**Lab Assignment 2**

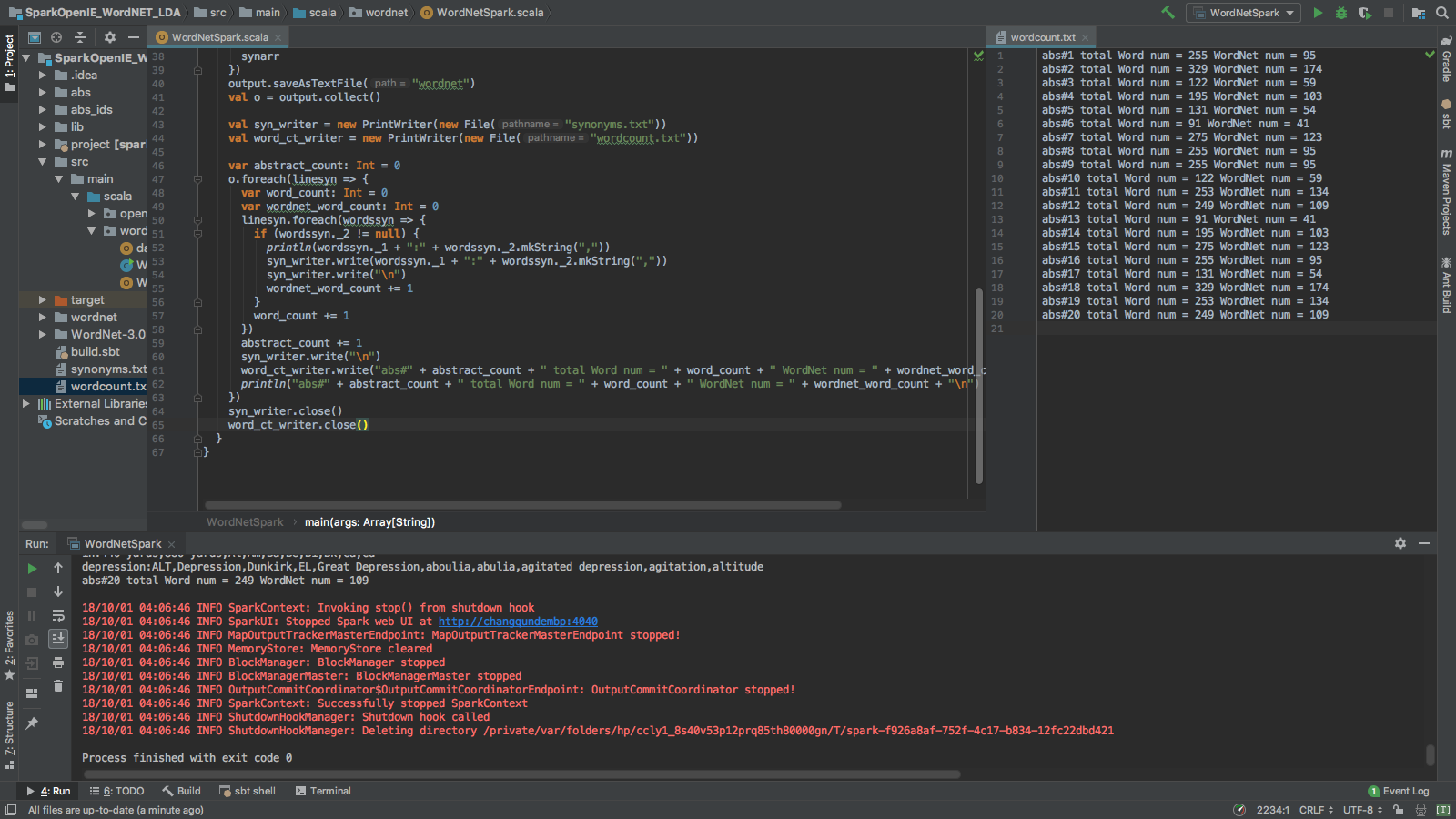
Yangfan Sun 16207318

Consider 20 abstracts, write spark code for the following:

