텍스트 마이닝 기초
(품사 판별기 라이브러리 사용 및 키워드 추출)

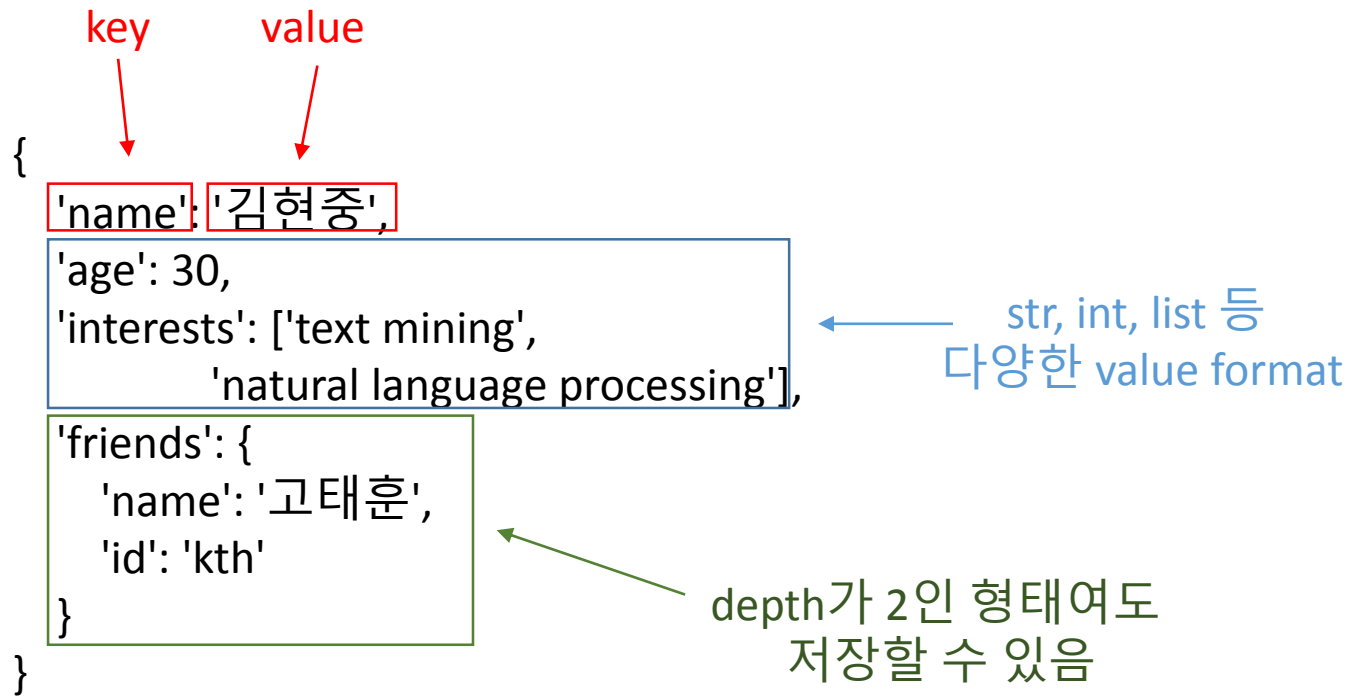
김현중 (soy.lovit@gmail.com)

Data format

- File format
- Data representation

JSON format vs txt format

- **JSON** (JavaScript Object Notation)**은 dictionary 형식으로 되어 있는 파일 포멧으로, depth가 다르거나 어떤 attributes가 필요한지 정하기 어려운 상황에서 유용한 포멧**



JSON format vs txt format

- **Python은 JSON 형식으로 파일을 손쉽게 입출력 할 수 있음**
  - [실습자료] Day1_JSON.ipynb

- **그러나 일단 string parsing을 해야 하기 때문에 여러 개의 JSON 파일로부터 필요한 attributes를 가져오는 작업은 느릴 수 있음.**
  이 때에는 txt 포멧을 추천

