中文作家写作风格识别

15307130194

冯梓源

zyfeng15@fudan.edu.cn

摘要

在文学领域，作家的写作风格是一个相对抽象的概念。在中文领域，怎样让计算机理解并识别作家的写作风格，是一个极具挑战性的命题。本文试图在几篇经典论文的思路上做进一步探索，充分利用课上所学的NLP相关概念与方法，结合以FudanNLP、NLTK为代表的高性能工具包，将中文作家的写作风格以合适的方式编码，借助机器学习甚至深度学习的技术，在一定问题范围内训练分类器，有效解决中文作家写作风格识别的问题。对于新的文本，分类器能够以令人满意的准确率识别可能的作者。

关键词：写作风格、句法依存树、功能词、LSTM

问题背景

古今中外，著名作家的优秀作品总是被人们口耳传颂，除了作品本身的主题以外，作家独具一格的遣词造句能力，也是优秀作品的不可或缺的要素。跟其他任何艺术创作相同的是，作家在创作过程中必然会在作品中留下自己的文字风格。如何定义并解读作家的写作风格，无论对文学界还是对语言学界都是极富价值的研究命题。

从经验上看，要区别甚至识别作家的行文风格是很困难的。写作风格实际上是一种个人的行为方式，作家在创作过程中会不知不觉地将其个性和个人社会背景融入或体现于作品中 (胡壮麟, 2000)。

问题分析

对于写作风格(writing style)，采用 (Sebranek, Kemper, & Meyer, 2006)的定义：writing style is the choice of [words](https://en.wikipedia.org/wiki/Word), [sentence](https://en.wikipedia.org/wiki/Sentence_(linguistics)) structure, and [paragraph](https://en.wikipedia.org/wiki/Paragraph) structure, used to convey the meaning effectively.

根据该定义，词语、句子和段落结构的信息能够表达作家的写作风格。接下来会围绕这三方面设计和计算特征。值得注意的是，在这种定义下的写作风格是与主题无关的(subject independent)，即作品的主题以及内容相关信息不应该作为写作风格的衡量。

在一定条件下，识别某位作家的写作风格的问题，就转化为特征选取和多元分类的问题。

问题的难点在于两方面。一、怎样在词、句、段落层面上构建合适的特征。二、怎样选择合适的分类器。

****数据收集****

从五六文学网([www.56wen.com](http://www.56wen.com))下载一批中文作家的作品，以txt形式保存。

选取的作家包括：鲁迅、周作人、林语堂、三毛、刘慈欣、王小波、史铁生

每个作家采用收集至少20万字的文本

其中70%作为训练集，10%作为交叉验证集，20%作为测试集

预处理

1．将文本编码为UTF-8

2．使用正则表达式匹配，清除文本中的广告。

3．手动清除非作家本人编写的内容（如编者写的序言、注释）。

4．划分句子：以现代汉语句子的概念为标准，寻找句号、叹号、问号、冒号、分号、省略号，以此划分每一个句子（特殊情况另外处理），使得每个句子单独占据一行。

5．分词和词性标注：使用jieba分词的Python包，对每个句子分词，分词的结果附带词性标注，词性标注与“NLPIR汉语分词系统”兼容。

特征选择

一般来说，从文本中提取特征会考虑以下几个方面：

**词汇特征**是基于字符和词语的特征，主要包括词性、词汇丰富度和高频词等 (Abbasi A, 2005)。词汇特征在传统英文文学作品作者识别中效果较好；但是由于作品词汇的选择与主题高度相关，词汇特征在跨主题的文本风格分析中效果会受影响。

**语法特征**指的是功能词、标点符号和Ngram等。有研究 (Zhao Y, 2007)表明，在英文语料中功能词能够有效表征作家在写作时的个人语言习惯，对文本识别类型的问题有一定帮助。不同于英语、法语等黏着语，在动词上添加词缀来增强表意能力，汉语属于孤立语（又称分析语），汉语的虚词承担了大部分的语法任务和语义功能 (黄进, 2006)。虚词主要包括：介词、副词、助词、叹词、连词等。