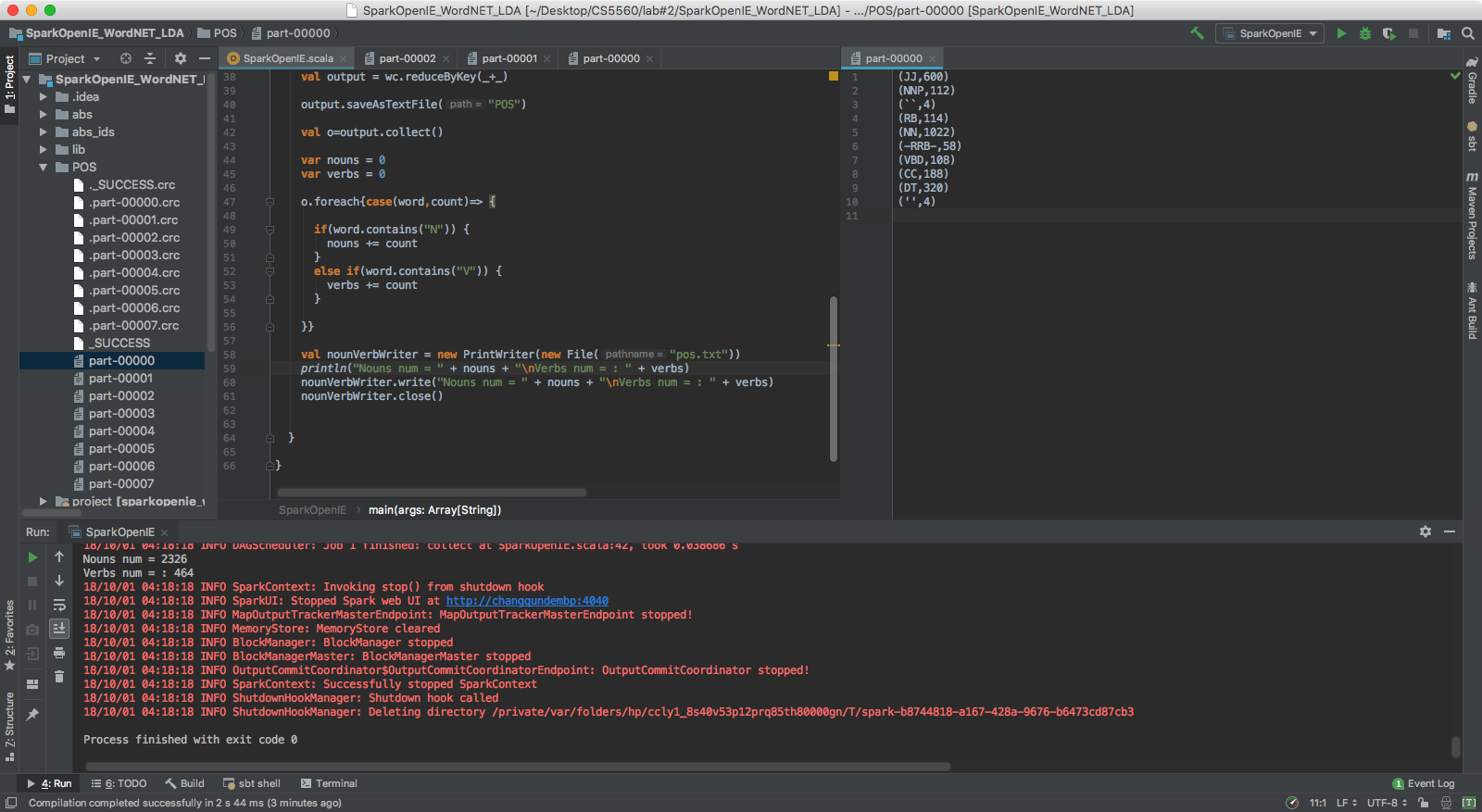
1. ***Report Data Statistics as Lab Assignment 1 Based on Spark Code.***

This question is quiet similar to ICP4, two sections will be included, “word count” section is applied “WordNet” code and “part of speech” section is applied “coreNLP.java”.

Part 1. In this section, 20 abstracts as input and implement work count one by one, recognized by WordNet. In the meantime, synonyms also are found by it. Total number of words and synonyms included in each abstract as output, shown in txt file.

