Time-Aware Topic Modeling

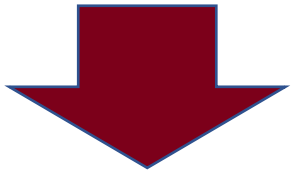
LDA와 다중 변화점 모형을 기반으로

1. Time aware Topic Model

프로젝트 목적

기존 연구 및 문제인식:

- LDA와 같은 기존 문서 군집화 모형들은 시간에 따른 초모수의 변화를 감지할 수 없음
- 특히나 뉴스나 기사 같은 어떤 사건에 영향을 받는 매체의 경우 시간에 영향을 많이 받음.



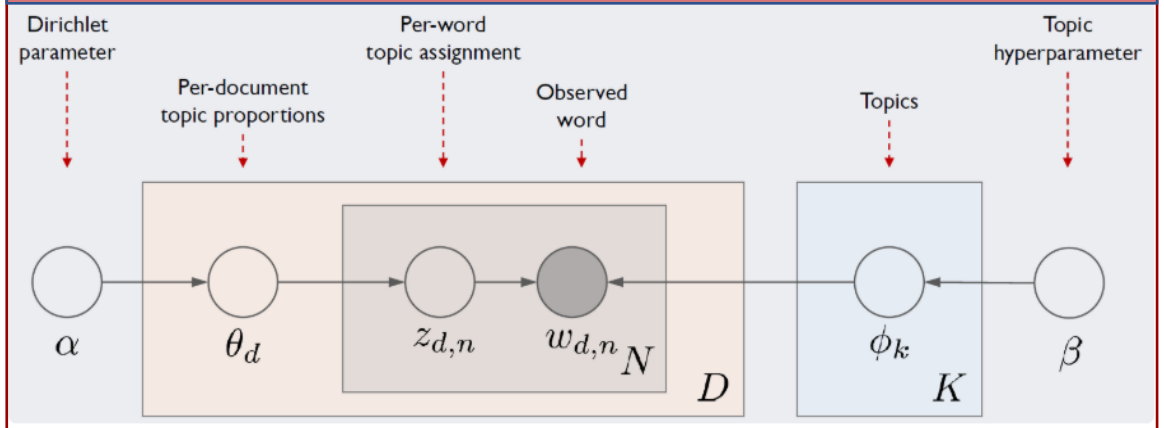
해결 방안:

