Homework #5
406.424 인터넷응용
2016년도 봄학기

제출기한: **2016년 6월 9일 23:59까지**

1. (100점) **MovieReview** 데이터를 활용하여, 아래의 작업을 수행하시오.
(1-1의 경우 문제에서 요구하는 arff 파일을 함께 제출하시오. 1-2, 1-3, 1-4의 경우, 결과 비교 및 의견 서술 내용만을 MS-Word 문서 파일 형태로 작성하여 제출하시오. 모든 문제에서, 명시한 것 이외의 옵션들은 기본값으로 유지하고 진행하시오.)

1-1. (10점) Weka에서 StringToWordVector filter를 사용하여, 원본 데이터를 word vector 형태로 변환하고 이를 **word-vector-data.arff** 파일로 저장하시오. 이 과정에서 아래의 옵션을 설정하고 진행하시오.
- **outputWordCounts: True, stemmer: SnowballStemmer, stopwordsHandler: Rainbow, tokenizer: WordTokenizer**

1-2. (30점) 1-1에서 설정한 옵션 조합을 아래와 같이 변경하면서, 최소 4가지 조합 이상의 옵션들을 각각 적용하여 데이터를 다시 변환한 뒤, Naïve Bayes classifier를 사용하여 sentiment classification을 수행하고 분류 정확도를 서로 비교 분석하시오. 테스트 시에는, 아래의 테스트 옵션을 설정하고 진행하시오.
- **outputWordCounts: {True, False}, stemmer: {SnowballStemmer, NullStemmer}, stopwordsHandler: {Rainbow, Null}, tokenizer: {WordTokenizer, NGramTokenizer}**
- **Test options: Percentage split (%70)**

1-3. (30점) 1-1에서 변환하여 얻은 데이터에 대하여, 커널을 변경하면서 SVM classifier를 사용하여 sentiment classification을 수행하고, 분류 정확도를 서로 비교 분석하시오. 이 과정에서, 아래의 커널 옵션 및 테스트 옵션을 설정하고 진행하시오.
- **kernel: {PolyKernel(exponent=1.0), PolyKernel(exponent=2.0), RBFKernel}**
- **Test options: Percentage split (%70)**

1-4. (30점) 1-1에서 변환하여 얻은 데이터에 대하여, K값을 2~10 사이의 값으로 조정하면서 K-means clustering을 적용하여 sentiment classification을 수행하고, 분류 정확도를 서로 비교 분석하시오.
- **Cluster mode: Classes to clusters evaluation**

ETL 사이트에 업로드된 예시 코드를 응용하시오. 과제는 MS-Word나 PDF의 파일 형태로, txt 파일 및 Java 코드 결과물과 함께 압축하여 ETL 사이트에 제출하시오.