扩展：Excel能做什么

> 编写：钟钧儒
>
> 最后修改日期：2018年3月28日

我估计很多人在自己的记忆开始之初，就知道了微软的Office系列；Office三剑客也成为Windows乃至Mac用户的装机必备软件。大家都知道Word是文字处理，PowerPoint是演示文稿，那Excel能做什么呢？

Excel电子表格

据维基百科，“电子表格”是一类模拟纸上计算表格的计算机程序。它会显示由一系列行与列构成的网格。单元格内可以存放数值、计算式、或文本。电子表格通常用于财务信息，因为它能够频繁的重新计算整个表格。

在我理解中，Excel就是一个典型的“电子表格”软件。所以在Excel的界面中，就只有行和列构成的网格，一切的操作就基于这些行和列，也就是说，他只对部分类型的数据有效。

Excel与大数据

“大数据”近几年是一个很火热的话题。因为我自己希望在未来投入这一块的研究，所以也关注了一些大数据相关的话题。

的确，Excel能做**数据处理**的工作。但是**大数据**、**数据处理**这些名词都是非常的宽泛，那Excel能在当今火热的数据产业中做点什么吗？

答案很可能是**不能**做很多。我们所说的大数据具备五种特性（这也是DST各专业的IT课程内容）：

- Volume 容量：这一点强调了大数据的“大”。在《大数据时代》一书中提到，大数据在统计中可以认为是“样本=总体”。也就是说，大数据需要对及其大量的数据进行处理。大数据现在的处理一般是面向PB级别的数据量（1PB=1024TB）。
- Variety 多样性：这一点强调了大数据的“杂”。大数据中，大部分的数据均为*非结构化数据*，如文本。思考一下，你该如何处理文本？
- Velocity 速度：在大数据处理中，正因为有大量的类型多样的数据不断地产生，所以我们才需要以很快的速度处理数据。
- Variability 可变性：大数据处理中讲求实时可变。从这一点中我们可以推出，大数据处理中大部分情况需要机器自动化处理。
- Veracity 真实性：数据自己会说话。当你收集的数据足够的多，一幅*数据画像*自然会显现出来，人在其中能干预的东西就很少了。

Excel只能处理**结构化数据**。所谓的**结构化数据**就是说，用户能通过*键值*获得信息且数据的*结构*相对固定。考虑下面一张出自课上练习的表格：

Year	t(year)	Actual Population (P)
1988	0	81000
1989	1	84900
1990	2	89100
1991	3	93600
1992	4	98300
1993	5	103250
1994	6	108500
1995	7	114000
1996	8	119700
1997	9	125700
1998	10	132000

假设我想知道第10年的人口，我只需要观察**第二列**的内容为**10**的那一行，这一行中的**第三列**的数据即为我的目标数据132000。在此过程中，所谓的*键值*便是第二列的表头 `t(year)`。同样的，使用另一种典型的处理结构化数据的工具，*关系型数据库*（DBM学生IT课内容），要得到同样数据，输入的SQL代码应该是（假设表名为 `Population`，不保证这个代码能跑）：

```
SELECT `Actual Population (P)` FROM `Population` WHERE `t(year)` = 10
```

直接翻译这句代码其实就是上面的思考过程。这也是处理结构化数据的典型思路。

结构化 vs. 非结构化

并不是说处理结构化数据就不是大数据，我其实只是想说明Excel并不是处理大数据主流数据类型——非结构化数据的工具。

考虑下面一段从UIC官网复制的介绍，如何计算这段话的词数？

```
Situated in Zhuhai, Beijing Normal University-Hong Kong Baptist University United International College (UIC) was jointly founded by Beijing Normal University and Hong Kong Baptist University (HKBU). It is the first full-scale cooperation in higher education between the Mainland and Hong Kong. Its charter has been approved by the Ministry of Education with full support from local authorities.
```

这里提供一个这样的思路：英文中以空格作为单词的分割，从头开始*遍历*全文所有字符，直至找到一个空格，就可以计算其为一个单词。所以可以写成下列算法：

```
导入 文本
初始化 计数器
读取 文本中所有字符
    当 读取到一个空格
        计数器 + 1
    继续读取至文本结尾
循环结束
输出计数器
```

这就是最为广泛使用的大数据处理框架Hadoop的入门程序，词频计算。以下系其示例代码（已忽略细节，完整代码点这里）：

```
public class WordCount {

  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      // StringTokenizer 即为分割单词的关键工具
      StringTokenizer itr = new StringTokenizer(value.toString());
      // 循环计算单词数量
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        // 计数器 + 1
        context.write(word, one);
      }
    }
  }
}
```

反正我是想不出来怎么用Excel处理这些数据了……

怎么用Excel

其实我觉得，在我使用Excel的时候，实际上已经是有一点在编程的感觉了。况且Excel支持VBA脚本编写“宏”（听说有支持Python的计划）。所以能以计算机科学的一些思维去思考会比较好。

- 在开始之前，你需要收集数据
- 首先，用文字列出你的需求。统计？筛选？还是其他？
- 然后，在Excel尝试或者在Office支持中寻求相关的函数和功能，列出*算法*。
- 接着，在Excel中实现你的*算法*。
- 最后排版一下，美观一点。

Excel的替代品

Excel并不是第一个电子表格软件，但是确是搭乘了Windows普及的东风成为了当今当之无愧排名第一的电子表格软件。Excel并不免费，而且也并不是面面俱到，我这里列出了一些Excel的替代品：

- 电子表格软件
  - Calc： 免费开源的OpenOffice、LibreOffice包含
  - Google Sheets：在线的免费服务
  - Numbers：包含于苹果iWorks套件，对Mac用户免费
- 数据科学、统计软件：不是Excel的严格替代品，但这些能实现几乎相同或更高级的功能。
  - SPSS：IBM出品的统计软件，收费；开源社区推出的对应开源软件叫PSPP。
  - MATLAB：Mathworks出品的数据科学、计算机科学编程语言及相关软件，收费；对应开源软件为GNU Octave。
  - 配合CSV格式的数据源，可以使用编程语言处理：
    - R：开源编程语言，关注数据分析和数据科学领域，可以使用各种*库*实现很多Excel的函数对应的功能。
    - Python：开源编程语言，被广泛应用在数据科学、人工智能等领域，同样可以使用*库*实现数据分析应用。配合Jupyter软件可以制作一份图文并茂的且可以直接运行Python代码的报告。
- Business Intelligence：微软PowerBI，桌面单机版免费，云服务收费。BI是一种实现高级数据可视化的工具。

参考资料

https://zh.wikipedia.org/wiki/%E9%9B%BB%E5%AD%90%E8%A9%A6%E7%AE%97%E8%A1%A8

https://en.wikipedia.org/wiki/Big_data

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html