Chapter. 03

텍스트 마이닝

텍스트 마이닝이란

FAST CAMPUS
ONLINE

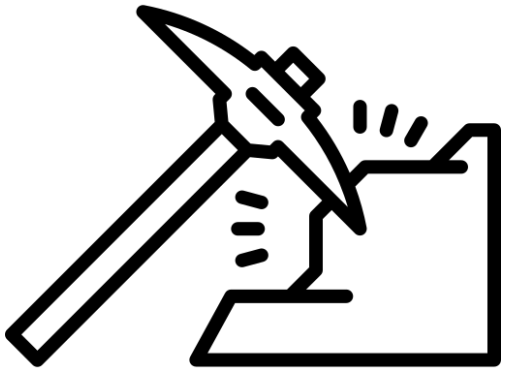
직장인을 위한 파이썬 데이터분석

강사. 윤기태

Chapter. 03

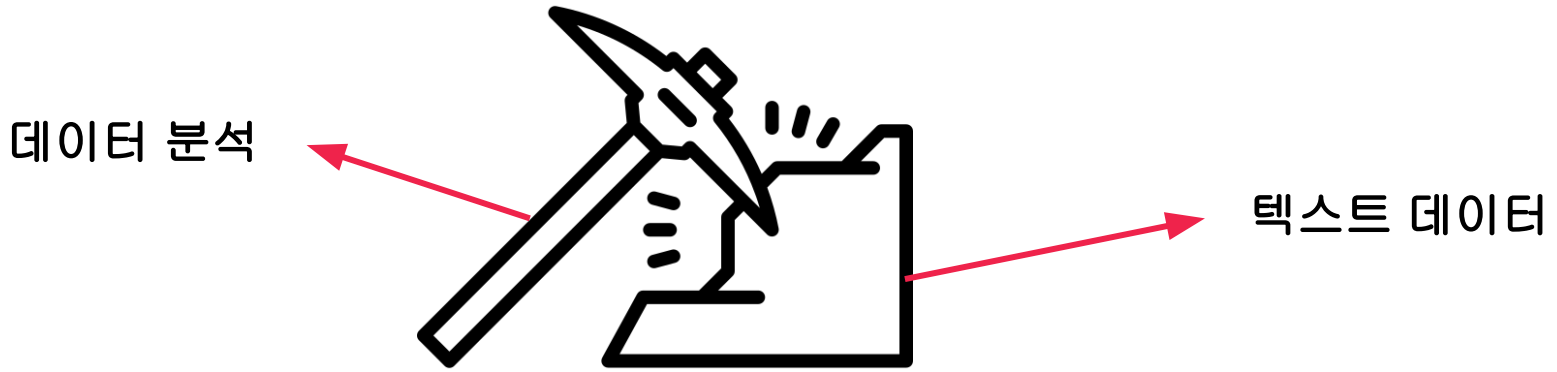
텍스트 마이닝이란

텍스트 마이닝이란



FAST CAMPUS
ONLINE
윤기태 강사.

텍스트 마이닝이란



텍스트 마이닝을 활용하는 방법들

FAST CAMPUS
ONLINE
윤기태 강사.

텍스트 마이닝을 활용하는 방법들

"선정적이고 재미없음"

"평범하고 연기도 그럭저럭"

"연출이 탁월하고 흥미진진"

텍스트 마이닝을 활용하는 방법들

“선정적이고 재미없음”

“평범하고 연기도 그럭저럭”

“연출이 탁월하고 흥미진진”

텍스트 마이닝을 활용하는 방법들

‘평범’

‘그럭저럭’

‘선정적’

‘재미없음’

‘탁월’

