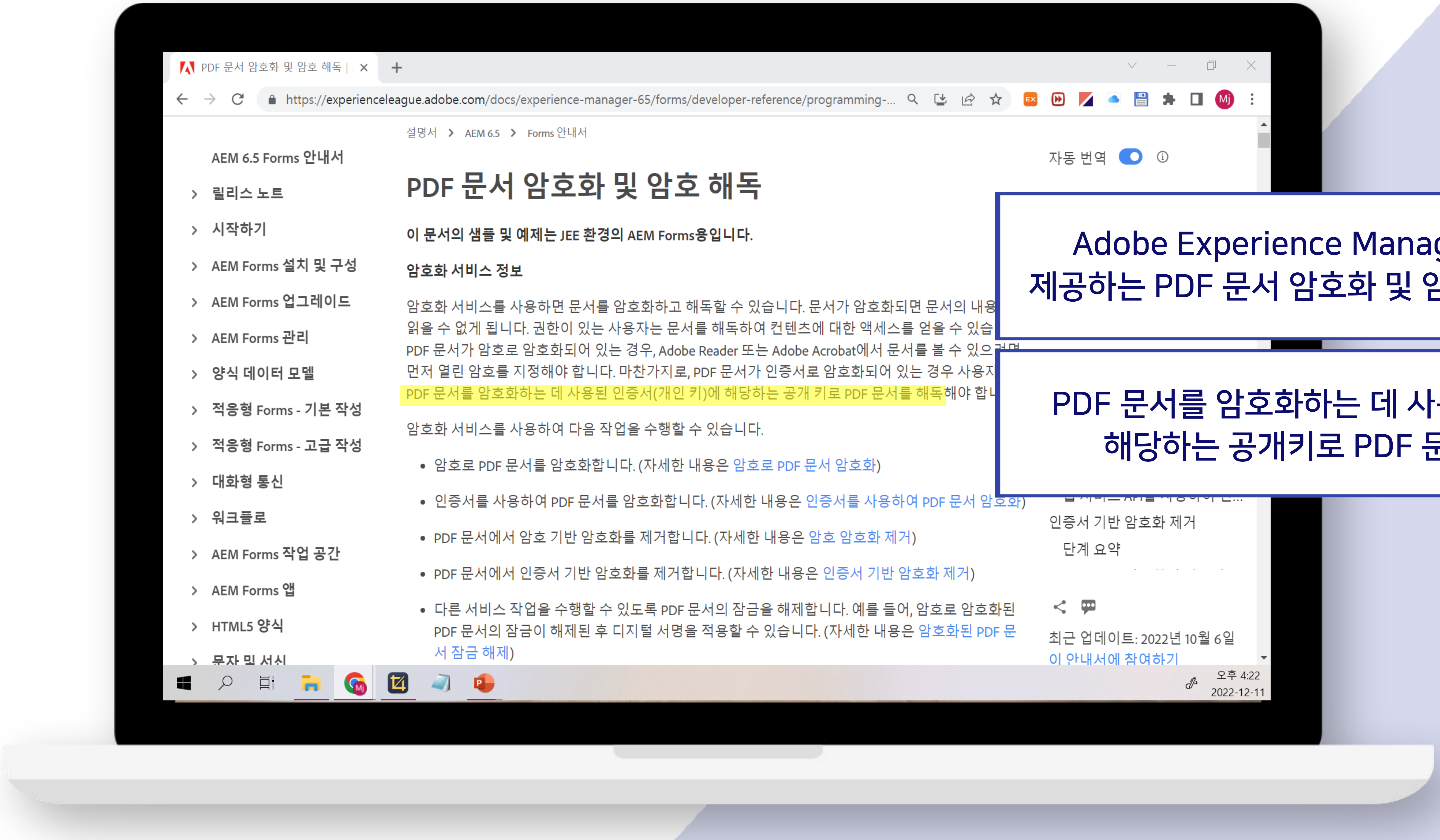
EWHA WOMANS UNIVERSITY, Department of Statistics

01

Introduction

1) 주제 선정 동기
2) MCMC Background
3) Cipher text Background

01 주제 선정 동기



PDF 문서 암호화 및 암호 해독

이 문서의 샘플 및 예제는 JEE 환경의 AEM Forms용입니다.

암호화 서비스 정보

암호화 서비스를 사용하면 문서를 암호화하고 해독할 수 있습니다. 문서가 암호화되면 문서의 내용 읽을 수 없게 됩니다. 권한이 있는 사용자는 문서를 해독하여 컨텐츠에 대한 액세스를 얻을 수 있습 PDF 문서가 암호로 암호화되어 있는 경우, Adobe Reader 또는 Adobe Acrobat에서 문서를 볼 수 있으 먼저 열린 암호를 지정해야 합니다. 마찬가지로, PDF 문서가 인증서로 암호화되어 있는 경우 사용지 PDF 문서를 암호화하는 데 사용된 인증서(개인 키)에 해당하는 공개 키로 PDF 문서를 해독해야 합니

암호화 서비스를 사용하여 다음 작업을 수행할 수 있습니다.

- 암호로 PDF 문서를 암호화합니다. (자세한 내용은 암호로 PDF 문서 암호화)
- 인증서를 사용하여 PDF 문서를 암호화합니다. (자세한 내용은 인증서를 사용하여 PDF 문서 암호화)
- PDF 문서에서 암호 기반 암호화를 제거합니다. (자세한 내용은 암호 암호화 제거)
- PDF 문서에서 인증서 기반 암호화를 제거합니다. (자세한 내용은 인증서 기반 암호화 제거)
- 다른 서비스 작업을 수행할 수 있도록 PDF 문서의 잠금을 해제합니다. 예를 들어, 암호로 암호화된 PDF 문서의 잠금이 해제된 후 디지털 서명을 적용할 수 있습니다. (자세한 내용은 암호화된 PDF 문서 잠금 해제)

최근 업데이트: 2022년 10월 6일

이 안내서에 참여하기

01 Adobe Experience Manager 6.5에서 제공하는 PDF 문서 암호화 및 암호 해독 서비스

02 PDF 문서를 암호화하는 데 사용된 개인 키에 해당하는 공개키로 PDF 문서를 해독

02

MCMC Background

— Metropolis Algorithm

Step 1
Initial state

Step 2
Propose

Step 3
M Ratio

Step 4
Sample value

Choose an initial state $X^{(0)} \in X$

Propose a new state $X^{*} \in X$ from symmetric proposal density

Calculate acceptance probability

$$R\left(X^{(t)}, X^{*}\right) = min\{1, (\pi(X^{*})/\pi(X^{(t)}))^{p}\}$$

* p: scaling parameter

Sample a value for

$$X^{(t+1)} = \begin{cases} X^{*} & U_t < min\{R(X^{(t)}, X^{*}), 1\} \\ X^{(t)} & otherwise \end{cases}$$

03 Cipher Background

Simple substitution cipher	plain text의 원래 문자를 다른 문자로 대체하는 방식으로, 각 문자로는 알파벳을 사용
accuracy	$\frac{m_s}{n_s} = \frac{\text{the number of letters correctly revealed}}{\text{the number of available letters in the plain text}}$
$score(f)$	$\prod_{t=1}^{N-1} M_{f(c_i)f(c_{i+1})}$
$M_{f(c_i)f(c_{i+1})}$	pair(c_i,c_{i+1}) 즉, 문자c_i 바로 뒤에 문자 c_{i+1}가 오는 확률
Target function	$\pi(f) = \frac{score(f)}{\sum_g score(g)}$

02

본론1 – Decryption using MCMC

1) Metropolis Algorithm in Decryption
2) Decipher Algorithm

01 Metropolis Algorithm in Decryption

Step 1
Initial state

Step 2
Propose

Step 3
M Ratio

Step 4
Sample value

Choose an initial decryption key f, and fixed scaling parameter $p > 0$

Pick two letters uniformly at random from f and switch the values call proposal key f^*

