고려대학교
빅데이터 연구회

KU-BIG

중간 점검

- 알고리즘 스터디 -

조원 : 박정진 이길재 정석원 최지인

0. 들어가기 전에

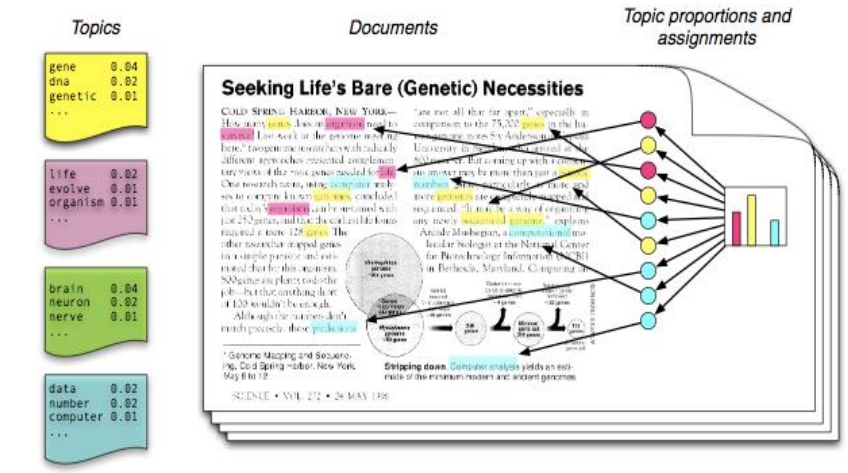
$$\begin{bmatrix} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{bmatrix}$$

<문서 - 단어 행렬 (DTM)>

- 우리 조는 Review data에 대해 DTM을 구축하여 모델링을 진행하고 있다.
- 구체적으로 진행하고 있는 지도학습 감정 분석 외에도, DTM을 기반으로 수행되는 알고리즘은 다양하다.
- 이번 시간에는, DTM 기반 알고리즘을 몇 가지 소개할 예정.

CONTENTS

토픽 모델이란?



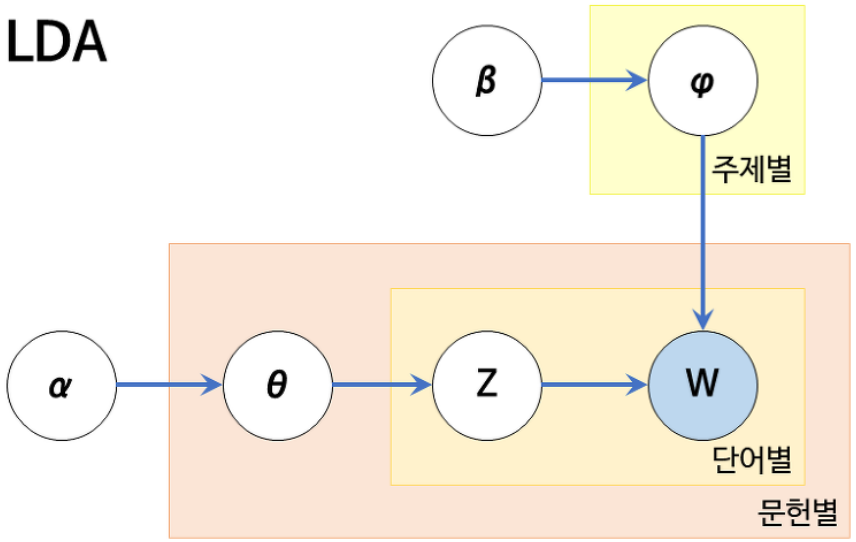
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

- 토픽 모델(Topic model)이란 문서 집합의 추상적인 "주제"를 발견하기 위한 통계적 모델

- 텍스트 본문의 숨겨진 의미구조를 발견하기 위해 사용되는 텍스트 마이닝 기법 중 하나

- 예시로, 개와 관련된 주제에는 '개'와 '뼈다귀'가, 고양이와 관련된 주제에는 '고양이'와 '생선'이 등장할 확률이 더 높음.

