**Lab 1 Advanced - Apache Spark**

Используя *spark-shell* или *pyspark (https://spark.apache.org/docs/latest/quick-start.html)*

Напишите набор transformations and actions, который бы для файла файл со статьями с Wikipedia [http://alaska.epfl.ch/~dockermoocs/bigdata/wikipedia.dat](http://alaska.epfl.ch/%7Edockermoocs/bigdata/wikipedia.dat)

1. Подсчитывали сколько раз встречается каждый из языков

"JavaScript", "Java", "PHP", "Python", "C#", "C++", "Ruby", "CSS", "Objective-C", "Perl", "Scala", "Haskell", "MATLAB", "Clojure", "Groovy"

1. Для каждого из языков в скольких из статей он упоминается

Для выделение отдельной статьи можете использовать такую структуру

class WikipediaArticle(title: String, text: String)

def parse(line: String): WikipediaArticle = {

val subs = "</title><text>"

val i = line.indexOf(subs)

val title = line.substring(14, i)

val text = line.substring(i + subs.length, line.length-16)

WikipediaArticle(title, text)

}