**보건사회연구원 형태소 데이터 분석 (Graph VAE)**

**안승환**

* **인용 논문**

1. KIPF, Thomas N.; WELLING, Max. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.
2. SALHA, Guillaume; HENNEQUIN, Romain; VAZIRGIANNIS, Michalis. Keep it simple: Graph autoencoders without graph convolutional networks. *arXiv preprint arXiv:1910.00942*, 2019.
   1. 모형의 간소화와 해석의 용이성을 위함

* **모형 가정**

1. 1개의 기사 = 1개의 undirected, unweighted graph
2. 1개의 기사에 등장한 keywords = 서로 fully connected되어 있음

* **데이터 전처리**

1. 전체 월별 데이터로부터 교집합 keywords 1856개에서 임의로 300개 선택 = 300 keywords (계산상의 용이함을 위함)
2. Train: 월별로 10000개의 Graph(기사) sampling = 전체 10만개의 Graph  
   (이는 모든 월별 데이터에 대한 모형의 적합에서 latent space를 공유하기 위함)
3. Test: 월별로 100개의 Graph(기사) sampling
4. 기사를 sampling할 때 300개의 keywords 중 적어도 3개를 포함하는 기사를 선택 (minimum degree)

* **활용 방향(모형 해석)**

1. Test data를 latent space에 embedding한 결과를 시각화하여 월별로 embedding이 구분되는지 확인(10개의 색깔로 구별)
   1. 실제로 사용된 keywords들도 같이 비교
   2. 월별로 따로 시각화도 추가
2. Embedding vector(mean parameter)들을 시각화 비교
3. 중심 (0, 0)으로부터 거리가 먼 keyword = dominate word in latent space (즉, 해당 keyword는 다른 keyword들과 graph에서 높은 빈도로 connect되는 경향이 있음)

* **개선 사항 및 향후 연구 방향**

1. 인접행렬이 1인 확률에 가중치 beta 부여 = weighted binary cross entropy
2. 차원이 큰 sparse한 행렬(인접행렬)들의 행렬곱으로 인한 속도 저하
3. 기사별 감성 정보 사용 (node들의 prior에 가정으로 사용)
4. 가중치가 있는 edge를 사용하는 경우의 모형 적합
5. 모형 저장은 가중치 저장으로 대체하여 해결
6. Adjacency matrix 원소들 사이에 dependency를 부여하기