Lu Heng

How do you know what you know?

用GibbsLDA做Topic Modeling

June 24, 2011, 4:00 pm

Topic Modeling是一种文本挖掘的方法。将文本视作一个由许多许多词组成的数据库，就能通过分析哪些词经常在一起出现，哪些词出现的多，等等手段来探测文本中隐含的主题。比如，有一万篇文章，人工对这些文章进行分类，提取主题是十分浩大的工程。但是如果能用计算机来实现的话，对于社会科学研究者，尤其是注重对文本分析的传播学者来说，是个非常值得应用的工具。

简单来说，Topic Modeling就是干这么个活。把一堆文字放进去，你告诉计算机，你要多少个主题（比如，5），每个主题包含多少个字（比如，10），然后让计算机跑跑跑，过一会儿，计算机就告诉你5个主题，每个主题10个字分别是什么。

听起来有点玄乎，但是如果你能明白传统的因子分析（Factor Analysis），基本上就能明白Topic Modeling。核心思想无非就是降维度。对于因子分析来说，可以从100个变量，降维提出5个因子。对于Topic Modeling来说，可以通过100篇文章（可能包含100，000个字/词），降维提出5个主题。

具体到对Topic Modeling的操作，那就千变万化了。计算机科学（Computer Science）领域专门有人做这个，发展出各种算法。

我只介绍一种，叫做GibbsLDA。这其实就是Gibbs Sampling（一种抽样方法）跟Latent Dirichlet Allocation（LDA，一种算法/模型）的结合。这玩意儿太深奥了。我也解释不清楚。反正如果你google的话，有人用这两种东西的结合GibbsLDA写了文章，发了，貌似影响不小。是可行的，靠谱的。LDA最早是由David Blei提出的。

D. Blei, A. Ng, and M. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research (2003).

更多的文章可以看看这里。也可以google下列文章：

l Finding scientific topics.

l The author-topic model for authors and documents.

l Bibliometric Impact Measures Leveraging Topic Analysis.

l Detecting Topic Evolution in Scientific Literature: How CanCitations Help?

这个GibbsLDA有很多的软件版本，比如C++版，Java版，Python版，MatLab版。各种版本对输入数据的要求可能还不一样。就我使用的情况来看，C++版本最稳定，运算速度也最快。

但是呢，C++版本一般在Linux上运行，如在Windows下运行，还得按照个Visual Studio。工程浩大。

Windows上装个Linux其实不难，搞个Wubi，在C盘上分出一个空间（10G左右），傻瓜都能搞定。这个Wubi给装的Linux是Ubuntu版本。不难用。Wubi其实是给你的电脑上傻瓜式地装上了一个虚拟Linux系统。开机时会让你选进Linux还是Windows。进了Linux也不怕，搞个天翻地覆，也就在你分给它的那点硬盘（比如10G）里，坏不了大事。当然，你舍不得搞复杂你的Windows的话，可以想办法搞个Linux的机器来玩玩。

第三步：解压缩及安装。对于没用过Linux的同学来说，没有右键解压缩这个故事是很痛苦的。好吧，慢慢来。比如你这个狗屁文件放到了/home/user/LDA/下面。而你甚至连你在什么文件夹下都不知道。你可以这样。在Terminal（也就是一个黑屏幕，只能输入命令的那种）里面输入（下面的$表示一行命令的开始，不用输入）

$ cd /home/user/LDA/

就行了。然后，解压缩。输入

$ gunzip GibbsLDA++-0.2.tar.gz（这个gunzip后面是你刚下载的文件的文件名，我下的是GibbsLDA++-0.2）

$ tar -xf GibbsLDA++-0.2.tar

然后进到你刚解压出来的那个文件夹（假设你现在还是在/home/user/LDA/下面）。输入

$ cd \GibbsLDA++-0.2

现在，你已经在/home/user/LDA/GibbsLDA++-0.2/ 这个文件夹下面了已然后安装GibsLDA。输入

$ make clean

$ make all

到目前为止，你已经大功告成了。安装完成。

第四步：准备你要让计算机去做Topic Modeling的文件。在C++的环境里，Topic Modeling需要这样的一个文件。文件格式是dat。这是最原始的txt文件。你也可以用任何软件存成txt文件之后，直接把后缀改成dat就行。比如，你的文件包含1，000篇文章。那你的文件就是这样的

第1行是你总共的文章篇数，在我们的例子里面是1000

第2行到第1001行就是你的那些文章，每篇文章占一行。对于英文来说，每个词之间已经用空格分开了，但是中文不行，所以你要先对文章进行切词。切词这事儿，就不归我这篇小臭长文管了。

第五步：运行GibbsLDA++，得到你要的结果。

将你要跑的文件，比如就叫test.dat吧。将文件放到/home/user/LDA/ 下面，也就是/home/user/LDA/test.dat

然后进入到你装了GibbsLDA++的文件夹，也就是/home/user/LDA/GibbsLDA++-0.2/，然后运行指令。其实就是在Terminal里面输入

