**Assignment 1**

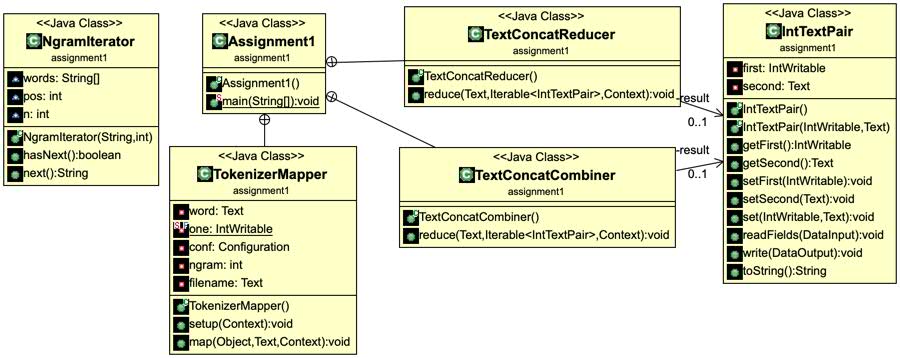


Figure1: Class Diagram

**Additional Classes**

1. *NgramIterator*: a java class for creating ngram-iterator from a string
2. *IntTextPair*: output value class implemented writable, this class contain pair of an integer and a string. The integer represents number of ngram-word and the string represents filenames of this word is found.

**Mapper Class**

1. Setup method: to get ngram parameter from configuration and filename from filesplit class.



1. Map method: creae a NgramIterator class to split ngram-words from document. Then, iterate through each token and form a key value pair. Key is a ngram-word and value is a IntTextPair object which is a pair of Intwritable(1) and it’s filename.



**Combiner and Reducer Class**

These two class is almost the same but the reducer class has a filter of minimum count for each ngram-word.

1. Only reducer class read *min\_count* parameter from configuration.



1. Create a set of string to contain list of filenames. Then, iterate through all values with respect to a key, sum up all of value from integer in a value object and add filename from string in value object to set of string.



1. Only reducer class will terminate itself if the sum is lower than min\_count.



1. Sort the set of string by parsing them to a treeset object and combine them into a string.



1. Form the final key/value pairs result for each word using context.

