텍스트 전처리 : 코퍼스 데이터를 토큰화 / 정제화 / 정규화 하는 과정을 통해서 해당 데이터를 사용하기 알맞은 형태로 가공하는 과정

1. 토큰화
2. 정제 및 정규화
3. 어간 추출 및 표제어 추출
4. 불용어 추출
5. 정규표현식

1, 단어 토큰화 :
토큰의 기준을 word로 하는 경우, 단어 토큰화라고 함.. 단어는 단어 단위 외에도 단어구, 의미를 갖는 문자열로 간주되기도 함.
구두점, 특수문자 제거가 보통 필수적이나 가끔 구두점이나 특수문자를 전부 제거하면 의미를 잃어버리는 경우도 많음, 영어는 띄어쓰기를 기준으로 토큰을 구별하기 쉬우나 한글은 그렇지 않음

주의해야 할 점
1) 구두점, 특수문자를 단순 제외해서는 안됨. 특히 숫자나 날짜, 다양한 단위를 다룰때 쓰이는 콤마(,)가 중요한 역할을 할 떄가 있음
2) 줄임말과 단어 내에 띄어쓰기가 있는 경우, 영어의 경우 (')와 줄임표현이 굉장히 많이 쓰임
3) 한글의 경우 "자립 형태소"와 "의존형태소"가 공존하고 있음. 어절이 아닌 형태소를 기준으로 토큰화가 이뤄져야 하는데 이 경우 모호한 기준에 적합한 결과를 만들어내지 못 할 경우가 있음
4) 띄어쓰기가 영어보다 잘 지켜지지 않는 경우가 많기 때문에(학술 논문에서는 그렇지 않을 확률이 높겠지만..)

KoNLP 사용
- extractNoun() : 명사추출
- Pos() : 형태소 분리
- SimplePos09() SimplePos22() : 형태소 분석
- Kospacing() : 띄어쓰기 문제 어느정도 해결해 주는 라이브러리
…

2. 정제 및 정규화
정제 : 노이즈 데이터 삭제
정규화 : 표현방법이 다른 단어를 통합시켜 같은 단어로 만들어주는 과정

정제 > 토큰화 과정에 앞서 일부 필요하기도 함

정규화 > 100% 해내는 것은 매우 힘들기 때문에 어느정도의 합의점을 설정해 두고 진행해야 할 필요가 있음.

<정규화>
1. 규칙에 기반한 표기가 다른 단어 통합
- 어간추출(stemming)
- 표제어 추출(lemmatization)
2. 대 소문자 통합
3. 불필요 단어제거
- 등장 빈도가 너무 작은 단어
- 길이가 지나치게 짧은 단어(영어 한정)

정규표현식 : 노이즈 데이터의 특징을 잡아내어 정규표현실을 통해 이를 제거.\

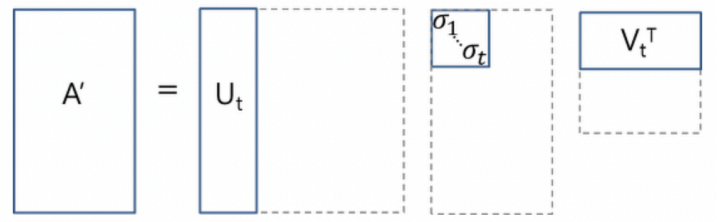
3, 정규화 - 어간추출 및 표제어 추출

두 방법 모두 특정 단어 집합을 하나의 단어로 일반화 시킬 수 있다면, 하나의 단어로 표현하여 최종적으로 문서 내 단어를 줄이는 과정에 해당함.

표제어 추출(Lemmatization) :
주어진 단어의 "기본 사전형 단어"를 찾아내는 것
어간과 접사를 구별하여

LSA에 대한 전반적 이론 설명(블로그 포스트)
https://wikidocs.net/24949
LDA에 대한 전반적 이론 설명(블로그 포스트)
http://bigdata.emforce.co.kr/index.php/2020072401/
https://bab2min.tistory.com/567?category=673750
LSA, LDA에 대한 전반적 이론 설명(영문)
https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

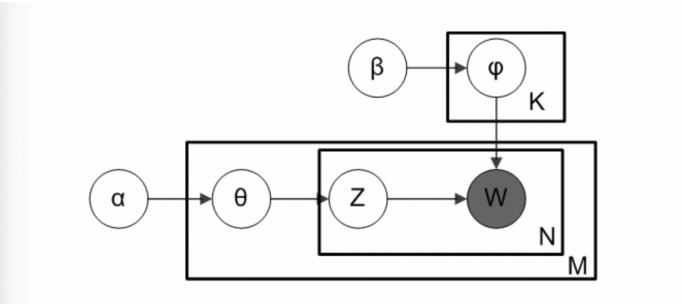
LSA
- 단어들의 확률을 이용하여 문서집합 내의 잠제된 토픽 도출, 문헌-용어 행렬(DTM)에 주성분 분석과 특이값 분해(SVD)를 실시, 하이퍼파라미터 값을 설정하여 SVD 행렬을 절단한 뒤 (데이터 차원 축소) 잠재의미 분석실시.
- 제작한 문서 벡터와 단어 벡터를 통해
- 다른 문서의 유사도
- 다른 단어의 유사도
- 단어로부터의 문서 유사도 추출
- 새로운 개선한 기법으로 LDA가 존재하며 토픽모델링의 주요 기법으로 사용되고 있음
- 이미 계산된 LSA에 새로운 데이터를 넣으면 처음부터 다시 계산해야함 > 새로운 정보 업데이트 어렵다는 단점



LDA : 잠제 디리클레 할당
- 디레클레 분포를 적용하여 문서, 단어 등 관찰된 변수를 통해 문맥, 문서의 구조 등 보이지 않는 변수를 추론하는 방법으로 전체 문서집합의 주제, 각 문서별 주제 비율(문헌별 주제분포), 각 단어들이 각 주제에 포함될 확률(주제별 단어분포) 등을 파악하여 주제 추출.
- 새로운 문헌이 주어져도 분석할 수 있음
- 설정한 파라미터 값에 따라 결과가 달라지므로, 적절한 파라미터 값을 설정하지 않으면 적합한 결과 얻을 수 없다는 한계점 존재

M: 문헌의 갯수
K: 주제의 갯수
N: 문헌에 속한 단어의 갯수
θ: 문헌의 주제 분포(주제 분포는 문헌마다 다를테니, 총 **θ_1~θ_M**까지 M개의 주제 분포가 있음.)
φ: 주제의 단어 분포(단어 분포는 주제마다 다를테니, 총 **φ_1~φ_K**까지 K개의 단어 분포가 있음.)
Z: 해당 단어가 속한 주제의 번호
…

HDP
LDA에선 사전에 정해진 주제 수 K값으로 분포를 형성하는 디레클레 분포를 사용함
DP에서는 직접 K값을 사전에 설정하지 않아도 모분포에 따른 임의의 주제 개수를 가진 표본의 분포를 생성할 수 있다. (자세한 내용은 너무 복잡해서.. 정리하질 못하겠습니다 ㅠ)


β
φ
K
α
θ
Z
W
N
M