LDA (잠재 디리클레 할당, Latent Dirichlet Allocation)



- LDA 모형은 문서들을 말뭉치, 문서, 단어의 위계로 분류하여 분포를 부여하는 통계적 모델링
- 단어들의 모음은 주제를 나타낼 수 있고, 문서는 여러 주제로 이루어짐
- 이해를 위해서는 몇 가지 수학적 이론에 대한 간략한 이해가 필요.

베이즈 추론 - 1

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(X|\theta)p(\theta)}{p(X)}$$

<베이즈 정리 식>

- 사전 확률(prior probability, 선험적 확률): 관측에 의거하지 않고 가정하는 특정 사건이 일어날 확률
- 관측: 사건이 실제로 발생한 것
- 사후 확률(posterior probability): 실제 관측된 사건을 가지고 해당 사건이 일어날 확률을 더 정확하게 계산한 것.
- 가능도(우도, likelihood) : 관심 있는 사건에 대해 실제 사건 X가 얼마나 일어날 수 있는지를 보여주는 값

베이즈 추론 - 2

$$p(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(X|\theta)p(\theta)}{p(X)}$$

<베이즈 정리 식>

- 위의 식에서는, P(θ)는 사전 확률, P(θ|X)는 사후 확률, P(X|θ)는 가능도 라고 할 수 있음
- 사후 확률 ∝ 가능도 × 사전 확률
- 관측된 사건이 많아질수록 사후 확률(추정량)이 실제 확률(사전 확률)에 가깝게 개선됨
- 즉 베이즈 추론은 사후 관측된 값을 가지고 추정량을 개선해 가는 과정임.

켤레 사전 분포

	가능도	켤레 사전 분포
가짓수=2	Binomial	Beta
가짓수>2	Multinomial	Dirichlet

$$f(x_1,x_2,\cdots,x_k\,;p_1,p_2,\cdots,p_k) = f(X\,;P)$$

〈 P(X|θ) 〉

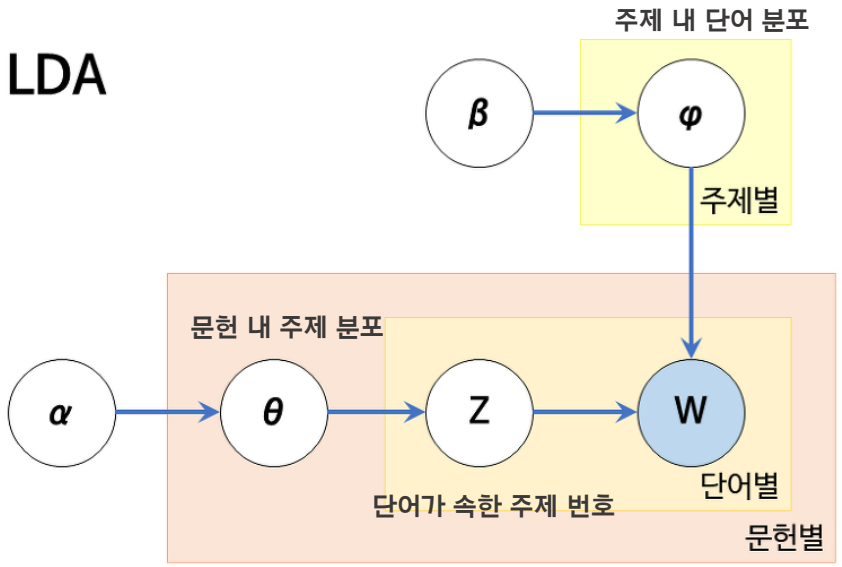
↓

$$Dir(p_1,p_2,\cdots,p_k\,;\alpha_1,\alpha_2,\cdots\alpha_k) = Dir(P;\alpha)$$

〈 P(θ) 〉

- 간단한 확률의 경우, 관측된 사건들의 가능도를 통하여 실제 확률들을 계산할 수 있으나, 이것이 조금만 복잡해질 경우에도 계산이 어려워짐
- 특정 관측 사건에 대해서는 미리 정해진 사전 확률 분포를 사용하면 계산이 매우 간편해짐
- 관측 사건이 다항 분포인 경우, 사전 확률을 다항분포의 켤레 사전 분포인 디리클레 분포로 사용하면 손쉽게 사후 확률을 계산 할 수 있음

