1. (1) 토큰화: **형태소** 단위로 잘라서 정리

   (2) 정규화: 대/소문자를 소문자로만 바꾸는 것처럼 **일정한 형태**를 만들어 줌 (tolower)

   (3) 불용어제거: Actually나 the처럼 **중요하지 않은데 빈번한 단어**들이 존재할 수 있으니 이것들을 제거해줌으로써 분석의 정확도 올려줌

   (4) 스테밍: 단어의 핵심인 어근을 추출함

2. TDM은 단어를 중심으로 파악하며, 이 단어가 각각 문서에서 몇 번 나왔는지를 확인하기에 용이해 단어의 중요도를 파악할 수 있다.

   DTM은 문서중심으로, 각 문서에 어떤 단어들이 들어있는지를 파악하기가 용이하다.

3. 기존의 전처리는 문장에서의 순서를 고려하지 않아, 단어의 중요도를 따지지 않았다. 또한 같은 단어라도 스펠링이 틀리면 다른 단어로 파악하고, 유의어를 파악하지 못했다.

4. 코사인 유사도의 분모는, 각각을 제곱해서 모두 더한 후 루트를 씌운 것(벡터)끼리의 곱이며,

   분자는 분석하는 두 문서 간의 각각의 곱을 모두 더한 것이다.

   유사도가 1에 가까울수록 두 문서는 유사하다는 뜻이 된다.

5. TF는 단어가 문서에서 몇 번 나오는지를 파악하는 것으로, 단어의 중요도&빈도를 의미한다. DF는 이 단어가 전체 문서 중 몇 개의 문서에서 나타나는지를 말하는데, DF가 높으면 오히려 흔한 단어가 되어 중요도는 떨어지게 된다. 그렇기에 두 개는 상충되는 개념이기에 DF가 아닌 IDF를 사용하는 것이다.

6. IDF에 로그를 씌우는 이유는, 전체 문서의 개수인 분자가 너무 커지거나, 분모인 DF의 값이 너무 작아지면 숫자가 커져 각각의 차이가 많이 벌어지게 되는 스케일이펙트가 발생할 우려가 있기에 이를 방지하기 위함이다. 또한 분모에 1을 더하는 이유는, DF가 0이 나와 분모가 0이되면 제대로 된

분석을 할 수 없기에 이를 방지하기 위함이다.

7. 전치행렬은 기존 행렬에서 행과 열의 위치를 바꾼 행렬이다. 단위행렬은 대각선의 값이 1이고 나머지가 모두 0인 행렬이다. 역행렬은 기존행렬을 단위행렬로 만들어주는 행렬을 말한다. 직교행렬은 전치행렬과 역행렬이 같을 때를 의미한다. 대각행렬은 대각선의 값이 1이 아닌 다른 숫자도 존재할 수 있고, 나머지의 값이 0인 행렬이다.

8. 절단된SVD를 사용하는 이유는, 불필요한 정보들을 한 번 더 제거해줌으로써 노이즈를 제거하고, 계산비용을 줄이기 위함이다.

9. 토픽의 수를 크게 잡으면, 분석의 다양성은 올라가지만 계산해야 하는 값이 늘어나 시간과 비용이 더 많이 들어가며, 노이즈가 더 생겨날 우려가 있다.

10. 첫 번째는 직교행렬로, 설정한 K값(토픽 수)만큼을 열로 가지는 행렬이며 단어별로 각 토픽에 어느정도 영향을 주는지 알아보기 용이하다. 두 번째는 대각행렬로, 대각선에만 값이 존재하며 이는 내림차순 정렬이다. 세 번째는 전치행렬로 행과 열의 위치가 바뀌며 각 토픽들이 문서에 어느정도 영향을 주는지를 알아볼 수 있다.

11. LSA는 잠재의미분석으로 단어들의 의미적인 내용을 분석해서 유사한 단어들끼리 집단을 만든 후, 내부적으로 유의어를 찾기 위해 연산처리하는 것을 말한다. **행렬분해**를 해야 하는데 매트릭스가 커질수록 처리시간이 오래 걸린다. 행렬 중에 빈값이 많으면 처리가 어렵고, 결과가 마이너스 값으로 나오면 분석하기가 애매하다는 단점도 존재한다. 그에 비해 LDA는 행렬분해가 필요없으며, **다의어**도 처리할 수 있다는 장점이 있다. 또한 내가 가진 데이터보다는 전세계 모든 문서의 특징을 이용한다는 것에 가깝다.