Calculate acceptance probability

$$R\left(X^{(t)}, X^*\right) = \{score(f^*)/score(f)\}^p$$

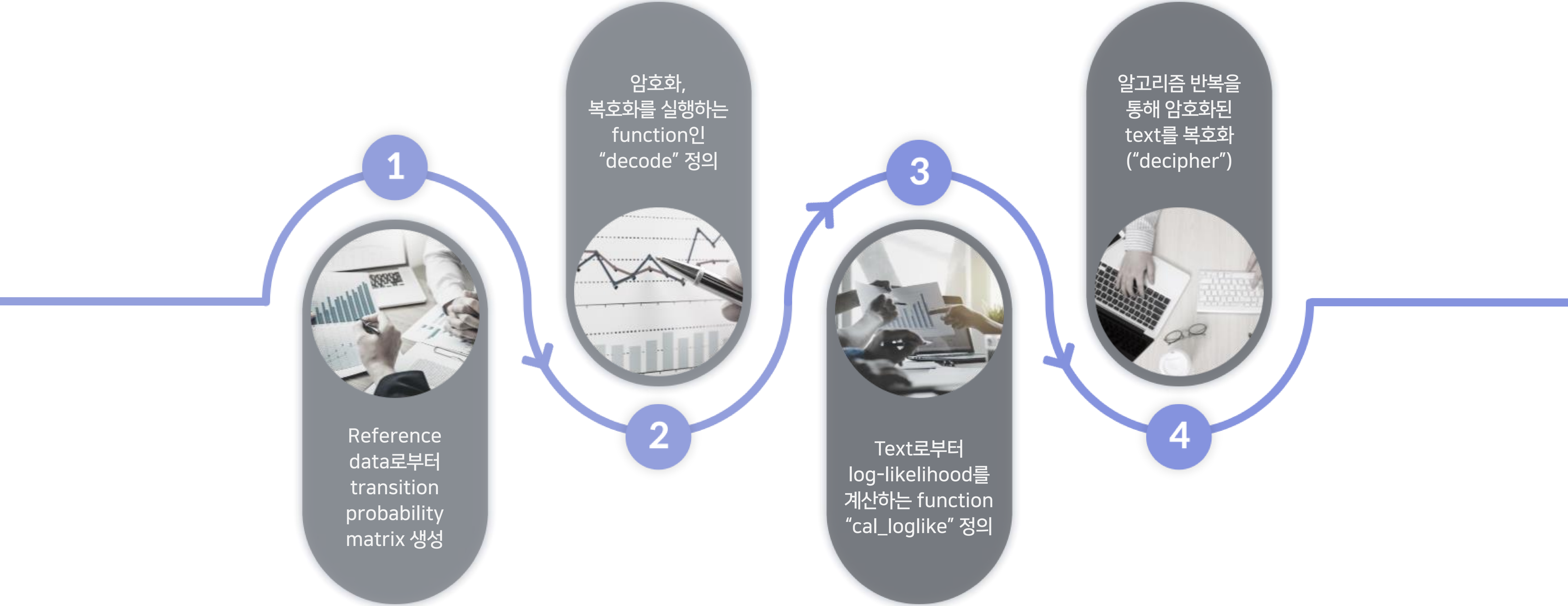
* $score(f^*)$: proposal prob

* $score(f)$: current prob

Using key function

$$f^{(new)} = \begin{cases} accept\ f^* & U_t \le min\{\,R(f, f^*), 1\} \\ remain\ f & otherwise \end{cases}$$

02 Decipher Algorithm



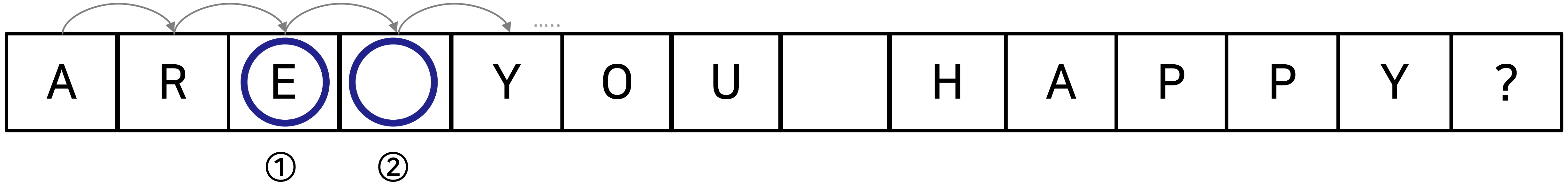
1. Reference data로부터 transition probability matrix 생성
2. 암호화, 복호화를 실행하는 function인 “decode” 정의
3. Text로부터 log-likelihood를 계산하는 function “cal_loglike” 정의
4. 알고리즘 반복을 통해 암호화된 text를 복호화 (“decipher”)

02

Decipher Algorithm

— Step 1: Transition Probability Matrix

(예시) Reference data를
한 줄 씩 나눈 데이터



A	R	E		Y	O	U		H	A	P	P	Y	?
		①	②										

① \ ②	A	B	…	P	…	Y	Z	' '
A	0	3		1		8	0	0
B	4	0		2		5	1	2
…								
E	9	2		1		6	3	2
…								
Y	0	3		3				1
Z	2	5		2		0	7	9
' '	6	2		9		2	5	2

Transition Matrix (27X27)

	A	B	…	P	…	Y	Z	' '
A	0.2	0.2		0.23		0.18	0.06	0.62
B	0.1							
…								
P	0.01	0.3		0.01		0.25	0.28	0.3
…								
Y	0.12	0.05		0.23		0.3	0.2	0.02
Z	0.01	0.2		0.13		0.2	0.1	0.31
' '	0.03	0.02		0.16		0.5	0.2	0.3

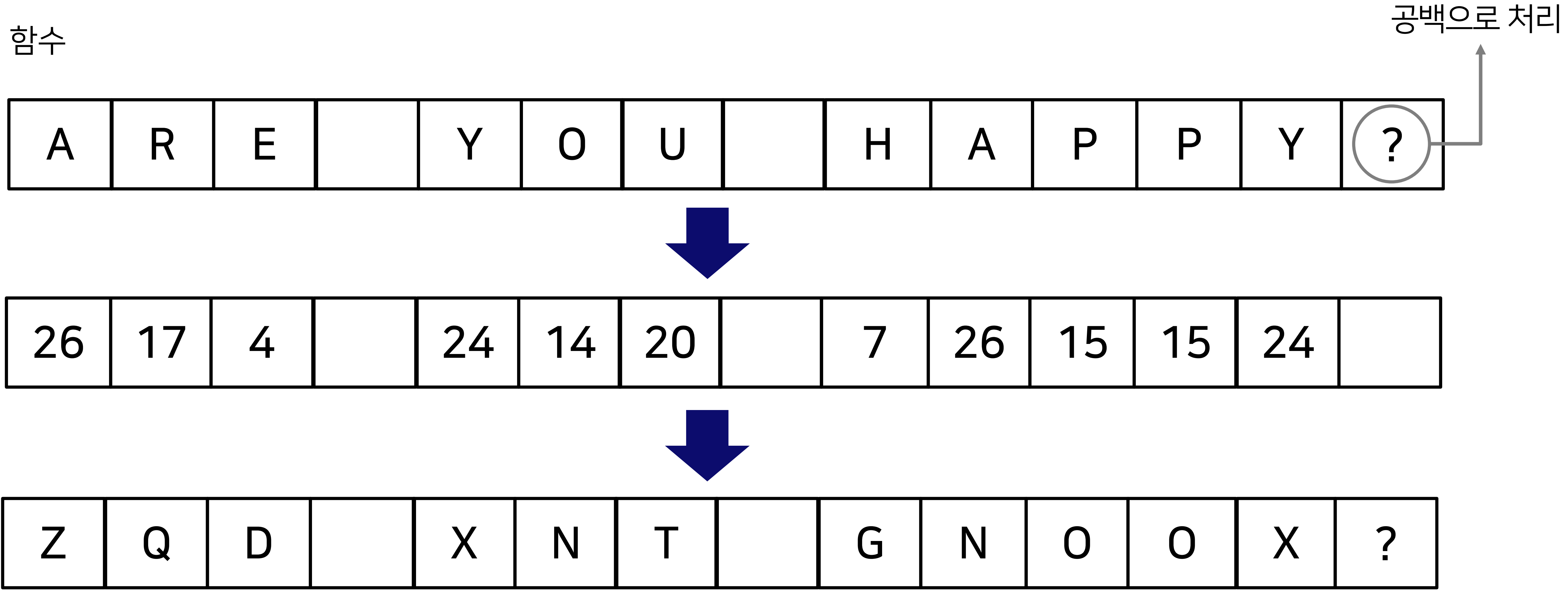
Transition Probability Matrix
= (각 행렬의 값) / (각 행의 합)

02

Decipher Algorithm

— Step 2: “decode” function

decode : 암호화, 복호화를 진행하는 함수



공백으로 처리

	A	R	E		Y	O	U		H	A	P	P	Y	?
mapping order	26	17	4		24	14	20		7	26	15	15	24	
LETTERS	Z	Q	D		X	N	T		G	N	O	O	X	?

LETTERS

A, B, C, D, E, F, G, H, I, J, K, L, M, N,
O, P, Q, R, S, T, U, V, W, X, Y, Z

: 알파벳 ‘A’-‘Z’를 나열한 배열

mapping

B, C, D, E, F, G, H, I, J, K, L, M, N,
O, P, Q, R, S, T, U, V, W, X, Y, Z, A

: LETTERS를 랜덤으로 재배열

02

Decipher Algorithm

— Step 3: “cal_loglike” function

암호화된 문장

Z	Q	D		X	N	T		G	N	O	O	X	?

① ② ③

	A	B	···	Q	···	Y	Z	‘ ’
···	0.2	0.2	···	[illegible]	···	0.18	0.06	0.62
D	0.1	0.2	···	[illegible]	···	0.2	0.03	0.15
···	···	···	···		···	···	···	···
X	0.01	0.3	···	[illegible]	···	0.25	0.28	0.3
···	···	···	···		···	···	···	···
Y	0.12	[illegible]	[illegible]	[illegible]	···	0.3	0.2	0.02
Z	[illegible]	[illegible]	[illegible]	0.13	···	0.2	0.1	0.31
‘ ’	0.03	0.02	···	0.16	···	0.5	0.2	0.3

Z → Q

Transition Probability Matrix

① 현재 문자가 알파벳인 경우

② 현재 문자가 알파벳이 아니고, 이전 문자가 공백이 아닌 경우

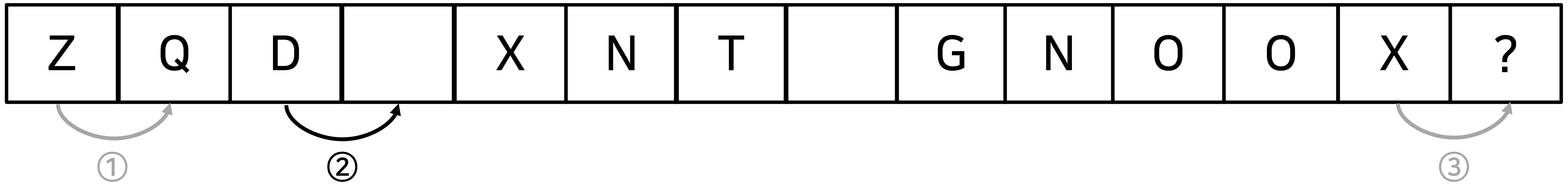
③ 문장이 알파벳으로 끝나거나 특수문자로 끝나는 경우

Transition Probability Matrix에서 해당하는 값에 로그를 취함
⇒ log-likelihood

02

Decipher Algorithm

— Step 3: “cal_loglike” function



암호화된 문장

Z	Q	D		X	N	T		G	N	O	O	X	?

