

텍스트 데이터 분석

(단어사전 구축하기)

우리는 앞의 내용을 살펴보면

- 01. 텍스트 마이닝은 무엇인지?
- 02. 텍스트 분석을 위한 전체적인 절차
- 03. 텍스트 모델 구축하기

1-1 텍스트 마이닝이란?

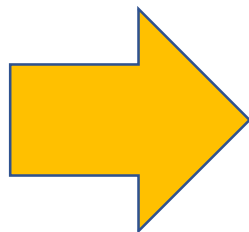
비정형 데이터

From today's featured article

"**X-Cops**" is the twelfth episode of the [seventh season](#) of the American science fiction television series *The X-Files*. Directed by [Michael Watkins](#) and written by [Vince Gilligan](#), the installment originally aired on the Fox network in February 2000. In this episode, [Fox Mulder](#) ([David Duchovny](#)) and [Dana Scully](#) ([Gillian Anderson](#)), special agents for the [Federal Bureau of Investigation](#), are interviewed for the Fox network reality television program *Cops* during an X-Files investigation. Mulder, hunting what he believes to be a werewolf, discovers that the monster terrorizing people craves the fear it provokes. While Mulder embraces the publicity of *Cops*, Scully is uncomfortable about appearing on national television. "X-Cops" is one of only two *X-Files* episodes that was shot in [real time](#). The episode has been thematically analyzed for its use of [postmodernism](#) and its presentation as reality television. It has been named among the best episodes of *The X-Files* by several reviewers, for its humor and format. ([Full article...](#))

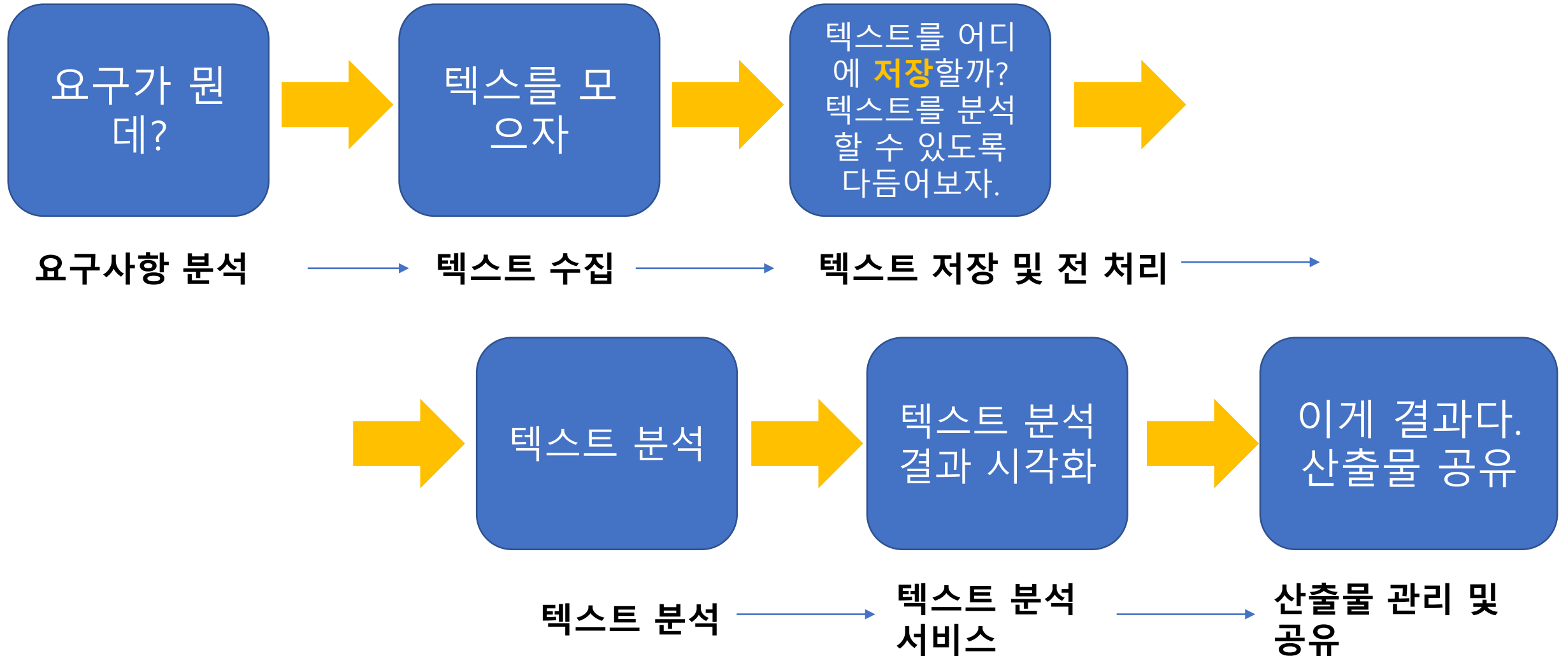
Recently featured: [Battle of Winterthur](#) · [38th \(Welsh\) Infantry Division](#) · *[New Worlds](#)* (magazine)

