Лабораторные работы: пояснения

К.Т.Н., ЗАВ. ЛАБ. ИТ ГРИЦЮК С.В.

Правила

1. Отчет, как итог каждой лабораторной работы (в электронном виде);
2. Минимум кода в основном теле программы (метод main или тело класса-наследника App);
3. Код должен быть логически разделен на функции;
4. С умом используем mutable/immutable структуры;
5. Стараемся использовать функциональный подход;
6. Если пишем в коде комментарии, то только по-английски;
7. Не отформатированный код даже не показываем.

Используем случайность

```
Seq.fill(n)(Random.nextInt)
```

```
Random.nextInt(100)
```

> res1: Int = 58

```
val lst = List(1, 2, 3, 4, 5)
val lst2 = Random.shuffle(lst)
```

Коллекции (1)

```
val lst1 = List(1, 2, 3, 4, 5)
lst1.take(3)
lst1.drop(3)
```

```
val lst2 = List("Petr" -> 324, "Ivan" -> 123, "Vasya" -> 425)
lst2.sortBy(_._2)
```

> res2: List[(String, Int)] = List((Ivan,123), (Petr,324), (Vasya,425))

Коллекции (2)

```
val lst3 = List("a" -> 5, "b" -> 2, "b" -> 6, "a" -> 2, "c" -> 8)
```

> lst3 : List[(String, Int)] = List((a,5), (b,2), (b,6), (a,2), (c,8))

```
val m = lst3.groupBy(_._1)
```

> m : Map[String,List[(String, Int)]] = Map(b -> List((b,2), (b,6)), a -> List((a,5), (a,2)), c -> List((c,8)))

```
m.map(x => (x._1, x._2.map(y => y._2).sum))
```

> res3: Map[String,Int] = Map(b -> 8, a -> 7, c -> 8)

Читаем данные

Из сети

```
// load URL contents as text
def get(url: String): String = {
  val content = Source.fromURL(url)
  content.mkString
}
```

Из файла

```
val fileName = "c:/my-file.txt"
val file = Source.fromFile(fileName)
val lines = file.getLines.map { line =>
  ...
}.toList
```

Из ресурсов аналогично

Подключаем библиотеки к проекту

Maven проект – заготовку проекта (Scala IDE) ищите среди файлов

Файл pom.xml, раздел dependencies

```
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.11</artifactId>
    <version>1.6.2</version>
</dependency>
```

Формат входных данных

Проверяйте, чтобы данные (текст) были в правильном формате и кодировке (UTF-8)

Иначе не удивляйтесь, что в консоль ничего не выводится

Spark

Чтобы запустить Spark-проект в режиме «все-на-одном-узле» на Windows вам, вероятно:

1. Потребуется найти и скачать файл winutils.exe
2. Расположить его где-либо на своем компьютере
3. Прописать путь к нему (как путь к Hadoop-окружению) в коде своего приложения:

```
System.setProperty("hadoop.home.dir", "<ваш путь>/")
```

Stemming

Если ничего своего не нашли, используйте код Snowball libstemmer (Java версия):

http://snowball.tartarus.org/

Файлы добавляйте в свой проект

Очистка данных

Очистка данных подразумевает, что Вы:

1. Удалите из текста все ненужные и спец. символы
2. Решите проблему с переносом слов (если она присутствует для вашего варианта)
3. Удалите стоп-слова по самостоятельно подготовленному списку (будет отличаться для разных произведений)

Извлечение сущностей

Различные варианты, на Ваше усмотрение:

1. Найти готовую библиотеку/проект и использовать для своей работы
2. Обучить классификатор (или предложить другой подход) для выявления сущностей
3. Придумать набор правил и реализовать их в своем проекте

Используйте Google + запросы на английском языке!