	A	B	···	Q	···	Y	Z	' '
···	0.2	0.2	···	0.23	···	[illegible]	[illegible]	[illegible]
D	0.1	0.2		[illegible]		0.2	[illegible]	0.15
···	···	···	···	···	···	···	···	···
X	0.01	0.3	···	0.01	···	0.25	0.28	0.3
···	···	···	···	···	···	···	···	···
Y	0.12	0.05		0.23	···	0.3	0.2	0.02
Z	0.01	0.2	···	0.13	···	0.2	0.1	0.31
' '	0.03	0.02	···	0.16	···	0.5	0.2	0.3

D→ ' '

Transition Probability Matrix

① 현재 문자가 알파벳인 경우

② 현재 문자가 알파벳이 아니고, 이전 문자가 공백이 아닌 경우

③ 문장이 알파벳으로 끝나거나 특수문자로 끝나는 경우

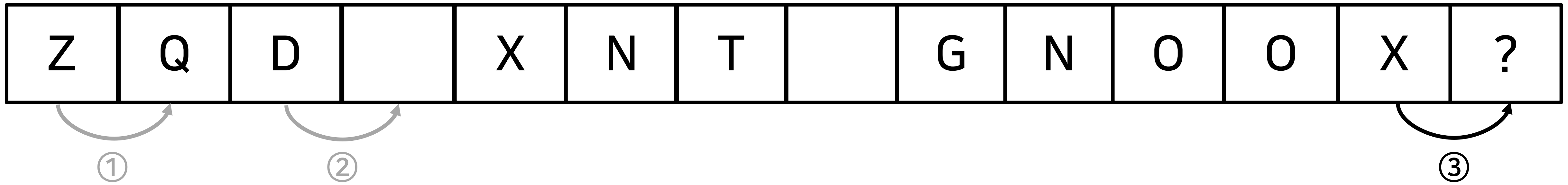
Transition Probability Matrix에서 해당하는 값에 로그를 취함
⇒ log-likelihood

02

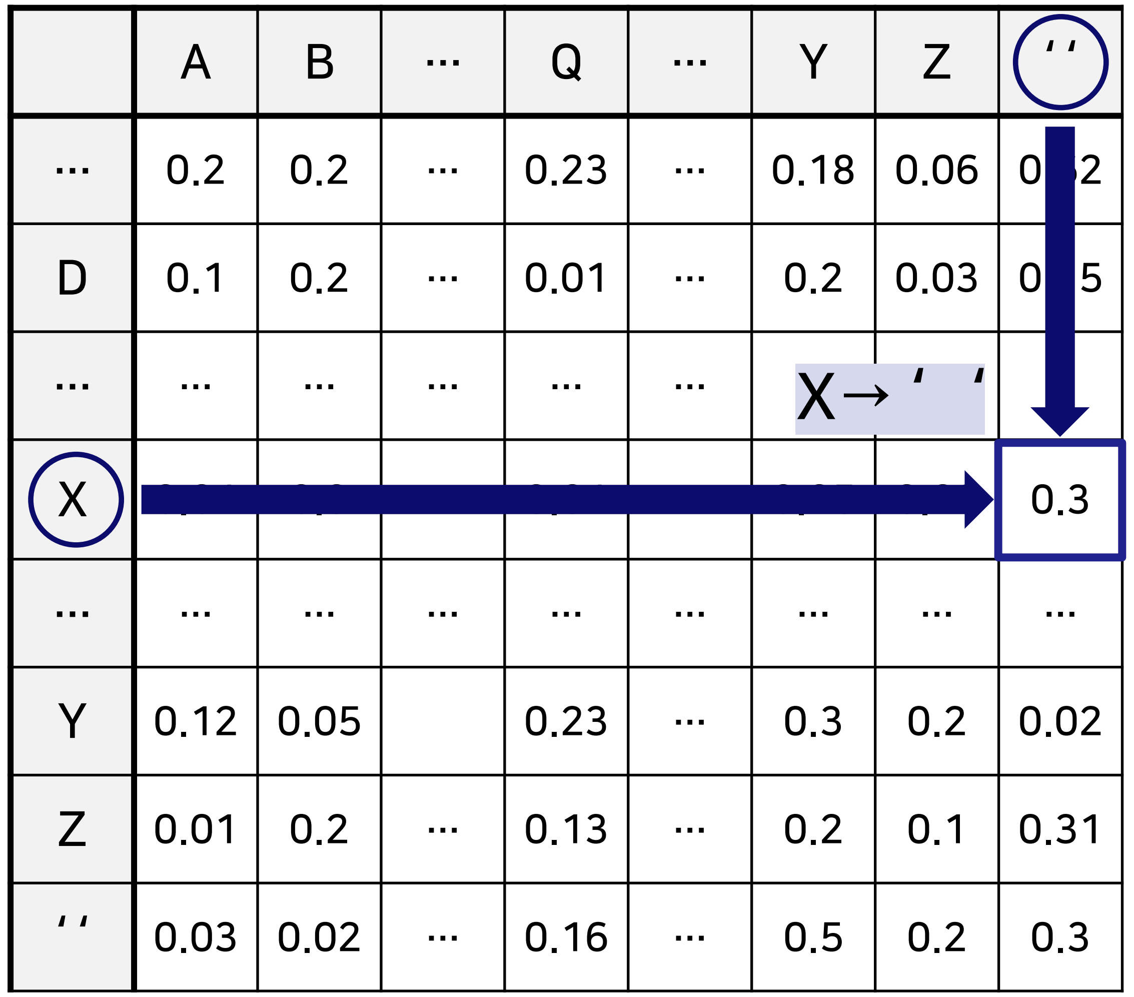
Decipher Algorithm

— Step 3: “cal_loglike” function

암호화된 문장



Z	Q	D		X	N	T		G	N	O	O	X	?



	A	B	…	Q	…	Y	Z	‘ ’
…	0.2	0.2	…	0.23	…	0.18	0.06	0.[illegible]2
D	0.1	0.2	…	0.01	…	0.2	0.03	0.[illegible]5
…	…	…	…	…	…			
X	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.3
…	…	…	…	…	…	…	…	…
Y	0.12	0.05		0.23	…	0.3	0.2	0.02
Z	0.01	0.2	…	0.13	…	0.2	0.1	0.31
‘ ’	0.03	0.02	…	0.16	…	0.5	0.2	0.3

Transition Probability Matrix

① 현재 문자가 알파벳인 경우

② 현재 문자가 알파벳이 아니고, 이전 문자가 공백이 아닌 경우

③ 문장이 알파벳으로 끝나거나 특수문자로 끝나는 경우

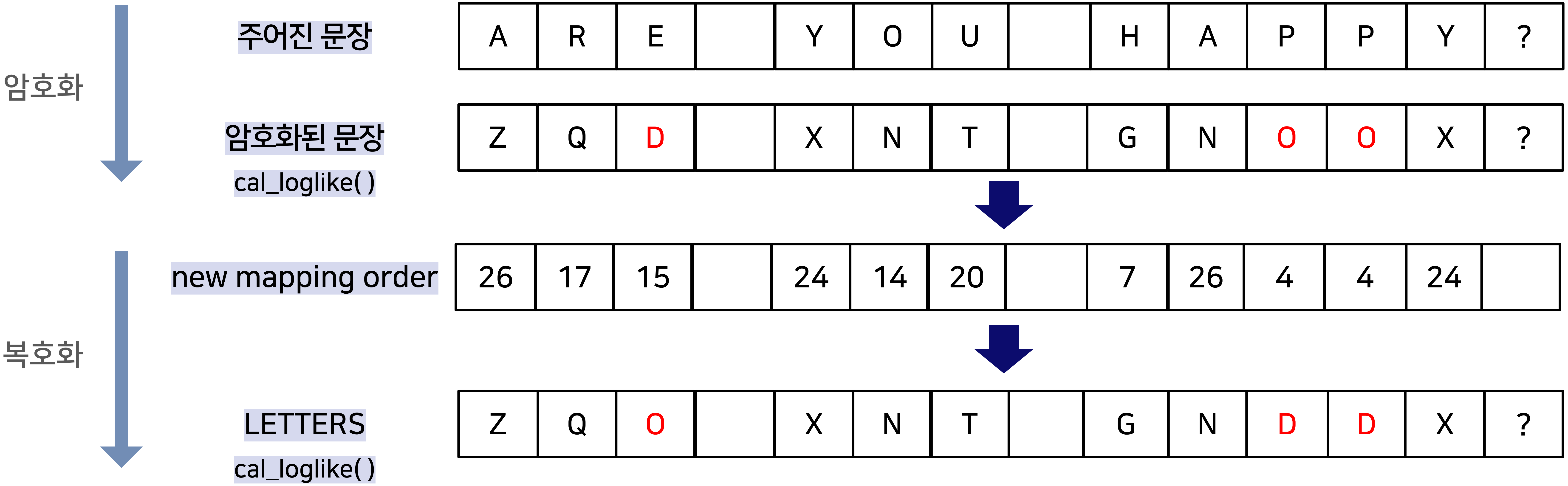
- 특수문자의 경우 공백으로 간주

Transition Probability Matrix에서 해당하는 값에 로그를 취함
⇒ log-likelihood
⇒ (① + ② + ③)에서 log-likelihood를 구한 값이 최종 log-likelihood

02

Decipher Algorithm

— Step 4: "decipher" function



주어진 문장	A	R	E		Y	O	U		H	A	P	P	Y	?
암호화된 문장 cal_loglike( )	Z	Q	D		X	N	T		G	N	O	O	X	?
new mapping order	26	17	15		24	14	20		7	26	4	4	24	
LETTERS cal_loglike( )	Z	Q	O		X	N	T		G	N	D	D	X	?

