컴퓨테이셔널 방법론 부트캠프

06 | NLP (Natural Language Processing)

서울대학교 사회과학대학

박기범

목차

강의 내용

- 형태소 분석기를 활용하여 wordcloud 만들기
  - 기초 자연어처리 방법론 학습
  - 형태소분석기 설치 및 활용
  - 데이터를 형태소분석기를 통해 처리
  - Wordcloud tool을 활용하여 결과 시각화
- TF-IDF를 활용한 문서 검색, Word2Vec을 활용한 유사 단어 검색
  - 기계학습 기반 자연어처리 방법론 학습
  - TF-IDF의 개념 및 동작 원리 학습
  - 네이버 영화 리뷰 감성 데이터셋(NSMC) 탐색
  - TF-IDF 를 활용하여 문서 검색

강의 내용

- Word2Vec을 활용한 유사어 탐색
  - Word2Vec의 개념 및 동작 원리 학습
  - Word2Vec을 활용한 단어 유사성 탐색 실습
- BERT를 활용한 감정 분석
  - 딥러닝 기반 자연어처리 방법론 학습
  - BERT 모델 설치 및 감정 분석 실습

세션1: NLP 기초

- 형태소 분석기를 활용하여 wordcloud 만들기
  - 기초 자연어처리 방법론 학습
  - 형태소분석기 설치 및 활용
  - 데이터를 형태소분석기를 통해 처리
  - Wordcloud tool을 활용하여 결과 시각화

자연어처리
NLP(Natural Language Processing)

- 인간의 언어 현상을 컴퓨터와 같은 기계를 이용해서 모사 할 수 있도록 연구하고 이를 구현하는 인공지능의 주요 분야
- 정보검색, QA 시스템, 문서 자동 분류, 신문기사 클러스터링, 대화형 Agent 등
- NLP의 역사
  - 초기 – 규칙 기반 NLP(1950s~1990s)
  - 중기 – 통계적 NLP(1990s~2010s)
  - 현재 – 뉴럴 NLP(2010s~)

NLP vs. Text mining

- 비슷하지만 또 조금은 다름
  - NLP: 컴퓨터가 인간의 언어 현상을 이해하고 모사하는 것이 목적
  - Text mining: 텍스트에서 의미 있는 정보를 추출
- 최근에는 그 경계가 점점 희미해지는 추세

Text Mining	Natural Language Processing
Aim of text mining is to extract useful insights from structured & un-structured text.	Aim of NLP is to understand what is conveyed in speech.
Text Mining can be done using text processing languages like Perl, statistical models, etc.	NLP can be achieved using advanced machine learning models, deep neural networks, etc.
Outcome: • Frequency of words • Patterns • Correlations	Outcome: • Semantic meaning of text • Sentimental analysis • Grammatical structure

NLP 주요 과제

- 텍스트, 음성 프로세싱
  - OCR, 음성 인식, TTS(text-to-speech) 등
- 형태적 분석
  - 표제어 추출, 형태 분리, 품사 태깅, 어간 추출 등
- 어휘 의미 분석
  - 개체명 인식, 감정 분석, 전문용어 추출 등
- 고수준 NLP
  - 자동 요약, 기계 번역, 자연언어 이해, 자연언어 생성, 질의응답 등

일반적인 NLP 처리 프로세스

- 전처리 (Preprocessing)
  - 데이터를 학습에 적합하도록 정리
  - 클렌징, 토큰화, 품사 태깅, 어근 추출, 표제어 추출, 불용어 제거 등
- 벡터화 (Feature Vectorization)
  - 각각의 데이터를 컴퓨터로 처리할 수 있도록 수치화
- 모델 학습 (ML Modeling)
  - 수치화된 데이터에서 의미를 학습하고 추론

형태소 분석
Morphological analysis

- 단어를 형태소 단위로 분리해내는 과정
- 자연어 처리의 가장 기초적인 절차
- 형태소 분석 이후 구문 분석이나 의미 분석을 수행
- 한국어의 경우 이 단계에서부터 이미 난이도가 높은 편
  - 형태적 변형이 다양
    - 노랗다, 누렇다, 누르스름하다, 누리끼리하다
  - 어휘 형태소와 문법 형태소의 결합과정에서 형태적 변형 발생
    - (하늘을) 날다 → (하늘을) 날/는 → (하늘을) 나는
  - 형태소만으로 정의 불가한 것들이 많음
  - 한글은 음소 단위로 결합하므로 음절 단위의 글자를 음소 단위로 나누어야 함

토큰화
Tokenization

- 토큰(token): 더 이상 나눌 수 없는 언어 요소 단위
- 토큰화: 말뭉치(Corpus)로부터 토큰을 추출해내는 것
  - 문장 토큰화: 한 문장 단위로 추출
  - 단어 토큰화: 단어 단위로 추출
- 언뜻 간단해 보이지만, 현실은...
  - "아버지가방에들어가신다"
  - "6억원 이상 9억원 미만은 최대 0.5%, 9억원 이상은 최대 0.9% 이내"
  - "너무햌ㅋㅋㅋㅋㅋ"
  - "ㅇㄱㄹㅇㅂㅂㅂㄱ"

