2021.11.09

피클 파일 저장방식 python 3.6으로 바꿔서~ 다시 보내드리기

<https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-18-%EC%84%A0%ED%98%95%ED%8C%90%EB%B3%84%EB%B6%84%EC%84%9DLDA>

PCA 말고 LDA로 함 보자.

1. 기업개요 추가 크롤링
2. ## Dart 기업 공시 서비스 페이지 에서 내부 API 를 이용해 크롤링
3. ## Fnguide 페이지에서 남은 3줄 정도 추가 크롤링 해서 해보기
4. Stop words 나 기타 vectorizing 할 때 더 추가할 수 있는 사항들 시험해보기

* 지금은 명사만 하고 있는데, 다른 품사에 속한 단어들도 포함시켜서 morph 컬럼 다시 만들기
* 실제 토큰화 된 단어들 보면서 Stop words 추가하기
* 토큰화를 더 잘 할 수 있게, (ex. 하이닉스반도체 -> 하이닉스 반도체)
* 영어 포함 미포함, ( ex. Dram, vs D램)

1. 주가변동 데이터 클러스터링

전처리 : 전날 종가~오늘 종가: 상승, 하락 폭 (%) column 만들어야 함 (정규화)

!! 시계열 데이터 DTW(dynamic time warping) 방식

Tslearn 패키지에 있는 Kmeans 사용시 (metric = ‘dtw’ 로 설정하면 가능한 듯)

((종가 데이터에도 NA/n 값 있음))

1. PER 을 어떻게 쓸 것인가?

* PER 값이 na/n 으로 되어있는 친구들 왜 그런지
* 다른 재무지표(PBR, EPS,,,) 들 추가해서 해보기?

1. (2번대안) Clustering 시 변수의 중요도 설정이 가능한가? -메일 답장 오는 거 보고나서

1번 안 : 각각의 클러스터링 진행후, 결과 분석

2번 안 : 모든 변수들을 하나로 합쳐 클러스터링 진행

대안: 클러스터링 결과를 가지고, one-hot encoding 형식으로 바꾸어 벡터로 만든 후, 다시 클러스터링 -> 조교님한테 메일???