Lecture 1

텍스트 마이닝 기초

텍스트 마이닝 (Text Mining)이란?

- Wikipedia: the process of **deriving high-quality information from text**.
- High-quality information is typically derived through the devising of **patterns and trends** through means such as **statistical pattern learning**.
- The overarching goal is, essentially, to **turn text into data for analysis**, via application of natural language processing (NLP) and analytical methods.
- 텍스트 데이터에서 자연어처리(Natural Language Processing, NLP) 기술을 바탕으로 유의미한 패턴 또는 지식을 추출하는 과정

텍스트 마이닝 (Text Mining)이란?

- Turn unstructured text into structured data,
  - 일정한 길이 (sparse or dense) 의 vector로 변환 (임베딩)
- and use the structured data in various tasks such as
  - text classification, clustering, sentiment analysis, document summarization, translation, prediction, etc.
  - 변환된 vector에 머신러닝 (딥러닝) 기법을 적용

텍스트 마이닝 유형

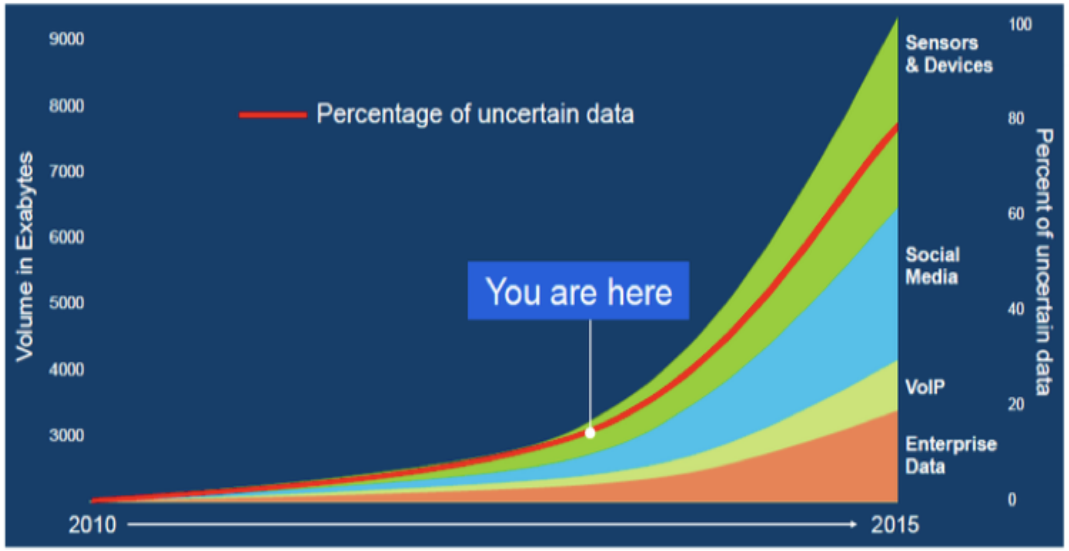
1. 설명적 마이닝(descriptive mining): 텍스트 집합에 있는 의미나 개념을 찾아내거나 이해를 돕는 형태 (분류, 검색, 여론조사 등)
2. 예측적 마이닝(predictive mining): 텍스트에 내포된 정보를 의사결정에 활용하는 형태 (질문 자동답변, 구매 예측, 주가예측, 스팸분류 등)