**结构特征**指的是与文本组织和布局相关的特征，如段落数目、段落长度、字体、字号等。显然，这类特征受编辑者的影响很大，同一个作家的作品在不同出版商的手上可能会有不同的字体、字号、缩进等，而且文本的段落设计往往由根据文章的主题需求而定（例如议论文的分段往往带有内容上的逻辑变化，而涉及人物的小说仅仅用分段来表示对话）。所以这类特征在写作风格识别的问题中很难起作用。

**语义特征**是与内容相关的特征，一般来说写作风格识别不应该与内容相关，但是有研究表明，在短文本的文体风格中引入内容相关特征，能让模型表现更好，但是该结论缺乏进一步的验证，在长文本中的表现也未知。

综上所述，选取的文本特征如下：

1. 词性比例——不同词性的词的数量与总词数之比
2. 词汇丰富度——不同的词数与总词数之比
3. 单现词比例——只出现一次的词的数目与总词数之比
4. 高频虚词（功能词）——介词、连词、叹词、结构助词、语助词、方位词的前若干个高频项，
5. 平均句长——所有句子所含字数（包括标点）的算术平均
6. 短句比例——短于平均句长一半的句子数量与总句数之比
7. 长句比例——长于平均句长两倍的句子数量与总句数之比
8. 疑问句比例——以问号结尾的句子与总句数之比
9. 感叹句比例——以感叹号结尾的句子与总句数之比

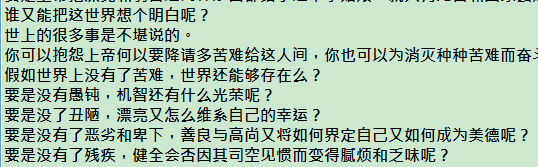
10．句法依存关系——使用FudanNLP计算每个句子的句法依存关系。

数据分析

列表分析选取的各种特征在不同作家的文本上的表现：

一、疑问句和感叹句比例

上图统计的是各个作家使用的疑问句和感叹句的比例。可以发现，史铁生使用疑问句的比例远比其他作家高。熟悉史铁生作品的人应该知道，他的散文无论是偏议论还是偏叙事，很多时候会用平静的问句来推动人物的对话或思考的逻辑。例如：

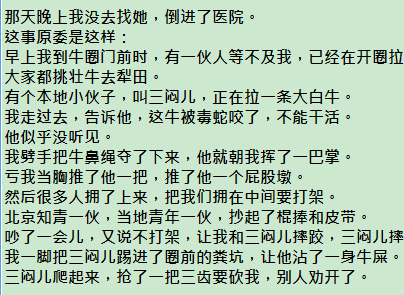


对问句的高频使用可以说是史铁生作品的一大风格。与之相对的，史铁生对感叹句的使用非常少，可以理解为他的作品极少抒发强烈的情感，更多的是传达平淡而隽永的人生哲理。

二、 平均句长

以上是各作家作品的平均句长。显然周作人的句子较长，王小波的句子较短。然而这只是一种很粗糙的统计，算法使用句号等标点符号显式切分句子，但是中文文学的句子会以逗号表示意义的承接，用句号表示逻辑的结束。所以这里的一个句子可能包含多个顺次承接的短句。这个特点在周作人的作品中表现更明显。

王小波的作品擅长以短句嬉笑怒骂。例如：



三、长短句比例

以下是长短句比例的信息。除了王小波，其他作家的长短句比例大致相似。

四、 词汇丰富度与单现词比例

词汇丰富度和单现词比例信息如图。刘慈欣的作品在这两个指标上都比较低，这也是“硬科幻”作家被诟病的原因之一——过于注重技术细节和人文伦理的讨论，而淡化人物塑造和情景描写。因为人物面谱化严重，对人物的描写上用词重复就会导致新鲜的词汇较少。与此相对，林语堂的文学语言中词汇较丰富，艺术味道较浓。