품사 태깅
POS(Part-of-speech) Tagging

- 형태소 분석을 통해 나온 결과 중 가장 적합한 형태의 품사를 표시
- '나는' 이라는 어절에 대한 형태소 분석이 다음과 같다면,
  - 나는 → 나 + 는 (ex. 너와 나는 ...)
  - 나는 → 날 + 는 (ex. 하늘을 나는 새는 ...)
- 다음과 같이 적절한 품사를 부착하는 것
  - 나는 오늘 학교에 갔다' → '나(대명사)+는(조사) 오늘 학교+에 가다+았+다'
  - 하늘을 나는 새를 보았다' → '하늘+을 날(동사)+는(관형형어미) 새+를 보다+았+다'

품사 태깅
POS(Part-of-speech) Tagging

한국어와 영어의 품사

한국어에는 9품사가 있고 영어에도 9품사가 있습니다. 품사 수는 동일하지만 품사 구성은 다릅니다.

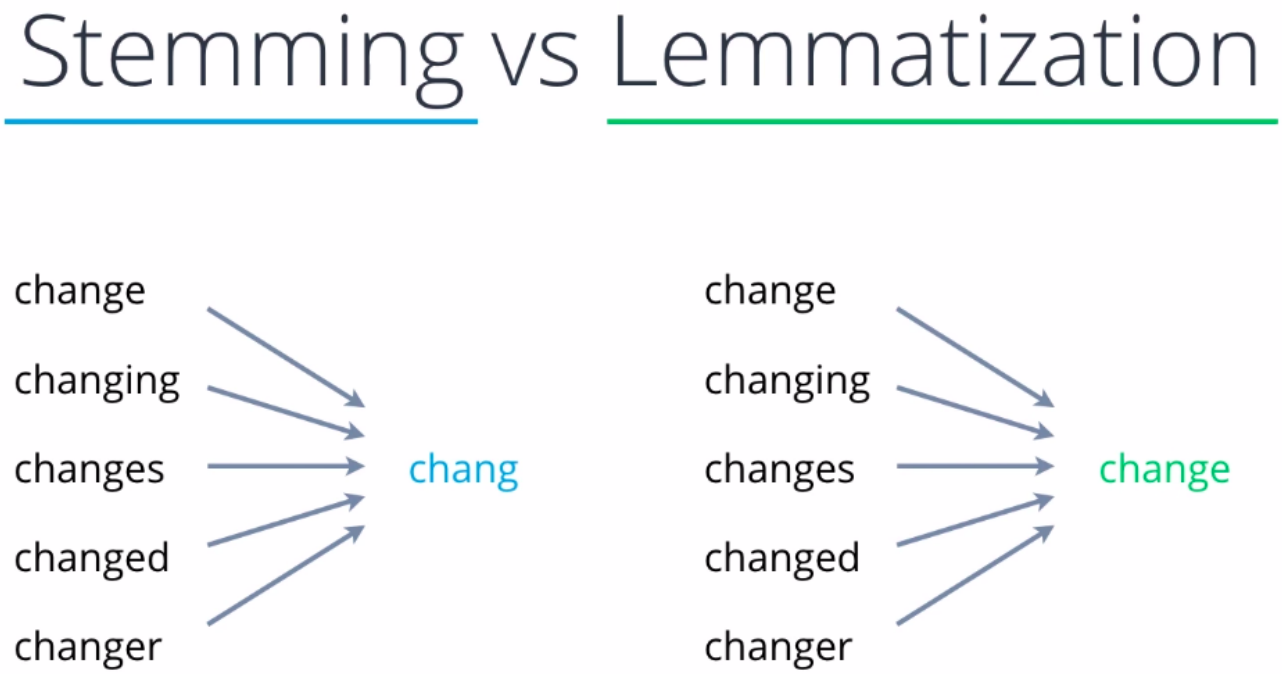
순서	품사명	한국어	영어
1	명사(noun)	O	O
2	대명사(pronoun)	O	O
3	동사(verb)	O	O
4	형용사(adjective)	O	O
5	관형사		
6	수사	O	
7	부사(adverb)	O	O
8	조사	O	
9	감탄사	O	
10	관사(article)		O
11	전치사(preposition)		O
12	접속사(conjunction)		O
13	감탄사(Interjection)	O	O

어간 추출
Stemming

- 어형이 변형된 단어로부터 접사 등을 제거하고 그 단어의 어간을 분리해 내는 것
- 관련이 있는 단어들이 일정하게 동일한 어간으로 맵핑되게 하는 것
- 어간은 반드시 어근과 같아야 할 필요는 없음
- 단어를 보고 대략적으로 어미를 잘라서 어간을 추출함
  - works, worked, working → work
  - plays, played, playing → play
  - amuses, amused, amusing → amus (?)