JSON format vs txt format

- **Txt 포멧은 컬럼 내용이 정해지거나 mysql과 같은 RDB의 형태의 데이터에 적합**

  - 항목 값에 '\t, '\n'이 들어가지 않도록 유의할 것.
  - 물론 package 쓰고, str 좌우로 " "으로 감싸면 되지만, 혼동될 수 있음

Movie id	Movie name	Movie eng. name	Grade	Running time	Number of rating
10001	시네마 천국	Cinema Paradiso, 1988	전체 관람가 PG	124	3508
10002	빽 투 더 퓨쳐	Back To The Future, 1985	12세 관람가 PG	120	2748
10003	빽 투 더 퓨쳐 2	Back To The Future Part 2, 1989	12세 관람가 PG	107	926
10004	빽 투 더 퓨쳐 3	Back To The Future Part III, 1990	전체 관람가 PG	117	640

JSON format vs txt format

- **이번 강의에서 이용할 scraping된 네이버 뉴스 코퍼스는 일단 JSON으로 파일을 저장한 뒤, 필요한 텍스트 부분만 txt 파일로 만들어 사용**

  - JSON에 새로운 정보를 추가할 가능성이 있기 때문
  - JSON 데이터는 chrome extension을 이용하여 손쉽게 볼 수 있음

JSON format vs txt format

- 명심하세요! 데이터를 수집할 때에는 필요 없을 것 같다고 함부로 버리지 마세요. Step by step으로 데이터를 저장하면 분석 시간을 줄여줍니다.

Vector space representation

- **One hot representation (Bag of Words model)**

  - BOW model라고 부르기도 하며, 한 개의 row는 문서를, 각 컬럼은 하나의 단어를 표현하는 방법으로, 벡터의 값은 그 문서에 등장한 단어의 빈도수

	기계	학습	은	텍스트	마이닝	는
Doc 1	3	2	5	0	0	0
Doc 2	0	0	0	3	5	5
…	…	…	…	…	…	…

Doc 1 = [(0, 3), (1, 2), (2, 5)]
Doc 2 = [(3, 3), (4, 5), (5, 5)]

	0	1	2	3	4	5
Doc 1	3	2	5	0	0	0
Doc 2	0	0	0	3	5	5
…	…	…	…	…	…	…

Vector space representation

- **One hot representation (Bag of Words model)**

	0	1	2	3	4	5
Doc 1	3	2	5	0	0	0
Doc 2	0	0	0	3	5	5
…	…	…	…	…	…	…

  - 단어의 종류는 많지만 하나의 문서에 등장하는 단어의 종류는 적기 때문에, 문서 벡터의 대부분의 값은 0을 지님. 이는 sparse vector라 부름. 백터 공간의 차원은 단어 개수 |V|

  - 오래전부터 이용되는 방법이지만, 한 문서에 어떤 단어가 등장하는지 해석할 수 있으며, 사용하기에 따라서는 여전히 유용한 방법.