복호화 후 새로운 log-likelihood와
이전 log-likelihood의 차이가 임계치를
넘을 시 변수 업데이트

→ 이 경우를 count하여 정해진 값에
도달하기 전까지 "복호화 " 과정 반복

LETTERS

A, B, C, D, E, F, G, H, I, J, K, L, M, N,
O, P, Q, R, S, T, U, V, W, X, Y, Z

new mapping

B, C, O, E, F, G, H, I, J, K, L, M, N,
D, P, Q, R, S, T, U, V, W, X, Y, Z, A

기존 mapping에서 임의의 문자 2개('D', 'O')를 swap

EWHA WOMANS UNIVERSITY, Department of Statistics

03

본론2 – Optimize and Apply

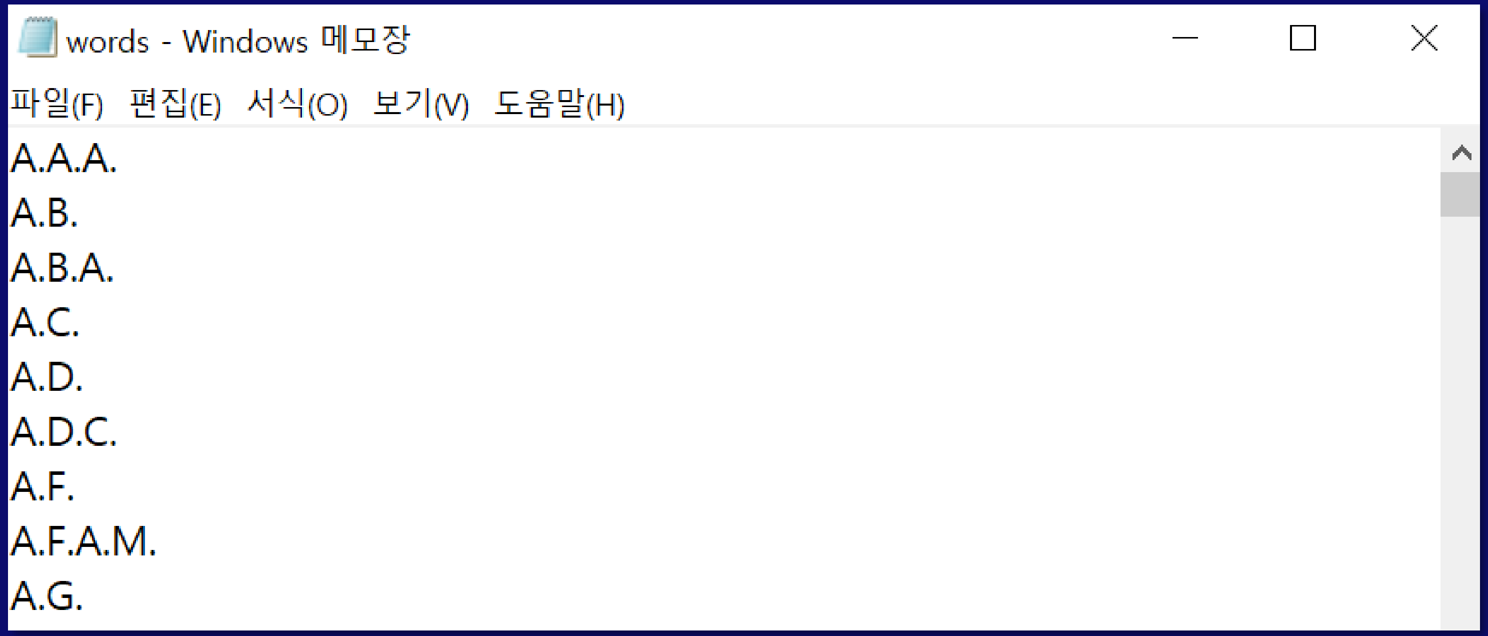
1) Tuning Parameter
2) Apply to Exercise

01

Tuning Parameter

1) Reference Data

words.txt



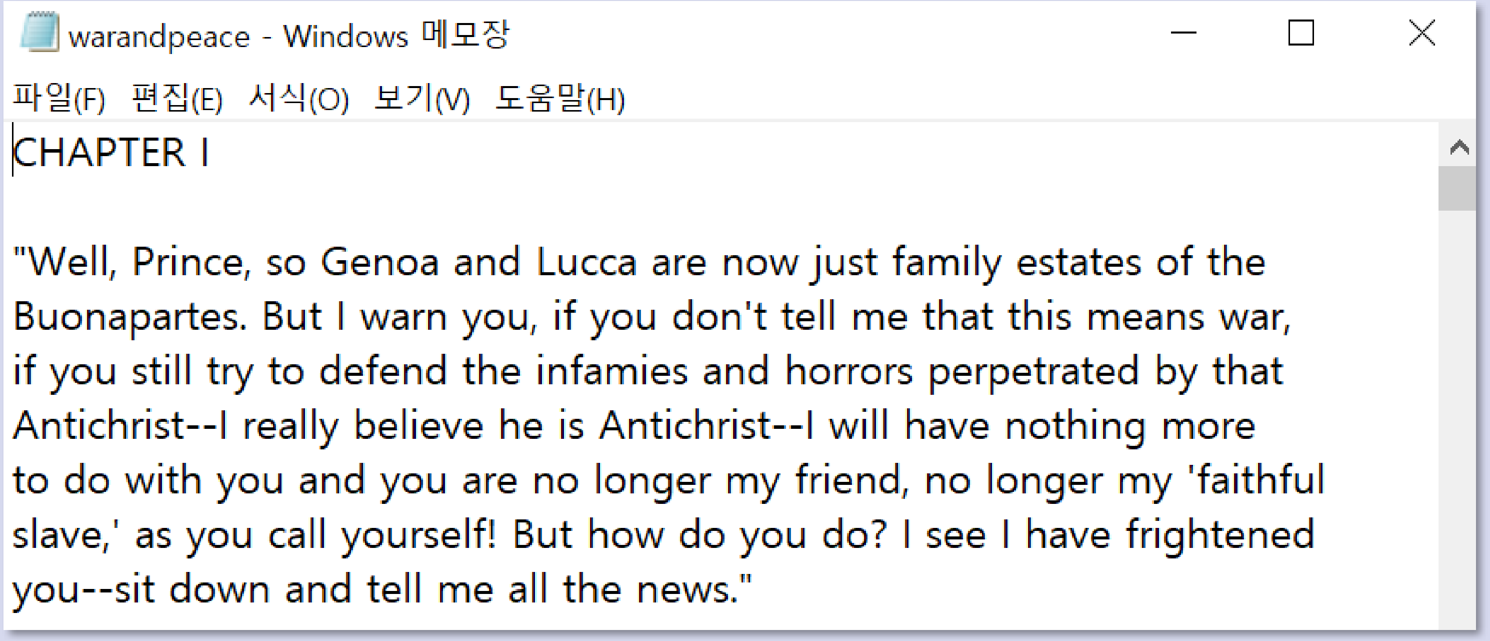
```
A.A.A.
A.B.
A.B.A.
A.C.
A.D.
A.D.C.
A.F.
A.F.A.M.
A.G.
```

4,396,442 letters | 466,553 words

문맥이 없고 약자가 포함된 단어장

https://github.com/dwyl/english-words/blob/master/words.txt

warandpeace.txt


warandpeace - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

CHAPTER I

"Well, Prince, so Genoa and Lucca are now just family estates of the
Buonapartes. But I warn you, if you don't tell me that this means war,
if you still try to defend the infamies and horrors perpetrated by that
Antichrist--I really believe he is Antichrist--I will have nothing more
to do with you and you are no longer my friend, no longer my 'faithful
slave,' as you call yourself! But how do you do? I see I have frightened
you--sit down and tell me all the news."

3,138,459 letters | 564,428 words

문맥이 있고 약자가 거의 없는 소설

Leo Tolstoy, *War and Peace* (1869)

