LDA主题模型介绍

<https://www.jianshu.com/p/fa97454c9ffd>

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层[贝叶斯](https://links.jianshu.com/go?to=https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF" \t "_blank)概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

****1 LDA的目的就是要识别主题，即把文档—词汇矩阵变成****

****文档—主题矩阵（分布）****

****主题—词汇矩阵（分布）****

语料库中的每一篇文档与T个主题的一个多项分布相对应，将该多项分布记为θ

即模型对该文档给出的分属各个主题的概率分布，即我们生成的文本衍生特征。

每个主题又与词汇表（vocabulary）中的V个单词的一个多项分布相对应，将这个多项分布记为φ。即该主题的词汇描述

Tp1 (1, '0.068\*"手机" + 0.037\*"贷" + 0.036\*"我" + 0.036\*"借款" + 0.035\*"腾讯" + 0.034\*"百度" + 0.033\*"视频" + 0.032\*"你" + 0.031\*"支付宝" + 0.030\*"微信"')

2 LDA整体流程

先定义一些字母的含义：文档集合D，主题（topic)集合T

D中每个文档d看作一个单词序列<w1,w2,...,wn>，wi表示第i个单词，设d有n个单词。（LDA里面称之为wordbag，实际上每个单词的出现位置对LDA算法无影响，也就是说LDA没有引入上下文信息）

D中涉及的所有不同单词组成一个大词袋VOC，LDA以文档集合D作为输入，希望训练出的两个结果向量然后聚类（设聚成k个topic，VOC中共包含m个词）

对每个D中的文档d，对应到不同Topic的概率θd<pt1,...,ptk>，其中，pti表示d对应T中第i个topic的概率。计算方法是直观的，pti=nti/n，其中nti表示d中对应第i个topic的词的数目，n是d中所有词的总数。

因每个主题还是由词构成（一般10个词来描述这个主题），故的nti是计算了该文档中的词在该主题中命中了几个，n是该文档词的数目

Tp1 (1, '0.068\*"手机" + 0.037\*"贷" + 0.036\*"我" + 0.036\*"借款" + 0.035\*"腾讯" + 0.034\*"百度" + 0.033\*"视频" + 0.032\*"你" + 0.031\*"支付宝" + 0.030\*"微信"')

对每个T中的topict，生成不同单词的概率φt<pw1,...,pwm>，其中，pwi表示t生成VOC中第i个单词的概率。计算方法同样很直观，pwi=Nwi/N，其中Nwi表示对应到topict的VOC中第i个单词的数目，N表示所有对应到topic的单词总数。

LDA的核心公式如下：

p(w|d)=p(w|t)\*p(t|d)

直观的看这个公式，就是以Topic作为中间层，可以通过当前的θd和φt给出了文档d中出现单词w的概率。

其中p(t|d)利用θd计算得到， p(topic|docu) = θd

p(w|t)利用φt计算得到。 p(word|topic) = φt

实际上，利用当前的θd和φt，我们可以为一个文档中的一个单词计算它对应任意一个Topic时的p(w|d)，然后根据这些结果来更新这个词应该对应的topic。然后，如果这个更新改变了这个单词所对应的Topic，就会反过来影响θd和φt。

LDA主题模型在短文本上的应用

数据：地址文本

目标：通过主题一致性，困惑度等指标找到，最佳语料以及最佳主题数的边界，通过程序自动对最优主题寻优

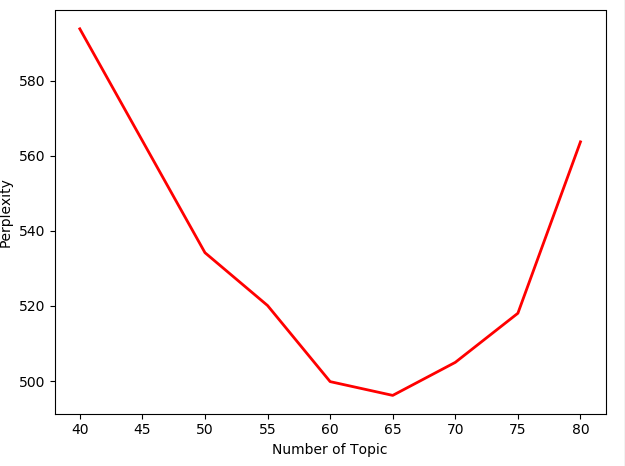
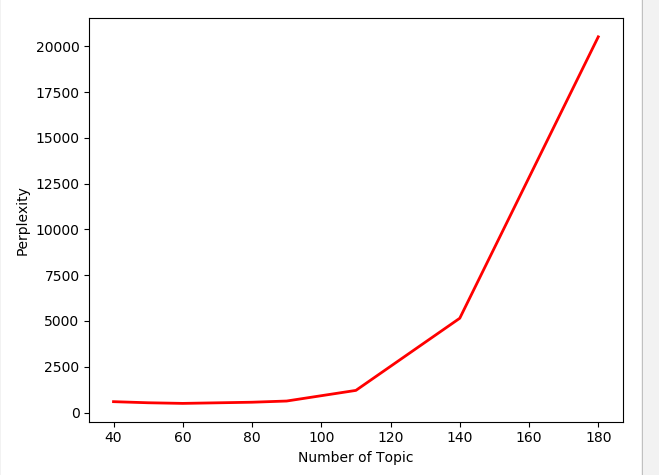
困惑度：困惑度可以理解为对于一篇文本，所训练出来的模型对文本属于哪个主题有多不确定，这个不确定程度就是困惑度。困惑度越低，说明聚类的效果越好。