12. 어떤 문서에는 보통 한 가지 이상의 크고 작은 주제(토픽)들이 존재한다. 우선은 문서의 토픽 중요도를 확률분포로 파악한 후 그 안에서 룰렛을 돌려 하나의 토픽을 선정한다. 선정된 토픽 안에는 또 여러 가지의 단어들이 존재할텐데, 그 토픽 안에 있는 단어들의 중요도를 파악한 후 룰렛을 돌려 하나의 단어를 선택한다. 이러한 과정을 계속 반복하며 새로운 문서를 만

들어낸다.

13. LDA추론 과정은 문서 생성과정을 역으로 바라보는 것이다. 우선 특정 단어의 확률분포를 찾고, 토픽의 확률분포를 찾아 분석하며, 문서 내에서의 토픽의 분포&중요도&영향력 등을 알아볼 수 있다.

14. Word Embedding기법이란, 텍스트/문서들에서 토픽을 뽑아내거나 문서를 생성하는 등의 텍스트 데이터 처리 방안을 의미한다. 원 핫 벡터가 아닌, 연속된 숫자들의 표현을 통해서 각 단어(요소)들간의 유사도&중심성등을 알아볼 수 있으며, TF-IDF&LSA 뿐만 아니라 최근에는 딥러닝에 많이 사용되고 있는 기법이다.

15. 모수는 계산과정에서 각 값에 더해주게 되는데, 이를 통해 값이 0이 되는 것을 방지해준다. 또한 모수가 커질수록 속성들 간의 **쏠림**이 줄어들어, 토픽 내에 있는 단어들의 구성이 다양해질 수 있다.

16. 우선은 토픽의 개수를 정한 후, 각 단어별로 토픽을 임의로 할당한다. 랜덤이기 때문에 같은 단어라도 다른 토픽이 할당될 수 있다. 그 후에 자기 자신이 잘못되었다고 가정하고 마스킹한 후 동일 문서 내에서 토픽의 비율을 파악한다. 그 후에 전체 문서에서 자신이 어느 토픽에 속해있는지를 파악하고 이 두개를 같이 고려하여 토픽을 결정한다. 이러한 과정을 더 이상 계산이 안될 때까지 반복하며 단어들의 토픽을 정한다.

17. 7개의 다리를 단 한 번씩만 건너면서 원래의 위치로 돌아올 수 있느냐하는 문제이다. 오일러 정리에 의해서 불가능하다는 것이 증명되었고, 7개의 다리는 선(아크), 4개의 구역은 점(노드)으로 표현할 수 있다.

18. 케빈 베이컨 6단계 법칙은, 아무리 멀게 느껴지는 상대방이더라도 6단계 안에 접촉할 수 있다는 법칙이다. 예를 들어 페이스북에 있는 15억명의 회원들을 대상으로 통계를 내 본 결과, 평균적으로 3.57명 내로 만날 수 있었다고 한다.

19. 연결망은 종형 연결망과 역함수 법칙 분포 연결망이 존재한다. 종형 연결망은 평균이 많고 극소/극대가 적은 모양으로, 대부분 평균값에 가까운 값을 가진다. 이는 균등하게 분포 되어있는 고속도로 연결망을 예시로 들 수

있다. 그에 비해 역함수형 연결망은 평균값과 가까운 값들이 적고, 극소값이 아주 많으며, 극히 많은 링크를 가지고 있는 소수의 허브가 존재한다. 이는 허브가 존재하는 비행기 항공 연결망을 예시로 들 수 있다.

