7/23

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。  
  
    0.0263    0.0295    0.0141    0.0236    0.0161    0.0562  
    0.0498    0.0115    0.0460    0.0359    0.0258    0.0733  
    0.0607    0.0072    0.0026    0.0070    0.0228    0.0358  
    0.0499    0.0055    0.0080    0.0166    0.0421    0.0360  
    0.0649    0.0063    0.0109    0.0077    0.0223    0.0114  
    0.0396    0.0221    0.0208    0.0205    0.0238    0.0538  
  
经营范围上的6个话题分别是：  
开发 项目 务 和 不含 商业 与 物业 兴办 投资 另行 业 房地产 供销 医用 物资 实业 控 进出口业 申报 销售 管理  
投资 的 开发 及 房地产 项目 咨询 许可证 实业 和 工程 管理 凭 销售 涉及 规定 除 专项 建设 许可 证券 贸易 设施  
销售 和 制造 技术 机械 化工 电子 务 开发 制品 项目 原料 橡胶 电器 外汇 许可证 机构 按 许可 集装箱 咨询  
服务 销售 汽车 和 咨询 制品 管理 开发 务 剂 运输 计算机 电子 通信 软件 系统 物业 以 材料 制造 进出口  
及 技术 销售 材料 业务 国家 玻璃 务 服务 贸易 企业 本 除外 禁止 进出口 零售 管理 三来一补 机械 代理 限定 进料加工  
投资 材料 管理 开发 服务 房地产 销售 建筑 工程 物业 技术 安装 和 系统 电力 咨询 五金 租赁 电 高新技术 设施  
  
股价上的6个话题这边还没法列举。

8/13

这两天跑了一下数据。参见附件。其中，WF表示股价涨跌幅表示法，PD表示正负零收益表示法；IND表示经营范围描述全文，IDK表示经营范围描述中的关键字。  
  
股价上，似乎是最简单的用K-means效果最好，用了LDA之后不同板块的股票就混杂得多了。比如  
1(4)    600000浦发银行      600015华夏银行      600016民生银行      600036招商银行  
2(4)    600019宝钢股份      000959首钢股份      000709河北钢铁      600569安阳钢铁  
3(4)    600085同仁堂60     600129太极集团      600422昆明制药      000423东阿阿胶  
这是用Kmeans对股价涨跌幅进行聚类的结果，  
1(4)    600085同仁堂60     600129太极集团      600422昆明制药      000423东阿阿胶  
2(4)    600000浦发银行      600015华夏银行      600016民生银行      600036招商银行  
3(4)    600019宝钢股份      000959首钢股份      000709河北钢铁      600569安阳钢铁  
这是用Kmeans对正负零收益表示法进行聚类的结果。  
  
用上LDA以后效果都会变差，比如（正负零收益）  
1(7)    600000浦发银行      600015华夏银行      600016民生银行      600036招商银行      600129太极集团      600422昆明制药      600019宝钢股份  
2(4)    600085同仁堂60     000959首钢股份      000709河北钢铁      600569安阳钢铁  
3(1)    000423东阿阿胶  
（股价涨跌幅）  
1(6)    600016民生银行      600036招商银行      600019宝钢股份      000959首钢股份      000709河北钢铁      600569安阳钢铁  
2(2)    600085同仁堂60     000423东阿阿胶  
3(4)    600000浦发银行      600015华夏银行      600129太极集团      600422昆明制药  
  
  
  