01

Tuning Parameter

1) Reference Data

— Result Table —

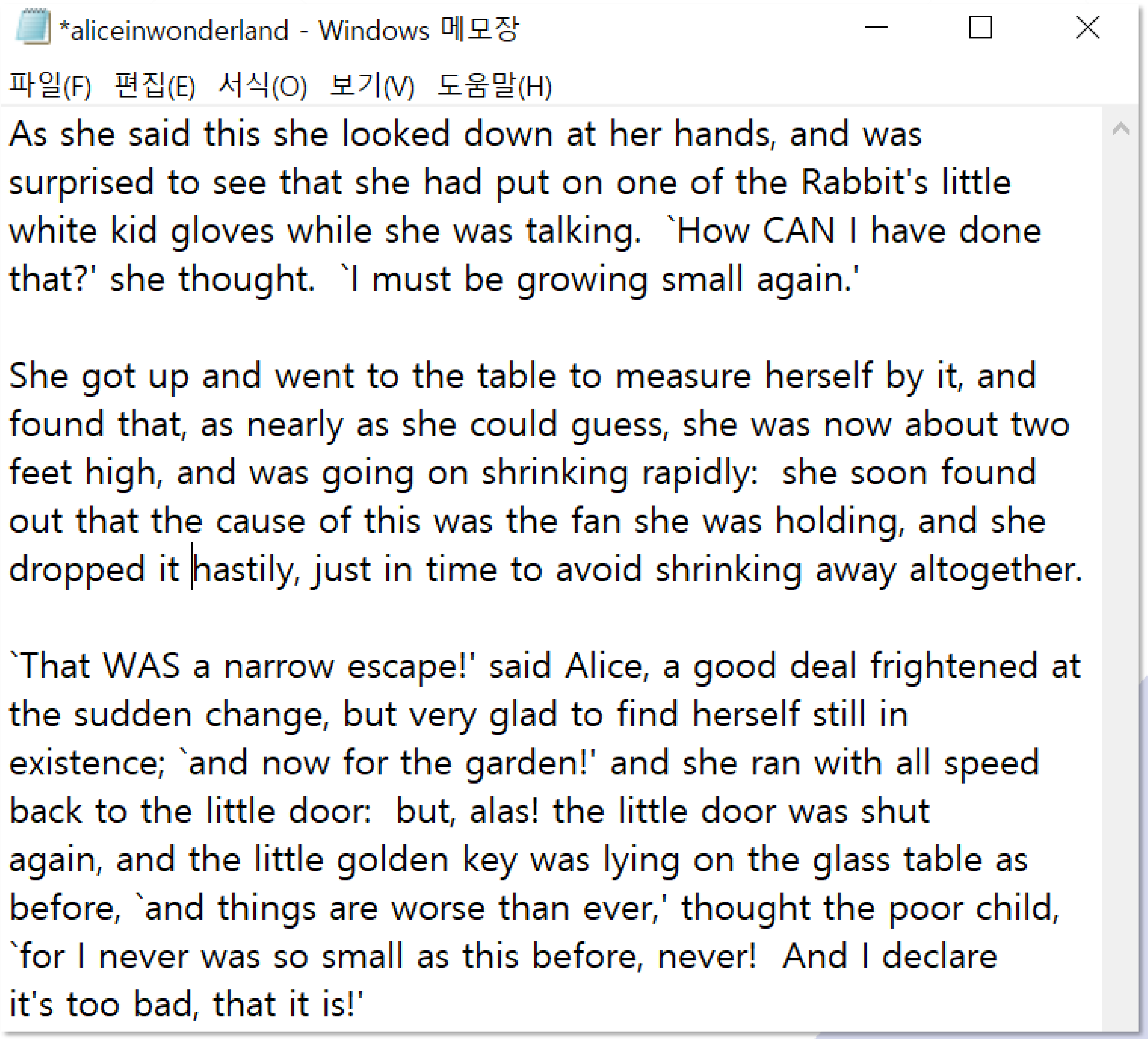
p=1 | # of iteration: 2000 | text length: 558		
Reference Data	**Accuracy**	**System time**
words.txt	0.9215247	37m 40s
warandpeace.txt	0.9977578	37m 1s

* Decipher function을 다섯번 돌려서 나온 평균값

01

Tuning Parameter

2) Cipher text length


*aliceinwonderland - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

As she said this she looked down at her hands, and was surprised to see that she had put on one of the Rabbit's little white kid gloves while she was talking. `How CAN I have done that?' she thought.  `I must be growing small again.'

She got up and went to the table to measure herself by it, and found that, as nearly as she could guess, she was now about two feet high, and was going on shrinking rapidly: she soon found out that the cause of this was the fan she was holding, and she dropped it hastily, just in time to avoid shrinking away altogether.

`That WAS a narrow escape!' said Alice, a good deal frightened at the sudden change, but very glad to find herself still in existence; `and now for the garden!' and she ran with all speed back to the little door: but, alas! the little door was shut again, and the little golden key was lying on the glass table as before, `and things are worse than ever,' thought the poor child, `for I never was so small as this before, never! And I declare it's too bad, that it is!'

문맥이 있고 약자가 거의 없는 소설

Lewis Carroll, *Alice in Wonderland* (1865)

233 letters (공백 포함) | 46 words

322 letters (공백 포함) | 63 words

470 letters (공백 포함) | 90 words

01

Tuning Parameter

2) Cipher text length

— Result Table —

p=1 | # of iteration: 2000 | reference data: warandpeace.txt

Text Length	Accuracy	System time
234	0.9572193	3m 51s
558	0.9976636	37m 40s
1030	0.8834207	3h 29m 48s

* Decipher function을 다섯번 돌려서 나온 평균값

01

Tuning Parameter

3) Number of iterations & Scaling parameter p

— Result Table —

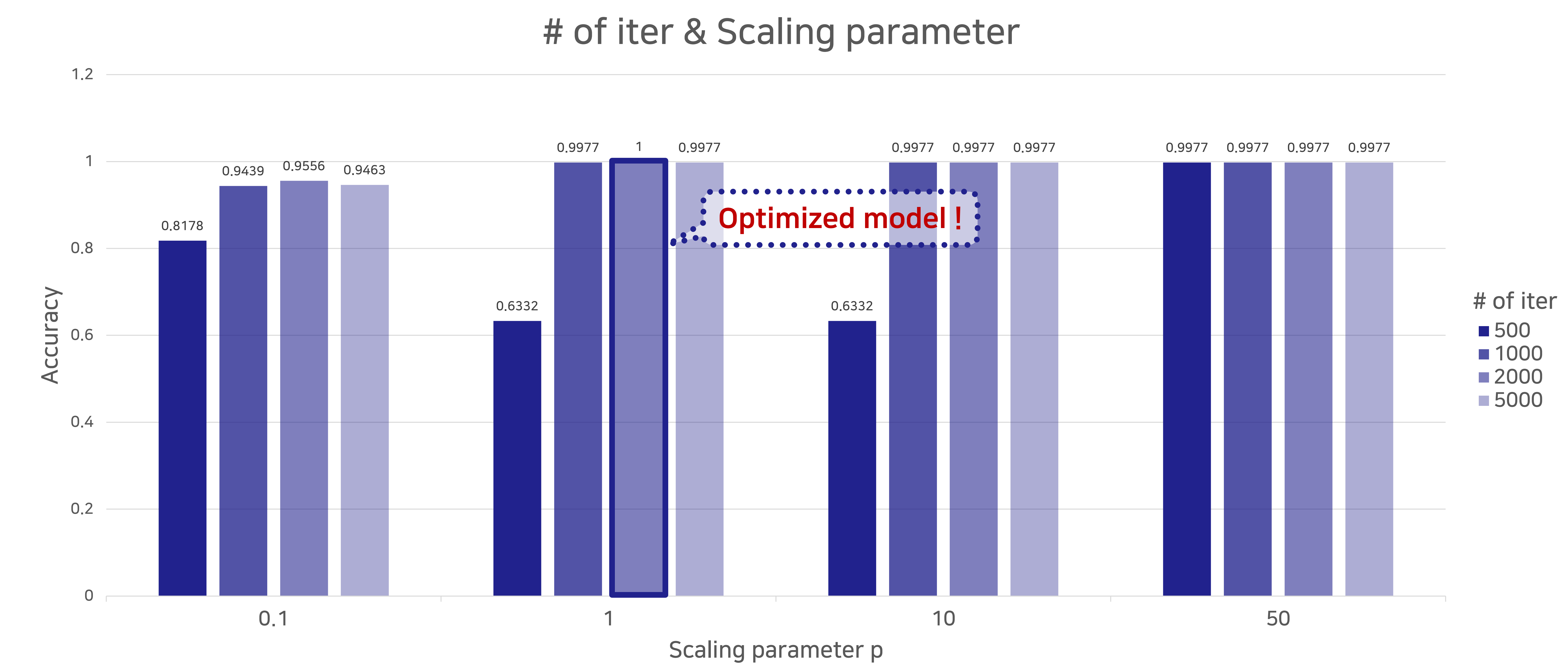
text length: 558 | reference data: warandpeace.txt

# of iter	500		1000		2000		5000	
scale parameter	Accuracy	Sys time	Accuracy	Sys time	Accuracy	Sys time	Accuracy	Sys time
0.1	0.8178	55s	0.9439	2m 41s	0.9556	5m 28s	0.9463	11m 54s
1	0.6332	11m 9s	0.9977	19m 8s	1.0000	37m 1s	0.9977	1h 44m 22s
10	0.6332	15m 59s	0.9977	30m 54s	0.9977	1h 6m 59s	0.9977	2h 53m 44s
50	0.9977	13m 52s	0.9977	30m 37s	0.9977	1h 8m 40s	0.9977	3h 15m 25s

