Optimization:

1. **Convert words to numbers, this number represent this word's order in frequency.**

   Example:
   Array(((0,20191124),Array(woman, stabbed, adelaide, shopping, centre)) => Array(((0,20191124),Array(12, 32, 19, 36, 29))

2. **Store key-value's value by title id instead of title content, create a look up table link title id => content**

   Example:
   Title id look up table Map(0 -> Array(12, 32, 19, 36, 29)) (title index, content)
   Key-value Array(Array((12,0,2019) (key, title index, date)

3. **Compute similarity in spark dataframe without collect them to memory, avoid memory out**

   Use user define function to operate spark dataframe.

   Example:

```
val convertUDF = udf(( index1:Int, index2:Int) => {
……
  intersectWords/unionWords
})
(dataframe).withColumn("sim", convertUDF(col("index1"), col("index2")))
```

   Calculate similarity in dataframe and append this column to dataframe