20. 페이지랭크 알고리즘은 구글의 기반이 되는 알고리즘으로, 페이지마다 조건에 따라 일정한 점수를 부여해 중요도를 파악하여 어떤 링크를 노출해야 할지 결정한다. 랭크가 높게 부여되는 경우는 보통 두 가지가 있는데, 첫 번째로는 많은 수의 링크가 본인을 가리키는 경우이다. 반면에 두 번째는 연결된 링크의 수가 많지 않더라도 랭크가 높게 설정될 수 있는데, 랭크가 높은 링크와 연결이 되는 경우이다. 랭크를 결정하는 점수에 영향을 주는 주요한 요인으로, 자신을 가리키는 링크의 점수와 그 링크가 가리키는 링크들의 숫자가 있다. 그렇기에 높은 점수의 링크가 자신만을 가리킨다면, 이에 영향을 받아 연결된 링크들이 많지 않더라도 높은 랭크를 받을 수 있는 것이다.

21. 컴퍼넌트는 연속적으로 연결되어 있는 하나의 그래프를 의미한다. 연결점은 연결을 하는 데에 주요한 역할을 하는 점으로, 이것이 없어진다면 컴퍼넌트가 하나가 아닌 여러 개로 분리되게 된다. 브릿지 또한 연결점과 비슷하게 없어진다면 컴퍼넌트가 분리되게 된다.

22. 비방향그래프에서 평균연결정도는 2L/g로 (연결된 라인의 수*2)/전체 점의 수이다.

23. 인디그리는 나를 가리키는 라인의 숫자, 아웃디그리는 내가 가리키는 라인의 숫자를 의미한다. 만약에 인디그리와 아웃디그리가 모두 0이라면 고립되어 있는 점이라 할 수 있다. 인디그리>0, 아웃디그리=0이라면 수신자, 인디그리=0, 아웃디그리>0이면 전달자라고 말한다. 인디그리와 아웃디그리가 모두 0보다 크다면 매개자라고 할 수 있다. 인디그리와 아웃디그리의 평균연결정도는 L/g로 (연결된 라인의 수)/전체 점의 수 이다.

24. 에고 네트워크란 전체 네트워크가 아닌 자신을 중심으로 연결된 네트워를 의미한다. 이를 로컬 인덱스라고 할 수 있으며, 전체 네트워크는 글로벌 인덱스라 부를 수 있다.

25. 사건 매트릭스는 사람, 소속집단, 사건 등의 개체로하며, 각각의 개체에 대해서 알아보기 용이하고 개체간의 특성이 다르다는 특징이 존재한다. 반면에 인접도 매트릭스는 같은 특성의 개체들로 형성되는데, 각각의 점수를 부여해 그래프를 형성하게 되면 가중치가 존재하는 계량그래프가 만들어지게 된다. 만약 c1=>c2가 3이라면, c2=>c1도 3이 된다.

26. 밀도란 그래프가 얼마나 촘촘한지를 알아보는 척도로, 1에 가까울수록 높은 값을 의미한다. 절대적 포괄성은 연결된 점의 수이며, 상대적 포괄성은 (연결된 점의 수)/g이다. 비방향 그래프에서의 밀도는 L/g(g-1)/2이며 방향 그래프에서는 L/g(g-1)이다. 계량그래프에서는 우선 최대 가중치를 정해주어야 한다. 분자는 실제로 그래프의 모든 가중치를 더해준 값이며, 분모는 최대 가중치에 나올 수 있는 최대 라인의 수를 곱한 값이 된다.

27. 중심성이란 네트워크 내에서 본인이 어느 정도로 중앙에 위치하는지를 나타내는 것으로, 보통 인디그리가 높아 인기/매력도가 높거나 아웃디그리가 높은 마당발인 경우 중심성이 높게 측정된다. 또한 점과 점을 이어주는데 주요한 역할을 하는 매개체가 되는 경우에도 중심성이 높다고 말할 수 있다.

28. 우선 내향중심성은 인디그리가 높은 것으로, 본인을 가리키는 점들이 많아 인기도/매력도가 높은 경우를 말한다. 외향중심성은 아웃디그리가 높아 다른 점들을 많이 가리키며, 여러 사람들과 관계를 맺는 마당발이라 할 수 있다.

    포인트 중심성은 특정한 점이 네트워크 내의 중심에 위치하는 정도를 표현한 지표로, 하나의 점에서 중심성을 바라보는 것이다. 그래프 중심성은 전체 네트워크가 가진 종합적인 중심으로의 응집도를 의미하며 집중도로 나타낼 수 있다.