[Archive](#) · [By email](#) · [More featured articles](#)



비정형 텍스트 데이터
로부터 유용한 정보를
추출하는 기술

1-3 텍스트 분석 절차

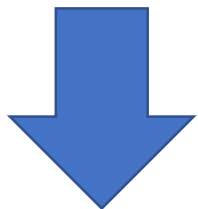


1-2 데이터 마이닝 vs 텍스트 마이닝

▶ 데이터 마이닝

이름	도시	나이
김철수	서울	45
김흥수	경기	20
임수하	원주	15

정형 데이터



유용하고 가치 있는 정보 추출

▶ 텍스트 마이닝

비정형 데이터

From today's featured article

"X-Cops" is the twelfth episode of the seventh season of the American science fiction television series *The X-Files*. Directed by Michael Watkins and written by Vince Gilligan, the installment originally aired on the Fox network in February 2000. In this episode, Fox Mulder (David Duchovny) and Dana Scully (Gillian Anderson), special agents for the Federal Bureau of Investigation, are interviewed for the Fox network reality television program *Cops* during an X-Files investigation. Mulder, hunting what he believes to be a werewolf, discovers that the monster terrorizing people craves the fear it provokes. While Mulder embraces the publicity of *Cops*, Scully is uncomfortable about appearing on national television. "X-Cops" is one of only two *X-Files* episodes that was shot in real time. The episode has been thematically analyzed for its use of postmodernism and its presentation as reality television. It has been named among the best episodes of *The X-Files* by several reviewers, for its humor and format. (Full article...)

Recently featured: Battle of Winterthur · 38th (Welsh) Infantry Division · *New Worlds* (magazine)

Are you subscribed by email · More featured articles



개체명(인명, 지역명 등), 패턴 혹은 단어-문장 관계 정보 추출

텍스트 분석 어떻게 해야 할까요?

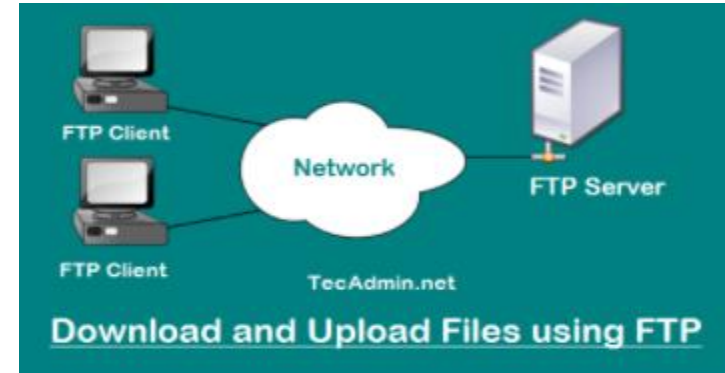
1-4 텍스트 수집



Crawling

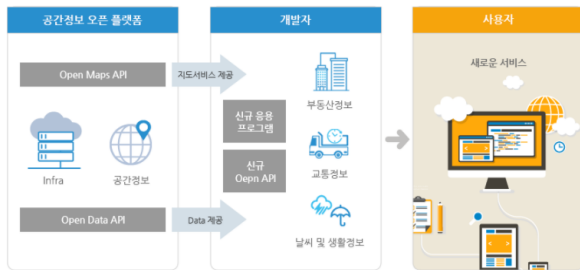


Scraping



FTP

공간정보 오픈플랫폼 개발자센터는 국가 공간정보의 개방, 공유, 참여를 통해 공간정보의 자율적이고 창조적인 다양한 애플리케이션을 개발할 수 있도록 2D/3D, 원격 오픈API 서비스와 기술을 제공합니다.