- 각 문서에 군집을 나타내는 잠재변수 도입
- 잠재 마르코프 모형을 이용하여 각 문서를 군집화

1. Time aware Topic Model

분석 전략

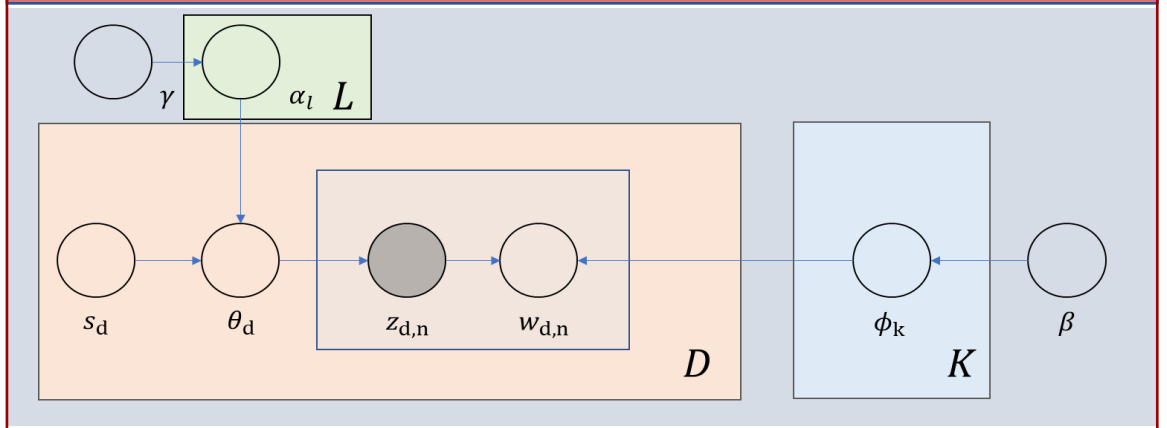
LDA



기존 모형은 문서 각각의 모수를 추정

VS

Proposed Model



문서 각각의 모수를 추정할 뿐만 아니라,
각 문서의 모수가 비슷한 문서를 군집화함

1. Time aware Topic Model

사용 데이터

출　　처 : BBC headline news

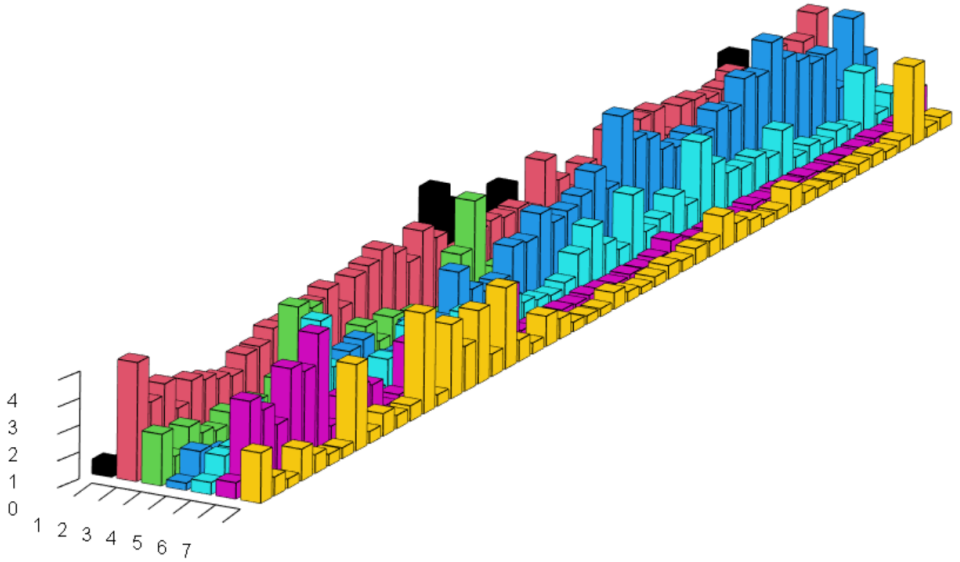
기　　간 : 2020.01.01 ~ 2020.05.31

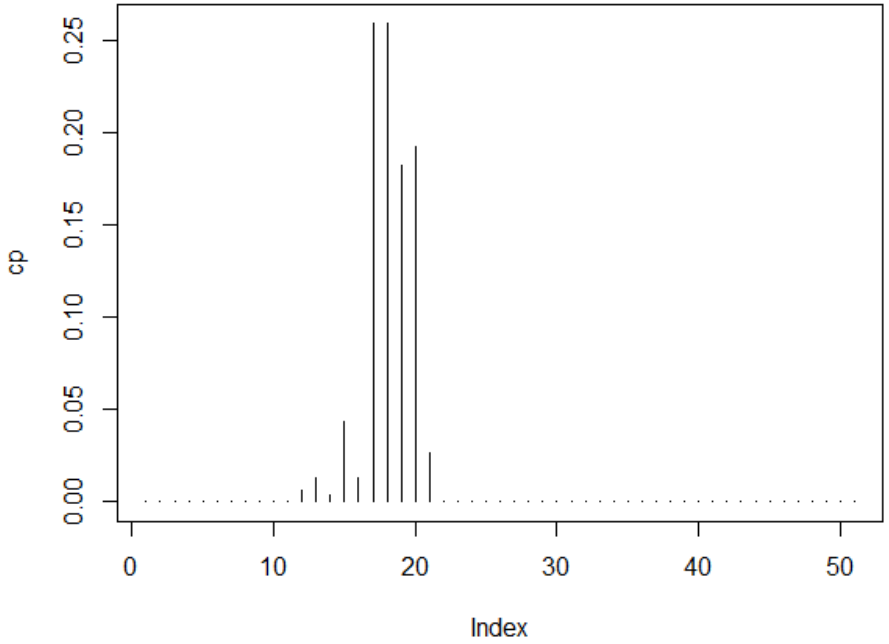
세부사항 : 연산 속도 및 모형을 고려하여 다음과 같은 제한조건을 둠

- 3일 간격으로 기사 수집
- 하루에 5개의 기사를 수집한 후, 5개의 기사를 하나의 문서로 취급
- 각 기사의 비중을 동일하게 하기 위해, 기사에서 100개의 단어를 추출