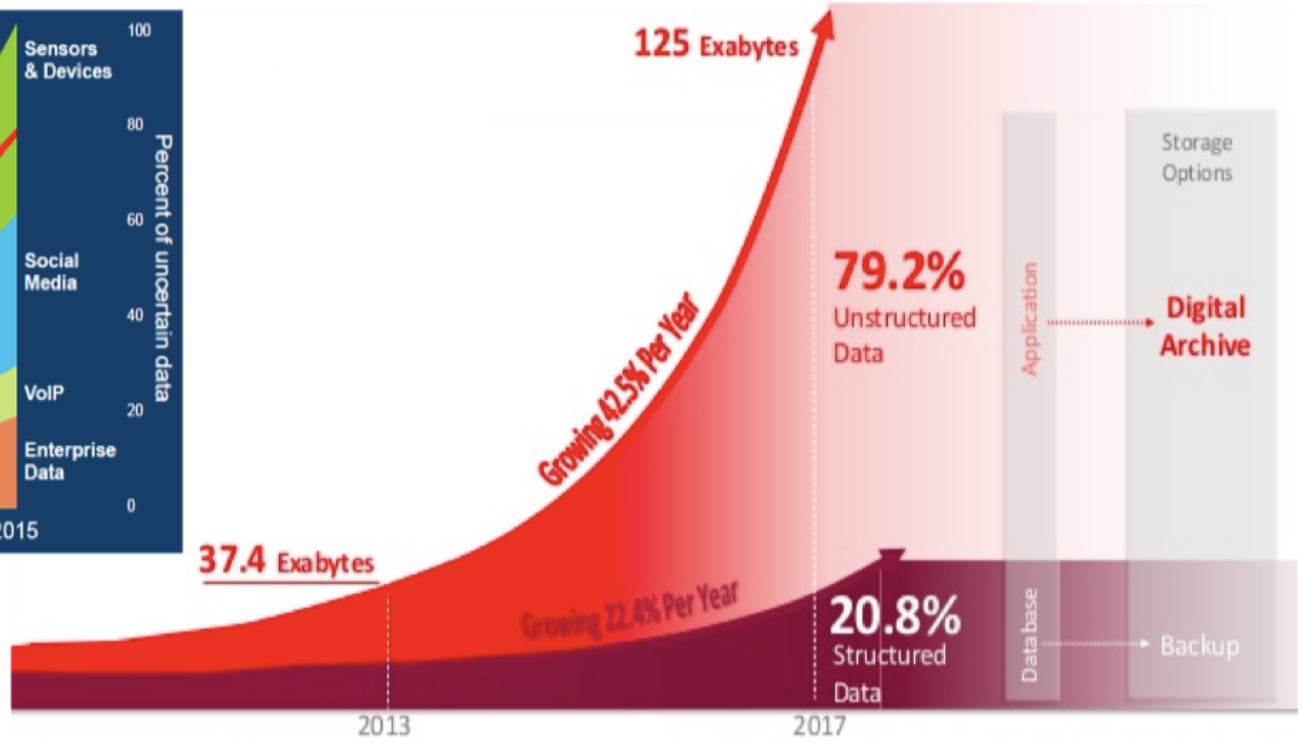
텍스트 마이닝 유형	활용분야	
	실무	연구
검색 (Information Retrieval)	스팸 필터링	사회동향 분석
분류 (Classification)	이슈 검출/트래킹	소셜미디어 분석
군집화 (Clustering)	정보검색	이슈 트래킹
웹마이닝 (Web Mining)	자살률 예측	온라인 행동 분석
정보추출 (Information Extraction)	주가 예측	연구분야 탐색
개념추출 (Concept Extraction)	소비자 인식 조사	질병관계 예측
~~자연어처리 (NLP)~~	경쟁사 분석	정책전략 수립

텍스트 마이닝이 중요한 이유

- 비정형 데이터의 폭발적 증가
  - 잠재적 가치를 포함하는 비정형 데이터(unstructured data)가 대규모로 생성됨과 동시에 비정형 데이터 속에서 미래의 의사결정에 관련된 유용한 정보를 찾아내어 활용하는 작업이 매우 중요해짐
  - 실제로 생산되는 데이터의 70~80 는 비정형 테이터에 해당함 (기사, 블로그, 문서, 보고서 등)
- 텍스트 데이터의 폭발적 증가
  - 소셜 네트워크 서비스(Social Network Service, SNS)를 통한 온라인 양방향 커뮤니케이션이 활성화됨
  - 4차산업혁명과 사물 인터넷(Internet of Things, IoT) 등 빅데이터 관련 기술이 급진적으로 발전함

