1장

1. 인트로덕션

✳ Text Mining
- Text mining = Data Mining + Text Data
- Unstructed text로부터 interesting, non-trivial (의미있는) 한 정보를 추출.
- Data Mining
  - Information Retrieval (정보 검색)
  - Natural Language Processing

✳ IR (Information Retrieval)
- large collection with 특정 information needs를 만족하는 문서 찾기.
  - 문서 뿐만 아니라, Unstructed data (text, image, video, ---)
- Ex) Web search engine
- F(q,d) 구하기
  ① q, d를 vectorization
  ② q, d 의 two vector matching.

✳ Classic IR Model
1) Boolean model.
- q, d를 binary vector (0 or 1)로 표현
- set theory.

2) Vector space model
- q, d를 vector로 표현, 두 vector 간의 similarity 정의 : sim(q,d)

3) Probabilistic model
- p(R|q,d) where R∈{0,1}. 확률모델로 d 순위화. Iterative.

4) Language model
- p(q|d). 주어진 d에 대해 q를 만들 확률로 순위화.

✳ Boolean Model

- q, d는 set of words로 표현.
  - 등복된 단어는 신경 X. 있으면 1, 없으면 0.
- Ranking의 의미는 없다. 일종의 document selection. ⁘) 차원 = vocab 수.

✳ Vector space model

- 모든 d는 concept space 상에서 표현됨.
  - 각 단어가 차원이 될 수도 있고, 다른 것도 가능.
- 두 document vector d_1, d_2 간의 유사도 $sim(d_1, d_2)$ 정의 필요.
  - 보통 Euclidean distance 나 inner product로 함.
  - Document 간의 ranking 가능.

✳ Probabilistic model

- 주어진 q에 대해 특정 문서 d가 이와 관련이 있을 확률.
- $f(q,d) = p(R=1 \mid q,d)$
  - $f(q,d_i)$를 모두 계산 후, ranking.

✳ Language model

- 각 document는 language model의 basis가 됨.
- 우리는 원래 주어진 q에 대해 d의 관련도를 측정함.
  - $p(d|q) = \frac{p(q|d) \cdot p(d)}{p(q)}$ 이므로 $p(q|d)$ 와 동일하다.
  - 문서 set에서 query를 생성할 확률. 높으면 관련성도 높다.
  - $p(q)$는 정해짐. $p(d)$는 constant하고, page Rank로 표현되기도 함.

✳Natural Language Model

- RNN 씀. Word Embedding.
- Word 2 Vec
  - 특정 단어에서 발생하는 주변 단어를 이용하여 그 단어를 표현.
  - 비슷한 의미면 비슷한 값을 가질것.

✳ 그 외 기법들

- document classification : supervised
- document clustering : unsupervised
- topic modeling
  - document에서 hidden topic 찾기.
  - document는 특정 topic에서 조합해서 생성되는 것임.

✳NLP

- AI + Linguistics = NLP
- Communication 개념.
  - NLP ⊂ NL Understanding : 귀
    NL Generating : 입. (more difficult)

✳ 전통 NLP

- Tokenization (쪼개기) → Tagging → Parsing (의미 구문)
  → Semantic analysis (구문 이해) → Pragmatic analysis (speech)

✳기타 기술

- Named Entity Recognition : 고유명사에 대해 category별 인식.
- Information Extraction : 객체 단위 검색
- Word Sense Disambiguation : 단어의 다양한 의미 이해.
- Coreference Resolution : 고유명사에 대한 관계 mapping.

- Machine Translation : seq2seq model
- Document Summarization : Extractive (문장 나열) or Abstractive (새로운 문장)
- QA
- Multimodal : Image captioning, Visual QA