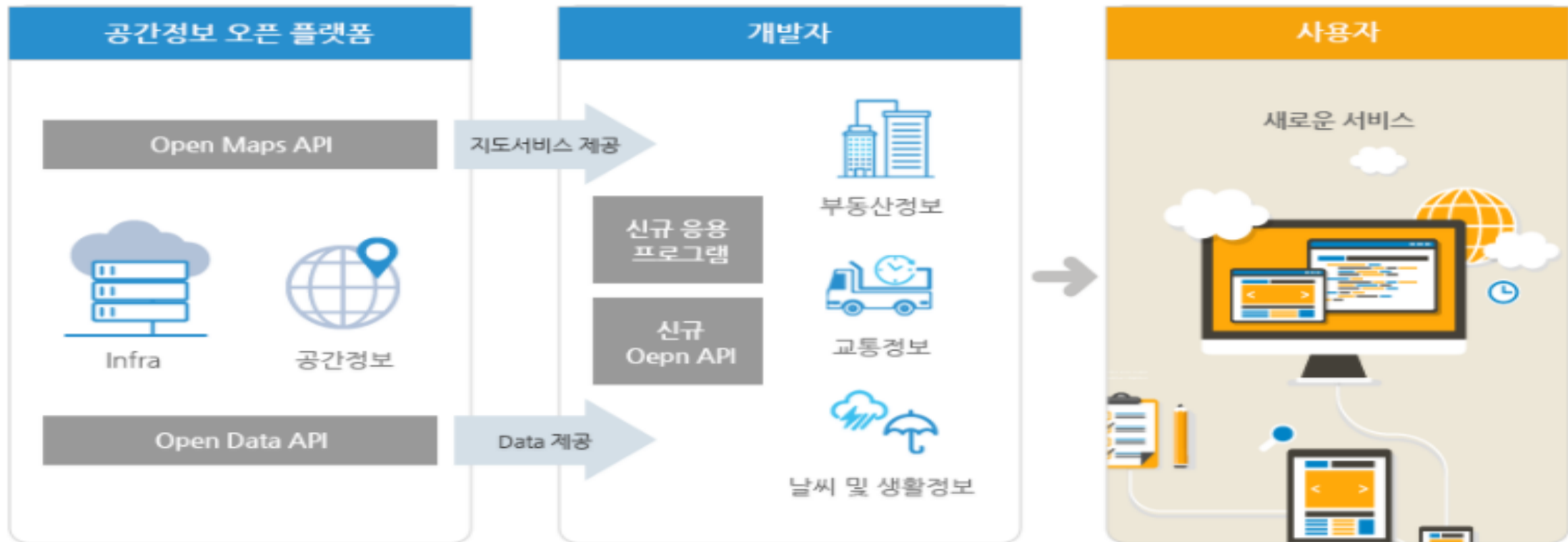


오픈 API

Ref : <https://www.promptcloud.com/blog/all-you-need-to-know-about-web-crawling>
<https://nocodewebscraping.com/web-scraping-for-dummies-tutorial-with-import-io-without-coding/>
http://dev.vworld.kr/dev/v4dv_apiuse_s001.do

1-5 텍스트 수집 – 오픈 API 이용하기

공간정보 오픈플랫폼 개발자센터는 국가 공간정보의 개방, 공유, 참여를 통해 공간정보의 자율적이고 창조적인 다양한 애플리케이션을 개발할 수 있도록 2D/3D, 검색 오픈API 서비스와 기술을 제공합니다.



오픈 API

Ref : http://dev.vworld.kr/dev/v4dv_apiuse_s001.do

텍스트를 분석하기 위해 텍스트를 어떻게
분리할 수 있을까요?

2-1 형태소 분석

- 주어진 텍스트를 단어와 문법적 특성에 맞추어 명사, 동사, 꾸밈어, 조사 등의 형태소로 분리할 수 있다.

2-1 형태소 분석

▶ 형태소란 무엇인가?(Morphology)

- **형태소란**, 의미가 있는 최소 단위로서 더 이상 분리가 불가능한 가장 작은 의미 요소. 즉 일반적으로 문법적, 관계적인 뜻을 나타내는 단어 또는 단어의 부분이다.
(매일경제 기획팀, 서울대 빅데이터 센터 참조)

2-1 형태소 분석

▶ 형태소 분석이란?

- 형태소 분석이란,

주어진 단어 또는 어절을 구성하는 각 형태소를 분리한 후, 분리된 형태소의 기본형 및 품사 정보를 추출.