******

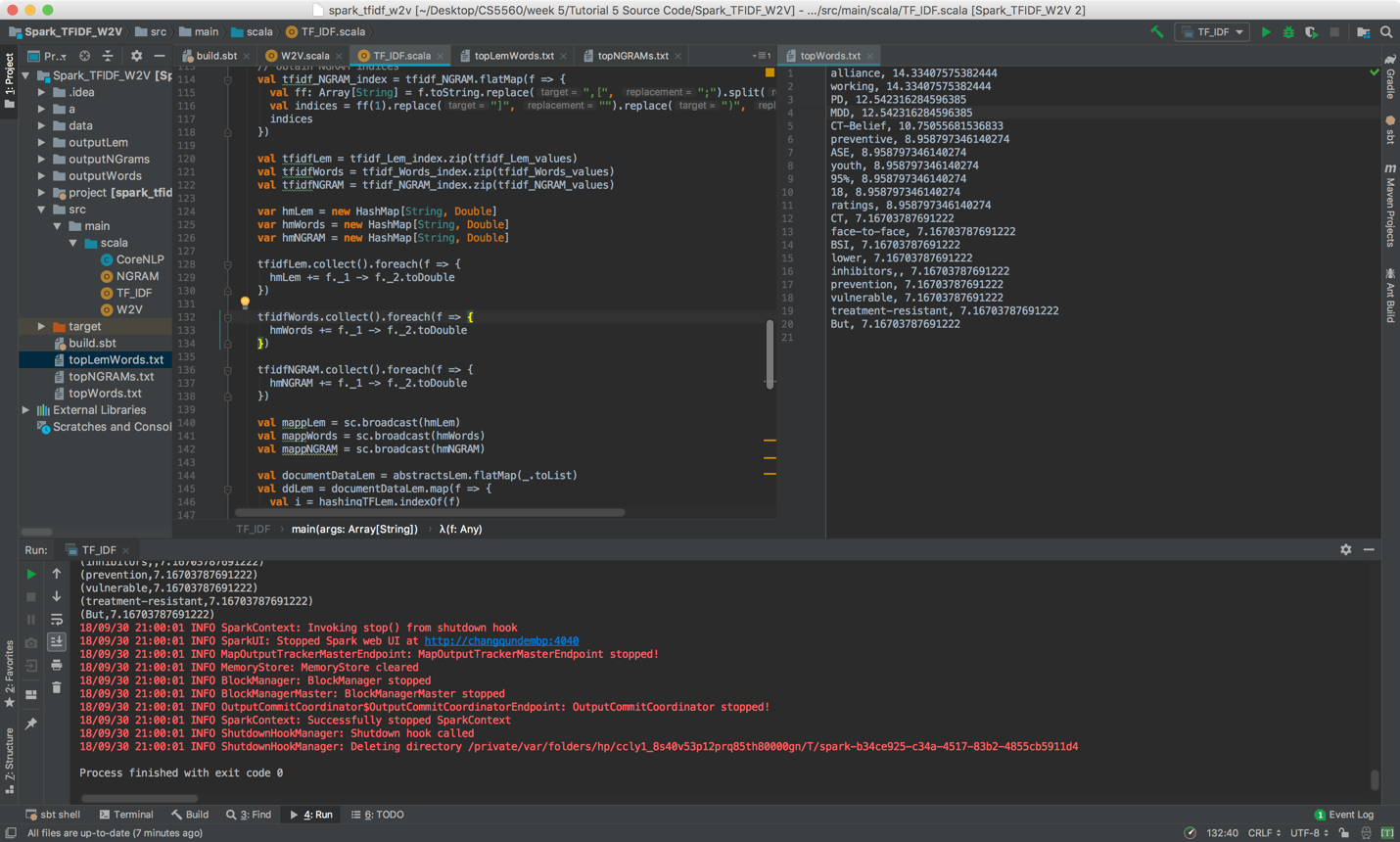
Part 2. The number of POS on each abstract as output and also the overall nouns and verbs will be statistic.