표제어 추출
Lemmatization

- 일반적으로 어간 추출보다 더 정확한 어근 단어를 탐색
- 단, 품사 정보가 함께 제공되어야 높은 성능



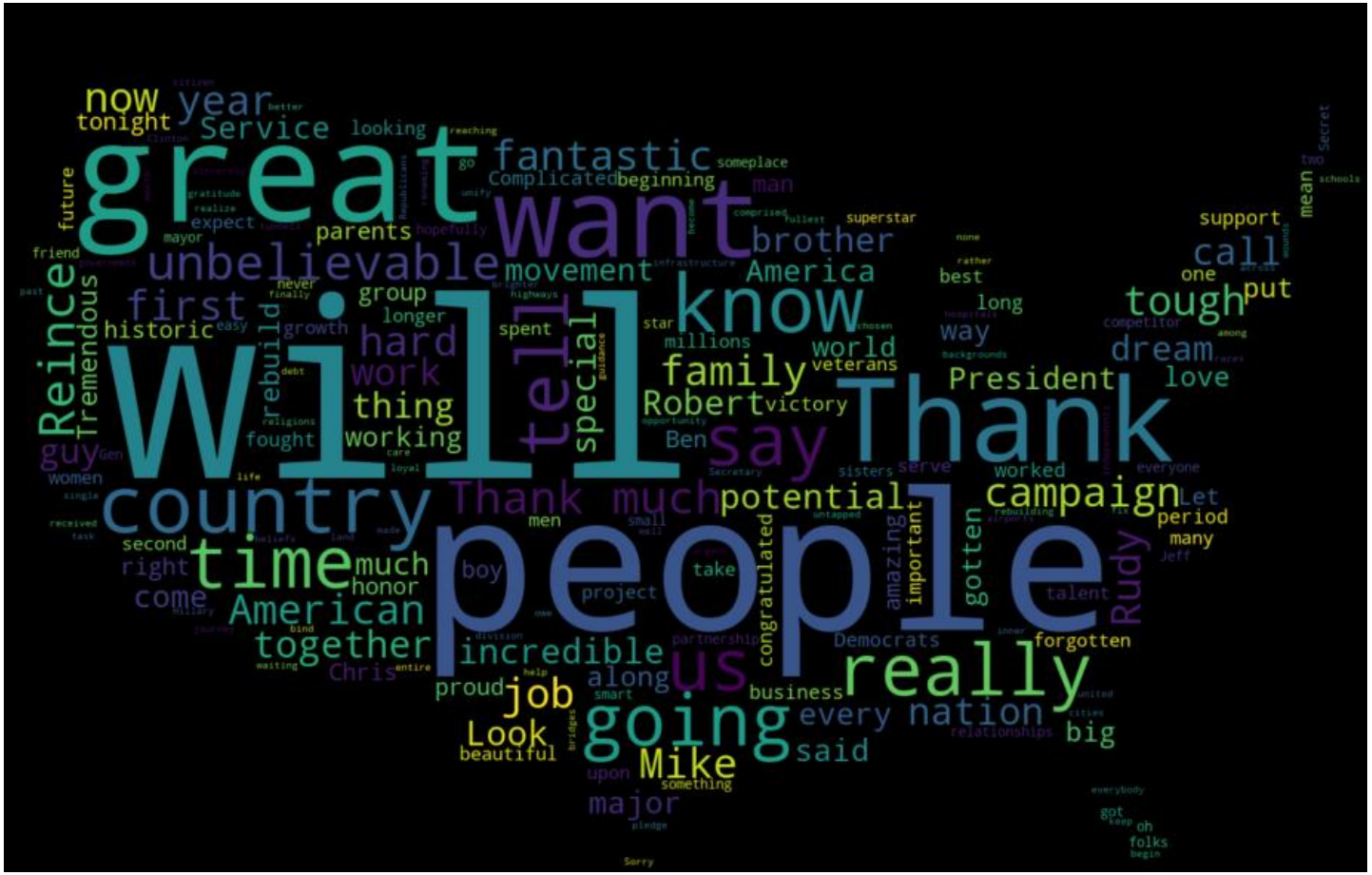
불용어
Stopwords

- 문장을 구성하지만 분석에 큰 의미가 없는 단어들
  - i, my, me, myself, we, our, ours, ...
- 의미는 없지만 말뭉치에 굉장히 빈번하게 나타나므로, 제거하고 분석하는 것이 분석에 효과적일 수 있음
- {'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'}
- https://www.ranks.nl/stopwords/korean

코엔엘파이
KoNLPy

- 한국어 정보 처리를 위한 패키지
- https://konlpy.org/
- 여러 종류의 형태소 분석기를 파이썬에서 사용 가능
- pip 등 파이썬 패키지 매니저를 통해 설치 가능
- 다양한 형태소 분석 기능 제공
  - .morphs() : 형태소 별 분리
  - .nouns() : 명사만 추출
  - .pos() : 형태소 태깅