经营范围描述方面，效果一直不理想。我看了一些网上别人的材料，用贝叶斯平均的方法提取了所有2215支股票经营范围描述的关键词，比如  
  
000001  监管 人民币 汇款 借款 放款 非贸易 有价证券 汇兑 信托业 外币 见证 资信 承兑 各项 存款 贴现 调查 票据 结算  
外汇 代理业 人民 保险 境内 买卖 允许 发行 有关 境外 办理  
002142  十一 十三 十二 金融债 中国银 公众 银行卡 信用证 发放 中期 款项 长期 收付 兑付 短期 吸收 债券 监督 保险业  
政府 承兑 银行 拆借 存款 担保 中国 贴现 保管 同业 票据  
600000  外汇 托管 保险箱 全国 离岸 保障 外币 借款 汇款 兑换 委员会 社会 银行业 见证 资信 中国银 拆借 结汇 股票 存款  
担保 公众 贴现 同业 信用证 发放 中期 款项 长期 收付  
600015  金融债 委员会 中国银 结汇 公众 银行卡 债券 信用证 发放 款项 中期 长期 收付 兑付 短期 政府 吸收 监督 承兑  
拆借 存款 担保 贴现 保管 同业 买卖 票据 结算 贷款 代理业  
600016  本行 十四 十一 十三 十二 可以 银行业 结汇 金融债 公众 银行卡 信用证 发放 中期 款项 长期 收付 兑付 短期 吸收  
监督 保险业 承兑 银行 拆借 债券 存款 担保 中国 政府  
这几个都是银行股  
  
000028  医用 区域性 救护车 口腔科 化验 缝合 一次性 灭菌 诊断 同化 第一 器具 激素 手术室 急救室 精神 抗生素 射线  
麻醉药 超声 敷料 附属 临床 诊疗 蛋白 毒性 疫苗 分析 消毒 合剂  
000423  膏剂 合剂 糖浆 口服液 保健 颗粒剂 胶囊 药品 批准 食品 范围 许可证 进出口业 商品 生产 销售  
002589  保存 常温 毒液 罂粟 助听器 隐形眼镜 同化 激素 体外 麻醉药 蛋白 毒性 疫苗 健身器 护理 诊断 抗生素 精神 配送  
三类 化学药 饮片 日用品 生化 试剂 中药材 制毒 生物制品 中成药 化妆品  
600085  营养液 老年病 乌鸡 作用 妇产科 儿科 梅花鹿 乌骨鸡 外科 冷食品 中医科 内科 马鹿 涂膜剂 同仁 皮肤科 供暖 定型  
皮肤 北京 诊疗 其中 股份 动植物 西药 饲养 有限公司 图书 保健 饮片  
600129  执业 中草药 旅馆 水产 西药 作业 二级 首饰 前不 副食品 保健 金银 土地 中成药 养殖 以下 工艺美术 维护 种植  
经济 印刷 不得 百货 医疗 旅游 器械 出租 自有 化学 包装  
这几个都是医药方面的  
  
但也有不好的例子，比如东华软件的这个  
002065  决定 国务院 机关 注册 选择 行政 自主 工商 登记 不得 活动 开展 批准 后方 法规 法律 审批 自营 规定 各类 限定  
许可 代理 禁止 进出口业 公司 管理 商品 除外 进出口  
完全没有体现出它经营范围是什么。  
  
用上关键词后，在12支股票的那个数据集上，除了000959首钢股份和000709河北钢铁混入了医药类以外，其他都分类正确，比用经营范围描述的全文效果要好（结果如下）  
1(1)    600085同仁堂60  
2(6)    600129太极集团      600422昆明制药      600019宝钢股份      000959首钢股份      000709河北钢铁      600569安阳钢铁  
3(5)    600000浦发银行      600015华夏银行      600016民生银行      600036招商银行      000423东阿阿胶  
  
总结报告上就先交待了一下数据集，后面的部分还没有写。

8/20

附件里的结果是两个视图用LDA，然后计算出A矩阵，再用A矩阵把一个视图上的LDA结果投影到另一个视图上去，两者叠加做K-means的结果。效果不太好，比原来单用LDA没有什么提高。  
用K-means的结果如下（在邮件结尾）  
  
  
然而用K-means，我这里先做了一个在S12-09上的结果：  
单独用股价，聚类结果是：  
     1     1     1     1     3     3     2     2     1     1     1     1  
单独用经营范围的关键字，结果是：  
     2     3     3     2     1     1     1     1     1     1     1     1  
  