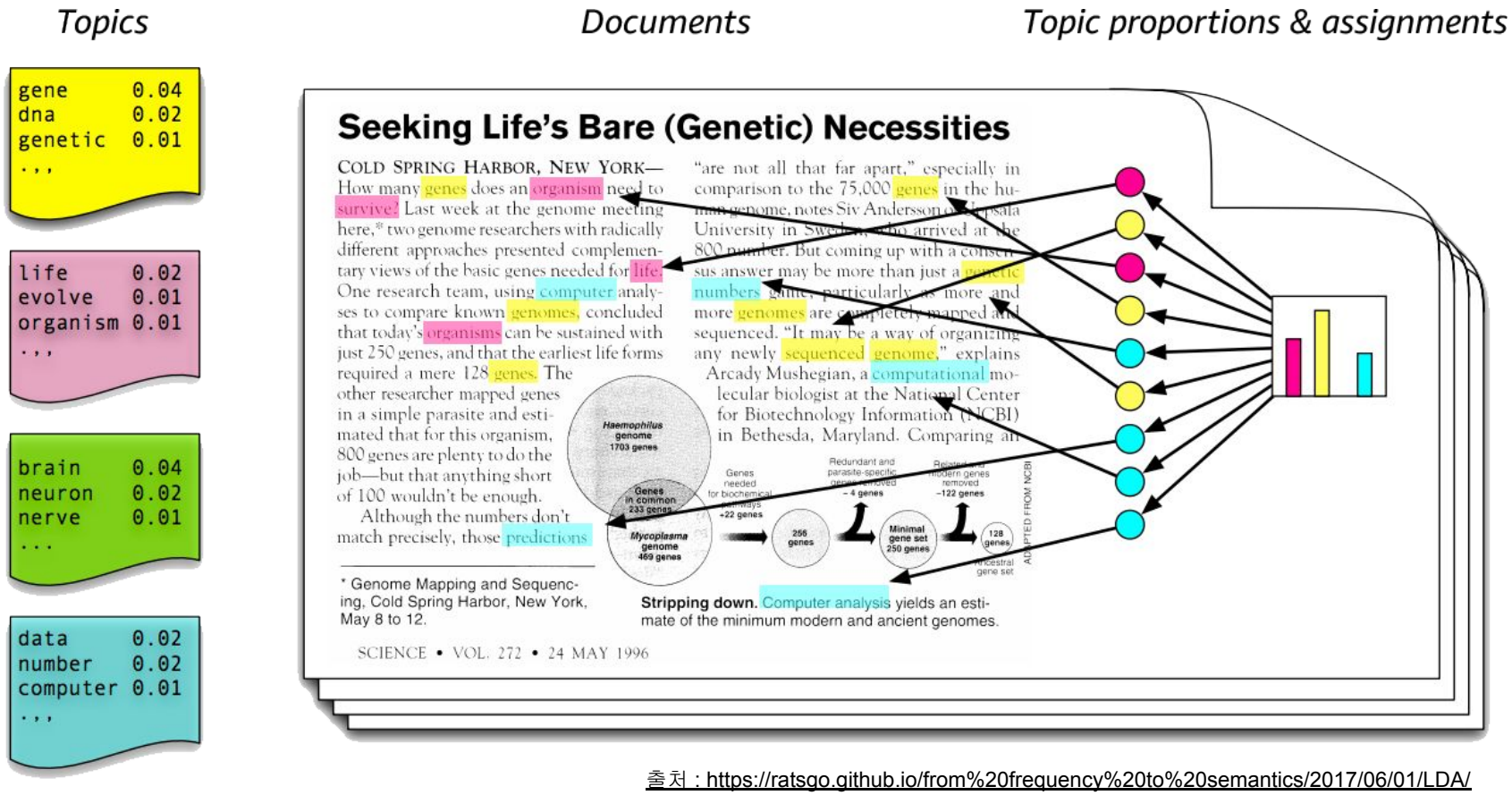
‘흥미진진’

“선정적이고 재미없음”

“평범하고 연기도 그럭저럭”

“연출이 탁월하고 흥미진진”

텍스트 마이닝을 활용하는 방법들



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

텍스트 마이닝을 활용하는 방법들

Semantic Analysis

Chapter. 03

텍스트 데이터의 처리 방법

FAST CAMPUS
ONLINE
윤기태 강사.

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."

FAST CAMPUS
ONLINE
윤기태 강사.

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."

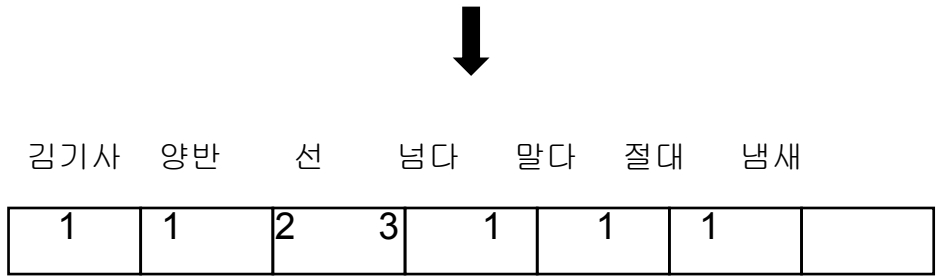
김기사	양반	선	넘다	말다	절대	냄새	
1	1	2 3	1	1	1		

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."



텍스트 데이터의 표현 방법

문서 단어 행렬 (Document-Term Matrix, DTM)

FAST CAMPUS
ONLINE
윤기태 강사.

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?”

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)

FAST CAMPUS
ONLINE
윤기태 강사.

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

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

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