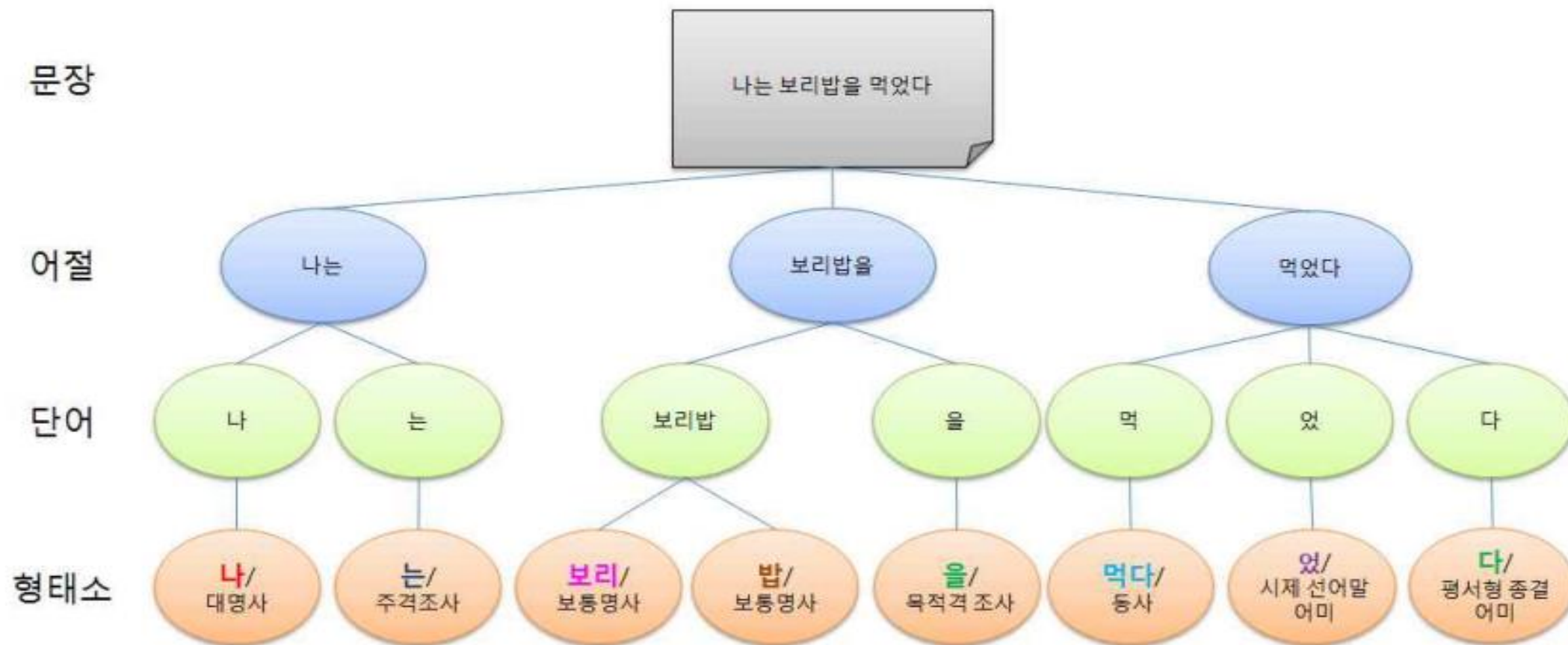


나는 보리밥을 먹었다

[그림 2-1] 형태소 예시

2-1 형태소 분석

▶ 문장, 어절, 단어, 형태소의 계층 구조



[그림 2-2] 문장, 어절, 단어, 형태소의 계층 구조

2-1 형태소 분석

- ▶ 형태소 개념 이해
- ▶ 텍스트 전처리(pre-processing) 이해
- ▶ 품사 태깅 이해

2-2 텍스트 전처리(text pre-processing)

- 텍스트 전처리 과정은 텍스트 분석을 위해 문장 분리, 불필요한 문장 성분을 제거하는 과정이다.
- 영미권에서는 예를 들면 대문자를 소문자로 변환하는 작업

2-2 텍스트 전처리(text pre-processing)



[그림 2-3] 전처리(pre-processing) 결과

2-3 품사 태깅

- 품사 태깅이란, (POS tagging, Part-Of-Speech tagging)
하나의 단어가 여러 품사를 갖는다. 따라서 품사의 모호성 (혹은 중의성)을 제거하는 과정이 필요. 이를 수행하는 과정을 품사 태깅이라 한다.

2-3 품사 태깅



[그림 2-4] 품사 태깅 예시

2-4 키워드 추출

▶ 가용어의 이해

불용어가 아닌 단어들

▶ 불용어의 이해

단어 성분 중에서 문서의 정보(의미)를 표현하지 못하는 단어.
즉, 문서와 관련이 없다.

▶ 키워드의 개념

가용어 중의 중심이 되는 단어

키워드 선정을 해 보자.
어떻게 해야 할까

2-5 키워드 선정

일반적으로 분석하고자 하는 목적 및 데이터 세트에 영향을 받지만,

문서 내에서 발생 빈도가 높은 단어들을 키워드로 선정한다.

2-6 불용어, 가용어, 키워드

▶ 불용어

한국어 : 조사 ('는', '을' 등)

영어 : '관사', '전치사' ('a', 'the', 'on', 'with')

▶ 키워드

불용어는 빈도가 높다고 키워드가 아니다.