텍스트 마이닝이 중요한 이유





텍스트 마이닝이 중요한 이유

- 가장 흔하고 접하기 쉬운 데이터
- 텍스트가 존재하지 않는 곳은 없으며, 다양한 서비스를 통해 수많은 텍스트 데이터가 생산됨
- 온라인 비정형 텍스트 데이터의 대부분이 SNS에서 발생함 (Twitter, Facebook, YouTube, 블로그, 커뮤니티 등)
- 웹에서 사용자들은 주로 텍스트 데이터를 활용해 콘텐츠를 생성하고 의사소통함
- 다양한 형태의 비정형 데이터(오디오, 비디오, 각종 센서 등)가 텍스트 형태로 변형되어 활용됨 → 음성-텍스트 변환(Speech to Text, STT)
- 웹 크롤링(web crawling), Open API(Application Programming Interface) 등의 활성화로 텍스트 데이터 수집 및 확보가 용이해짐

텍스트 마이닝이 어려운 이유

언어적 한계점

- 사람들이 작성한 문장은 맞춤법과 철자가 틀리고, 단어를 섞어 쓰고, 축약되는 등 규칙을 지키지 않음
- 동의어, 동형(동음) 이의어가 포함되거나 약어의 의미가 분야별로 다를 수 있음
- 문맥(context) 에 따라서 의미가 많이 달라지며, 애매한 표현이 많이 나타남 → 추상적 개념의 모호함

