

Chapter. 03

텍스트 마이닝

| 텍스트 마이닝이란

FAST CAMPUS
ONLINE

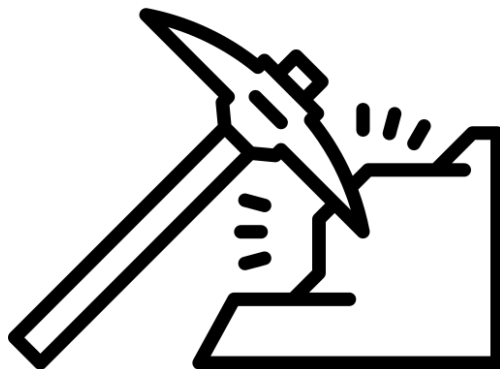
직장인을 위한 파이썬 데이터분석

강사. 윤기태

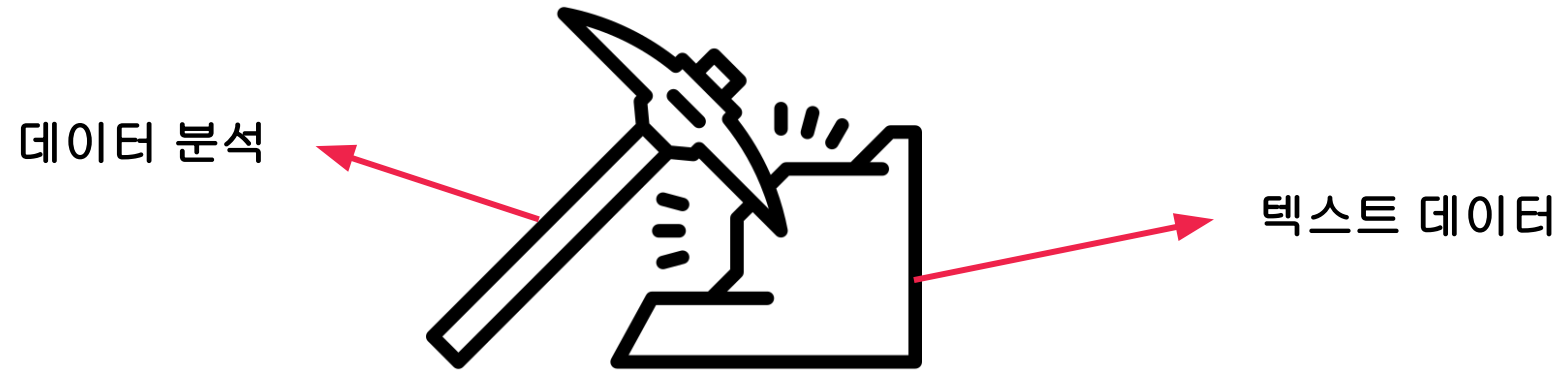
Chapter. 03

텍스트 마이닝이란

I 텍스트 마이닝이란



I 텍스트 마이닝이란





I 텍스트 마이닝을 활용하는 방법들

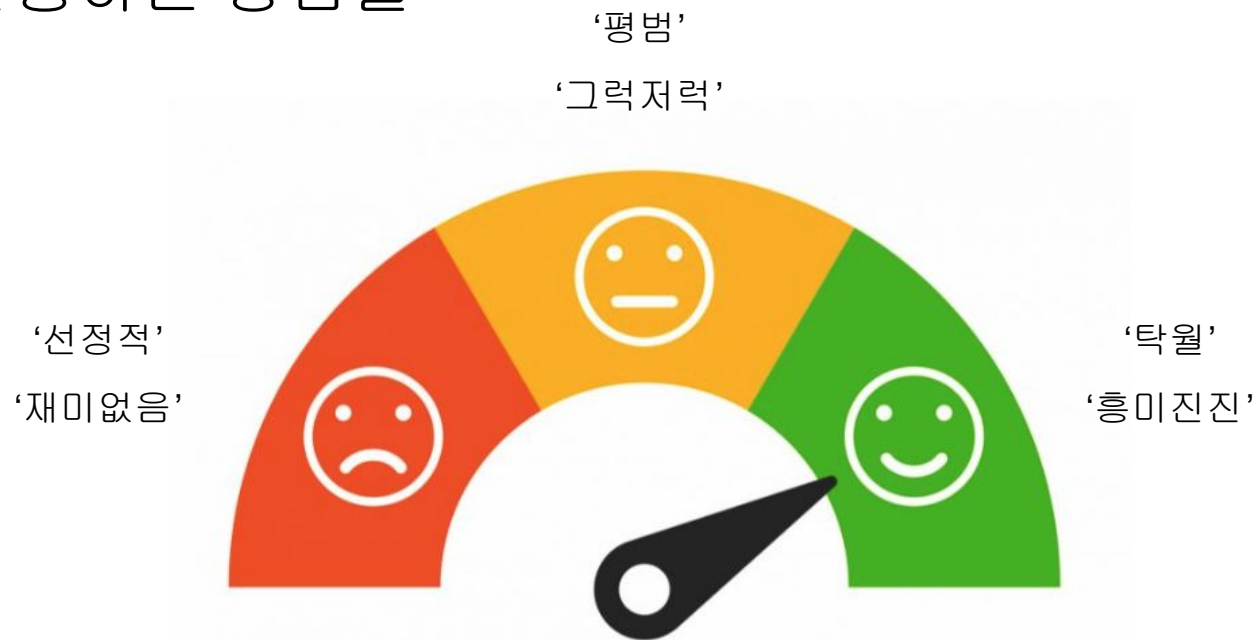
“선정적이고 재미없음”
“평범하고 연기도 그럭저럭”
“연출이 탁월하고 흥미진진”