다시, LDA…



$$P(Z|W) = \frac{P(W|Z)P(Z)}{P(W)}$$

- 우리가 실제로 관측 가능한 단어들을 통하여 **문헌 내 주제분포(θ)**와 **주제 내 단어분포(φ)**를 추론해야함 -> 베이즈 추론
- 이를 통해 각 단어가 어떠한 주제 번호에 속해 있는지(Z)를 구해낸다.

- 가정1 : 문헌 속의 단어 하나하나는 각각 주제를 나타냄

- 가정2 : 주제와 단어는 각각 디리클레 분포를 따름

깁스 샘플링

단어 m이 주제 j에 속하는 횟수

$$P(Z|W) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \cdot \frac{C_{dj}^{DT} + \alpha}{\sum_{j'} C_{dj'}^{DT} + T\alpha}$$

문헌 d의 단어들이 주제 j에 속하는 횟수

- 단어의 종류가 N개 라고 하면 N차 벡터에 관한 확률 분포를 계산해야 하는데, 이것은 매우 어려움
- **N차 자료를 1차 자료 N개가 모인 것**이라고 생각하고, 각각의 확률 분포를 구하여 이를 합치는 것이 깁스 샘플링의 아이디어!
- 즉 **단어 하나씩**을 보아가며 **전체 확률** 추론
- 연구에 따르면 이를 이용하면 문헌 d에 속하는 단어 m이 주제 j에 속할 확률은 주제 j에 속하는 모든 단어 중에서 단어 m이 차지하는 비중과 문헌 d에 속하는 모든 주제 중 주제 j가 차지하는 비중의 곱에 비례한다고 함.

LDA 수행

1. 하이퍼 파라미터 α, β, K 지정
   - α, β는 각각 문헌의 주제분포, 주제의 단어 분포를 결정함. [0,1]의 값을 지니며, 값이 클수록 더 많은 주제가 문헌에, 단어가 주제에 포함.
   - K는 나타날 수 있는 전체 주제의 숫자를 정함
   - 주어진 예시에서는 (α, β, K) = (0.1, 0.0001, 2)

2. 각각의 단어의 임의의 주제를 배정

W	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
Z	#1	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

LDA 수행

3. 문헌별 주제 분포, 주제별 단어 분포를 계산

θ	A	B	C	D	E	F
#1	1.1	2.1	0.1	2.1	2.1	2.1
#2	1.1	1.1	2.1	0.1	1.1	3.1

φ	cute	kit	eat	rice	cake	ham	bre	SUM
#1	1.001	0.001	2.001	1.001	3.001	0.001	2.001	9.007
#2	1.001	2.001	1.001	1.001	0.001	2.001	1.001	8.007

- 각각의 값이 정수가 아닌 이유는 모든 값에 α, β값이 합쳐져 있기 때문.

LDA 수행

4. 첫번째 단어를 골라 빼냄(여기서는 cute)

W	cute	kit	eat	rice	cake	kit	ham	eat	bre	rice	bre	cake	cute	ham	eat	bre	cake
Z	?	#2	#1	#2	#1	#2	#2	#1	#1	#1	#2	#1	#2	#2	#2	#1	#1

θ	A	B	C	D	E	F
#1	1.1	2.1	0.1	2.1	2.1	2.1
#2	1.1	1.1	2.1	0.1	1.1	3.1

φ	cute	kit	eat	rice	cake	ham	bre	SUM
#1	1.001	0.001	2.001	1.001	3.001	0.001	2.001	9.007
#2	1.001	2.001	1.001	1.001	0.001	2.001	1.001	8.007