단어구름
Wordcloud



세션1 실습. Wordcloud 만들기

- 실습 코드를 통해 진행합니다.
- https://colab.research.google.com/drive/1Kwmae_mKSycYJxwVkMBR6whRkZ5AYqb8?usp=sharing

세션2: NLP 심화

- TF-IDF를 활용한 문서 검색, Word2Vec을 활용한 유사 단어 검색
  - 기계학습 기반 자연어처리 방법론 학습
  - TF-IDF의 개념 및 동작 원리 학습
  - 네이버 영화 리뷰 감성 데이터셋(NSMC) 탐색
  - TF-IDF 를 활용하여 문서 검색

임베딩 (Embedding)

- 컴퓨터를 사용하여 처리하기 위해서는 수치화가 필수
- 주어진 텍스트를 대표하는 수치적 표현은?

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

단어가방
Bag of Words

- 문장을 토큰화하여, 문장에 나타난 토큰들로 문장을 대표해보자
  - I am a boy → [I, am, a ,boy]
  - You are a girl → [You, are, a, girl]
  - He is a teacher → [He, is, a, teacher]
- 모든 문장이 일정한 크기의 벡터로 나타낼 수 있도록, 전체 어휘 크기의 column size를 가지는 벡터로 위 문장들을 표현해본다면?

column	0	1	2	3	4	5	6	7	8	9
(meaning)	i	am	a	boy	you	are	girl	he	is	teacher
ex. I am a boy	1	1	1	1	0	0	0	0	0	0

카운트벡터
CountVector

- 문장에 출현한 단어 빈도를 활용하여 feature vector를 생성

	i	am	a	boy	you	are	girl	he	is	teacher
I am a boy	1	1	1	1	0	0	0	0	0	0
You are a girl	0	0	1	0	1	1	1	0	0	0
He is a teacher	0	0	1	0	0	0	0	1	1	1

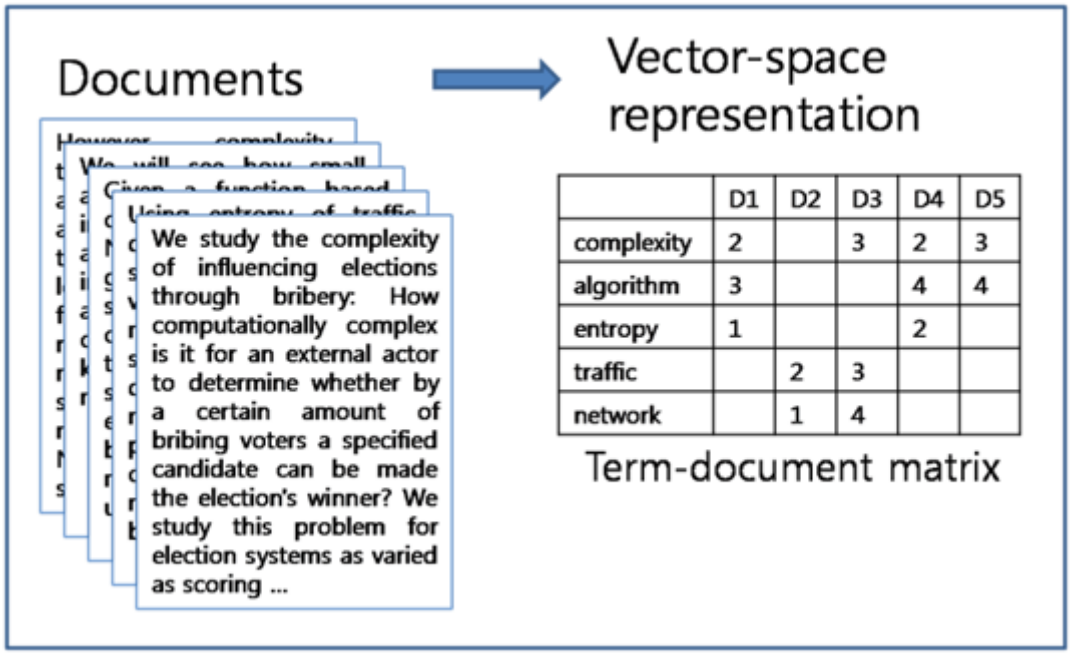
- 단어의 출현 순서나 문맥의 영향은 무시됨

	it	is	fun	at	not	good	bad	all
It is good, not bad at all.	1	1	0	1	1	1	1	1
It is bad, not good at all.	1	1	0	1	1	1	1	1

- 어휘 수가 증가함에 따라 각각의 수치적 표현은 매우 희박한 데이터 밀도를 가지게 됨 (해당되는 어휘 부분을 제외한 값의 대부분이 0)

단어 빈도
TF(Term-Frequency)

- 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값
- 해당 단어의 빈도가 높을수록 문서 내에서 중요할 수 있음
- TF는 주로 단어의 문서 내 출현 빈도를 사용



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

역문서빈도
IDF(Inverse Document Frequency)