Vector space representation

- **Distributed representation**

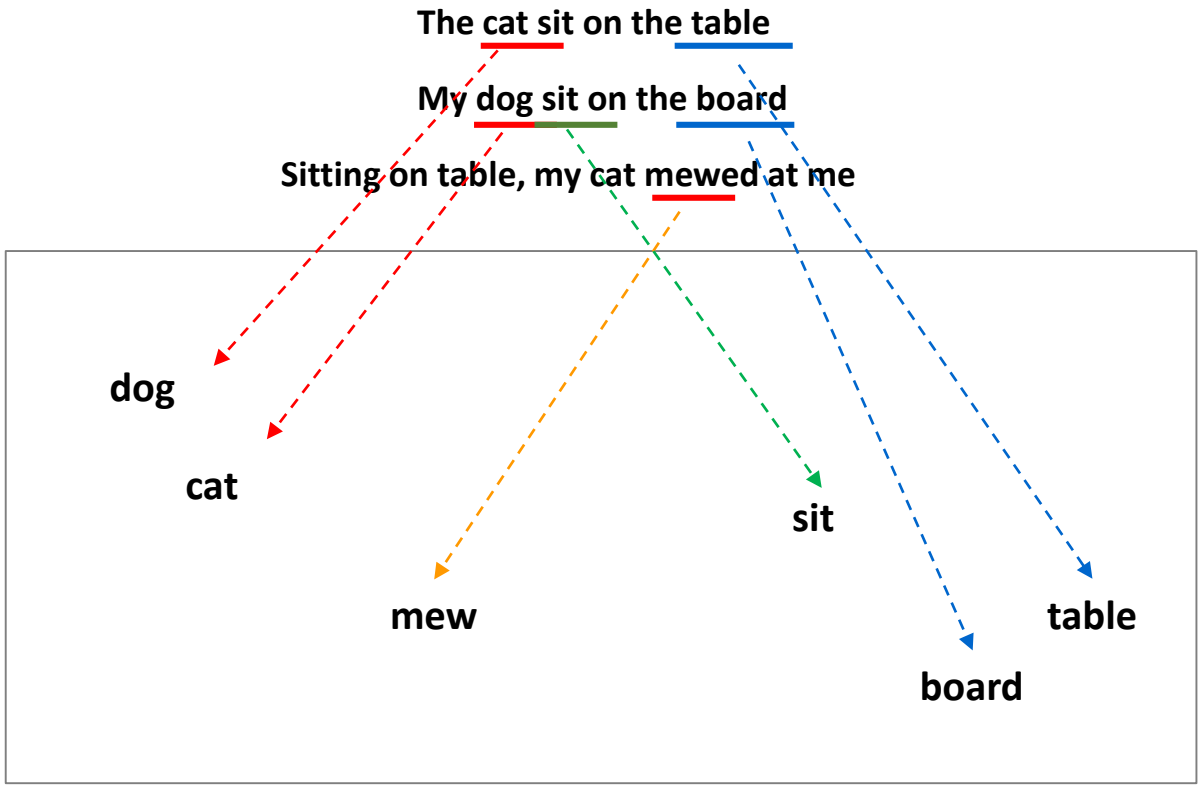
  - Word2Vec에 의해서 우리에게 친근해진 방법으로, 단어나 문서를 정해진 크기(d)의 공간 안에서 표현하는 방법.
  - 각 컬럼이 어떤 의미를 지니지는 않지만, 비슷한 단어/문서는 비슷한 벡터값을 지니도록 하는 방법

‘dog’=	[**0.31**, **0.21**, 0.01, 0.01, …]
‘cat’=	[**0.45**, **0.17**, 0.01, 0.01, …]
‘topic modeling’=	[0.01, 0.01, **0.22**, **0.54**, …]
‘dim. reduction‘=	[0.01, 0.01, **0.19**, **0.45**, …]

Vector space representation

- **Distributed representation**

  ◦ 의미 공간 속에 단어나 문서를 표현하는 방법



< 단어의 의미 공간 >

Why vector representation?

텍스트 프로세싱은 머신 러닝 알고리즘을 이용할 수 있도록 문서를 벡터로 표현하는 과정이다.

