实验结果说明

1 数据来源

① 中文：谭松波中文文本分类语料，不仅包含大的分类，例如经济、运动等等，每个大类下面还包含具体的小类，例如运动包含篮球、足球等等。能够作为层次分类的语料库，非常实用。

② 英文：20 Newsgroups，这个数据集由近 20000 个新闻文档组成，相应分为 20 个不同的新闻组。每一个组表示一个主题。如下所示

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

2 实验结果

① 使用谭松波中文文本分类语料中的 C5，C7，C11，C19，C31，C39 类（44MB）. 如图 1 左 所示，参数设置为如图 1 右 (10 个主题)

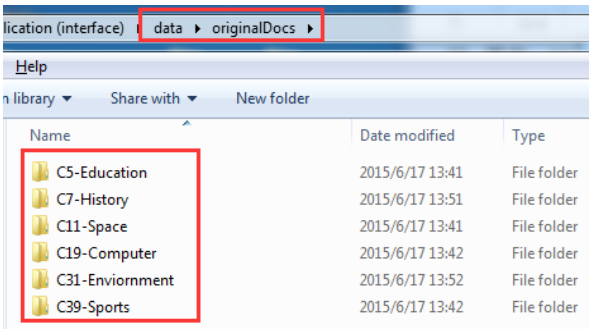

图 1 （左） 待分析文档

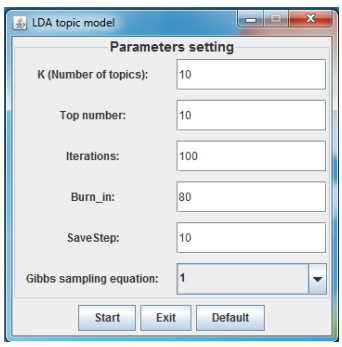

图 1 （右） 模型参数设置

值得注意的是，程序会输出四个结果文件，存放在程序 lda.jar 当前路径 data/result 目录下。其中 phi 是主题-词项分布，K*V 维的，每一项表示该词项 v 在主题 k 的概率是多少。final 就是根据概率从高到低选择 top number 个。theta 是文档-主题分布，M*K 维的，每一项表示主题 k 在文档 m 中的概率是多少。topic_assignment 文件表示词项被分配给了哪个主题。如图 2 右端所示，‘学报’在字典索引中为 1，最终被分配到主题 4。‘辐射’索引为 11，被分到主题 2.

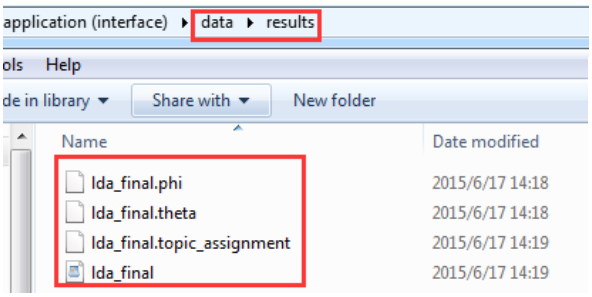

图 2（左） 程序结果输出

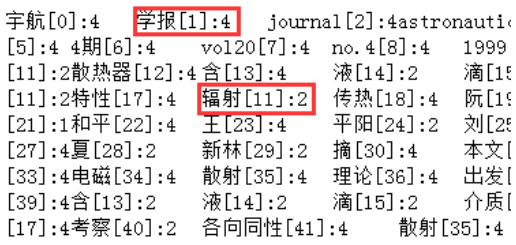

图 2（右）主题分配文件示例

```
topic 1 : 环境      0.026742319596622294 发展      0.013388740515237743 经济      0.009678257523321045 国家
          0.009344712260397084 美国      0.00905479830735323 我国      0.0084108948160832 计划      0.007794463776729729
          企业      0.007767276788679004 管理      0.00764994984576046 生产      0.007469110089299227 ,
topic 2 : 土壤      0.012830672701760481 污染      0.008871840256789183 影响      0.008684715964291349 环境
          0.008545414632246708 研究      0.00763018102952867 3 含量      0.0075068169286898155 水      0.007447660876234477
          表      0.006970810672462687 浓度      0.00646994688276452 2 变化      0.006380303203745125 ,
topic 3 : 系统      0.041474329485106015 控制      0.02884407010211216 式      0.01798606329140968 8 模型
          0.01647802246939623 2 参数      0.013629908921826837 方法      0.013493927415777724 问题      0.009633356420463892
          函数      0.008554577267832108 矩阵      0.00836604352877462 输出      0.008024062673772953 ,
topic 4 : 体育      0.06942982695020028 运动      0.02765755651104207 运动员      0.01809450702386 46 训练
          0.016900918749933144 比赛      0.010530073292909767 研究      0.010175894668589735 技术      0.009158224949341442
          动作      0.008660109525267135 成绩      0.008174550225053992 竞技      0.007266252379389233 5 ,
topic 5 : 浓度      0.014607748339094222 处理      0.01089762461218768 反应      0.01023042101295972 吸附
          0.01008914607564243 研究      0.009420620387377577 实验      0.009283487515112297 图      0.009071064748567632
          ph      0.008728979709155111 值      0.008563230021855887 影响      0.008122740098206365 ,
topic 6 : 历史      0.019017808102192953 中国      0.011770979974536501 文化      0.008089212360412644 这种
          0.005659524060513741 发展      0.005350126463366502 5 一种      0.005306454633275591 人      0.005300937127852569
          说      0.005206516089216163 5 研究      0.004970820028298871 文学      0.004946301141066025 ,
topic 7 : 图      0.014285153073568058 试验      0.010785313965165228 测量      0.009844663128475697 采用
          0.007876671485708798 分析      0.007487648159703287 4 温度      0.007205026854077005 影响      0.00698470008371139
          设计      0.006951240127769088 计算      0.00682663273940191 进行      0.006822149326083257 ,
topic 8 : 算法      0.02061861105404231 方法      0.014473287916650443 图      0.012002038006111288 表示
          0.01079552818622776 进行      0.010388579509772275 计算      0.009300099344261896 模型      0.008709002202802933
          特征      0.008525454028697779 问题      0.008325428037745827 研究      0.008060095334590463 ,
topic 9 : 系统      0.027900909419016393 数据      0.01901083654403899 对象      0.016751282312257386 用户
          0.014284272733040428 信息      0.014007920675900652 实现      0.011971367109608912 网络      0.011470703610060516
          应用      0.01057528835068246 进行      0.010258256869834401 数据库      0.009920730725578116 ,
topic 10 : 教育      0.03687064227507069 学生      0.026167583631883665 发展      0.017514426585996234 学校
          0.01401118528188915 教学      0.013614174238968992 课程      0.010919166221169486 活动      0.010188124629554924
          社会      0.009969552291169165 学习      0.00979749739736388 教师      0.009213079409736976 ,
```

