실습을 통해 간단한 wordcount를 진행했을 때, groupByKey가 다소 오래 걸리는 느낌을 받았다. 그래서 간단한 서치를 해보니 아래와 같은 특성이 존재했고, reduceByKey를 더 선호한다는 결과를 얻을 수 있었다.

groupByKey를 사용하면 더 좋은 케이스가 하나쯤은 있지 않을까 싶어서 찾아봤으나 실패했다.

reduceByKey

- 키를 기준으로 셔플링을 하기 전에 미리 각 파티션 내에 있는 데이터들을 먼저 combine (맵리듀스의 combiner를 사용하는 것과 동일한 역할)
- 따라서 네트워크를 통해 전송되는 데이터의 양을 최소화

groupByKey

- 모든 key-value 페어에 대해 셔플을 수행
- 따라서, 네트워크 자원을 더 많이 소모