01

Tuning Parameter

3) Number of iterations & Scaling parameter p

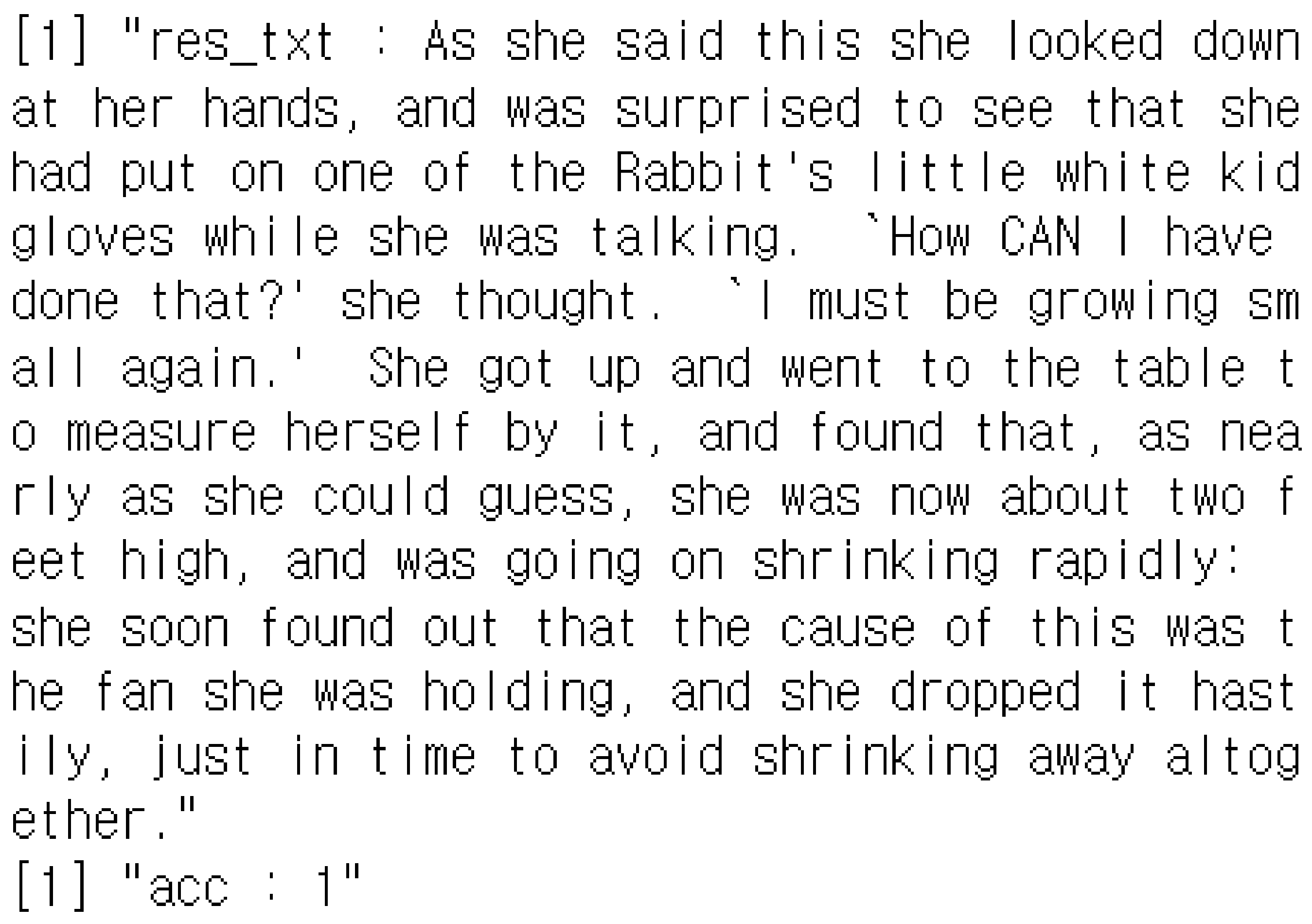


01

Tuning Parameter

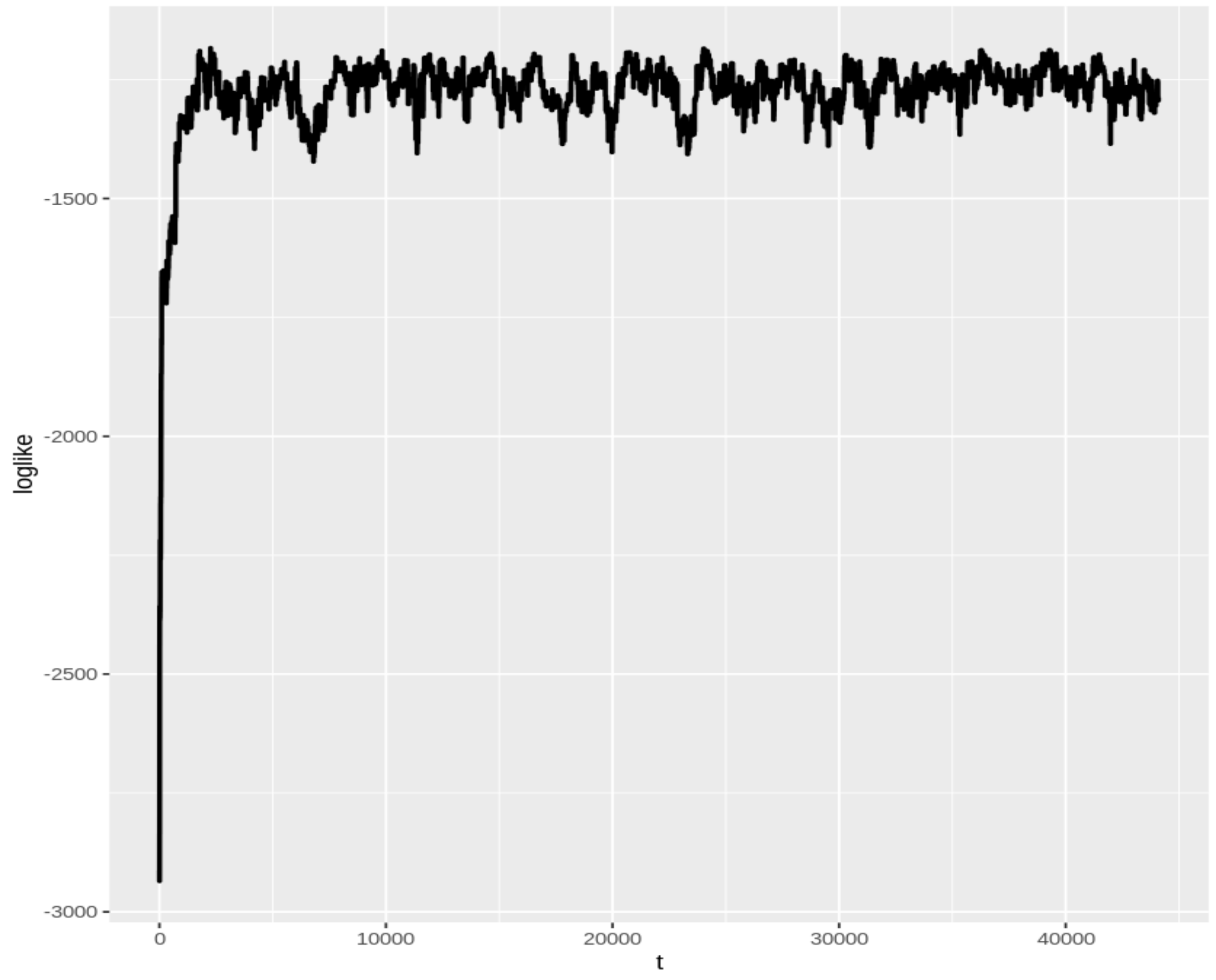
— Optimized model

Optimized model의 Decipher 결과



```
[1] "res_txt : As she said this she looked down
at her hands, and was surprised to see that she
had put on one of the Rabbit's little white kid
gloves while she was talking.  `How CAN I have
done that?' she thought.  `I must be growing sm
all again.'  She got up and went to the table t
o measure herself by it, and found that, as nea
rly as she could guess, she was now about two f
eet high, and was going on shrinking rapidly:
she soon found out that the cause of this was t
he fan she was holding, and she dropped it hast
ily, just in time to avoid shrinking away altog
ether."
[1] "acc : 1"
```

Optimized model의 log-likelihood



01 Tuning Parameter

— Optimized model

Reference data Text Length

warandpeace.txt
558 letters

+

# of iteration Scaling parameter

2000 iters
parameter p=1

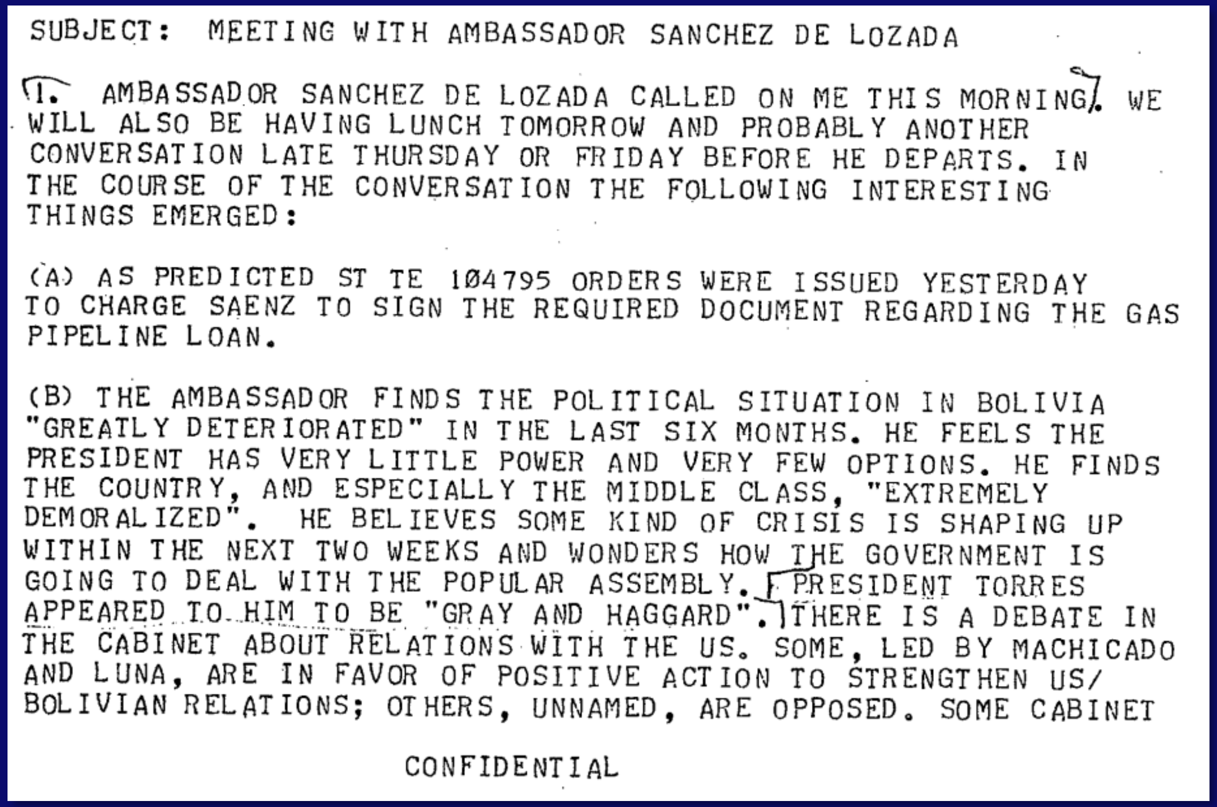
Optimized 된
model을
Exercise CIA
documents에 적용

02

Apply to Exercise

— CIA Documents

Meeting with Ambassador



SUBJECT: MEETING WITH AMBASSADOR SANCHEZ DE LOZADA

1. AMBASSADOR SANCHEZ DE LOZADA CALLED ON ME THIS MORNING. WE WILL ALSO BE HAVING LUNCH TOMORROW AND PROBABLY ANOTHER CONVERSATION LATE THURSDAY OR FRIDAY BEFORE HE DEPARTS. IN THE COURSE OF THE CONVERSATION THE FOLLOWING INTERESTING THINGS EMERGED:

(A) AS PREDICTED ST TE 104795 ORDERS WERE ISSUED YESTERDAY TO CHARGE SAENZ TO SIGN THE REQUIRED DOCUMENT REGARDING THE GAS PIPELINE LOAN.