I 텍스트 마이닝을 활용하는 방법들



“선정적이고 재미없음”
“평범하고 연기도 그럭저럭”
“연출이 탁월하고 흥미진진”

I 텍스트 마이닝을 활용하는 방법들



“선정적이고 재미없음”
“평범하고 연기도 그럭저럭”
“연출이 탁월하고 흥미진진”

I 텍스트 마이닝을 활용하는 방법들

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

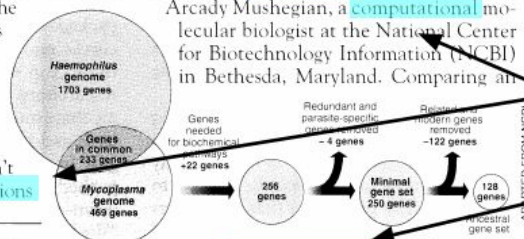
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments

출처 : <https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/>

I 텍스트 마이닝을 활용하는 방법들

Semantic Analysis



Chapter. 03

텍스트 데이터의 처리 방법

I 텍스트를 계산 가능한 데이터로 처리하는 방법 – BoW(Bag of Words)

"김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아.
근데, 냄새가 선을 넘지."

I 텍스트를 계산 가능한 데이터로 처리하는 방법 – BoW(Bag of Words)

"김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아.
근데, 냄새가 선을 넘지."



불용어가 아닌 형태소 추출

['김기사', '양반', '선', '넘다', '말다', '절대', '냄새']

I 텍스트를 계산 가능한 데이터로 처리하는 방법 – BoW(Bag of Words)

"김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아.
근데, 냄새가 선을 넘지."



불용어가 아닌 형태소 추출

['김기사', '양반', '선', '넘다', '말다', '절대', '냄새']



형태소의 등장 횟수

[1, 1, 2, 3, 1, 1, 1]

I 텍스트를 계산 가능한 데이터로 처리하는 방법 – BoW(Bag of Words)

"김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아.
근데, 냄새가 선을 넘지."



김기사	양반	선	넘다	말다	절대	냄새	
1	1	2	3	1	1	1	

I 텍스트를 계산 가능한 데이터로 처리하는 방법 – BoW(Bag of Words)

”Bag of Words 모델”

"김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아.
근데, 냄새가 선을 넘지."



김기사	양반	선	넘다	말다	절대	냄새	
1	1	2	3	1	1	1	

I 텍스트 데이터의 표현 방법

문서 단어 행렬 (Document-Term Matrix, DTM)

I 텍스트 데이터의 표현 방법

문장 1 : “아들아, 너는 계획이 다 있구나!”

문장 2 : "김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아. 근데, 냄새가 선을 넘지."

문장 3 : "가장 완벽한 계획이 뭔지 알아? 무계획이야."

문장 4 : “당신, 계획도 없지?”

I 텍스트 데이터의 표현 방법

-	아들	너는	계획	무계획	있다	김기사	양반	선	넘다	말다	절대	냄새	가장	완벽한	알다	당신	없다
문서 1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
문서 2	0	0	0	0	0	1	1	2	3	1	1	1	0	0	0	0	0
문서 3	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0
문서 4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1

문장 1 : “아들아, 너는 계획이 다 있구나!”

문장 2 : "김기사 그 양반. 선을 넘을 듯, 말 듯 하면서 절대 넘지 않아. 근데, 냄새가 선을 넘지."

문장 3 : "가장 완벽한 계획이 뭔지 알아? 무계획이야."

문장 4 : “당신, 계획도 없지?”

I 텍스트 데이터의 표현 방법 – TF-IDF

단어의 중요도를 계산하는 방법

Term Frequency-Inverse Document Frequency, TF-IDF)

I 텍스트 데이터의 표현 방법 – TF-IDF

TF : 특정 문서에서 특정 단어의 등장 횟수

DF : 특정 단어가 등장한 문서의 수

IDF : DF와 반비례 값을 가지는 수식

$$IDF(d, t) = \ln\left(\frac{n}{1 + DF(t)}\right)$$

TF-IDF : TF와 IDF를 곱한 값

I 텍스트 데이터의 표현 방법 – TF-IDF

A

B

-	아들	너는	계획	무계획	있다	김기사	양반	선	넘다	말다	절대	냄새	가장	완벽한	알다	당신	없다
문서 1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
문서 2	0	0	0	0	0	1	1	2	3	1	1	1	0	0	0	0	0
문서 3	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0
문서 4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1

IDF 값의 계산

A : $\ln(4 / (1 + 1)) = 0.6931$

B : $\ln(4 / (1 + 3)) = 0$

I 텍스트 데이터의 표현 방법 – TF-IDF

A

B

-	아들	너는	계획	무계획	있다	김기사	양반	선	넘다	말다	절대	냄새	가장	완벽한	알다	당신	없다
문서 1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
문서 2	0	0	0	0	0	1	1	2	3	1	1	1	0	0	0	0	0
문서 3	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	0
문서 4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1

TF-IDF 값의 계산

A : $1 * 0.6931$

B : $1 * 0$