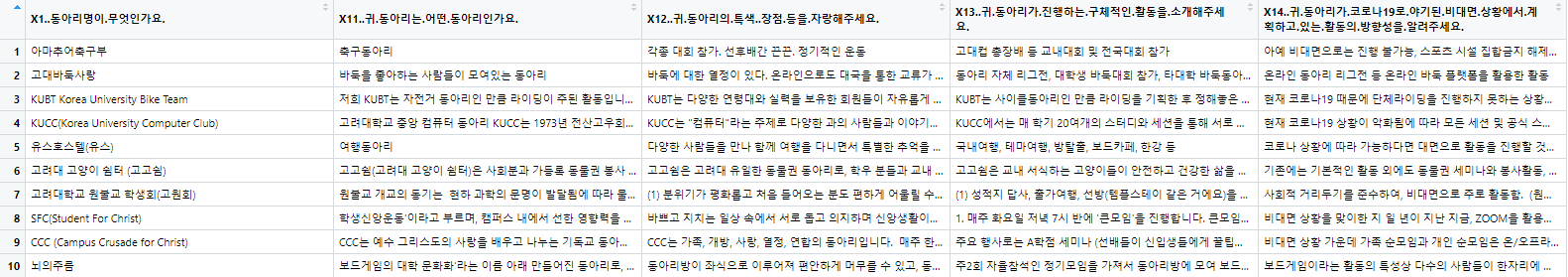
1. Corpus 구축

문서 중 동아리의 특성 분석에 필요한 대상 데이터를 선정한 후, 텍스트 파일의 형태로 R 프로그램에 데이터를 불러옴



(주어진 csv파일에서 중복된 행 제거하면 총 78개의 동아리 정보가 포함되어 있었음)

4가지의 소개글을 모두 합친 text를 분석함

tm패키지의 Vcorpus 함수를 이용하면 문서별로 meta와 content 두 요소를 가지는 PlainTextDocument 리스트 구조가 형성됨

1. 말뭉치 데이터 정제

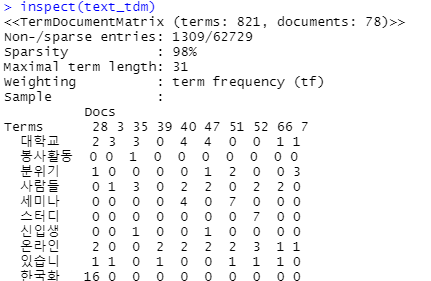
데이터 정제(전처리)는 공백, 문장부호, 불용어 등을 제거하여 좀 더 명확한 텍스트마이닝 결과를 얻을 수 있도록 텍스트를 다듬어 나가는 과정임

불필요한 텍스트 공백 제거, 문장부호 제거, 숫자 제거, 특수문자 제거, 불용어(‘동아리’) 제거 순서로 진행

이후, 한글 어근 추출을 진행함

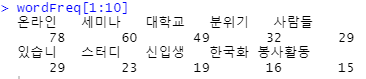
1. 빈도수 분석 및 시각화

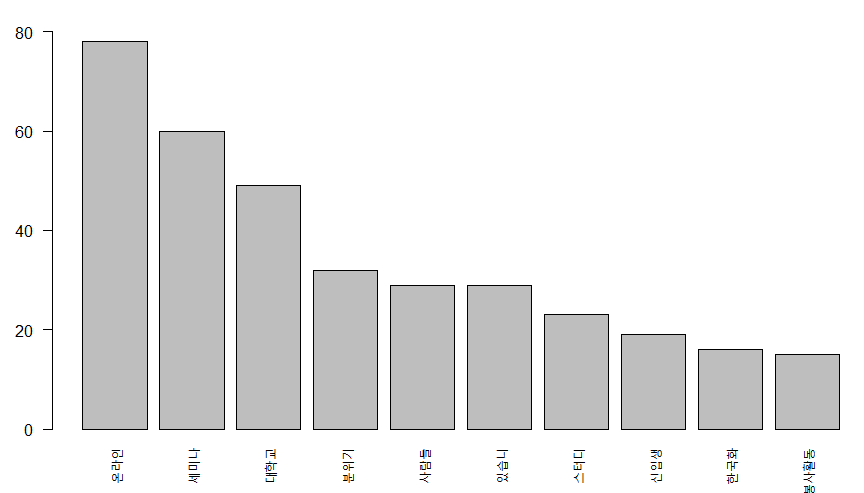
단어문서행렬(TermDocumentMatrix)이란 해당 단어가 문서에 몇 번 등장하는지를 행렬의 형태로 나타낸 것(단어 길이는 2 이상인 것을 선택하였음)



Sparsity가 98%라는 것은 tdm안에 0인 원소가 98%라는 것을 의미함

전체 문서에 등장한 단어의 빈도를 간단히 나타내기 위해서 행렬의 결과를 합해 내림차순으로 정렬하여 상위 10개 빈도 단어 추출





최소 빈도 3으로 한 wordcloud