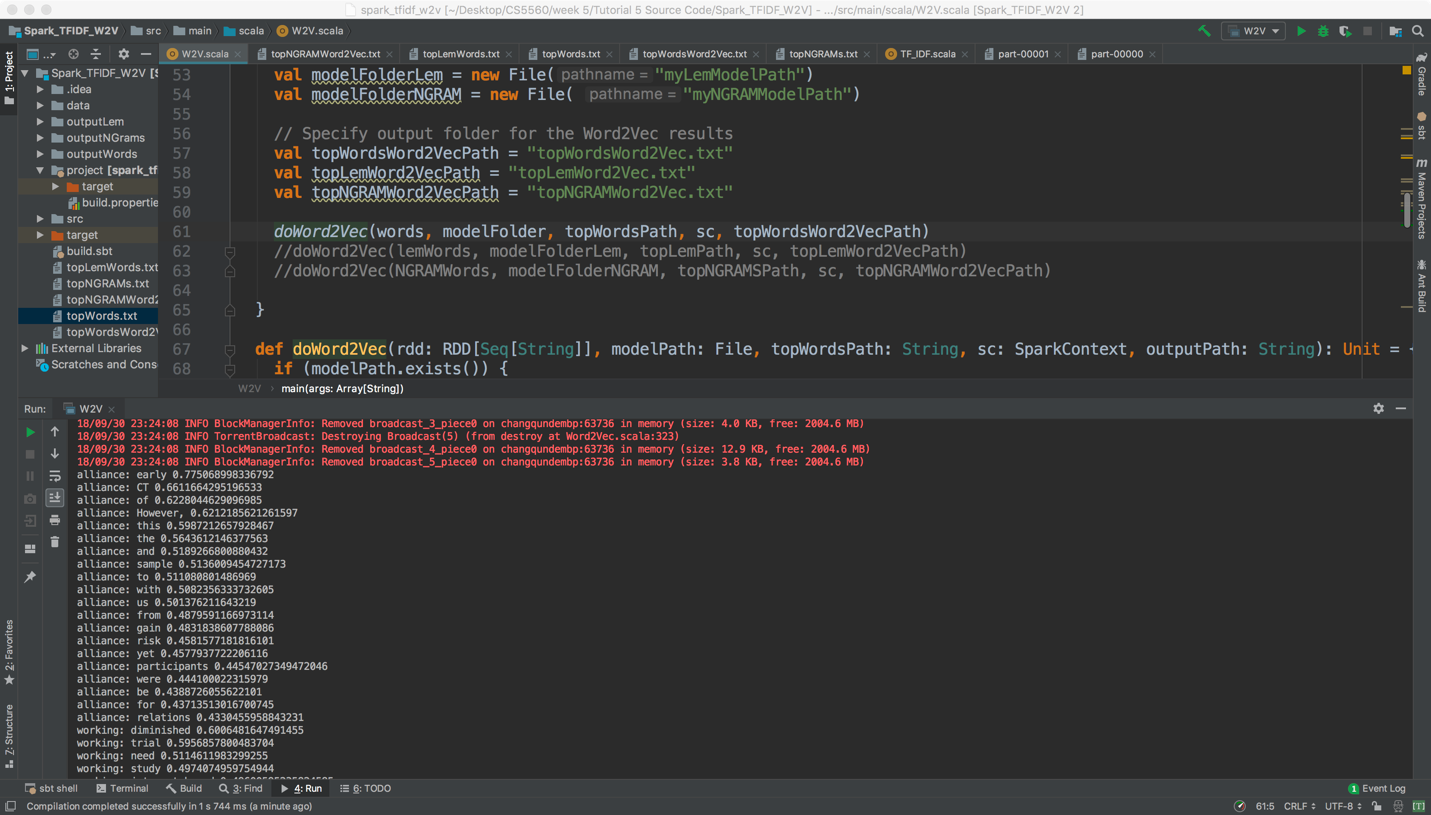
1. ***Word Statistics: Find out the top TF-IDF and corresponding Word2Vec word.***

This section is close to ICP5, extracting top 20 words and statistic their TF-IDF and work2vec.

Part 1. Extract top 20 words’ TF-IDF, same as ICP5 and we only picked up words without lemmatization or N-Gram as the input of next step.