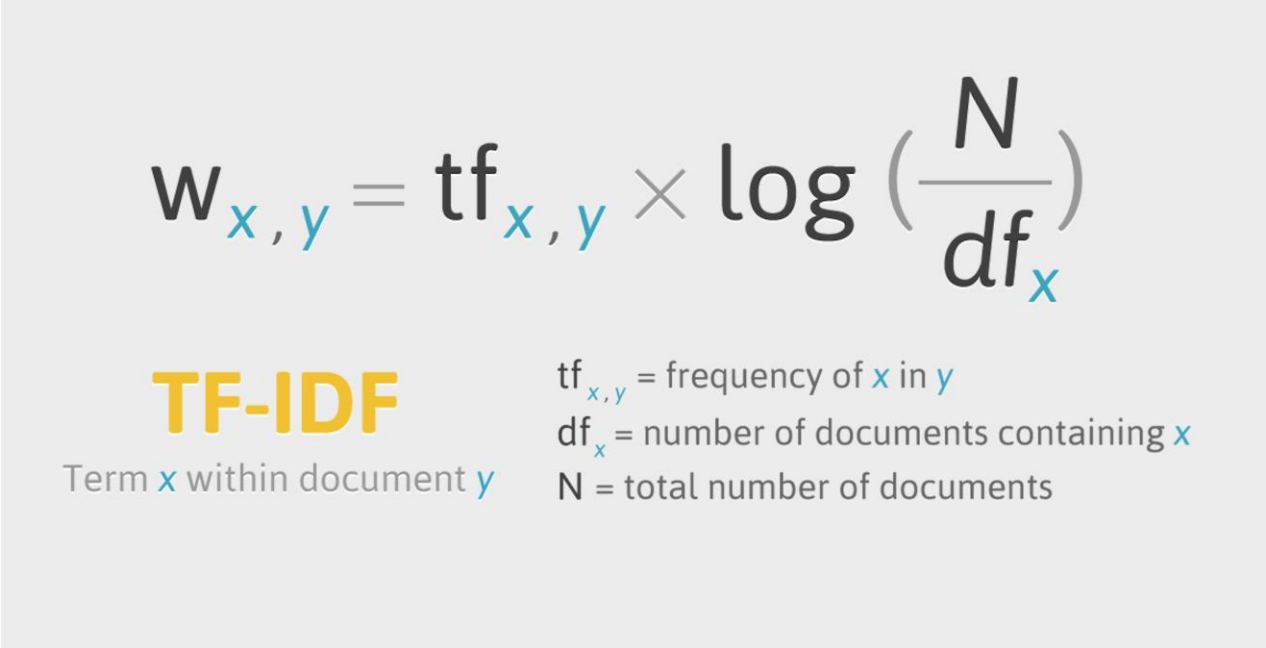
- 특정한 단어가 전체 말뭉치 내의 문서 중에서 얼마나 자주 등장하는지에 대한 역수 값
- 해당 단어의 전체 말뭉치 내 빈도가 높을수록 단지 흔하게 나타나는 단어일 수 있음 (중요하지 않은 단어일 수 있음)
- IDF는 주로 전체 문서의 수를 해당 단어가 출현한 문서의 수로 나눈 후 로그를 취해서 사용

Total Number of Documents	100,000,000

Term of Interest	Number of Documents Containing that Term
a	100,000,000
boat	1,000,000
mobile	100,000
mobilegeddon	1,000

TF-IDF

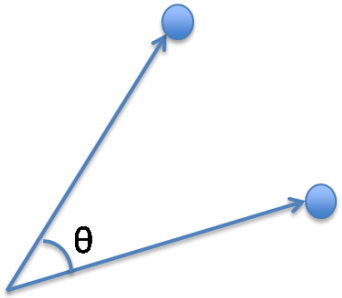
- TF와 IDF를 곱한 값 (TF×IDF)
- 문서 내에 단어의 빈도와 중요성의 관계(TF)와 말뭉치 전체에서 단어의 빈도와 중요성의 관계(IDF)를 동시에 고려