(B) THE AMBASSADOR FINDS THE POLITICAL SITUATION IN BOLIVIA "GREATLY DETERIORATED" IN THE LAST SIX MONTHS. HE FEELS THE PRESIDENT HAS VERY LITTLE POWER AND VERY FEW OPTIONS. HE FINDS THE COUNTRY, AND ESPECIALLY THE MIDDLE CLASS, "EXTREMELY DEMORALIZED". HE BELIEVES SOME KIND OF CRISIS IS SHAPING UP WITHIN THE NEXT TWO WEEKS AND WONDERS HOW THE GOVERNMENT IS GOING TO DEAL WITH THE POPULAR ASSEMBLY. PRESIDENT TORRES APPEARED TO HIM TO BE "GRAY AND HAGGARD". THERE IS A DEBATE IN THE CABINET ABOUT RELATIONS WITH THE US. SOME, LED BY MACHICADO AND LUNA, ARE IN FAVOR OF POSITIVE ACTION TO STRENGTHEN US/BOLIVIAN RELATIONS; OTHERS, UNNAMED, ARE OPPOSED. SOME CABINET

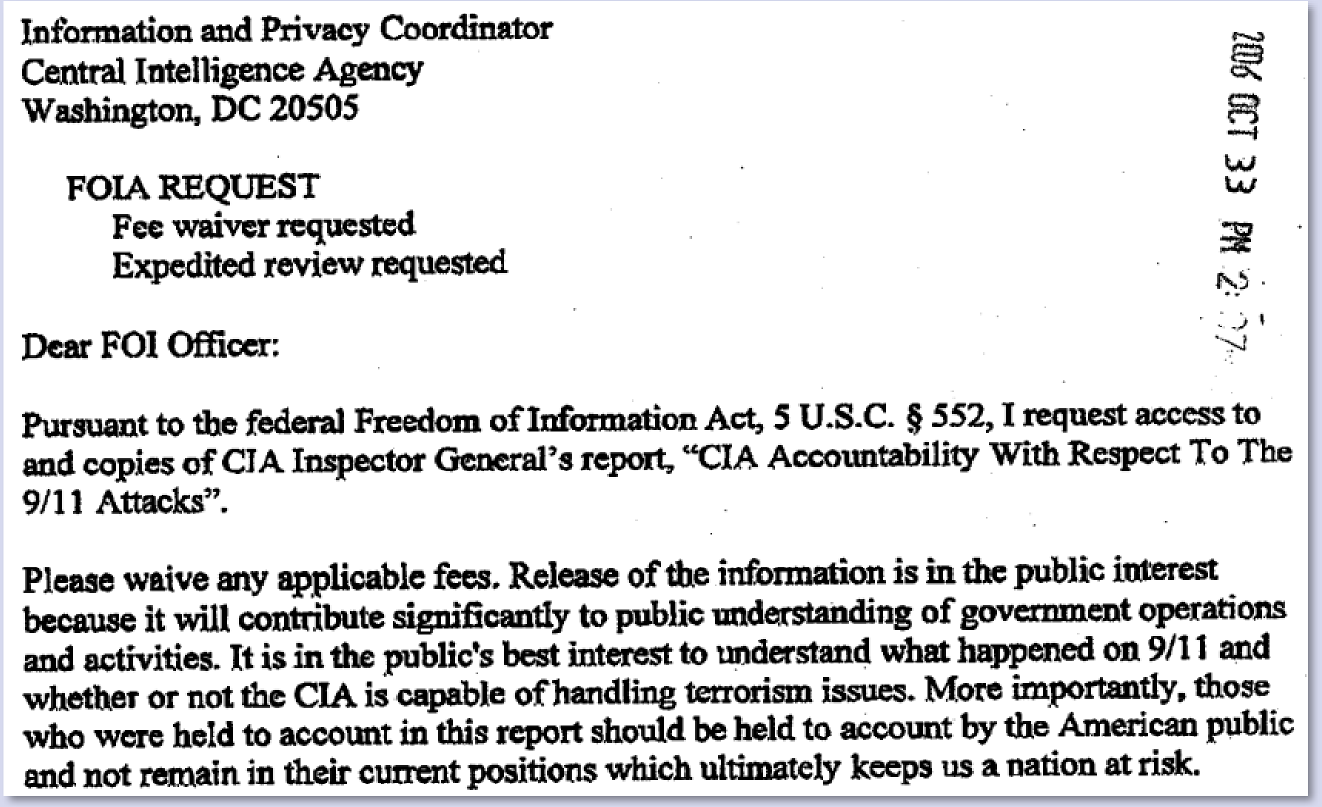
CONFIDENTIAL

517 Letters | 87 words (공백 포함)

문맥이 있는 문어체의 회의록에서 발췌

https://www.cia.gov/readingroom/collection/argentina-declassification-project-dirty-war-1976-83

Private Letter to CIA



Information and Privacy Coordinator
Central Intelligence Agency
Washington, DC 20505

2006 OCT 33 PM 2:27

FOIA REQUEST
Fee waiver requested
Expedited review requested

Dear FOI Officer:

Pursuant to the federal Freedom of Information Act, 5 U.S.C. § 552, I request access to and copies of CIA Inspector General's report, "CIA Accountability With Respect To The 9/11 Attacks".

Please waive any applicable fees. Release of the information is in the public interest because it will contribute significantly to public understanding of government operations and activities. It is in the public's best interest to understand what happened on 9/11 and whether or not the CIA is capable of handling terrorism issues. More importantly, those who were held to account in this report should be held to account by the American public and not remain in their current positions which ultimately keeps us a nation at risk.

545 Letters | 89 words (공백 포함)

문맥이 있는 구어체의 편지에서 발췌

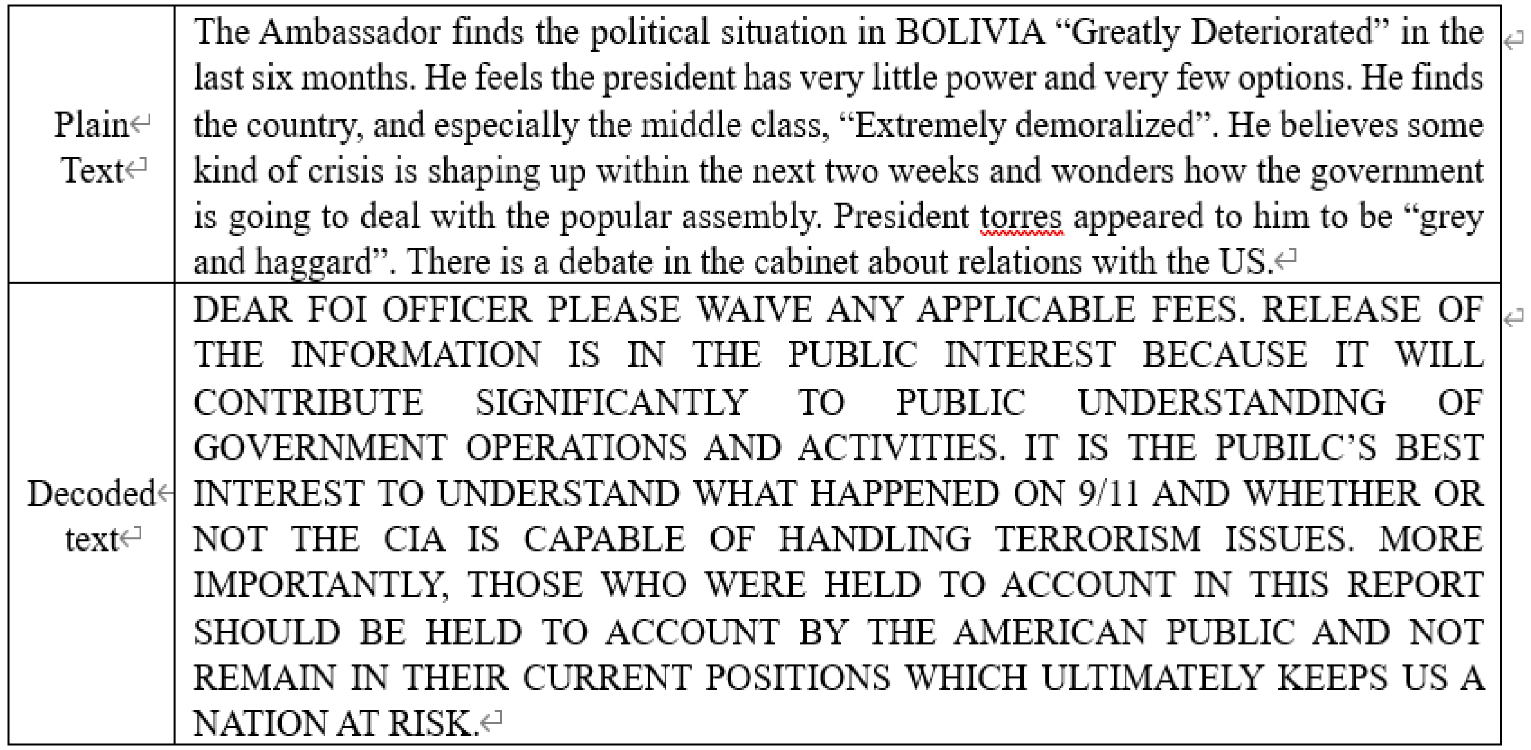
https://www.cia.gov/readingroom/document/0001500699