实践1 3410: 同信云盾数据 文本数80860 dict len 36014 DT len 80860

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 一致性 | 困惑度 | 测试集AUC |  |
| Tp40 | -12.073197 | 593.87015 |  |  |
| Tp50 |  | 534.20089 | 0.783270 |  |
| Tp60 | -13.031725 | 499.83862 |  |  |
| Tp80 |  | 563.74753 | 0.779730 |  |
| Tp90 | -13.021619 | 628.83943 |  |  |
| Tp110 |  | 1209.44703 | 0.747014 |  |
| Tp140 | -16.125137 | 5149.64382 |  |  |
| Tp180 | -16.647381 | 20515.11013 |  |  |
| Tp250 | -16.673315 | 24071.90215 |  |  |
| Tp300 | -16.673315 |  |  |  |
|  |  |  |  |  |

通过困惑度缩小范围

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 一致性 | 困惑度 |  |  |
| Tp55 | -12.073197 | 520.11836 |  |  |
| Tp65 | -12.073197 | 496.16712 | 0.795351 |  |
| Tp70 | -12.073197 | 504.97288 |  |  |
| Tp75 | -12.073197 | 518.06380 |  |  |
| Tp90 | -13.021619 | 628.83943 |  |  |



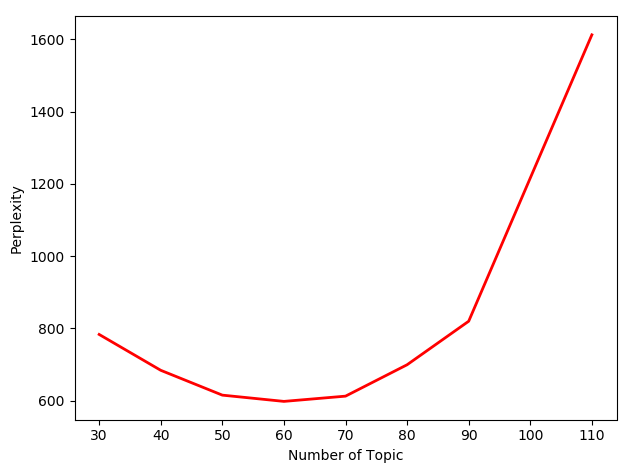
实践2 3410: 同信云盾数据 文本数172349 dict len 43739 DT len 172349

我们可以看到虽然文本数增加了不止一倍，但是词袋变化3.6--4.3，说明新增新词并不多，词袋再慢慢的变得饱和，随着文本的增加，很多词重复出现

如果模型可以随着文档的变化而自动的调整最佳的主题数，这样是最好的，因为虽然我们有困惑度可以确定最佳主题数，但是我们还没有找到最佳训练文档数，

其他大佬，融入了分层狄利克雷过程（HDP），构成一种非参数主题模型，好处就是不需要预先指定个数，模型可以随着文档的变化而自动的对主题个数进行调整，这个后面需要去关注一下

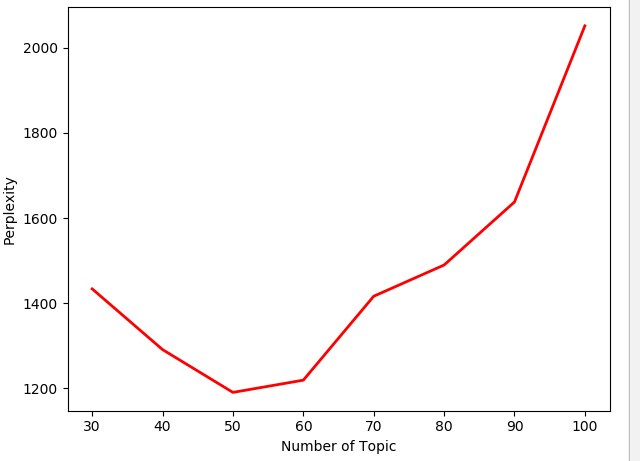
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 一致性 | 困惑度 | 测试集AUC |  |
| Tp30 | -11.346477 | 783.48359 |  |  |
| Tp40 | -12.376871 | 684.26496 |  |  |
| Tp50 |  | 615.55030 | 0.767938 |  |
| Tp60 | -13.702856 | 598.24407 | 0.770681 |  |
| Tp70 | -13.500750 | 612.76128 |  |  |
| Tp80 |  | 699.65867 | 0.757691 |  |
| Tp90 | -13.021619 | 820.04785 |  |  |
| Tp110 | -13.121567 | 1612.48937 |  |  |