    로컬 중심성은 본인과 직접적으로 연결된 네트워크만을 보는 것인데, 이는 에고 네트워크(=연결정도중심성)라고도 할 수 있다. 전체의 크기를 고려하지 않으며, 본인과 연결된 로컬 영역에 한해서만 중심성을 측정한다. 글로벌 중심성은 특정 한 점이 전체 네트워크 중심에 위치하는 정도로, 한 점

과 네트워크 전체 점들간의 거리에 의해 측정된다.(=근접중심성)

29. 우선 연결정도중심성은 그래프에서 점과 점들간에 직접적으로 연결되어 있는 정도를 기반으로 중심성을 파악하는 것이다. 절대적으로는 연결되어 있는 라인의 수를 세어주면 된다. 상대적으로는 (절대적 연결정도중심성)/g-1이다.

    근접 중심성은 한 점이 다른 점에 얼마만큼 가깝게 있느냐를 알아보는 것으로, 모든 점들과의 거리를 측정하여 파악할 수 있다. 절대적으로는 모든 점들과의 거리를 합한 것을 역수로 취해준 것이 된다. 역수를 취하는 이유는 거리의 값이 높을수록 좋은 것이 아니기 때문에, 이를 맞춰주기 위해 하는 것이다. 상대적은 절대적*(g-1)이 된다.

    마지막으로 매개중심성은, 점과 점간의 이동에 나를 얼마나 거쳐가야하는지를 통해 얼마나 주요한 매개체로 작용하고 있는지를 파악하는 것이다. 절대적으로는 모든 점들을 세어 계산해야하며, 상대적은 절대적 매개중심성값을 (g-1)(g-2)/2로 나눠준 것이다.

30. 근접/매개 중심성은 둘다 정보통제 측면에서 주요한 역할을 한다. 특히 매개 중심성은 더욱 그러한 특성을 가지는데, 애초에 점과 점을 이어주는 정도로 파악한 매개에 의존하는 측정방법이기 때문이다. 두 중심성은 소수의 특정 행위자들이 정보통제를 담당하는 정도를 포착한다는 특징이 있다.

31. 집중도란 네트워크 전체가 한 가지 중심으로 집중되는 정도를 표현하는 지표이다. 가장 중심적인 중심성 값과 다른 점들의 중심성 값 간의 차이를 보는 것으로, 네트워크 상의 실제 차이의 값과 이론적으로 가능한 최대 차이의 합 간의 비율로 계산된다.

    중심성은 한 점을, 집중도는 전체 네트워크의 입장을 가진다.

    밀도는 연결된 라인의 수를 통해 그래프가 얼마나 촘촘하게 연결되어 있는지를 파악한 평균의 개념이라면, 집중도는 최대값과 각 점들의 값의 차이를 통해 얼마나 차이가 있는지를 알아보는 분산의 개념이므로 반비례적인 관계를 가진다.

32. 연결정도집중도 계산식은 분자는 중심성이 가장 높은 점과 다른 모든 점들의 연결정도 차이의 합이며, 분모는 이론적으로 가능한 최대값으로

    (g-1)(g-2)가 된다.

    근접집중도의 분자는 네트워크 내에서 가장 높은 상대적 근접중심도와 다른 모든 점들간의 차이의 합이며, 분모는 (g-1)(g-2)/(2g-3)이 된다.

    매개집중도의 분자는 매개중심성 값이 가장 높은 점과 다른 모든 점들의 차이의 합이며, 분모는 [(g-1)^2(g-2)]/2가 된다.

33. 기계학습 기반 접근법은 기계에 긍정/부정 분류를 학습시킨 후에 진단(예측)하는 방식이며, 감성사전 기반 접근법은 학습시키는 것이 아니라 미리 감성사전에 점수를 부여한 후 그 사전을 기반으로 감성점수를 도출하는 방식이다.

34. 화장품산업을 예시로 들면, 우선은 타사의 경쟁 상품들의 감성점수를 계산하여 비교함으로써 각 상품들의 평이 어떤지를 알아볼 수 있다. 또한 상품을 정한 후 속성별(기능/향기/가격)로 비교할 수도 있으며, 특정 속성만을 선정해 비교해볼 수도 있다.