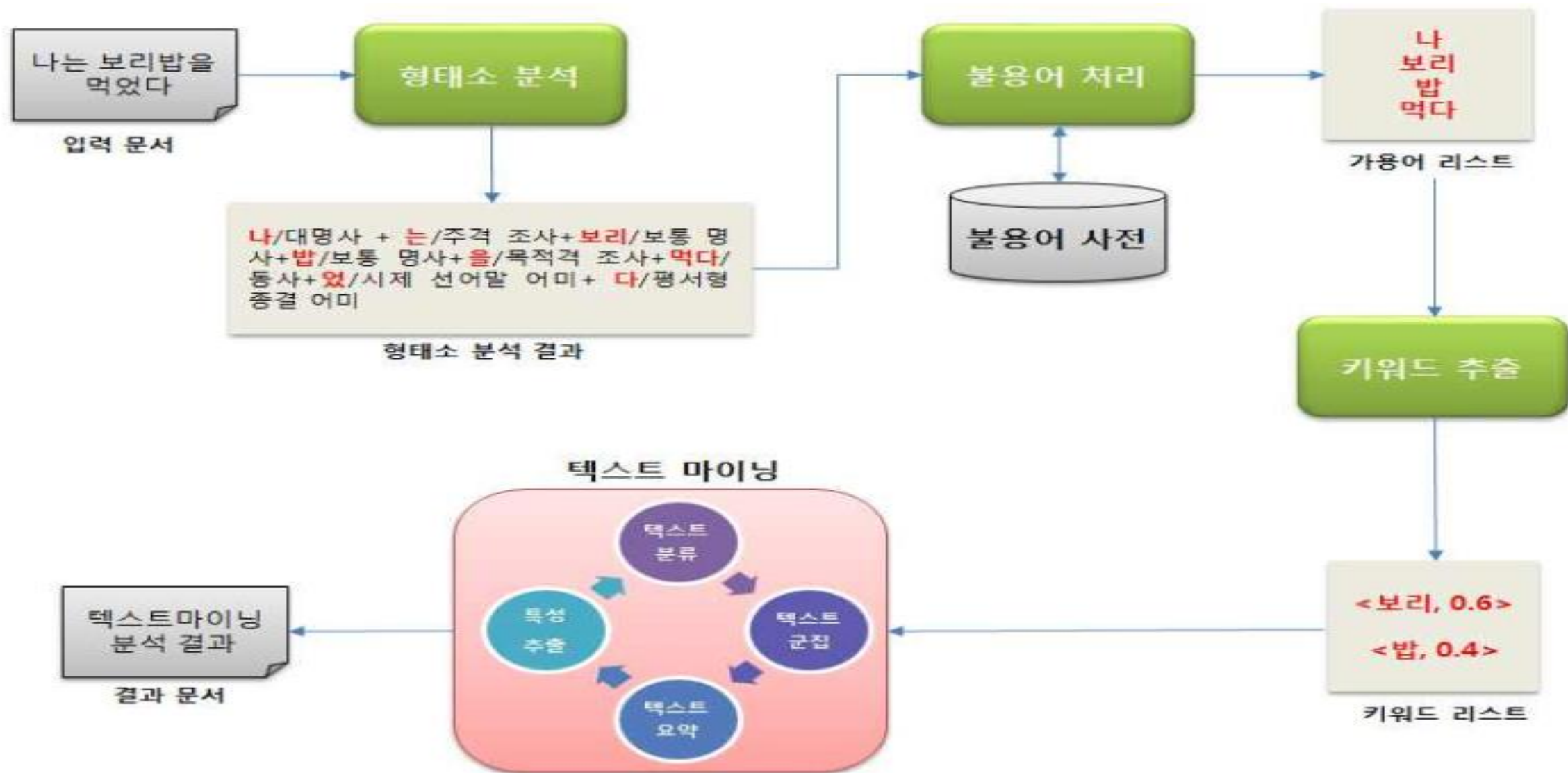
2-6 불용어 처리

- 불용어가 저장된 데이터 베이스를 참조하여 키워드 제거.

=> 형태소 분석 결과를 **불용어 사전에서 검색**하여 일치하는 내용이 나타나면 그 **내용을 삭제**한다.

2-6 키워드 추출(Text Wording) 절차

텍스트모델구축하기



2-7 말뭉치

- > 각 분야에서 필요로 하는 연구 재료로서 자료의 집합을 말한다.
- > 목적에 따라 작게는 소설 한편, 수십억 어절 이상의 말 또는 글로 표현된 자료 모음
- > 전산학적 말뭉치는 글로 표현된 자료에 대해 텍스트 정제, 통합, 변환의 절차를 거쳐서 구조화된 형태의 자료를 의미

(예제) 2개 이상의 언어로 만든 병렬 말뭉치. 영어 소설의 원본과 한국의 번역본의 데이터로 말뭉치를 만들면 **한영 소설어의 병렬 말뭉치**가 된다.

