빅데이터 자료분석.

Ⅰ. 텍스트 마이닝. ; 자연어로 구성된 비정형데이터에서 패턴 또는 관계를 추출하여 의미있는 정보를 얻는 기법.

⟷ NLP (인간의 언어 현상을 컴퓨터와 같은 기계를 이용하여 모사 할수 있도록 연구)

즉, 텍스트 마이닝은 데이터 마이닝과 다르게 비정형 데이터를 다루고, NLP를 적용하여 텍스트분석을 진행.

→ 웹 크롤링, 토픽 분석, 감성분석 + 교차·상관관계 분석, 인과분석, 토픽 모델링 등

* 워드 클라우드 ; 빈도수가 높은 키워드를 시각적으로 표시.

Ⅱ. 감성분석. sentiment analysis ; 텍스트를 작성한 사람의 태도, 의견, 성향과 같은 주관적 데이터 분석.

감정(Emotion) vs 감성(Sentiment) ; 감정은 행복·공포 등 복잡한 심리상태.

감성은 감정으로 부터 생성된 정신적 태도.

→ 감정을 정리된 형태로 사회적으로 표현.

∘ 감성점수 계산 ; 감성점수 = 긍정 단어 수 − 부정 단어 수

but. 역설(Irony)과 풍자(sarcasm)와 같은 문장은 여전히 모호.

Ⅲ. 소셜 네트워크 분석.

∘ 소셜 네트워크란? actor 간에 사회적으로 연결된 망으로 그래프 이론에 토대를 둠.

→ 사람 또는 객체들이 사회적 관계를 기반으로 상호작용을 통해 거미줄처럼 엮여 있다는 개념은 개인적 수준을 넘어 조직수준에서 발생하는 다양한 사회적 현상을 설명하는데 있어 기존의 관점과 다른 새로운 시각을 제공한다.

∘ 그래프 ; G 란 개체(Node)들의 집합 V와 개체를 연결하는 Edge의 집합 E의 쌍.

- 종류 : 방향 · 무방향 · 완전(모든 노드 연결) · 가중 그래프
- 특징 : degree(차수) − indegree, outdegree

  hub (허브) − 상대적으로 차수가 높은 노드

- 차수의 분포 : $P(k) = \frac{N(k)}{N}$, $N(k)$: 차수 k를 갖는 노드의 개수.

  { 랜덤 그래프 : 포아송분포 형태 / 척도없는 그래프 : 멱함수 법칙을 가짐.

- 그래프의 밀도 : $\frac{\text{현 그래프의 Edge의 수}}{\text{완전그래프의 Edge의 수}}$
- 노드의 중심성 ; 개체가 가지는 영향력을 분석하는데 사용.

  차수 중심성 $D_c(i)$ = i와 직접 연결된 노드 수 / i와 직·간접적 연결된 노드의 수

  근접 중심성 $C_c(i) = (N-1)/\sum_{i \neq j} d(i,j)$, $d(i,j)$: 최단경로 길이

  매개 중심성 $B_c(i) = \sum G_{jk}(i)/G_{jk}$, G_{jk} : 최단경로들의 수 , $G_{jk}(i)$ = 노드 i를 포함하는 최단경로.

⇒ 친구 관계도, 전염도, 투자관계도, 조직도 등 다양한 활용사례.

Ⅳ. 텍스트 클러스터링

- 텍스트 클러스터링이란? 텍스트의 개체들 간의 유사도를 계산한 후 이를 바탕으로 유사한 개체들을 묶어주는 작업.
- 텍스트 데이터의 수치표현.

  i) TF(Term Frequency) ; Term-document 행렬. 행은 단어, 열은 문서로 발생빈도 행렬.

  → 단점으로 별로 도움되지 않는 단어들의 빈도수가 높아 영향을 미침.

  ii) TF-IDF (TF-inverse document frequency) ;

  $IDF = \log_2\left(\frac{N}{DF}\right)$, DF : 단어가 나타난 문서의 수. → 단어의 중요도를 측정 가중치. (→ $\log_2$ 취하는 이유는 outlier 영향 최소.)

  $TF\text{-}IDF = TF(t,d) \times IDF(t)$ but. 문서의 길이가 길수록 출현빈도가 늘어남.

  ∴ 문서길이 또는 단어의 총 개수 보정 필요 (n_d : 문서 d의 단어 수.)

  → TF의 표준화 $\tilde{f}_{td} \leftarrow \frac{f_{td}-\bar{f}_d}{s_d}$ / TF의 정규화 $\tilde{f}_{td} \leftarrow \frac{1+\log_2 f_{td}}{n_d}$

∘ 유사도 측정.

  i) 유클리드 거리 ; 척도에 민감(표준화·정규화 필요), 희소행렬간 거리측정 문제.

  ii) 코사인 유사도. ; "각도"를 반영한 유사도.

∘ 클러스터링 방법 → #다변량 자료분석.

Ⅴ. 토픽 모델링

∘ 토픽모델링이란? 문서에서 특징적인 토픽을(주제) 찾아내는 기술.

* 토픽 ; 문서에서 자주 같이 등장하는 단어들이 묶여서 하나의 토픽을 이룬다고 가정. 문서와 단어는 관측변수. (일종의 잠재변수.)

즉, 한 문서에서 자주 같이 등장하는 단어들을 통계적으로 추출해 내는 모델링 기법.

∘ LDA(latent Dirichlet Allocation)

step1. 각 토픽 k에서 단어분포 $\phi_k \in R^W$는 Dirichlet 분포를 따름.

step2. D개의 문서의 각각은 K개 잠재 토픽들의 혼합으로 구성되어 있다고 가정. 토픽분포 $\theta_d \in R^K$는 Dirichlet 분포를 따름.

step3. 문서 d에 있는 n번째 단어에 대하여 토픽을 우선 선택.

토픽 $z_{d,n} \sim multinomial(\theta_d)$

토픽에 따라 단어 분포를 결정하고($\phi_{k=z_{d,n}}$) 단어를 선택.

$W_{d,n} \sim Multinomial(\phi_{z_{d,n}})$

i) 토픽 개수 K 결정.

ii) 각 토픽 별로 단어 분포 측정.

iii) 각 문장에 대한 토픽 분포를 측정.

D : 전체 문서 수	K : 전체 토픽 수
θ_d : 문서 d의 총 단어 수	ϕ_d : 문서 d의 토픽분포
$z_{d,n}$: 문서 d에서 n번째 단어 토픽	
$W_{d,n}$: 문서 d에서 n번째 단어	

* Dirichlet 분포란?

$f(x_1, \cdots x_k; \beta_1, \cdots, \beta_k) = \frac{1}{B(\beta)} \prod^K x_i^{\beta_i - 1}$,

$B(\beta) = (\prod^K \Gamma(\beta_i)) / \Gamma(\sum^K \beta_i)$는 다변량 베타함수.

* 다항분포.

$p(x_1, \cdots, x_k; \alpha_1, \cdots, \alpha_k) = \frac{n!}{x_1! \cdots x_k!} \prod^K p_i^{x_i}$

∘ LSA ; DTM을 차원축소하여 근접단어들을 토픽으로 묶음.