코사인 유사도
Cosine Similarity

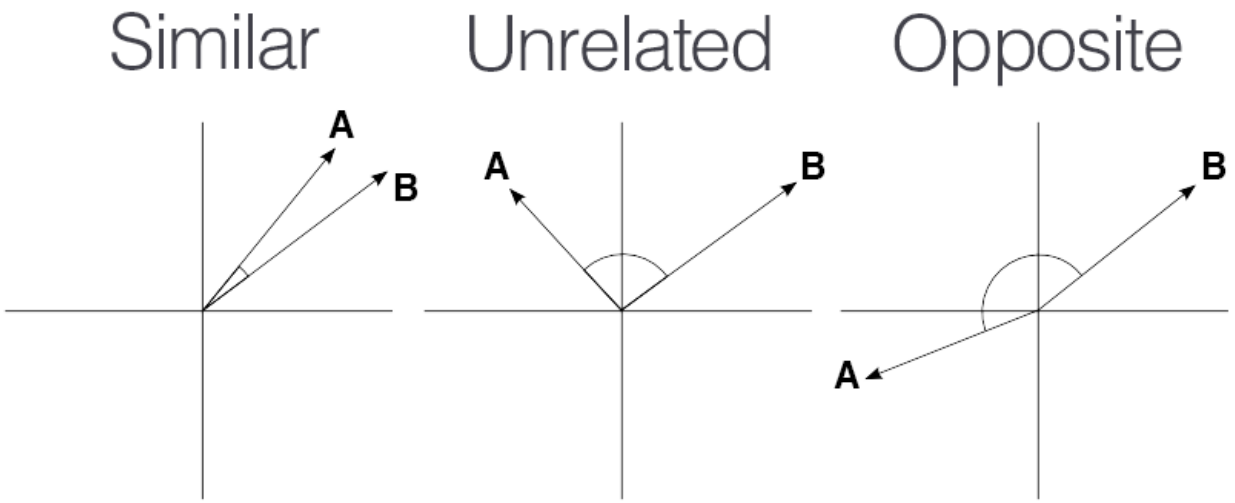
- 두 벡터의 유사한 정도를 판단하는 지표 중 하나
- 두 벡터 간 각도의 cosine 값을 이용하여 측정

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



문서 검색
Document Searching

- 문서(document)를 대표하는 TF-IDF 벡터와 검색어(query)를 대표하는 TF-IDF 벡터의 cosine similarity를 구함으로써 문서와 검색어의 유사한 정도를 계산할 수 있음



주요 라이브러리

- NumPy(넘파이): 수학과 과학 연산을 위한 파이썬 라이브러리
  - n차원 배열의 벡터 연산을 위한 기능 지원
- Pandas(판다스): Dataframe(데이터프레임) 형태의 자료 관리
- Scikit-learn(사이킷-런): 파이썬 머신러닝 라이브러리
  - 다양한 기계학습 관련 기능 제공
    - 분류(classification)
    - 군집화(clustering)
    - 특성 추출(feature extraction)
    - 모델 선택 및 평가(model selection & evaluation

데이터셋: NSMC

- Naver Sentiment Movie Corpus
- 네이버의 영화 페이지 댓글 및 긍정/부정 레이블 포함
  - document: 댓글 내용
  - label: 0/1 (부정/긍정)

```
id	document	label
9976970 아 더빙.. 진짜 짜증나네요 목소리	0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1
10265843	너무재밋었다그래서보는것을추천한다	0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.	0
7797314 원작의 긴장감을 제대로 살려내지못했다.	0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겟다 납치.감금만반복반복.
7156791 액션이 없는데도 재미 있는 몇안되는 영화	1
id	document	label
6270596 굳 ㅋ	1
9274899 GDNTOPCLASSINTHECLUB	0
8544678 뭐야 이 평점들은.... 나쁘진 않지만 10점 짜리는 더더욱 아니잖아	0
6825595 지루하지는 않은데 완전 막장임... 돈주고 보기에는....	0
6723715 3D만 아니었어도 별 다섯 개 줬을텐데.. 왜 3D로 나와서 제 심기를 불편하게 하죠??	0
7898805 음악이 주가 된, 최고의 음악영화	1
6315043 진정한 쓰레기	0
6097171 마치 미국애니에서 튀어나온듯한 창의력없는 로봇디자인부터가,고개를 젖게한다	0
8932678 갈수록 개판되가는 중국영화 유치하고 내용없음 폼잡다 끝남 말도안되는 무기에 유치한cg남무 아 그립다 동사서독같은
```

세션2 실습. TF-IDF

- 실습 코드를 통해 진행합니다.
- https://colab.research.google.com/drive/1xb7-LKx3EMQFL0dfj9iK8oeXDSl_m9rV?usp=sharing

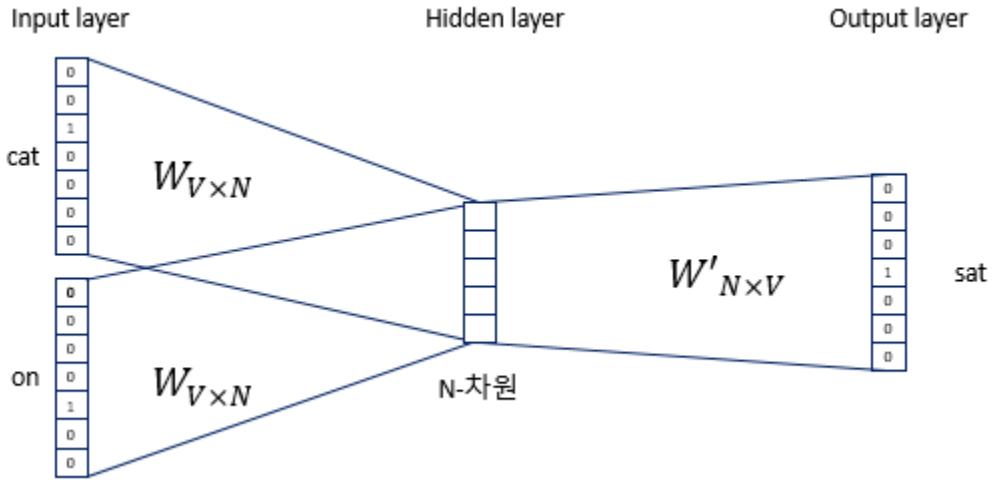
세션3: NLP 딥러닝

- Word2Vec을 활용한 유사어 탐색
  - Word2Vec의 개념 및 동작 원리 학습
  - Word2Vec을 활용한 단어 유사성 탐색 실습