- One hot (sparse) 혹은 distributed (dense) representation 모두 벡터로 단어/문서를 기술하는 표현 방법

- 기계학습 알고리즘들이 벡터 공간에서 만들어졌기 때문

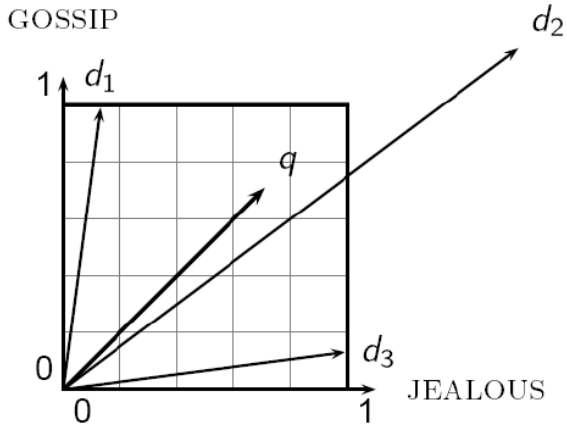
Document clustering

- **군집화 (Clustering)는 비슷한 데이터를 하나의 집합으로 묶어내는 작업**

  - Vector로 문서가 표현되었기 때문에 문서의 유사성은 Euclidean/Cosine 으로 계산할 수 있으나, 문서의 길이의 영향을 적게 받기 위해서 Cosine을 유사도로 이용
    - Logistic regression, Neural network와 같은 알고리즘도 원리는 cosine과 동일