▶ 말뭉치 개념

■ 말뭉치란?

- (1) 각 분야의 필요로 하는 연구 자료이다.
- (2) 언어의 본질적인 모습을 보여주는 자료의 집합

▶ 말뭉치 개념(전산학적)

■ 말뭉치란?(Corpus)

(1) 대규모 언어 데이터 베이스

(2) 인간의 음성 언어(문어, 구어)를 대용량 컴퓨터에 저장하고 이를 필요에 따라 가공하여 언어 연구에 사용

(3) 컴퓨터가 판독할 수 있는 형태(machine-readable form)

▶ 말뭉치 종류

■ 가공 방법에 따른 종류

(1) 원시 말뭉치(raw corpus)

(2) 가공된 말뭉치(tagged corpus)

■ 작성 방법에 따른 종류

(1) 샘플 말뭉치 : 텍스트를 일정량만 수집한 것으로 텍스트의 내용이 고정된 말뭉치

(2) 모니터 말뭉치 : 변화하는 언어의 실태 추적을 위한 것. 낡은 자료를 제외한 항상 새로운 정보를 수집 후, 최신 언어 정보를 데이터베이스화한 말뭉치

▶ 말뭉치 종류

- 기타

범용 말뭉치

특수목적 말뭉치

공시 말뭉치

통시 말뭉치

병렬 말뭉치....

2-8 단어와 문서 관계 표현

- > 말뭉치로부터 단어와 문서의 관계를 표현하기 위해 문서-단어 행렬 혹은 단어-문서 행렬을 작성할 수 있다.
- > 문서-단어 행렬 혹은 단어-문서 행렬로부터 단어의 빈도와 연관성등의 기본 집계 수행이 가능하다.

2-8 단어와 문서 관계 표현

▶ 단어-문서 행렬(문서-단어 행렬)

- (1) 각 단어는 문서의 의미를 나타내는 **가장 기본적인 단위**
- (2) 이러한 관계를 텍스트 마이닝에서 단어-문서 행렬로 표현. 이를 통해 **단어들 사이의 포함관계 쉽게 조망 가능**

2-8 단어와 문서 관계 표현

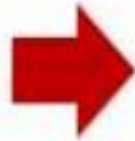
▶ 단어-문서 행렬예시

단어가 행을 구성,
문서가 열을 구성한다.
단어 포함(1), 불포함(0)

입력 텍스트 1: "파스타 먹방, 강남 파스타 데이트"

입력 텍스트 2: "강남 버스 파스타 맛집"

입력 텍스트 3: "강남 버스, 강남 파스타, 강남 맛집"



	입력텍스트 1	입력텍스트 2	입력텍스트 3	대상 텍스트 ←
파스타	1	1	1	
먹방	1	0	0	
강남	1	1	1	
데이트	1	0	0	
버스	0	1	1	
맛집	0	1	1	
...	

↑ 말뭉치에서 추출된 단어들

[그림 2-12] 단어-문서 행렬 예시

하나의 문서에서 단어의 중요성을 확인할 수 없을까?

2-9 하나의 문서에서 단어의 중요성 확인

텍스트모델구축하기

- ▶ TF(단어빈도) 특정한 단어가 문서 내에 얼마나 자주 등장하는가?
 - 이 값이 높을수록 문서에서 중요하다.

- ▶ DF(문서빈도, document frequency)

- 특정단어를 포함하고 있는 문서의 수
- df가 높다는 것은 많은 문서에서 나타난다.
검색에서 별로 중요한 단어가 아니다.
검색단어는 내가 필요한 문서에서 많이 나타나고 다른 문서에서 적게 나와야 정확도 높은 검색이 가능함.

파스타는 3개 문서 있으므로 3
먹방은 1개 문서 있으므로 1
강남은 3개 문서 있으므로 3

	입력텍스트 1	입력텍스트 2	입력텍스트 3	대상 텍스트 ←
파스타	1	1	1	
먹방	1	0	0	
강남	1	1	1	
데이트	1	0	0	
버스	0	1	1	
맛집	0	1	1	
...	

↑ 말뭉치에서 추출된 단어들

2-9 하나의 문서에서 단어의 중요성 확인

- ▶ IDF(역문서빈도) df(문서빈도)에서 역수를 취한 값.

$$\text{IDF} = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$



Log의 분모 부분이 0이 되면 안되기에 + 1

$$\text{IDF} = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수} + 1}\right)$$

2-9 TF-IDF

Term Frequency Inverse Document Frequency weighting scheme

- ▶ TF(단어빈도) * IDF(문서빈도의 역수)

$$\text{공식 : } \text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

- ▶ 문서 내에 단어들에 각각 부여한 중요도를 나타내는 숫자 값.