五、词性比例

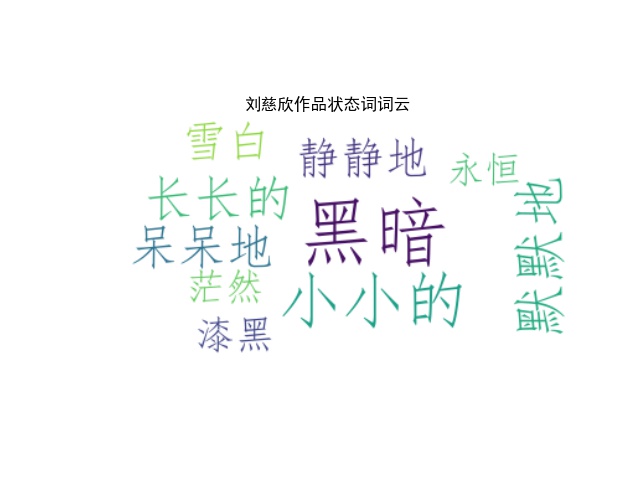
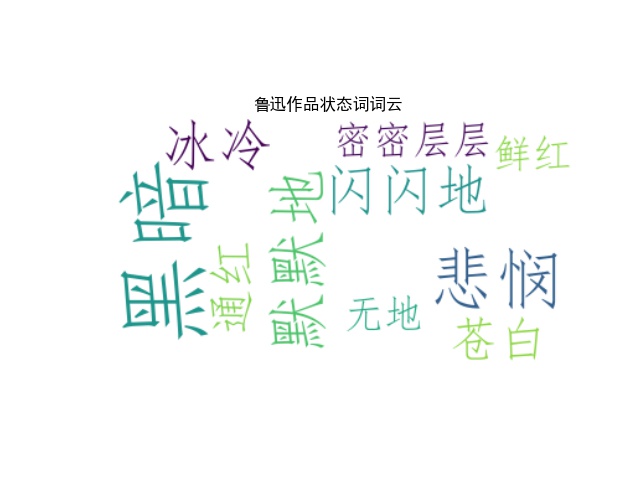
以下饼状图显示的是鲁迅和林语堂作品的词性比例。实际上中国现代作家的词性比例是相差不大的，原因是白话文普及之后，汉语的句法基本没有太大的变化。名词、动词和非语素字分别占约五分之一，助词、副词和代词分别占略低于10%，剩下各种词性比例少许。

至于作家之间的词性比例差异可以参考下面的三维图。可以看出，基本上不同作家的用词比例差异不大，除了几个比较明显的峰值：刘慈欣、周作人和林语堂使用名词的比例略高，王小波、三毛、史铁生使用动词的比例略高。这种差异比较细微，对于风格的解释意义不大。

六、高频功能词

高频功能词能够反映作家的写作习惯，使描述作家风格的有力特征。通过词云可以很好地可视化不同作家的高频功能词的差异。以下选取鲁迅、刘慈欣和周作人的副词、助词和状态词的词云分析：



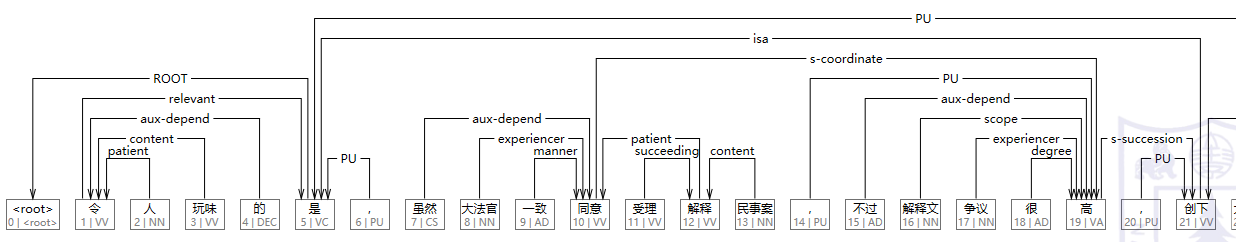


词云中字的大小代表频率。字越大出现频率越高。观察副词和助词的词云，可以发现被词云选中的高频词基本上是固定的几个，但是它们所构成的频率分布有显著的差异。例如，对于副词，鲁迅用“不”、“也”、“就”比较多，而刘慈欣用“都”、“也”、“就”比较多，周作人用“也”、“不”、“又”、“都”、“很”比较多。对于助词，鲁迅使用“等”的频率相对比周作人多，而使用“的话”的比例明显比刘慈欣少。