图 3 对应实验结果

程序跑了 401s，结果还是相当不错的，主题 1，2，5，7 对应 C31 环境类别，主题 3，8，9 对应 C19 计算机类别，主题 4 对应 C3 运动类别，主题 6 对应 C7 历史类别，主题 10 对应 C5 教育类别。

② 使用 20 Newsgroups 中的 comp.sys.mac.hardware，sci.electronics，sci.space，soc.religion.christian，talk.politics.guns 文档，剔除了其中小于 3KB 的文档，因为这是新闻文档，有固定的格式组成，如果文档很小，说明用户写的信息就小，会造成不小误差。所有剔除小文件后一共有 6.3MB，参数设置如图 1 右，程序跑了 48s. 结果如下。

topic 1 : space 0.030935821230434774 nasa 0.01133880823263434 launch 0.010744730488906229 sci.space 0.008963668598357424 1993 0.008030431242574073 shuttle 0.007522159387261111 data 0.0073145229511823555 technology 0.006353874689215802 program 0.0061931523372066016 mission 0.005929182838335432

topic 2 : god 0.048175524076195415 jesus 0.019701904881472942 christ 0.01491926000028172 church 0.013690123292204353 christian 0.011344651304621887 soc.religion.christian 0.010683393740160547 faith 0.008746017914417375 christian@aramis.rutgers.edu 0.008556819644593473 approved 0.00816923133681365 lord 0.007878253687109846

topic 3 : christian 0.017332118900049317 people 0.016820688126334818 paul 0.015991208384113748 christians 0.010614898585520425 law 0.010511320578648467 god 0.008995591933240697 love 0.008524417825740205 homosexuality 0.008281791842919621 bible 0.007275376300345255 soc.religion.christian 0.007023785729194903

topic 4 : earth 0.014120443313500588 moon 0.008683910576631848 sci.space 0.008075274274181917 space 0.008024780831001752 system 0.007317486044702594 sci.astro 0.00642277709056399 planet 0.006151049972989058 sun 0.0060731939893017214 sky 0.006037519301351466 solar 0.005830882639509193

topic 5 : cantaloupe.srv.cs.cmu.edu 0.030178207277853426 writes 0.030137903481233165 article 0.025931259630234238 apr 0.024904110198109064 lines 0.023874880557803076 organization 0.02304302613798489 path 0.02286115615834792 message-id 0.022319751128375663 newsgroups 0.0221162978615891 gmt 0.020357955988217057

topic 6 : people 0.026392521374996966 don 0.0242634762388851 writes 0.01219510273420791 time 0.012030977746240 84 question 0.011783796722373382 true 0.01172484994837094 reason 0.011505196467886555 truth 0.010707969393880558 subject 0.009256430087120196 article 0.009076133101621189

topic 7 : system 0.011359853941941857 software 0.010630011950477638 mac 0.010416057670371463 apple 0.010019722714361006 copy 0.008197546401309317 drive 0.007962282012979025 comp.sys.mac.hardware 0.006420311514898292 disk 0.006184331735148778 computer 0.006003799638801814 93 0.0054302441881297105

topic 8 : gun 0.02513849258439913 guns 0.012378760736155583 fbi 0.011601430310831625 talk.politics.guns 0.011492336463607459 people 0.010939073856604446 control 0.009086258291584461 weapons 0.008977421762040748 fire 0.008584084606122427 firearms 0.008175194740579412 file 0.00769066029129228

topic 9 : power 0.014774416980544146 subject 0.011755187037298753 sci.electronics 0.009106644130926236 ground 0.008587059198604896 lines 0.006300347502750337 wire 0.006099926454830878 current 0.006087829923284481 path 0.005987914278215276 circuit 0.005566516332369287 organization 0.005541263622892747

topic 10 : people 0.0097793302266669 rights 0.009104066471616703 militia 0.008865979413107595 government 0.007754495326578285 amendment 0.007452470089965696 1993 0.0069444092006685325 free 0.006927363272975532 political 0.0060320928535322815 bear 0.005309922946306203 arms 0.004790262685748265

图 4　对应实验结果

主题 1，4 对应 sci.space，表示太空类，主题 2，3 对应 soc.religion.christian，表示宗教类，主题 7 对应 comp.sys.mac.hardware，表示计算机类，主题 8，10 对应 talk.politics.guns，表示社会政治类，主题 9 对应 sci.electronics，表示电子类。