- ▶ 어디에 사용되는가?

TFIDF 값을 부여한 후, 코사인 유사도(Cosine Similarity)등을 이용해서 문서들의 유사도를 구하는데 흔히 사용됨.

TFIDF 값을 계산을 통해 문서 내에서 상대적으로 더 중요한 단어가 무엇인지를 알 수 있다.

2-9 TF-IDF (버스)

	입력텍스트 1	입력텍스트 2	입력텍스트 3
파스타	0	0	0
데이트	0.24	0	0
버스	0	0.044	0.029
맛집	0	0.044	0.029
***	***	***	***

입력 텍스트 1 : "파스타 먹방, 강남 파스타 데이트 "

입력 텍스트 2 : "강남 버스 파스타 맛집 "

입력 텍스트 3 : "강남 버스, 강남 파스타, 강남 맛집 "

TF(Term frequency)

입력 텍스트 1

$$\frac{0}{5} = 0$$

입력 텍스트 2

$$\frac{1}{4} = 0.25$$

입력 텍스트 3

$$\frac{1}{6} = 0.1667$$

$$TF = \frac{\text{문서내 단어수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$

IDF(inverse document frequency)

$$IDF = \log_{10}\left(\frac{3}{2}\right) = 0.1761$$

$$TF-IDF(D2) = 0.25 * 0.1761 = 0.044$$

2-9 TF-IDF (데이트)

	입력텍스트 1	입력텍스트 2	입력텍스트 3
파스타	0	0	0
먹방	0.0954	0	0
강남	0	0	0
데이트	0.0954	0	0
버스	0	0.044	0.029
맛집	0	0.044	0.029
...

$$TF = \frac{\text{문서내 단어수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$

입력 텍스트 1 : "파스타 먹방, 강남 파스타 **데이트**"

입력 텍스트 2 : "강남 버스 파스타 맛집 "

입력 텍스트 3 : "강남 버스, 강남 파스타, 강남 맛집 "

TF(Term frequency)

입력 텍스트 1

$$\frac{1}{5} = 0.2$$

입력 텍스트 2

$$\frac{0}{4} = 0$$

입력 텍스트 3

$$\frac{0}{6} = 0$$

IDF(inverse document frequency)

$$IDF = \log_{10}\left(\frac{3}{1}\right) = 0.477$$

$$TF-IDF(D1) = 0.2 * 0.477 = 0.0954$$

2-9 TF-IDF (먹방)

	입력텍스트 1	입력텍스트 2	입력텍스트 3
파스타	0	0	0
먹방	0.0954	0	0
강남	0	0	0
데이트	0.0954	0	0
버스	0	0.044	0.029
맛집	0	0.044	0.029
...

$$TF = \frac{\text{문서내 단어수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$

입력 텍스트 1 : "파스타 **먹방**, 강남 파스타 데이트 "

입력 텍스트 2 : "강남 버스 파스타 맛집 "

입력 텍스트 3 : "강남 버스, 강남 파스타, 강남 맛집 "

TF(Term frequency)

입력 텍스트 1

$$\frac{1}{5} = 0.2$$

입력 텍스트 2

$$\frac{0}{4} = 0$$

입력 텍스트 3

$$\frac{0}{6} = 0$$

IDF(inverse document frequency)

$$IDF = \log_{10}\left(\frac{3}{1}\right) = 0.477$$

$$TF-IDF(D1) = 0.2 * 0.477 = 0.0954$$

2-9 TF-IDF (먹방) – log2을 사용시(R)

텍스트모델구축하기

	Docs		
Terms	D1	D2	D3
강남	0.0000000	0.0000000	0.000000000
데이트	0.3169925	0.0000000	0.000000000
먹방	0.3169925	0.0000000	0.000000000
파스타	0.0000000	0.0000000	0.000000000
맛집	0.0000000	0.1462406	0.09749375
버스	0.0000000	0.1462406	0.09749375

입력 텍스트 1 : "파스타 **먹방**, 강남 파스타 데이트 "

입력 텍스트 2 : "강남 버스 파스타 맛집 "

입력 텍스트 3 : "강남 버스, 강남 파스타, 강남 맛집 "

TF(Term frequency)

입력 텍스트 1

$$\frac{1}{5} = 0.2$$

입력 텍스트 2

$$\frac{0}{4} = 0$$

입력 텍스트 3

$$\frac{0}{6} = 0$$

$$TF = \frac{\text{문서내 단어수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$

IDF(inverse document frequency)

$$IDF = \log_2\left(\frac{3}{1}\right) = 1.5849$$

$$TF-IDF(D1) = 0.2 * 1.5849 = 0.3169$$

2-9 TF-IDF – 영문서(blue)- relative tf-idf (R에서)

(실습 1) BLUE에 대한 TF-IDF를 구해보자.