θ	A	B	C	D	E	F
#1	**0.1**	2.1	0.1	2.1	2.1	2.1
#2	1.1	1.1	2.1	0.1	1.1	3.1

φ	cute	kit	eat	rice	cake	ham	bre	SUM
#1	**0.001**	0.001	2.001	1.001	3.001	0.001	2.001	8.007
#2	1.001	2.001	1.001	1.001	0.001	2.001	1.001	8.007

LDA 수행

θ	A	B	C	D	E	F
#1	**0.1**	2.1	0.1	2.1	2.1	2.1
#2	1.1	1.1	2.1	0.1	1.1	3.1

φ	cute	kit	eat	rice	cake	ham	bre	SUM
#1	**0.001**	0.001	2.001	1.001	3.001	0.001	2.001	8.007
#2	1.001	2.001	1.001	1.001	0.001	2.001	1.001	8.007

5. 첫 단어의 주제를 다시 정함

(문헌 A 내에 #1이 있을 확률) = 0.1 / (0.1+1.1) = 0.083

(#1 내의 단어가 cute일 확률) = 0.001 / 8.007 = 0.00012

따라서,

(A 안의 cute가 #1일 확률) = 0.083 * 0.00012 = 0.00008

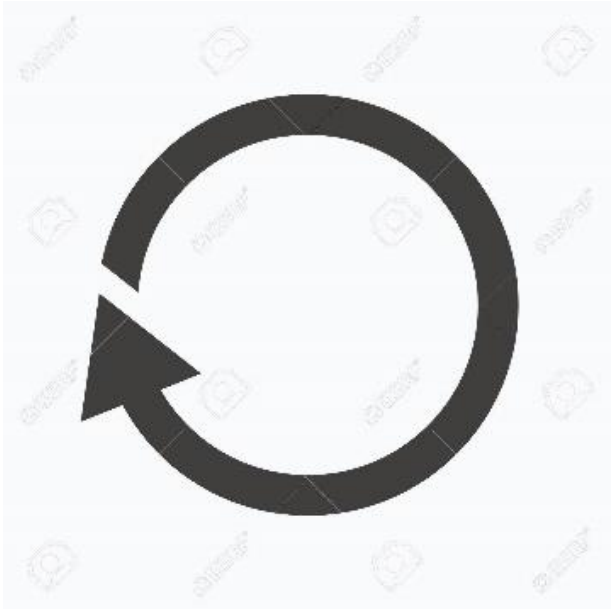
같은 방식으로 계산하면,

(A 안의 cute가 #2일 확률) = 0.916 * 0.125 = 0.114

이고, 높은 확률로 cute는 2번 주제에 속하게 됨

LDA 수행

6. 다음 단어에서 반복…



LDA 수행

7. 결과 값 도출 및 해석

θ	A	B	C	D	E	F
#1	0.1	3.1	0.1	2.1	3.1	3.1
#2	2.1	0.1	2.1	0.1	0.1	2.1

φ	cute	kit	eat	rice	cake	ham	bre	SUM
#1	0.001	0.001	3.001	2.001	3.001	0.001	3.001	11.007
#2	2.001	2.001	0.001	0.001	0.001	2.001	0.001	6.007