텍스트 마이닝이 어려운 이유

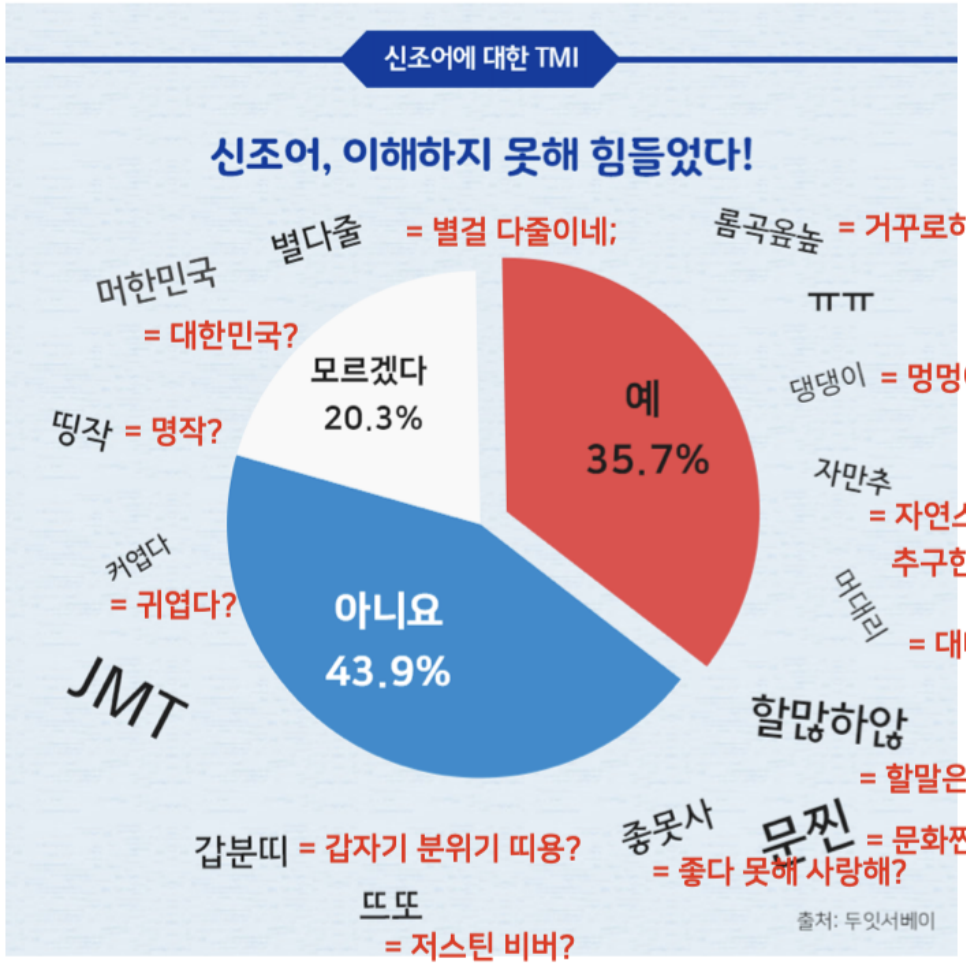
데이터적 한계점

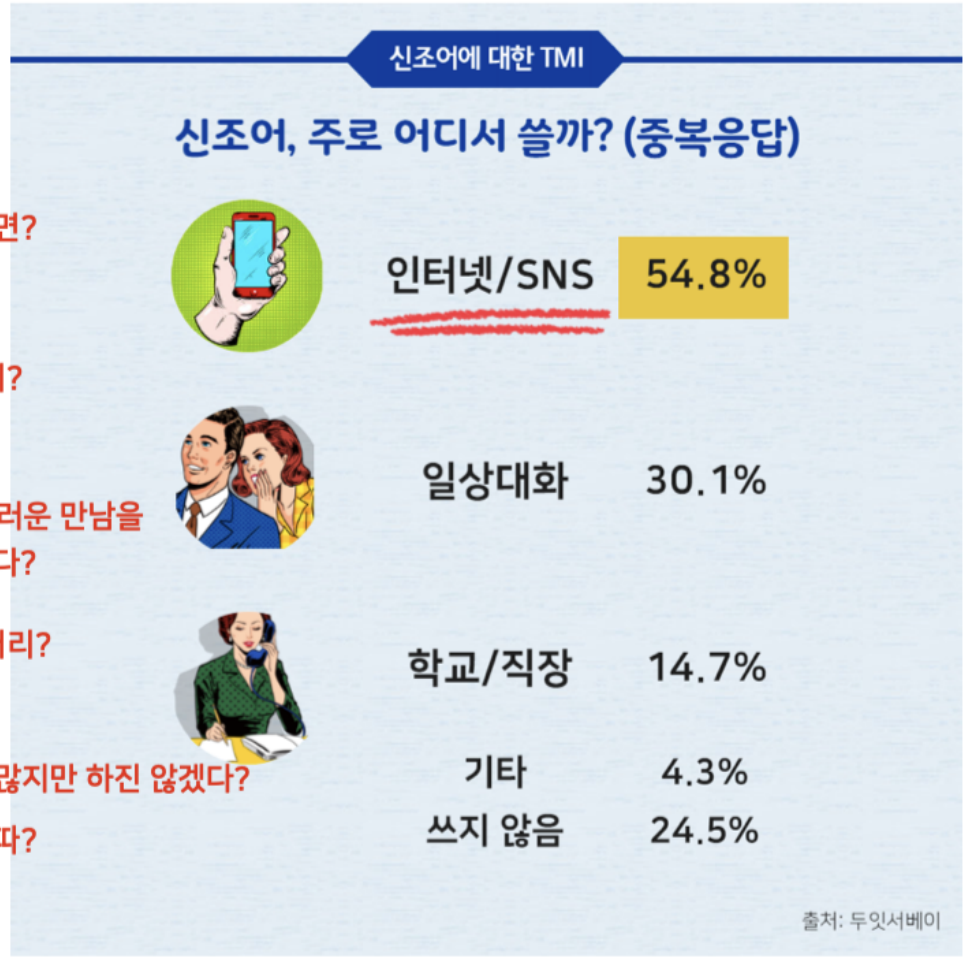
- 텍스트는 비정형 데이터로서, 일반적인 필드와 레코드 구조를 가지고 있지 않음 → 전처리 과정이 복잡하고 어려움
- 텍스트의 형태와 특징에 따라 전처리 과정과 분석방법에 대한 접근을 다르게 고려해야함
- 자연어처리에 대한 이해가 필요하고, 분석시간이 길어 잠재적 가치에도 불구하고 충분히 활용하지 못하고 있음
- 방대한 양, 데이터의 규모 증가, 그리고 그 형태의 비정형성으로 인하여 그 분석과 활용이 어려움