02

Apply to Exercise

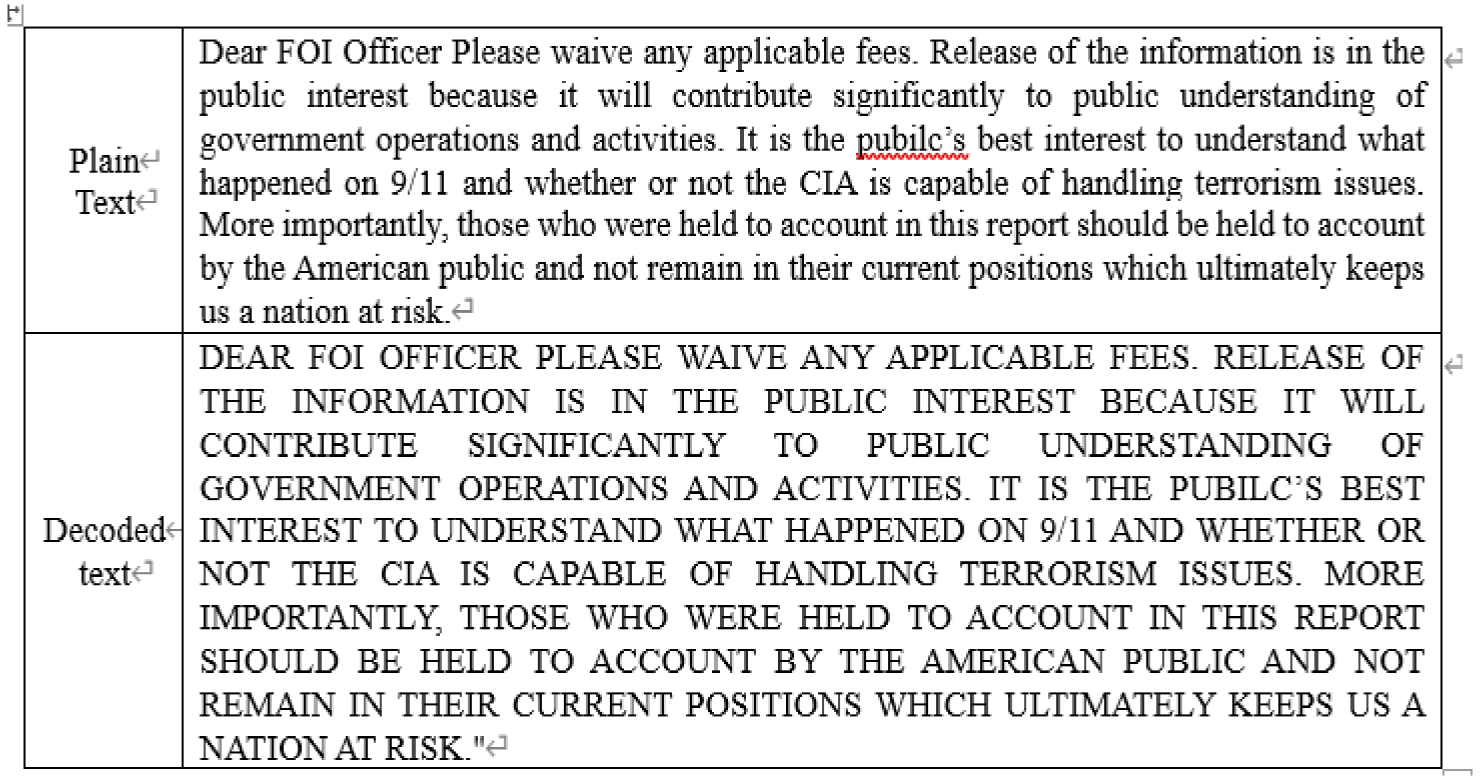
— CIA Documents

Meeting with Ambassador 실행 결과



Plain Text	The Ambassador finds the political situation in BOLIVIA “Greatly Deteriorated” in the last six months. He feels the president has very little power and very few options. He finds the country, and especially the middle class, “Extremely demoralized”. He believes some kind of crisis is shaping up within the next two weeks and wonders how the government is going to deal with the popular assembly. President torres appeared to him to be “grey and haggard”. There is a debate in the cabinet about relations with the US.
Decoded text	DEAR FOI OFFICER PLEASE WAIVE ANY APPLICABLE FEES. RELEASE OF THE INFORMATION IS IN THE PUBLIC INTEREST BECAUSE IT WILL CONTRIBUTE SIGNIFICANTLY TO PUBLIC UNDERSTANDING OF GOVERNMENT OPERATIONS AND ACTIVITIES. IT IS THE PUBILC’S BEST INTEREST TO UNDERSTAND WHAT HAPPENED ON 9/11 AND WHETHER OR NOT THE CIA IS CAPABLE OF HANDLING TERRORISM ISSUES. MORE IMPORTANTLY, THOSE WHO WERE HELD TO ACCOUNT IN THIS REPORT SHOULD BE HELD TO ACCOUNT BY THE AMERICAN PUBLIC AND NOT REMAIN IN THEIR CURRENT POSITIONS WHICH ULTIMATELY KEEPS US A NATION AT RISK.

Private Letter to CIA 실행 결과



Plain Text	Dear FOI Officer Please waive any applicable fees. Release of the information is in the public interest because it will contribute significantly to public understanding of government operations and activities. It is the pubilc’s best interest to understand what happened on 9/11 and whether or not the CIA is capable of handling terrorism issues. More importantly, those who were held to account in this report should be held to account by the American public and not remain in their current positions which ultimately keeps us a nation at risk.
Decoded text	DEAR FOI OFFICER PLEASE WAIVE ANY APPLICABLE FEES. RELEASE OF THE INFORMATION IS IN THE PUBLIC INTEREST BECAUSE IT WILL CONTRIBUTE SIGNIFICANTLY TO PUBLIC UNDERSTANDING OF GOVERNMENT OPERATIONS AND ACTIVITIES. IT IS THE PUBILC’S BEST INTEREST TO UNDERSTAND WHAT HAPPENED ON 9/11 AND WHETHER OR NOT THE CIA IS CAPABLE OF HANDLING TERRORISM ISSUES. MORE IMPORTANTLY, THOSE WHO WERE HELD TO ACCOUNT IN THIS REPORT SHOULD BE HELD TO ACCOUNT BY THE AMERICAN PUBLIC AND NOT REMAIN IN THEIR CURRENT POSITIONS WHICH ULTIMATELY KEEPS US A NATION AT RISK."

02 Apply to Exercise

— CIA Documents

— Result Table —

<table>
<tr><th colspan="3">p=1 | # of iteration: 2000 | reference data: warandpeace.txt</th></tr>
<tr><th>Exercise Data</th><th>Accuracy</th><th>System time</th></tr>
<tr><td><b>Meeting with Ambassador</b></td><td>0.9472</td><td>1h 56m 42s</td></tr>
<tr><td><b>Private Letter to CIA</b></td><td>1.0000</td><td>1h 23m 52s</td></tr>
</table>

* Decipher function을 다섯번 돌려서 나온 평균값

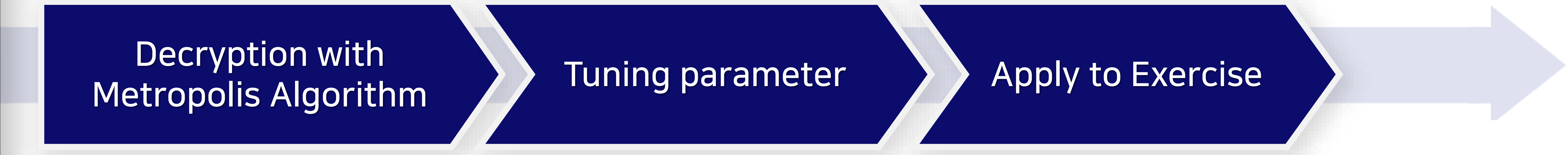
EWHA WOMANS UNIVERSITY, Department of Statistics

04

Conclusion

1) 결론 및 보완점
2) 참고 문헌

01 결론 및 보완점



Decryption with Metropolis Algorithm

Tuning parameter

Apply to Exercise

결론 1

Optimized model 생성

- 정확도와 실행 시간을 고려했을 때 scaling parameter가 1, iteration이 2000, text length가 500인 모델의 성능이 가장 좋음

결론 2

CIA Documents에 적용

- 구어체, 문어체로 구성된 과거 CIA Documents에 적용해봤을 때 정확도는 0.94, 1로 높았음
- 실행 시간이 너무 길어 효율적이지 못함

보완점

Random key의 한계

- Key를 랜덤 발생시키기에 initial key에 따라 local mode에 빠지기도 하고, 정확도가 낮아질 수 있으며 시간이 오래 걸림

문맥 예측의 한계

- 약자가 들어간 text의 경우 정확도 1을 구현하기 어려움

02 참고 문헌

① Chen, J., Rosenthal, J.S. Decrypting classical cipher text using Markov chain Monte Carlo. Stat Comput 22, 397–413 (2012).

② Diaconis, Persi. The markov chain monte carlo revolution. Bulletin of the American Mathematical Society 46.2 179-205 (2009).

EWHA WOMANS UNIVERSITY, Department of Statistics

감사합니다

Computational Statistics Final Project

— Using MCMC to break classical ciphers

2조 | 222STG24 김민지 | 222STG25 이다은 | 222STG26 이은효