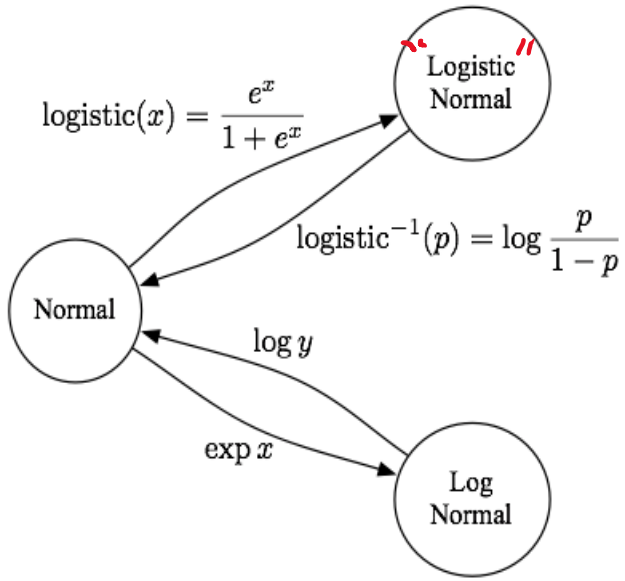
〈 최종 결과값 〉

- **계속된 반복을 통하여 최종적으로 안정된 값이 도출되며, 이를 통해 분류된 주제 번호에 알맞은 이름을 부여하면 됨.**

한계점

- 기본적으로 샘플링에 의존하기 때문에, 반복하여 값이 안정되어도 다른 값이 나올 가능성
- 단어가 희소할수록 랜덤 값에 의하여 결과가 도출 될 수 있음
- 단어의 분포만을 가지고 주제를 예측하는 것이기 때문에 실제 사람이 생각하는 것과 다른 결과가 나올 수 있음
- 주제간 연관관계 분석 어려움
- 적절한 K(주제의 개수)를 정하는 ISSUE

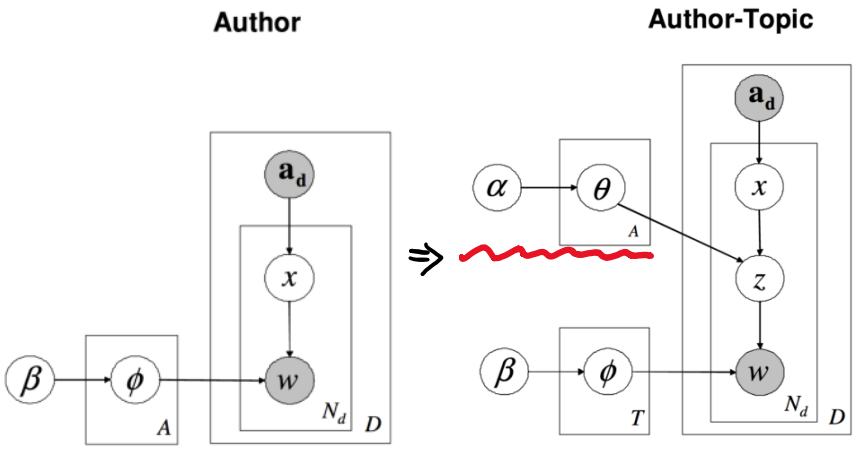
토픽 모형의 다양한 형태 – CTM (상관 토픽 모형)



〈로지스틱 정규 분포〉

- 상관 토픽 모형은 LDA와 이론적으로는 동일한 모형
- 사전확률분포로 디리클레 분포 대신 로지스틱 정규 분포를 사용
- 이름에서 알 수 있듯이, LDA에서는 제시해주지 않던 토픽 간의 Similarity, Dissimilarity에 관한 정보 또한 제시해 준다.

토픽 모형의 다양한 형태 – STM (구조 토픽 모형)



〈기존의 토픽모형 → STM〉

- 공학적 모델인 LDA, CTM과는 달리, STM은 문서의 정보에 기반한 통계적 분석을 가능케 한다.
- 기존의 토픽모형처럼 토픽에 관한 정보만이 아닌 문서들에 관한 정보 또한 모델링 과정에 반영되며, 이는 토픽의 분포에 대한 추가적 추론을 가능케 한다.
- 그러나, 정교한 만큼 모델의 적합에 비교적 오랜 시간이 걸린다는 단점을 지닌다.