Doc.	Word 1	Word 2	Word 3
Document 1	1	1	1
Document 2	3	3	3
Document 3	0	2	0

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2} \sqrt{\sum_k x_k^2}}$$

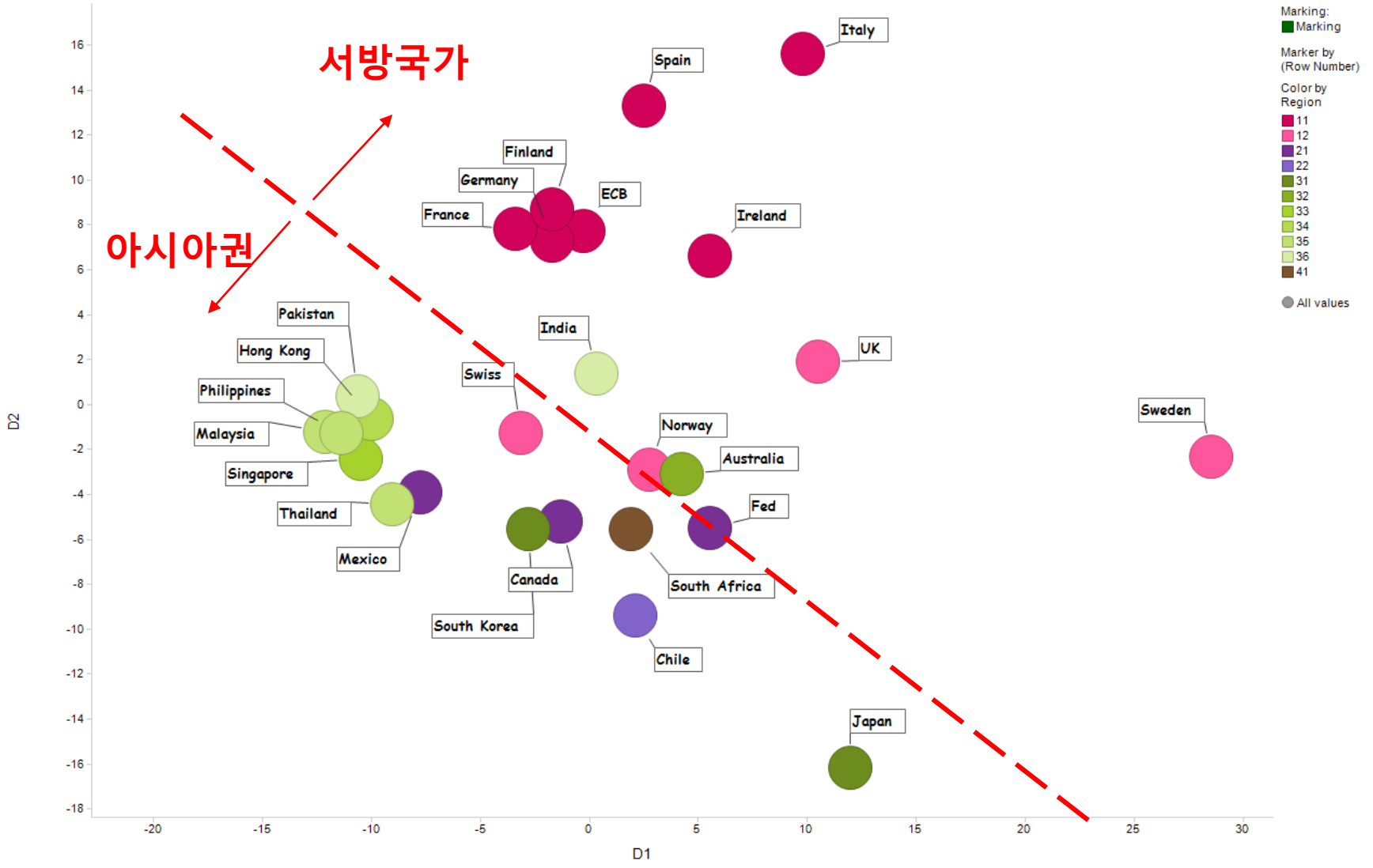


Document clustering

- 중앙은행 총재들의 연설문을 바탕으로 연설문 내용이 유사한 국가를 묶을 수 있음

Document classification

- 판별 (Classification)은 데이터를 구분하는 경계선을 긋는 것



Document classification

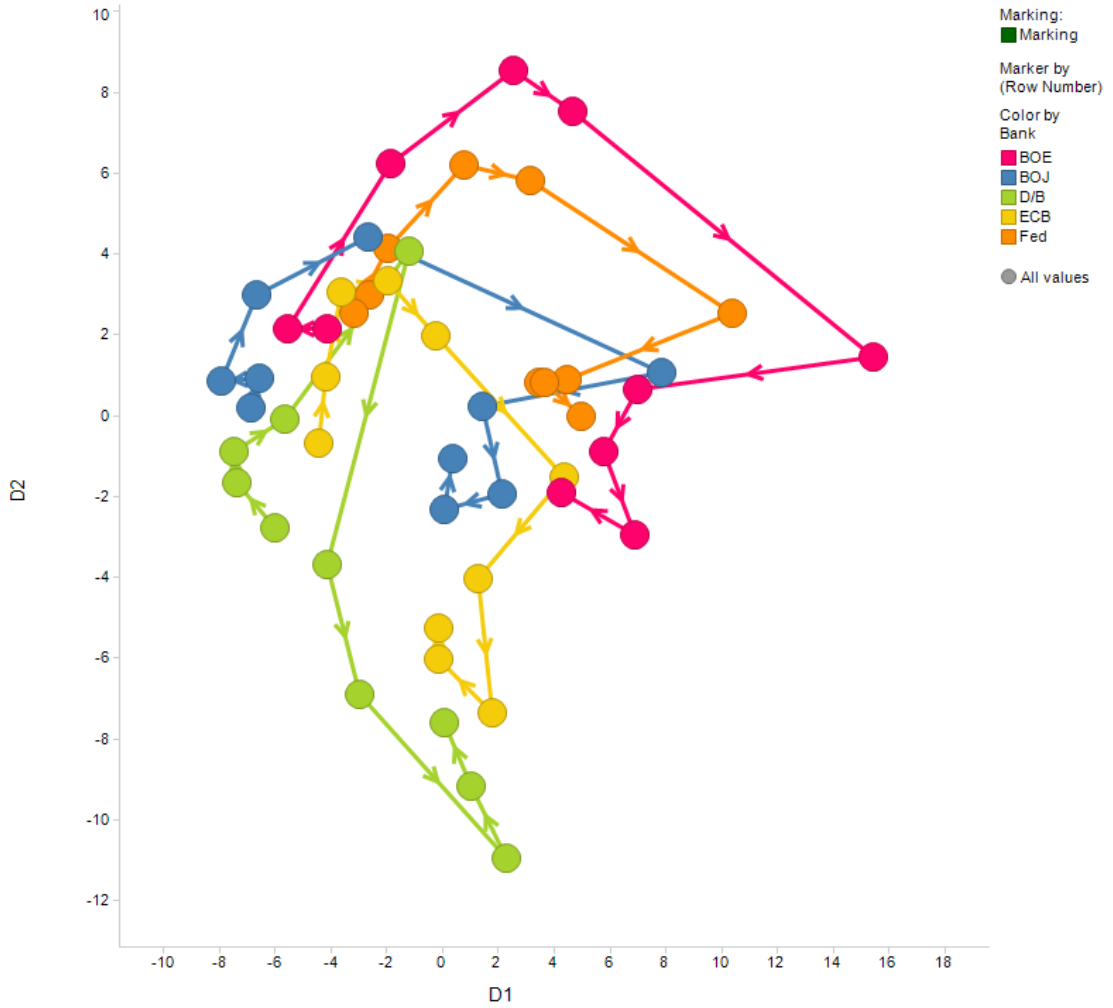
- **분류 (Classification)는 데이터를 구분하는 경계선을 긋는 것**