텍스트 마이닝이 어려운 이유

구분	내용
오탈자	"헝거게임 잼잇써요완전 대신 이전편꼭바여" "솔까 타노스 보석 하나도 못구했을때 다들 머했음? 3개 얻었을때도 그렇게 안쌔 뵈더만..."
동의어, 동음이의어	한혜진 : 1. 모델 한혜진 (달심), 2. 배우 한혜진 (기성용 부인), 3. 가수 한혜진 (트로트 가수) Close : 1. Opposite of open, 2. A preposition meaning not far IS : 1. Information System, 2. Islamic State, 3. International Standard
전처리	분석 데이터의 언어를 파악하고 언어의 특징 (교착어, 굴절어 등)에 맞는 전처리 작업 진행 댓글 단위로 분석할지, 문장 단위로 분석할지에 따라서 데이터 분리작업 진행
정보추출	해시 태그 (hash tag) 추출 : '#' + (문자) 핸드폰 번호 추출 '010' - (4자리 숫자) - (4자리 숫자)

텍스트 마이닝이 어려운 이유





텍스트 마이닝이 어려운 이유

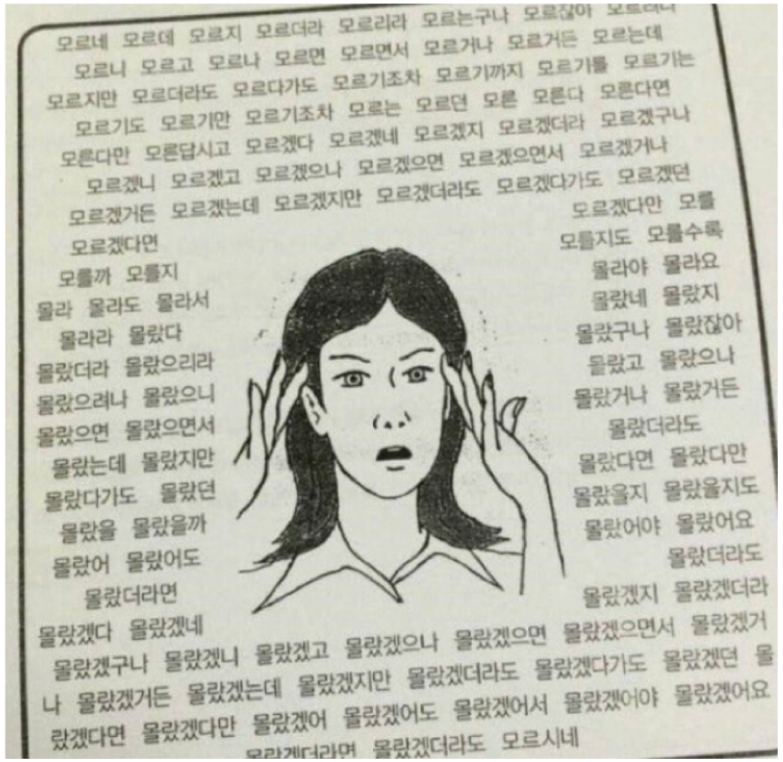
복잡한 한국어 텍스트 분석

- 한국어 텍스트 분석은 언어학적 특성으로 인해 전처리와 분석과정이 까다로움
    i. 초성, 중성, 종성의 조합이 하나의 음절을 형성함
    ii. 첨가어로 첨용과 활용의 케이스가 매우 많음 → 조사와 접사가 붙어 문법적 관계를 형성함
- 용언이 변하는 경우의 수가 매우 많고 그 과정이 하나의 개념으로 확인하기가 어려움
- 형태소 분석기의 한계 및 미비한 어휘사전 → 신조어, 미등록어, 새로운 용어의 조합을 반영하기 어려움

복잡한 한국어 텍스트 분석

한계점	예시
용언의 변형	모르다 → 모르네, 모르데, 모르지, 모르더라, 모르리라, 모르는구나, 모르니, 모르고, …
형태소 분석	"비비크림 빠빠빠~립스틱을 마마마" → 비/NNG + 비/NNG + 크림/NNG + 빠빠빠/UN + ~/SO + 립스틱/NNG + 을/JX + 마/NNG + 마마/NNG "황민현에게 트렌치코트는 정말 존멋♥" → 황민/NNG + 현/NNG + 에게/JKM + 트렌치/NNG + 코트/NNG + 는/JX + 정말/MAG + 졸/VV + ㄴ/ETD + 멋/NNG + ♥/SW "자다가 퇴근했음 좋게따 외냐면 내일 사랑니 째러 가야 되니까는…" → 자/VV + 다가/ECD + 퇴근/NNG + 하/XSV + 었/EPT + 음/ETN + 좋/VA + 게/ECD + 따/VV + 아/ECS + 외/NNG + 이/VCP + 냐/EFQ + 면/NNG + 게/ECD + 사랑니/NNG + 째/VV + 러/ECD + 가/VV + 아야/ECD + 되/VV + …
신조어 출현	지카 바이러스, 오백따리, 트둥이, 울와(우리 트와이스), 팬코(팬 코스프레 유저)