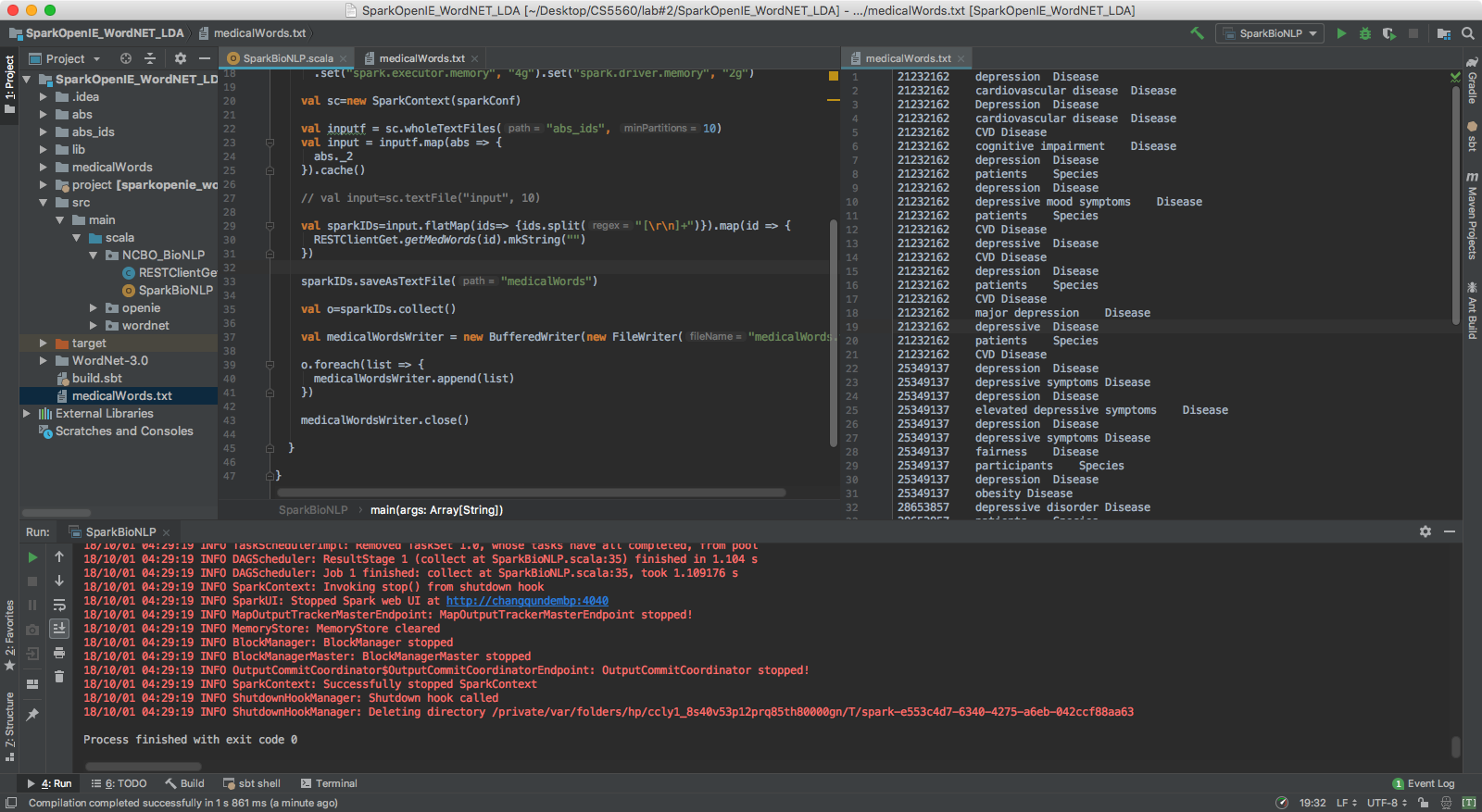
Part 2. Find word2vec by utilizing the TF-IDF we obtained from last part. Then the similarity of each word will be graded according to the discount of each transforming vectors from 20 top words.