实践3 3月份pdl 文本数6.7w dict len 44017 DT len 56879

文本特点：不同机构的地址数据准确度不同，有的机构地址数据杂乱无章

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 一致性 | 困惑度 | AUC |  |
| Tp30 | -12.653218 | 1433.79935 |  |  |
| Tp40 | -13.887622 | 1291.30237 |  |  |
| Tp50 |  | 1190.82393 | 0.639619 |  |
| Tp60 | -14.462109 | 1219.53827 |  |  |
| Tp70 |  | 1416.46896 |  |  |
| Tp80 |  | 1489.76287 |  |  |
| Tp90 | -12.877761 | 1637.69123 |  |  |
| Tp100 |  | 1612.48937 |  |  |



秒啦首贷app数据LDA主题建模评估报告

训练文本数据来源：

       1，秒啦首贷用户9W app数据

       2，金盘用户50Wapp数据

特征评估数据源： 秒啦首贷用户9W app数据

初步结论：用户app数据呈现的文本主题对于用户有一定的区分度，从目前的结果来看，在app文本环境下，在一定范围内，训练语料越充足，LDA子模型的预测效果可能越好

建模过程：

       1，基于用户app数据进行分词形成语料  
       2，利用全部语料训练LDA模型

       3， 通过预指定主题数迭代并保存模型

       4， 借鉴远江在淘宝数据上的应用，对用户app数据进行主题预测，利用预测的主题概率分布作为用户特征

评估结果

       1，基于秒啦首贷app数据，词袋数据8W，从[10,400]的区间取值作为预选主题数，利用lightGB评估特征

|  |  |  |  |
| --- | --- | --- | --- |
| Topic30 | Pd10 | AUC | KS |
| Train | 0.070177 | 0.719236 | 0.363084 |
| test | 0.07121 | 0.591022 | 0.147099 |
| oot | 0.072022 | 0.580061 | 0.125170 |

|  |  |  |  |
| --- | --- | --- | --- |
| Topic60 | Pd10 | AUC | KS |
| Train | 0.06996 | 0.818112 | 0.523493 |
| test | 0.07431 | 0.592581 | 0.135697 |
| oot | 0.07274 | 0.580091 | 0.122924 |

|  |  |  |  |
| --- | --- | --- | --- |
| Topic90 | Pd10 | AUC | KS |
| Train | 0.069845 | 0.837805 | 0.562507 |
| test | 0.07201 | 0.600809 | 0.15251 |
| oot | 0.073135 | 0.594774 | 0.140387 |

|  |  |  |  |
| --- | --- | --- | --- |
| Topic150 | Pd10 | AUC | KS |
| Train | 0.070408 | 0.797205 | 0.463967 |
| test | 0.06908 | 0.601394 | 0.144803 |
| oot | 0.07124 | 0.589584 | 0.134825 |

|  |  |  |  |
| --- | --- | --- | --- |
| Topic250 | Pd10 | AUC | KS |
| Train | 0.069932 | 0.837664 | 0.530586 |
| test | 0.0698 | 0.616040 | 0.176073 |
| oot | 0.07284 | 0.600602 | 0.156731 |

|  |  |  |  |
| --- | --- | --- | --- |
| Topic300 | Pd10 | AUC | KS |
| Train | 0.07097 | 0.811886 | 0.510554 |
| test | 0.06890 | 0.600527 | 0.146083 |
| oot | 0.069362 | 0.604660 | 0.146901 |

|  |  |  |  |
| --- | --- | --- | --- |
| Topic350 | Pd10 | AUC | KS |
| Train | 0.071015 | 0.848392 | 0.547280 |
| test | 0.06930 | 0.603552 | 0.149974 |
| oot | 0.069217 | 0.598420 | 0.139162 |