복잡한 한국어 텍스트 분석



I just got here.
상기 문장은 영어로 "나 막 도착했어" 가 된다. 자연스럽게 위 문장을 바꿀 수 있는 경우는

I have just arrived 하나 정도다.

한국어에서 저 just라는 표현은 대체 수십 가지로 가능하다.

나 막 왔어.
나 방금 왔어.
나 지금 왔어.
나 금방 왔어.
나 온 지 조금/좀 됐어. (조금에 강세)
나 온 지 별로/얼마 안 됐어.
나 이제 왔어.
나 바로 막 왔어.

게다가 위의 모든 표현의 '왔어'를 "도착했어"로 바꿔도 말이 된다.

- 시발ㅋ, 시발 ㅋㅋ : 웃김
- 오 시발 : 놀라움
- 아 시발 : 아쉬움
- 시발... : 슬픔
- 시발! : 분노
- 시발; : 어이없음
- 시발ㅜㅜ : 격한슬픔
- 시발;; : 당황스러움
- 시바ㄹ : 급함
- ㅅㅂ : 더욱 급함
- tlqkf : 정말로 급함

텍스트 마이닝이 어려운 이유

사생활 침해와 보안 (Privacy)

- 트위터, 페이스북, 블로그 등의 텍스트는 개인의 정보와 생각을 그대로 반영
- 자칫 무분별한 개인정보의 수집 및 활용으로 사생활 침해 등의 문제를 야기 할 수 있음
- 데이터 분석 전 데이터에서 반드시 개인정보를 제거 또는 마스킹(masking) 처리하는 전처리 과정이 필요함

정확도 측정과 평가 (Accuracy & Validation)

- 정확도, 신뢰도 등 분석결과를 정량적으로 평가하기 어려움
- 정형 데이터를 활용하는 데이터 마이닝에 비해 분석결과가 충분하지 않거나 정확성이 떨어지는 경향이 있음
- 오피니언 리더들의 영향력이 과도하게 작용해 분석결과가 편향될 가능성이 있음

텍스트 마이닝의 패러다임 변화

- 카운트 기반의 문서 표현
  - Vector Space Model에 기반
  - Bag of Words, TFIDF
  - 주로 통계적 기법을 사용
- 시퀀스 기반의 문서 표현
  - 단어의 시퀀스로 문서를 표현
  - 각 단어를 임베딩하고 문서를 임베딩된 단어의 시퀀스로 표현
  - 주로 딥러닝 기법을 사용
- (기타)문서를 일정 길이의 벡터로 직접 임베딩
  - 보통 단어 시퀀스로부터 출발

텍스트 마이닝의 이해를 위한 기본요구지식

- 자연어 처리 기본 도구
  - tokenize, normalize, POS-tagging 등
- 통계학 & 선형대수
  - 조건부 확률, 벡터, 선형결합 등
- 머신러닝
  - 회귀분석의 개념
  - 머신러닝의 다양한 기법(나이브 베이즈, 로지스틱 회귀, Decision Tree, SVM, ...)
- 딥러닝
  - 딥러닝의 개념
  - 딥러닝의 다양한 기법(CNN, RNN, ...)

텍스트 마이닝 방법

- NLP(Natural Language Processing) 기본도구
  - Tokenize, stemming, lemmatize
  - Chunking
- 카운트 기반의 문서 표현과 활용
  - BOW, TFIDF – sparse representation
  - Naïve Bayes, Logistic egression, Decision tree, SVM
- 딥러닝 기반의 문서 표현과 활용
  - Embedding(Word2Vec, Doc2Vec) – dense representation
  - RNN(LSTM), Attention, Transformer, BERT, GPT