- BERT를 활용한 감정 분석
  - 딥러닝 기반 자연어처리 방법론 학습
  - BERT 모델 설치 및 감정 분석 실습

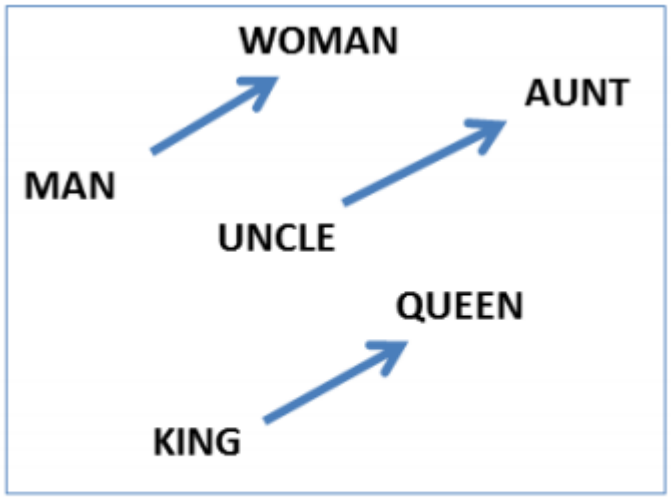
Word2Vec

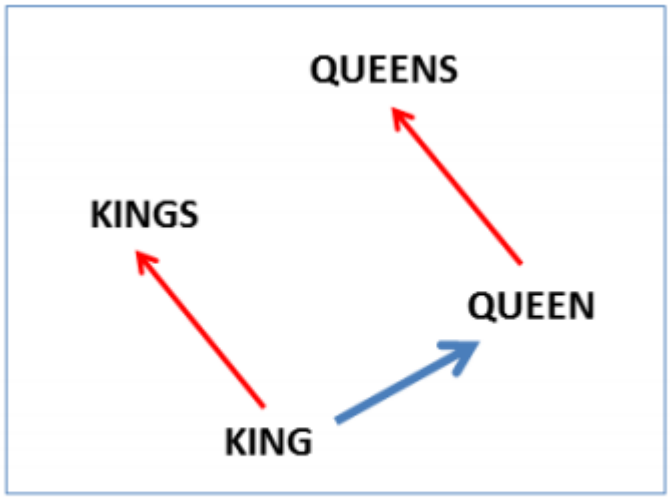
- 얕은 신경망 구조
- 분산된 표상(Distributed Representation)
- 비슷한 위치에 나타나는 단어는 비슷한 의미를 가진다는 가정
- 반복되는 학습을 통하여 각 단어를 일정한 차원의 벡터로 표현



Word2Vec의 장점

- 단어의 유사성을 cosine similarity로 계산 가능!
- 단어가 벡터로 변환되었으니 단어 간 연산이 가능!

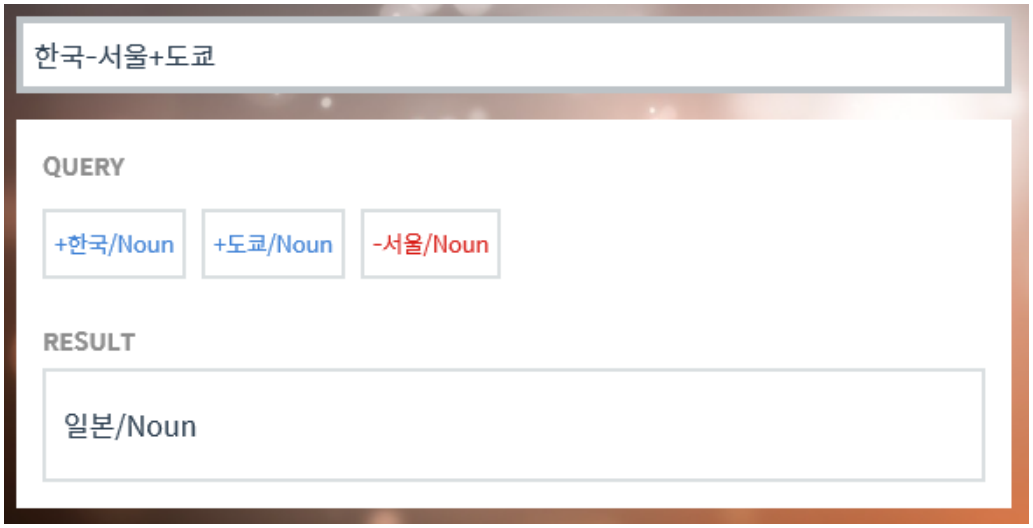




(Mikolov et al., NAACL HLT, 2013)

한글 Word2Vec

- https://word2vec.kr/search/



- 직접 몇 가지 만들어 입력해보세요
  - 한국 - 서울 + 파리 = ?
  - 컴퓨터공학 - 자연과학 + 인문학 = ?
  - 사랑 + 이별 = ?