有趣的是，标注系统把“黑暗”认为是状态词，而且都在三人的作品中较多出现。但除了“黑暗”，鲁迅使用的“悲悯”、“冰冷”，周作人使用的“般若”、“隐逸”、“匆匆”，都能在一定程度上说明他们各自作品的风格特点。

七、依存句法

依存文法 (Dependency grammar., 2017)最早由法国语言学家L.Tesniere在其著作《结构句法基础》中提出。依存语法分析基于一个的基本假设：句法结构本质上是词间关系。这种关系称为依存关系 (Dependency Relations)。一个依存关系连接两个词，分别是核心词 (Head) 和修饰词 (Dependent)。依存关系可以细分为不同的类型，表示两个词之间的句法关系 (Dependency Relation Types)。因此自然语言文本可以从序列形式转化为树状结构，从而刻画句子内部词语之间的句法关系。和其他句法分析形式如短语结构句法分析相比，依存句法分析具有形式简单、易于标注、分析效率高等优点。而且，依存句法更适合于表达非连续的、远距离的结构，这对于作家写作风格的研究非常重要。以这种方式对句子编码并使用DependencyViewer可视化如下：



在上图中，每一个词以唯一的依存关系指向它的父节点。即使把所有位置上的词都更换为相同语法功能的另一个词，也不会改变这棵树的结构。所以句法依存树可以编码一个句子的逻辑架构。为了方便计算机处理，我们可以将句子编码为依存关系的序列。例如图中的句子，可以表示为

令/relevant, 人/patient, 玩味/content, 的/aux-depend, 是/root, ，/PU, 虽然/aux-depend…

在后面的实验中将以这种方式利用句法依存信息。

在中文领域，研究者采用的句法依存关系大体如下。饼状图刻画各种关系在中文句子中的统计占比。

这也是FudanNLP采用的版本。

基于以上的特征抽取，我们得到了描述一个作家写作风格的48个特征（30个句法依存特征+11个功能词特征+其他句法词汇特征7个）。因此，每个文本样本可以表示为一个48维向量。

相似度衡量：

对作家所有作品的文本特征向量取平均，再以欧氏距离和余弦距离进行两两相似度衡量，得到相似度矩阵如下。

欧氏距离相似度：归一到[0, 1]区间，0表示最近，1表示最远

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 |
| 鲁迅 | 0.000 | 0.324 | 0.645 | 0.873 | 0.138 | 0.649 | 0.829 |
| 周作人 | 0.324 | 0.000 | 0.212 | 0.982 | 0.342 | 0.482 | 0.521 |
| 林语堂 | 0.645 | 0.212 | 0.000 | 0.493 | 0.898 | **1** | 0.767 |
| 三毛 | 0.873 | 0.982 | 0.493 | 0.000 | 0.453 | 0.233 | 0.452 |
| 王小波 | 0.138 | 0.342 | 0.898 | 0.453 | 0.000 | 0.454 | 0.546 |
| 刘慈欣 | 0.649 | 0.482 | 1 | 0.233 | 0.454 | 0.000 | 0.545 |
| 江南 | 0.829 | 0.521 | 0.767 | 0.452 | 0.546 | 0.545 | 0.000 |