1. Time aware Topic Model

분석 결과





오른쪽 그림은 문서의 주제 구성이 변화할 확률을 나타내는 그림으로 2월 중순에서 높은 확률을 가짐

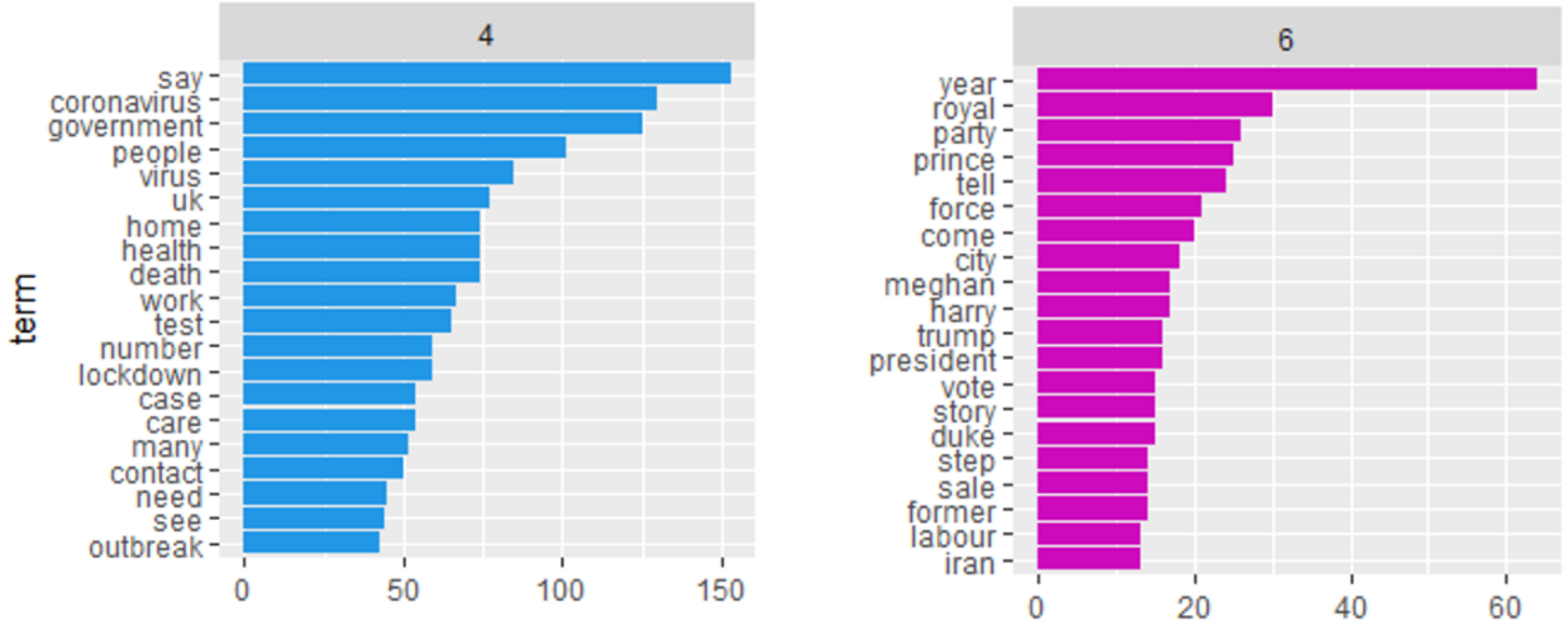
왼쪽 그림은 실제 문서 주제의 구성비율로 짧은 축은 주제, 긴 축은 시간을 나타냄.

오른쪽 그림과 마찬가지로 2월 중순에서 주제 구성비율이 바뀐 것을 확인할 수 있음

2월 중순이후 파란색 주제가 급등한 것을 확인 가능함.

1. Time aware Topic Model

분석 결과



2월부터 미국에서 코로나에 대한 관심이 급증했다는 사실과

1월에 영국의 해리왕자가 선임멤버로 물러나겠다는 사실을 고려했을 때,

군집화가 잘된 것을 알 수 있음