我用K-means聚类的中心代替LDA里面每个词在某一topic下的概率分布，把K-means聚类的结果拉成0/1矩阵，比如股价的结果对应如下矩阵：  
     1     0     0  
     1     0     0  
     1     0     0  
     1     0     0  
     0     0     1  
     0     0     1  
     0     1     0  
     0     1     0  
     1     0     0  
     1     0     0  
     1     0     0  
     1     0     0  
  
用上面这些数据计算A矩阵，得到：  
    0.3192    0.1687    0.1632  
    0.2142         0    0.0003  
    0.1487         0         0  
  
然后把第一个视图（股价）上的聚类结果0/1矩阵，通过A矩阵投影到第二个视图（经营范围），得到  
    0.4903    1.2591    0.2506  
    0.4903    0.2591    1.2506  
    0.4903    0.2591    1.2506  
    0.4903    1.2591    0.2506  
    2.0000         0         0  
    2.0000         0         0  
    1.9985         0    0.0015  
    1.9985         0    0.0015  
    1.4903    0.2591    0.2506  
    1.4903    0.2591    0.2506  
    1.4903    0.2591    0.2506  
    1.4903    0.2591    0.2506  
  
最后把它做K-means聚类，得到的结果是  
     2     2     2     2     1     1     1     1     3     3     3     3  
  
确实比单独用K-means做的结果要好些。  
在其他几个数据集上跑出来的结果如下：  
  
  
SET: S12-09  
1(4)    600019宝钢股份      000959首钢股份      000709河北钢铁      600569安阳钢铁  
2(4)    600085同仁堂60     600129太极集团      600422昆明制药      000423东阿阿胶  
3(4)    600000浦发银行      600015华夏银行      600016民生银行      600036招商银行  
  
SET: S50-09  
1(1)    600559老白干酒  
2(20)   000709河北钢铁      000717韶钢松山      000898鞍钢股份      000932华菱钢铁      000959首钢股份      600019宝钢股份      600022济南钢铁      600282南钢股份      600569安阳钢铁      000423东阿阿胶      600422昆明制药      000568泸州老窖      000596古井贡酒      000799酒鬼酒       000858五粮液       002304洋河股份      600519贵州茅台      600702沱牌舍得      600779水井坊       600809山西汾酒      （酒＋钢铁）  
3(8)    000938紫光股份      000948南天信息      002065东华软件      002073软控股份      002090金智科技      002230科大讯飞      600271航天信息      600570恒生电子      （软件）  
4(11)   000629攀钢钒钛      000028一致药业      002589瑞康医药      600085同仁堂60     600129太极集团      600511国药股份      600713南京医药      600833第一医药      601607上海医药      002063远光软件      600756浪潮软件      （一个钢铁，两个软件，八个医药）  
5(10)   000001深发展A0     002142宁波银行      600000浦发银行      600015华夏银行      600016民生银行      600036招商银行      601009南京银行      601166兴业银行      601169北京银行      601288农业银行      （银行）  
  
  
SET: S50-10  
1(6)    000596古井贡酒      600519贵州茅台      600559老白干酒      600809山西汾酒      002063远光软件      600756浪潮软件  
2(15)   000629攀钢钒钛      000717韶钢松山      000959首钢股份      600019宝钢股份      600022济南钢铁      600282南钢股份      000423东阿阿胶      600422昆明制药      000568泸州老窖      002304洋河股份      600702沱牌舍得      600779水井坊       002065东华软件      002073软控股份      002230科大讯飞  
3(15)   000001深发展A0     601288农业银行      000709河北钢铁      000898鞍钢股份      000932华菱钢铁      600569安阳钢铁      600085同仁堂60     600129太极集团      000799酒鬼酒       000858五粮液       000938紫光股份      000948南天信息      002090金智科技      600271航天信息      600570恒生电子  
4(8)    002142宁波银行      600000浦发银行      600015华夏银行      600016民生银行      600036招商银行      601009南京银行      601166兴业银行      601169北京银行  
5(6)    000028一致药业      002589瑞康医药      600511国药股份      600713南京医药      600833第一医药      601607上海医药  
  
  
剩下的几个数据集上没有跑出结果来，可能是程序的问题，我还要调一下。