**预测峰值对应的主题数范围在250--350之间，AUC表现稳定在0.6左右**

2， 基于金盘用户app数据，词袋数据15W，从[10,400]的区间取值作为预选主题数，利用lightGB评估特征

|  |  |  |  |
| --- | --- | --- | --- |
| Topic 10 | Pd10 | AUC | KS |
| train | 0.07020661 | 0.779322 | 0.458662 |
| test |  | 0.589421 | 0.458662 |
| oot | 0.071926 | 0.574794 | 0.114753 |
| Topic 20 | Pd10 | AUC | KS |
| train | 0.06942 | 0.845051 | 0.570215 |
| test |  | 0.584677 | 0.128859 |
| oot | 0.074538 | 0.599533 | 0.128859 |
| Topic 45 | Pd10 | AUC | KS |
| train | 0.070582 | 0.770401 | 0.453158 |
| test |  | 0.609315 | 0.178152 |
| oot | 0.0706684 | 0.609315 | 0.154927 |
| Topic 60 | Pd10 | AUC | KS |
| train | 0.0702644 | 0.779437 | 0.465035 |
| test |  | 0.618220 | 0.182083 |
| oot | 0.071732 | 0.601385 | 0.150762 |
| Topic 70 | Pd10 | AUC | KS |
| train | 0.0712324 | 0.837654 |  |
| test |  | 0.626203 |  |
| oot | 0.0684918 | 0.617197 |  |
| Topic 80 | Pd10 | AUC | KS |
| train | 0.0706689 | 0.832868 |  |
| test |  | 0.611443 |  |
| oot | 0.070378 | 0.614228 |  |
| Topic 90 | Pd10 | AUC | KS |
| train | 0.070510 | 0.825880 | 0.526645 |
| test |  | 0.623023 | 0.184657 |
| oot | 0.070910 | 0.617809 | 0.178808 |
| Topic 120 | Pd10 | AUC | KS |
| train | 0.070625 | 0.840934 | 0.542027 |
| test |  | 0.616560 | 0.180450 |
| oot | 0.0705233 | 0.601049 | 0.148882 |
| Topic 180 | Pd10 | AUC | KS |
| train | 0.07074 | 0.840517 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| test |  | 0.599955 |  |
| oot | 0.070136 | 0.605056 |  |
| Topic 300 | Pd10 | AUC | KS |
| train | 0.07007 | 0.821268 | 0.490411 |
| test |  | 0.599088 | 0.149712 |
| oot | 0.072361 | 0.603842 | 0.162424 |

**对比秒啦语料，从整体看模型的效果得到了提升，验证测试集表现相对稳定，app语料的丰富对模型的预测效果有一定的提升作用，**

**当主题数[70,90]的范围内，测试集上的AUC的峰值均曾达到0.62**

选用topic\_num=70作为主题数：

        columns  feature\_importances

        tp68             0.040525

        tp1               0.039692

        tp33             0.030923

        tp18             0.030038

        tp19             0.027367

        tp68

        (68, '0.057\*"平安" + 0.043\*"银行" + 0.041\*"生活" + 0.038\*"信用卡" + 0.037\*"管家" + 0.027\*"吧" + 0.026\*"买单" + 0.023\*"口袋" + 0.020\*"金融" + 0.019\*"掌上"')

        tp1

        (1, '0.068\*"手机" + 0.037\*"贷" + 0.036\*"我" + 0.036\*"借款" + 0.035\*"腾讯" + 0.034\*"百度" + 0.033\*"视频" + 0.032\*"你" + 0.031\*"支付宝" + 0.030\*"微信"')

        tp33

        (33, '0.102\*"贷" + 0.101\*"借款" + 0.044\*"你" + 0.040\*"我" + 0.040\*"拍拍" + 0.031\*"金融" + 0.030\*"借钱" + 0.030\*"贷款" + 0.027\*"分期" + 0.027\*"支付宝"')

        tp18

        (18, '0.074\*"钱包" + 0.031\*"花" + 0.024\*"贷" + 0.022\*"秒" + 0.021\*"借" + 0.020\*"花花" + 0.019\*"钱" + 0.018\*"贷款" + 0.015\*"借钱" + 0.015\*"分期"')

        tp19

        (19, '0.070\*"哔哩" + 0.039\*"视频" + 0.036\*"音乐" + 0.035\*"淘宝网" + 0.035\*"易云" + 0.030\*"求生" + 0.029\*"绝地" + 0.021\*"王者" + 0.021\*"荣耀" + 0.019\*"抖音"')

        从排名靠前的主题来看，基本都刻画了一个借贷相关的主题

  后续工作：

        1，融合金盘app数据，构造全量的app语料库训练主题模型，再一次进行评估。

        2， 就app数据本身还可以衍生出很多变量，比如指标的映射，通过app分类挖掘不同用户app潜在的关联模式等。