Word Feature Vectors: Traditional Methods

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Data-driven Methods

- Thesaurus 기반 방식은 사전에 대해 의존도가 높으므로, 활용도가 떨어질 수 있음

- 데이터에 기반한 방식은 (데이터가 충분하다면) task에 특화된 활용이 가능

TF-IDF

- 텍스트 마이닝(Text Mining)에서 중요하게 사용

- 어떤 단어 w가 문서 d 내에서 얼마나 중요한지 나타내는 수치

- TF(Term Frequency)
  - 단어의 문서 내에 출현한 횟수
  - 숫자가 클수록 문서 내에서 중요한 단어
  - 하지만, **'the'와 같은 단어도 TF값이 매우 클 것**

$$\text{TF-IDF}(w, d) = \frac{\text{TF}(w, d)}{\text{DF}(w)}$$

- IDF(Inverse Document Frequency)
  - 그 단어가 출현한 문서의 숫자의 역수(inverse)
  - 값이 클수록 'the'와 같이 일반적으로 많이 쓰이는 단어

TF-IDF를 feature로 사용할 수 없을까?

- TF-IDF는 문서에서 해당 단어가 얼마나 중요한지 수치화
- 중요한 문서가 비슷한 단어들은 비슷한 의미를 지닐까?
- 각 문서에서의 중요도를 feature로 삼아서 vector를 만든다면?

TF-IDF Matrix

- 단어의 각 문서(문장, 주제)별 TF-IDF 수치를 vector화
  - Row: 단어
  - Column: 문서

예제: 각 단어별 주제에 대한 TF-IDF 수치

단어	정치	경제	사회	생활	세계	연예	스포츠
문재인	높음	높음	높음	낮음	중간	낮음	낮음
BTS	낮음	낮음	낮음	낮음	높음	높음	낮음
류현진	낮음	낮음	낮음	낮음	높음	중간	높음
날씨	낮음	높음	중간	높음	낮음	낮음	높음
주식	높음	높음	중간	낮음	높음	낮음	낮음
버핏	낮음	높음	낮음	낮음	높음	낮음	낮음

Based on Context Window (Co-occurrence)

- 함께 나타나는 단어들을 활용

- 가정:
  - 의미가 비슷한 단어라면 **쓰임새가 비슷**할 것
  - 쓰임새가 비슷하기 때문에, 비슷한 문장 안에서 **비슷한 역할**로 사용될 것
  - **따라서 함께 나타나는 단어들이 유사**할 것

- Context Window를 사용하여 windowing을 실행
  - window의 크기라는 hyper-parameter 추가
  - **적절한 window 크기**를 정하는 것이 중요

Example

각 단어별 context window 내에 함께 나타난 빈도

	문재인	박근혜	이명박	BTS	싸이	방탄	주식	KOSPI	양적완화
문재인		높음	높음	낮음	낮음	낮음	높음	높음	낮음
박근혜	높음		높음	낮음	중간	낮음	높음	높음	낮음
이명박	높음	높음		낮음	낮음	낮음	높음	높음	중간
BTS	낮음	낮음	낮음		높음	높음	중간	낮음	낮음
싸이	낮음	낮음	낮음	높음		높음	중간	낮음	낮음
방탄	낮음	낮음	낮음	높음	높음		낮음	낮음	낮음
주식	높음	높음	높음	중간	중간	낮음		높음	높음
KOSPI	높음	높음	높음	낮음	낮음	낮음	높음		높음
양적완화	낮음	낮음	중간	낮음	낮음	낮음	높음	높음	

주요 단어만 feature로 활용하는 것도 한 방법

Summary

- Thesaurus 기반 방식에 비해 코퍼스(or 도메인) 특화된 표현 가능

- 여전히 sparse한 vector로 표현됨
  - PCA를 통해 차원 축소를 하는 것도 한 방법