余弦距离相似度

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 |
| 鲁迅 | 0.000 | 0.358 | 0.661 | 0.880 | 0.181 | 0.672 | 0.800 |
| 周作人 | 0.324 | 0.000 | 0.230 | 0.960 | 0.384 | 0.490 | 0.473 |
| 林语堂 | 0.645 | 0.212 | 0.000 | 0.484 | 0.932 | **1** | 0.746 |
| 三毛 | 0.873 | 0.982 | 0.493 | 0.000 | 0.482 | 0.270 | 0.476 |
| 王小波 | 0.138 | 0.342 | 0.898 | 0.453 | 0.000 | 0.424 | 0.566 |
| 刘慈欣 | 0.649 | 0.482 | 1 | 0.233 | 0.454 | 0.000 | 0.545 |
| 江南 | 0.829 | 0.521 | 0.767 | 0.452 | 0.546 | 0.545 | 0.000 |

由上表可知，刘慈欣跟林语堂的写作风格相差最大。这是容易理解：刘慈欣擅长刻画叙事而不擅长描写（特别是人物描写）和抒情，林语堂的小品文恰恰主要是抒情和描写。除此以外，我们惊奇地发现鲁迅和王小波的写作风格比其他作家更相似，具体的原因还需要进一步分析。

机器学习

机器学习方法——多元分类问题

数据增强：将每个作家的所收集的全部作品切分为约1000份数据，切分的依据是保持段落的完整性并控制每个样本字数范围在500-600，在段间进行切分。在这1000份数据中，使用70%作为训练集，10%作为验证集，20%作为测试集。

性能评价指标：

本实验采用以下指标对结果进行评定：查全率(Recall)，查准率(Precision)，正确率(Accuracy)，错误率(Error)，F值(F-Measure)。对于每项指标，把不同作家的文本数目（训练样本）作为权重求加权平均值。

决策树

决策树算法是一种使用实例逼近离散值函数的归纳学习方法，其学习过程是从一组没有次序和规则的实例中推理出以树形式表示的分类规则。决策树算法在自然语言识别的问题上已经有不少成功的先例： (Frery J, 2014)采用分类回归树的算法，以基尼不纯度作为分类判断的指标、后剪枝技术避免过拟合，能够有效识别作者的身份。在本次实验中，决策树的分裂采用“信息增益”标准，使用“后剪枝”避免过拟合。得到的模型评价如下表：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 | 加权平均值 |
| 准确率 | 0.894 | 0.887 | 0.796 | 0.896 | 0.886 | **0.898** | 0.859 | 0.824 |
| 召回率 | **0.843** | 0.824 | 0.801 | 0.832 | 0.841 | 0.803 | 0.815 | 0.803 |
| F值 | **0.868** | 0.854 | 0.798 | 0.863 | 0.863 | 0.848 | 0.836 | 0.827 |
| 正确率 | 0.891 | 0.881 | 0.795 | 0.887 | 0.877 | 0.897 | 0.854 | 0.861 |
| 错误率 | 0.109 | 0.119 | 0.205 | 0.113 | 0.123 | 0.103 | 0.146 | 0.137 |

准确率最高可以达到89.8%，对每位作家的识别都可以达到约80%左右的准确率，错误率基本维持在10%-20%。该结果是reasonable的，但不算令人满意。有理由怀疑，训练数据量有限不能使模型很好地学习到数据中的规律，即该模型有欠拟合的可能。下面采用一种对少样本友好的模型。

支持向量机

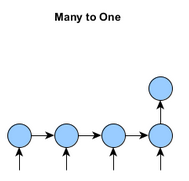
支持向量机算法的复杂度和样本维数无关，因此非常适合少样本高维数的情况。而且学习效率和准确率较高，是文本识别研究中常见的算法。经典的例子是 (施建军, 2011)利用支持向量机计算出《红楼梦》前后80回的风格差异，认为不是出自同一作者之笔。另外，支持向量机可以很好地求解有限样本的泛化问题，在本次实验中，作者的文本数据是非常有限的，训练的效果如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 | 加权平均值 |
| 准确率 | 0.915 | 0.907 | 0.931 | 0.898 | **0.939** | 0.841 | 0.893 | 0.903 |
| 召回率 | 0.769 | 0.762 | 0.779 | **0.879** | 0.733 | 0.804 | 0.821 | 0.792 |
| F值 | 0.836 | 0.828 | 0.848 | **0.888** | 0.823 | 0.822 | 0.855 | 0.843 |
| 正确率 | 0.913 | 0.903 | 0.927 | 0.890 | 0.935 | 0.832 | 0.888 | 0.898 |
| 错误率 | 0.087 | 0.097 | 0.073 | 0.110 | 0.065 | 0.168 | 0.112 | 0.102 |

