Explanation:

In this case each paragraph was considered to be a different document new line characters were introduced between each paragraph. The mapper function reads from stdin which is the iub_wiki.txt file and for each line in the file creates a new document id and strips and finds all words in the line and then prints out the word, a tab, then the document id or in this case the paragraph id, then a colon, and then the location or the index of the word in the line. The reducer reads the output and then after returns a data structure consisting a dictionary with the word as the index and the value as a dictionary containing the document id as the key and the value as an array with a number for index 0 representing the count of that word in that document and an array at index 1 containing all of the locations or indexes of that word in the given document. An example output line is:

and {'doc2': [6, ['69', '24', '86', '55', '75', '21']], 'doc3': [6, ['71', '66', '11', '51', '15', '27']], 'doc1': [2, ['8', '45']], 'doc4': [5, ['70', '45', '86', '20', '5']], 'doc5': [3, ['80', '31', '57']]}

From that line it can be interpreted that the word and is the 69th, 24th, 86th, 55th, 75th, and 21st word in document 2 and the same can be done for the other documents. From this you are able to tell which word is in which document, the number of times it is seen in that document, and the locations of the word in that document.

The one improvement I would make is to be able to read multiple text files and tag them as different documents rather than each paragraph being considered a new document.