Write a scala program using Apache Spark to find Top N bi-gram (https://en.wikipedia.org/wiki/N-gram) words (with its frequencies) from a document whose frequencies are above a certain number.

For example, for the following raw text, we want to apply bi-gram analysis.

“Alice is testing spark application. Testing spark is fun”.

Applying bi-gram will generate token like this:

[(Alice, 'is'), ('is', 'testing'), ('testing', ‘spark’), (‘spark’, 'application.'), ('testing', ‘spark’), (‘spark’, 'is'), (‘is’, 'fun')].

So, if we want to find top 2 frequent bi-gram tokens whose frequencies are above 1, the output will be:

('testing', ‘spark’)   2

In this program, you have also remove the stop words (http://www.textfixer.com/resources/common-english-words.txt) and stemming (use nltk) (http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html) before applying your bi-gram logic for your document. For example,

“Testing spark is fun” will look like “test spark be fun”.

After stemming, your output of bi-gram will be like following:

Document:

“Alice is testing spark application. Testing spark is fun”.

After removing stop words, stemming and applying bi-gram will generate tokens like this:

[(Alice, 'be'), ('be', 'test'), ('test', ‘spark’), (‘spark’, 'application.'), ('test', ‘spark’), (‘spark’, 'be'), (‘be’, 'fun')].

So, if we want to find top 2 frequent bi-gram tokens whose frequencies are above 1, the output will be:

('test', ‘spark’)   2