불용어 처리 후, 입력 문서

입력 텍스트 1 : "The sky blue"

입력 텍스트 2 : "The sun bright"

입력 텍스트 3 : "The sun sky bright"

$$TF = \frac{\text{문서내 단어수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$

TF(Term frequency)

입력 텍스트 1 입력 텍스트 2 입력 텍스트 3

$$\frac{1}{3} = 0.333$$

$$\frac{0}{4} = 0$$

$$\frac{0}{4} = 0$$

IDF(inverse document frequency)

$$IDF = \log_2\left(\frac{3}{1}\right) = 1.5849$$

$$TF-IDF \Rightarrow 0.33333 * 1.5849 = 0.5283$$

**R에서 IDF를 구할 때, \log_{10} 이 아닌 \log_2 를 사용한다.
이를 relative tf-idf라고 한다.**

2-9 TF-IDF – 영문서(bright)

(실습 2) bright에 대한 TF-IDF를 구해보자.

불용어 처리 후, 입력 문서

입력 텍스트 1 : "The sky blue"

입력 텍스트 2 : "The sun **bright**"

입력 텍스트 3 : "The sun sky **bright**"

$$TF = \frac{\text{문서내 단어수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$

TF(Term frequency)

입력 텍스트 1

$$\frac{0}{3} = 0$$

입력 텍스트 2

$$\frac{1}{3} = 0.3333$$

입력 텍스트 3



$$\frac{1}{4} = 0.25$$

IDF(inverse document frequency)

$$IDF = \log_2\left(\frac{3}{2}\right) = 0.5849$$

$$TF-IDF \Rightarrow 0.33333 * 0.5849 = 0.1949$$

$$TF-IDF \Rightarrow 0.25 * 0.5849 = 0.1462$$

Terms					
Docs	blue	sky	the	bright	sun
D1	0.5283208		0	0.0000000	
D2	0.0000000		0	0.1949875	
D3	0.0000000		0	0.1462406	

2-9 TF-IDF – 영문서(bright)

(실습 1) 불용어 처리된 입력 문서는 다음과 같다. sky에 대한 단어의 TF-IDF를 구해보자.(밑수가 2인 \log_2 를 사용해서 구해보자.)

불용어 처리 후, 입력 문서

입력 텍스트 1 : "The **sky** blue"

입력 텍스트 2 : "The sun bright"

입력 텍스트 3 : "The sun **sky** bright"

2-10 TF, DF, IDF

- ▶ TF(단어빈도) 한 문서내에서 term이 몇 번 나왔는가?
- ▶ DF(문서빈도, document frequency)
특정 단어가 포함된 문서는 몇 개인가?
- ▶ Inverse Document Frequency = Inverse(역수) => 시간에 따라 변경 발전됨.

● $idf = 1/df$ => 처음의 개념에서 점점 발전

● $idf(t, D) = \log\left(\frac{\text{전체 문서의 수}(D)}{\text{특정단어}(t)\text{포함문서 수}}\right)$

● $idf(t, D) = \log\left(\frac{\text{전체 문서의 수}}{\text{특정단어}(t)\text{포함문서 수} + 1}\right)$

D는 전체 문서의 수
t는 특정 단어

2-10 코사인 유사도(Cosine Similarity)

- ▶ TF(단어빈도) * IDF(문서빈도의 역수)

$$\text{tf-idf}(\text{문서}, \text{단어}) = \text{tf}(\text{문서}, \text{단어}(t)) * \text{idf}(\text{단어}(t))$$

- ▶ 문서 내에 단어들에 각각 부여한 중요도를 나타내는 숫자값.
- ▶ 어디에 사용되는가?

TFIDF 값을 부여한 후, 코사인 유사도(Cosine Similarity)등을 이용해서 문서들의 유사도를 구하는데 흔히 사용됨.

2-10 TF-IDF의 정리

- ▶ TF(단어빈도) : 특정한 단어가 문서 내에 얼마나 등장할까?
- ▶ DF(문서빈도) : 단어 자체가 문서들에서 얼마나 자주 사용될까?
- ▶ 어디에 사용되는가?

TFIDF 값을 부여한 후, 코사인 유사도(Cosine Similarity)등을 이용해서 문서들의 유사도를 구하는데 흔히 사용됨.