PART. Ⅱ 감정 분석

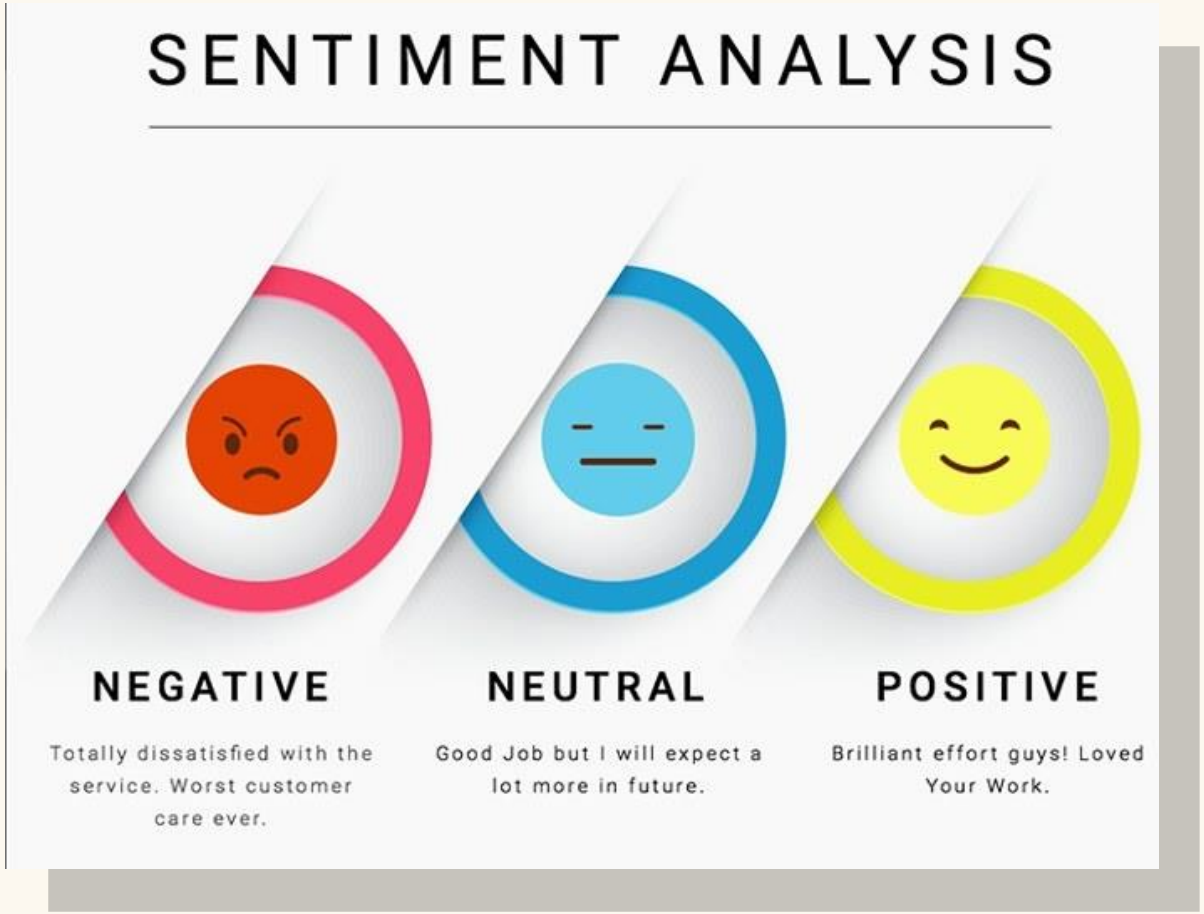
1. 정의
2. 감정 어휘 사전
3. 감정사전 예시
4. 한계

정의

감정분석?