세션3-1 실습. Word2Vec

- 실습 코드를 통해 진행합니다.
- https://colab.research.google.com/drive/1b04puQHvh_vWE3pftMcq_WnpgLc56qCF?usp=sharing

세션3-2 실습. BERT 기반 긍정/부정 분류기

- 실습 코드를 통해 진행합니다.
- https://colab.research.google.com/drive/1MjP1y1yp_hg2mIR8cmXBkqbGq6BKU1t5?usp=sharing
- 런타임 유형을 GPU로 실행하시기 바랍니다.
- 실행 시간이 꽤 오래 걸리기 때문에 미리 "모두 실행" 하시면 좋습니다.

BERT
Bidirectional Encoder Representations from Transformers

- 구글이 2018년 11월 발표
- Pre-training + Fine-tuning 구조
- 당시 NLP competition의 거의 모든 영역 State-of-the-Art 석권
- Human Performance를 능가하는 성능
- 현재도 다양한 BERT 파생 모델이 만들어지고 있음

BERT Pre-training

- 엄청난 크기의 말뭉치(Corpus)로 모델을 선 학습
  - 영어 기반 모델의 경우 BooksCorpus(800M words), Wikipedia(2,500M words)를 통해 학습
  - 다국어 모델의 경우 Wikipedia 상위 100개의 언어 자료로 학습
- 두 가지 방법으로 학습
  - Masked Language Model: 문장 내 랜덤한 단어를 마스킹 또는 랜덤한 단어로 치환한 상태에서 모델이 해당 단어를 추측하는 학습
  - Next Sentence Prediction: 두 문장을 주고 두 번째 문장이 첫 문장의 바로 다음에 오는 문장이 맞는지 여부를 판단하는 학습
- 학습에 큰 시간과 자원이 소요 – 미리 학습 완료된 모델을 배포함

BERT Fine-tuning

- 선 학습된 모델을 바탕으로 특정 목적에 특화된 학습을 수행하는 것
- 적은 데이터만으로 빠른 학습이 가능하며, 그 결과 또한 수많은 데이터를 모아 학습한 것과 비등하거나 더 뛰어남
- "이미 한글을 통달한 사람에게 한글 법조문을 가르친다면, 한글을 모르는 사람에게 한글 법조문을 가르치는 것 보다 더 적은 예시만으로도 더 빨리 가르칠 수 있고 그 결과도 더 좋을 것이다"

- 실습에서는 Fine-tuning만 시도해봅니다.

SQuAD

- The Stanford Question Answering Dataset
- https://rajpurkar.github.io/SQuAD-explorer/

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978

SQuAD Example

" . . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."

- Question 1: "Which laws faced significant opposition?"
  Plausible Answer: later laws
- Question 2: "What was the name of the 1937 treaty?"
  Plausible Answer: Bald Eagle Protection Act

CoQA

- A Conversational Question Answering Challenge
- https://stanfordnlp.github.io/coqa/

Leaderboard

Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* (Reddy & Chen et al. TACL '19)	89.4	87.4	88.8
1 Sep 05, 2019	RoBERTa + AT + KD (ensemble) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	91.4	**89.2**	**90.7**
1 Apr 22, 2020	TR-MT (ensemble) *WeChatAI*	**91.5**	88.8	**90.7**
2 Sep 05, 2019	RoBERTa + AT + KD (single model) *Zhuiyi Technology* https://arxiv.org/abs/1909.10772	90.9	**89.2**	90.4

CoQA Example

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

- Q1: Who had a birthday?
  A1: Jessica
  R1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

CoQA Example

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

- Q2: How old would she be?
  A2: 80
  R2: she was turning 80

CoQA Example

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

- Q3: Did she plan to have any visitors?
  A3: Yes
  Q4: How many?
  A4: Three
  Q5: Who?
  A5: Annie, Melanie and Josh

마무리

- 재미와 관심 - 형태소 분석기를 활용하여 wordcloud 만들기
- 통계적 방법 - TF-IDF를 활용한 문서 검색
- 최신 기술 - BERT를 활용한 감정 분석

- 텍스트 데이터가 많은 영역에서 큰 힘을 가진 도구가 될 수 있습니다.
- 개발하기는 어렵지만 가져다 쓰는 것은 그보다 훨씬 쉽습니다.
- 연합전공 정보문화학 많이 사랑해주세요.
- 긴 시간 수고하셨습니다!