可以发现，支持向量机能够把最高准确率提高到93.9%，而且相比决策树在准确率、F值和正确率方面都有或大或小的提升。可以认为，支持向量机能够更好地解决作家文本风格识别的问题。

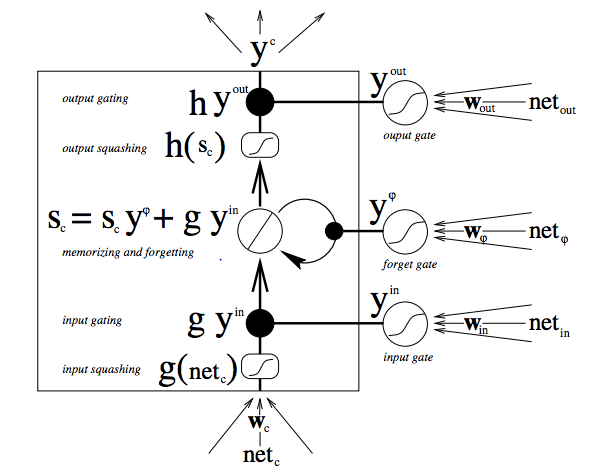
深度学习方法

一般来说，深度学习对数据量的要求比较高，训练一个合格的深度模型至少也需要千万级别的数据作为“燃料”。但是在这个问题中，数据是比较缺乏的。一位作家即使再高产，也未必能达到如此巨大的数据量。而且不同作家之间的作品产量相差很大，对深度模型而言这是一个“非平衡样本”的问题。为了缓解这种问题，在深度模型的训练样本上与之前的机器学习有些许不同——我们不再使用原来抽取的特征，而使用FudanNLP得到的整个文本的句法依存关系序列作为对原始文本的一种逻辑编码。LSTM(Long-Short-Term-Memory)分类器能够动态挖掘结构数据以序列形式呈现的变化。把句法依存关系构成的序列作为LSTM的输入，输出一个表示作家分类的标签，得到的就是一个简单的LSTM分类器，即一个Many-to-One形式的RNN。



图片来源：https://deeplearning4j.org/lstm.html

本次试验采用的模型架构如右图。



图片来源：https://deeplearning4j.org/lstm.html

输入是句法依存关系的序列，按照以下顺序编码为0到29。

*{ "之字结构", "时态", "构式", "修饰", "话题", "关联", "同位", "数量", "顺承", "定语", "主语", "限定","从属", "介宾", "标签", "连动", "总括", "标点", "的字结构", "Root", "宾语", "语态", "并列", "得字结构", "地字结构", "核心词", "疑问连动", "感叹","补语", "状语" }*

输出是10维向量，代表识别作家的分布

*｛鲁迅，周作人，林语堂，三毛，王小波，刘慈欣，江南｝*

实验使用L1正则化，激活函数使用softsign函数（可以有效避免饱和现象），损失函数是均方误差函数(MSE)，求解方法是随即梯度下降中的AdaGrad算法，使用early stopping的方法自动结束训练。