- 텍스트에 담겨져 있는 **태도**를 추정하는 텍스트 데이터 분석
- 텍스트가 내포하는 **감정**을 추정
- 제품 후기 같은 온라인 리뷰, 트위터나 페이스북 같은 SNS를 통해 수집된 텍스트의 감성이 긍정인지 부정인지 혹은 중립인지 알아내는 데 사용

정의



- **예시 :**
  - 'dissatisfied', 'worst' **단어 사용**

    ➔ 부정적 감정의 텍스트
  - 'brilliant', 'loved' **단어 사용 ➔**

    긍정적 감정의 텍스트

정의

토픽 모형과의 차이점

토픽 모형	감정 사전
• 텍스트 데이터의 토픽이 무엇인지 판단 • 텍스트 자체 분석 ➔ 결과 제시	• 텍스트 데이터에 담긴 감정이 무엇인지 판단 • 보조자료 필요 [사전, 인간의 판단 등]

감정 어휘 사전

감정 어휘 사전?

- 어떤 단어가 긍정적 감정을 드러내는 데 주로 사용되고 어떤 단어가 부정적 감정을 드러내는데 주로 사용되는지를 모아 놓은 데이터
- 이미 구축된 감정사전을 이용하거나 연구목적에 맞게 감정사전을 구축하여 사용

감정 어휘 사전

감정 어휘 사전의 종류

- AFINN : **감정 어휘들에 대해 -5(가장 부정적)부터 +5(가장 긍정적)의 점수 부여**
- Opinion Lexicon : **어휘들을 수치 점수가 아닌 긍정과 부정 두 가지로만 분류**
- EmoLex : **긍정, 부정은 물론 분노, 공포, 놀람, 등과 같은 인간의 정서 정보 제공, 크라우드소싱으로 정서에 대한 정보축적**

감정사전 예시

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 3.4.4
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
#AFINN 감정어휘 사전 호출
get_sentiments("afinn")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## # A tibble: 2,476 x 2
##    word       score
##    <chr>      <int>
##  1 abandon       -2
##  2 abandoned     -2
##  3 abandons      -2
##  4 abducted      -2
##  5 abduction     -2
##  6 abductions    -2
##  7 abhor         -3
##  8 abhorred      -3
##  9 abhorrent     -3
## 10 abhors        -3
```

- 앞에 소개한 감정사전들은 R의 tidytext 라이브러리에 저장되어 있음
- 단어와 감정을 나타내는 점수로 구성

한계 1

- 긍정적 감정과 부정적 감정을 동시에 느끼는 단어의 구분
  - 같은 단어에 대하여 사전마다 다른 감정으로 구분할 수 있음
  - 사전에서 구분한 감정과 인간의 판단에 따른 감정의 차이
  - 예를 들어 [Ambivalence, 양가감정]의 경우

    ➔ Opinion Lexicon에서 부정적 감정으로 분류

한계 2

- 문맥에 따른 뉘앙스를 파악하기 쉽지 않음
  - 긍정적인 단어가 주를 이뤄도 부정 어휘(not, never, no 등)와 함께 쓰인 경우

    ➔ 부정적인 감정을 지닌 텍스트이지만 긍정적 텍스트로 분류
- 감정에 대한 문서와 문서의 감정을 구분하는 것이 어려움
  - 감정 자체에 대한 객관적 분석을 하더라도 해당 감정을 지닌 텍스트로 분류

한계 3

- 결정론적 관점
  - 단어를 사전에 정해진 감정으로 결정한 후 분석에 사용
    - 사전에서 정의되지 않은 단어는 (신조어 등) 사전을 이용한 분석 불가
    - 의도적, 비의도적 오탈자에 대한 분석 불가

프로젝트 적용 방안

- 감정 어휘 사전에 포함된 단어들을 사용

➔ 크롤링 완료한 텍스트 데이터에 적용

프로젝트 적용 방안

- AFFIN 사전 사용시 점수가 -5에 가까울 수록 rotten, +5에 가까울 수록 fresh일 것으로 추정
- Opinion Lexicon 사전 사용시 부정으로 분류되면 rotten, 긍정으로 분류되면 fresh일 것으로 추정
- EmoLex 사전을 사용시 영화 장르에 따른 차이도 분류 가능 예상


Do you have any
question?
Thank you
for your attention.

“감사합니다.”