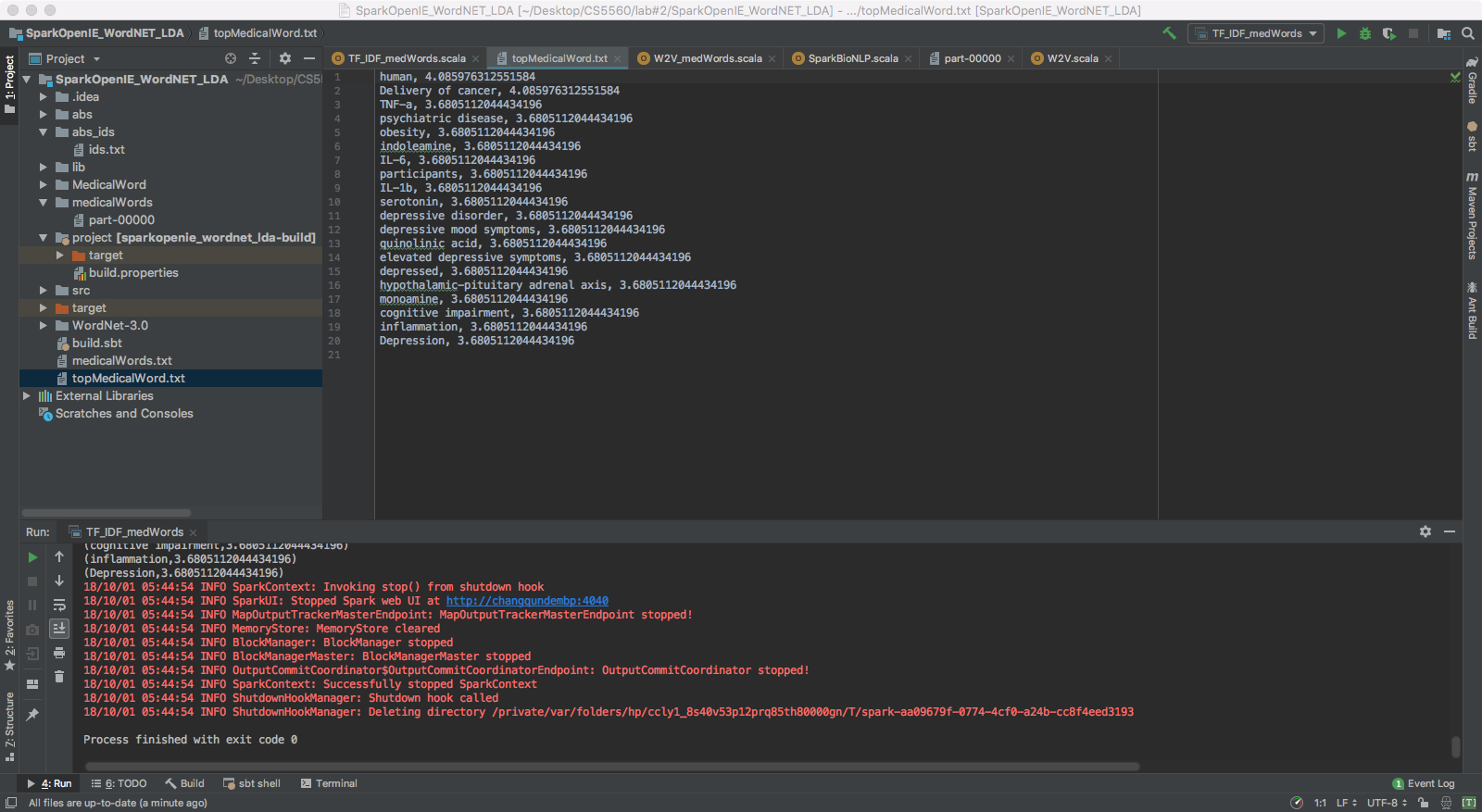
1. ***Medical Word Statistics: Find out the top TF-IDF and corresponding Word2Vec words.***

Same as last section, only need extract the matching medical words firstly. Each word that was found within the following bio-concepts: Gene, Disease, Chemical, Species, Mutation.

Part 1. Extract medical words, by applying SparkBioNLP and extract from PMID.



Part 2. Find the TF-IDF of medical words extracted.



Part 3. Find the word2vec of each medical word extracted.