  - 경계선이 직선일 경우 linear model, 곡선일 경우 non-linear 모델이라 부름
    - Linear model: Logistic regression, Support Vector Machine, Decision Tree
    - Nonlinear model: Neural network, Support Vector Machine with Kernel, Deep learning 계열

  - 예시
    - 주어진 보험 청구서가 허위인가 / 사실인가?
    - 해당 제품의 리뷰는 긍정적인가 / 부정적인가?

Word/Document Visualization

- 중앙은행들의 연설문을 연도별로 표현한 뒤, 2차원으로 시각화하면 각 나라의 금융정책이 어떤 방향으로 변화하는지를 시각적으로 표현 가능



Keyword extraction

- 키워드 추출은 해당 문서/집합을 요약

Year	Common Concern	Federal Reserve System	European Central Bank	Bank of England	Deutsche Bundesbank	Bank of Japan
2004	Sustainability Credibility	Corporate Gover-nance Scandals	Parliaments Growth and Job	Low Inflation	Government-Deficit	QE Deflation
2005	China Inflation	Oil/Natural Gas Basel II	Domestic Inflat-ionary Pressure	Repo Rate Crystallising	Debt Levels	Private-Consumption
2006	Competitive Global Imbalance	Risk Management Creditworthiness	Monetary-Expansion	Exchange Rates China / India	Future Inflation Two Pillar Strategy	Domestic and-External Demand
2007	Subprime-Mortgage	Foreclosures	Risk to Price Stability	Growth of Money and Credit	Local Currency Bond Market	Price Stability
2008	Financial Turmoil Commodity Prices	Primary Dealers Foreclosures	Price Stability Supply of Liquidity	Failing Banks Spare Capacity	Resilience of-Financial System	Securitized-Product
2009	Financial Crisis Lehman Brothers	TALF SCAP	Non-standard Macroprudential	Asset Purchase	Expansionary-Monetary Policy	Outright Purchase Credit Bubble
2010	Recovery Reform	Unemployed SCAP	Macroprudential Excessive Deficit	VAT Depreciation of £	Microprudential Reform of Basel II	Overcoming-Deflation
2011	Sovereign Debt Basel III	Job Growth Dodd-Franc Act	Economic Governance	Real Incomes PRA	No Bail Shadow Banking	After Earthquake Monetary Easing
2012	Europe Deleveraging	Maturity Extension Forward Guidance	OMT / SSM Fragmentation	FLS	Liability Rescue Package	European Debt
2013	Real Economy Price Stability	(At least as long as) Unemployment	Fragmentation SSM / SRM	FLS PRA	Liability SSM	QQE