模型在测试集的识别效果如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 预测分类|实际分类 | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 |
| 鲁迅 | 252 | 2 | 0 | 1 | 0 | 1 | 2 |
| 周作人 | 3 | 417 | 0 | 0 | 0 | 0 | 0 |
| 林语堂 | 0 | 2 | 309 | 1 | 0 | 0 | 0 |
| 三毛 | 0 | 1 | 0 | 436 | 1 | 0 | 0 |
| 王小波 | 1 | 0 | 0 | 2 | 393 | 2 | 1 |
| 刘慈欣 | 0 | 0 | 1 | 3 | 2 | 144 | 2 |
| 江南 | 5 | 2 | 1 | 4 | 0 | 0 | 316 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 | 加权平均值 |
| 准确率 | 0.890 | 0.952 | **0.969** | 0.980 | 0.968 | 0.837 | 0.938 | 0.933 |
| 召回率 | 0.854 | 0.965 | 0.957 | 0.934 | **0.968** | 0.954 | 0.969 | 0.943 |
| F值 | 0.872 | 0.959 | 0.963 | 0.956 | **0.968** | 0.892 | 0.953 | 0.937 |
| 正确率 | 0.890 | 0.952 | 0.969 | 0.980 | 0.968 | 0.837 | 0.938 | 0.933 |
| 错误率 | 0.110 | 0.048 | 0.031 | 0.020 | 0.032 | 0.163 | 0.062 | 0.067 |

效果十分明显，以大概率完成正确识别不同作家的写作风格。

应用

LSTM分类器可以很好地划分作家的写作风格。对于模仿风格的作品，也能很好地识别。

以下是网友模仿鲁迅风格的一段文字。以它为输入，LSTM分类器可以很好地识别出鲁迅的风格。

风雨渐作，乃至及于冬日的时分，广州的天色总是阴晦的。街口只有几盏残灯，没了人流的道上静悄悄的。不必竖起耳朵，就能听到啮嘴的小鼠在角落里发作，和细虫一齐鼓噪。夜已深了，明凯兀自坐在摊前，黑沉沉的脸上带着些蜡黄，须发似乎许久没有整理。素日锐利的眼里，此时也没了神采。良久，他搓了搓手，攥在手里的肉松已捏出了絮。他是在廿日的傍晚，与友人看完了一场新派电影，才听说这场风波的。……（省略后文约600字）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 鲁迅 | 周作人 | 林语堂 | 三毛 | 王小波 | 刘慈欣 | 江南 |
| 分类概率 | **0.419** | 0.132 | 0.039 | 0.046 | 0.248 | 0.031 | 0.049 |

结论

在文学领域，作家的写作风格是一个相对抽象的概念。在中文领域，怎样让计算机理解并识别作家的写作风格，是一个极具挑战性的命题。本文试图在几篇经典论文的思路上做进一步探索，充分利用课上所学的NLP相关概念与方法，结合以FudanNLP、NLTK为代表的高性能工具包，将中文作家的写作风格以合适的方式编码，借助机器学习甚至深度学习的技术，在一定问题范围内训练分类器，有效解决中文作家写作风格识别的问题。对于新的文本，LSTM分类器能够以极高的可信度识别出作者。这项技术除了在作家文本风格识别中得到应用，还能用于作者身份建模、匿名文本识别、文学研究当中。

完整代码： <https://github.com/FengZiYjun/NLP-Lab>

# 参考文献

Abbasi A, C. H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, pp. 67-75.

*Dependency grammar.* (2017, November 10). Retrieved from Wikipedia, The Free Encyclopedia: https://en.wikipedia.org/w/index.php?title=Dependency\_grammar&oldid=809707273

Frery JC, Juganaru-Mathieu MLargeron. (2014). UJM at CLEF in Author Verification ased on optimized classification trees[C]. CLEF.

Sebranek, P., Kemper, D., & Meyer, V. (2006). *A Student Handbook for Writing and Learning.* Wilmington: Houghton Mifflin Company.

Zhao YJZobel. (2007). Searching with style: Authorship attribution in classic literature[C]. In Proceedings of the 30th Australasian Computational Science Conference, 页 59-68.

胡壮麟. (2000). 理论文体学[M]. 北京: 外语教学与研究出版社.

黄进. (2006). 现代汉语功能词的语义语法学研究[D]. 南京: 南京师范大学.

诺姆-乔姆斯基. (1979). 句法结构. 中国社会科学出版社.

施建军. (2011年5月). 基于支持向量机技术的《红楼梦》作者研究. 红楼梦 学刊, 页 35-52.

武晓春黄萱菁,吴立德. (2006). 基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 页 63-70.