**Weekly Report**

姓名：朱恬骅

时间： 2012-7-16

1. 上周计划及其完成情况：
   1. 工作总结

写了几个帮助数据分析的小程序，包括选择部分股票生成输入数据，LDA输入数据格式和普通矩阵输入格式的转换，TFIDF的计算工具。

我对这个模型的理解是这样的：有两组不同的话题W和Z，这两组话题间两两可能存在关联（就像一个二分图），而能够观察到的两个view中的信息（X和Y）分别是由W、Z中选中的某个话题w、z独立产生的，X和Y之间的关联是由于w和z之间有关联所致。

那么这样的话，是不是可以把w生成X的过程看作是一个普通的topic产生观察向量的过程，z产生Y的过程同样。所以在E = E[XY'] =

PAQ' 这个式子中，P、Q就是W、Z两组话题中每一个话题产生每一次词的概率所组成的一个矩阵？所以就有E = \sum{w, z} P(xi | w)P(w, z)P(yj | z)=PAQ'。

这样的话，实际P、Q可以用LDA或者类似的算法直接估计出来，而A只要是所有元素的和为1的一个矩阵，并且让E-PAQ'的每个元素的平方和最小就可以了。

由于话题模型本身的特点（把一个文档看作若干词的组合，而不管它们之间的顺序），所以用来表示有时间性的股价信息可能有点问题。目前我采取的是这样一种方法：建立一个词典“第一天跌1%”、“第一天涨1%”、“第二天跌1%”、“第二天涨1%”……，这样就可以把股价当作词频来处理了。计算E[XY']的时候要求的X、Y实际是频率而不是一个词出现的次数。这一点我之前也搞错了，所以得不到有意义的结果。

刘隽给了我到六月底的所有股价数据，不过我发现选择一段相同的时间段，里面的数据条目数是不一样的。所以先是写了一个小程序来选择一些股票某一时间段的数据（还写了一些数据预处理的小工具），如果缺失就认为涨跌幅为0。我选择的是上市时间最长的100只股票，对它们的经营范围和股价涨跌幅，各用LDA聚成六个话题，以此计算出P和Q。然后再算出了A。

* 1. 主要成果

在钢铁、银行、医药领域各选择了12只股票，选择2005年7月15日到2007年7月15日这段时间的数据：

b1-600000 b2-600015 b3-600016 b4-600035.

p1-600085 p2-600129 p3-600422 p4-000423.

s1-600019 s2-000959 s3-000709 s4-000581.

用3个话题做了LDA，在经营范围方面的结果如下：

代理 业务 政府 服务 金融债 批准 拆借 管理 监督 经营 保险业 咨询 投资 自营 技术 制造 进出口业 各类 开发 内燃机 燃油 测试 农用 用车 汽车

技术 经营 加工 保健 包装 服务 销售 开发 中成药 制造 营养液 不含 西药 零售 普通 货运 制剂 转让 咨询 机械 进出口业 种植 中药材 货物 劳务

技术 服务 经营 机械 安装 销售 开发 咨询 化工 进出口 除外 国家 钢铁 材料 制造 限定 禁止 汽车 业务 冶金 冶炼 转让 加工 管理 设计

和预期相符。

计算A矩阵，结果如下：

0.0855 0.1301 0.1177

0.0855 0.1301 0.1177

0.0855 0.1301 0.1177

其中i行j列应代表价格走势的第i个话题和经营范围的第j个话题共现的概率。从上述结果来看，与价格走势无关。在价格上用12个话题、经营范围上用3个话题，得到的A矩阵如下：

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

0.0196 0.0235 0.0391

提示价格走势无法区分这3个话题。

再改用2007年7月到2009年7月的数据，价格分6个话题，得到的A矩阵如下：

0.0197 ***0.1136*** 0.0404

0.0325 0.0744 0.0627

0.0357 0.0654 0.0641

0.0334 0.0702 0.0584

0.0370 0.0691 0.0591

***0.0444*** 0.0466 ***0.0679***

说明在这段时间内，第二个话题（制药）和其余话题（银行和钢铁）走势上有所不同，以第1种为主。

* 1. 技术报告
  2. 阅读文献

David M. Blei and Michael I. Jordan. Modeling annotated data. In Proceedings of the 26th ACM SIGIR conf., pages 127–134, New York, NY, USA, 2003

David M. Blei, Andrew Y. Ng and Michael I. Jordan. Latent Dirichlet allocation. The Journal of Machine Learning Research, pages 993-1022, Vol.3, 2003

Xuerui Wang , Andrew Mccallum: Topics over time: a non-Markov continuous-time model of topical trends, In SIGKDD ’06

1. 下一周详细安排计划：
   1. 目标

着手考虑如何完善对模型的描述；主要考虑如何更有效地表达走势（和先后次序相关的信息如何运用话题模型）。

* 1. 详细安排
  2. 阅读文献

Xuerui Wang , Andrew Mccallum: Topics over time: a non-Markov continuous-time model of topical trends, In SIGKDD ’06

1. 补充说明