$ cd /home/user/LDA/GibbsLDA++-0.2/

$ lda -est [-alpha <double>] [-beta <double>] [-ntopics <int>] [-niters <int>] [-savestep <int>] [-twords <int>] -dfile <string>

这句话"$ lda -est [-alpha <double>] [-beta <double>] [-ntopics <int>] [-niters <int>] [-savestep <int>] [-twords <int>] -dfile <string>"里面其实是GibbsLDA进行估算的各种参数设计，你实际输入的指令可能是：

$ src/lda -est -alpha 0.5 -beta 0.1 -ntopics 100 -niters 1000 -savestep 100 -twords 20 -dfile /home/luheng/LDA/test.dat

这意思是，参数alpha是0.5（这个可以先不管），参数beta是0.1（这个也可以先不管），产生100个topic，运算迭代1000次，每迭代100次之后的结果都保存出来，每个topic包含20个词，要运算的文件是/home/luheng/LDA/test.dat

其中最直接的是.twords文件。这个文件里面就是你要的n个topic，以及每个topic下面包含的具体的字词。

.others里面是各种你设置的参数

.theta里面是每篇文章对应你设置的n个topic的"因子载荷"（factor loading）

.phi里面是每个topic对应每篇文章的"因子载荷"（factor loading）

.theta 和 .phi 里面的数据其实是一回事，互为转置罢（transpose）了。

.tassign是个啥玩意儿我暂时还没去搞明白。反正除此以外的那几个文件里面的东西已经够我用了。

一些提醒：

计算机很聪明也很笨，你给什么它都帮你算，所以你准备的文件一定要想清楚。比如，是不是所有字词都放心去，那些a, the, of啊，那些华丽丽的形容词啊，是拿掉还是放心去。或者是不是只放名词进去。这事儿只能自己决定，计算机帮不了你。

你选多少个主题，每个主题要多少字，迭代多少步。这玩意儿没有一定的规定。你就试吧。撞大运吧。虽然GibbsLDA靠谱，可终究还是跟因子分析一样，充满了arbitrary！

Category: Statistics 1 2 3 | Comment (RSS) | Trackback

15 Comments

1. *tangyyyyy* says:
   June 25, 2011 at 9:31 am

   看起来GibbsLDA的算法像个黑盒子，各种参数的设置对结果的影响和很难理解，期待下篇算法大揭秘

2. *bigrat911* says:
   July 6, 2011 at 7:04 am

   我看博主是学文的，也用LDA啊？

   - *luheng* says:
     July 19, 2011 at 9:51 am

     确实是学文的。只是没办法，还是要多学点科学技术啊。光扯淡是不行的。

3. *cinderella* says:
   March 23, 2012 at 11:21 am

   博主您好，我刚刚编译了GibbsLDA ，并且用它做出了model。但是在newdocs.dat.twords，文件中每个主题下面是具体字词，但是在字词后还有一个double（实数）类型的参数，这个参数是什么含义，在文档中没有说明，特此请教博主。先谢谢了

   - *Jack* says:
     March 31, 2013 at 3:37 pm

     double类型的参数，是指这个字在主题上的分布概率，也可以说重要程度

4. *Xu Liheng* says:
   March 27, 2012 at 8:53 am

   呵呵　看到楼主的博文写得太好了　让我这个专学计算机的感到汗颜那

博主你好，每次处理的文件数有限制吗，我同时导进去10000个文本就会出错，1000个就能运行

- *luheng* says:
  June 2, 2012 at 6:40 am

  我还真没试过10000个文本文件。但应该是没关系的，只要你的内存够大。

- *Huyi* says:
  October 17, 2012 at 8:32 am

  我也遇到这样的情况….可能是因为你的文档之间有空行的原因。

7. *wang* says:
   November 19, 2012 at 7:02 am

   hello
   博主，你用GibbsLDA做过中文的文档没。我试试的，貌似总是报错，说我的文件是空文件。

   - *terry* says:
     April 5, 2013 at 7:44 am

     同问，我是按照楼主的博文来配置的，发现不能处理中文文档。报错：
     no available document.
     failed to read training data.

     我的中文文档是第一行是文档数，编码为UTF-8。

     - *luheng* says:
       June 19, 2013 at 1:58 am

       这个问题中英文都可能出现。我也遇到过多次，还是编码的问题，show all symbols，确保文档的第一行没有特殊符号之类。

8. *Thomas* says:
   January 24, 2013 at 10:39 am

   楼主好厉害，学习了。

9. *[补遗]零基础小白使用LDA模型 – Huyi's* says:
   June 10, 2013 at 2:33 pm

   [...] 1.LuHeng的博客 [...]