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。本文从经典的研究方法出发，做了进一步探索，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。对于未知文本，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、风格识别、句法依存树、功能词、LSTM

项目代码： https://github.com/FengZiYjun/Writing-Style-Recognition

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)。作家的写作风格不仅受所处社会背景、创作时的个人状态、作品主题的影响，还跟当时的语言演变和发展情况密切相关。

我们通常说某篇文章有鲁迅的风格，实际指的不是该文章的主题跟鲁迅的某一篇文章的主题相似，而是遣词造句的方法跟鲁迅的某些作品很像。这里所说的遣词造句的方法其实可以视为作家风格的代表。而关键就是，怎样用语言学的手段表示作家的写作风格。

问题分析

为了简化问题，这里研究的中文作家限制在从新文化运动至今这段时期的作家，即不考虑白话文出现之前的文章。选取的作家可以是任何文学形式的作家，包括小说、散文、杂文等。

本文所指的写作风格(writing style)，统一采用 (Sebranek, Kemper, & Meyer, 2006)的定义：Writing style is the choice of words, sentence structure, and 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。

在限定条件下，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。

数据收集和预处理

从五六文学网(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 10 万字的文本
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

根据研究经验，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。

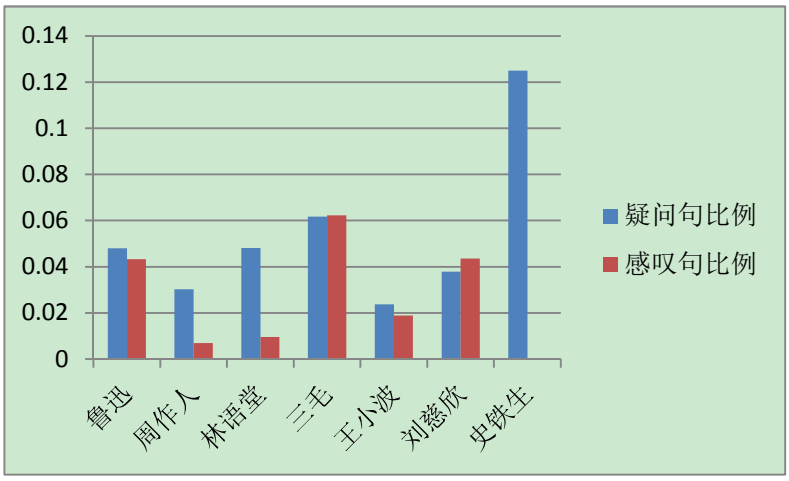
综上所述，选取的文本特征如下：
1. 词性比例——不同词性的词的数量与总词数之比
2. 词汇丰富度——不同的词数与总词数之比
3. 单现词比例——只出现一次的词的数目与总词数之比
4. 高频虚词（功能词）——介词、连词、叹词、结构助词、语助词、方位词的前若干个高频项，
5. 平均句长——所有句子所含字数（包括标点）的算术平均
6. 短句比例——短于平均句长一半的句子数量与总句数之比
7. 长句比例——长于平均句长两倍的句子数量与总句数之比

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。

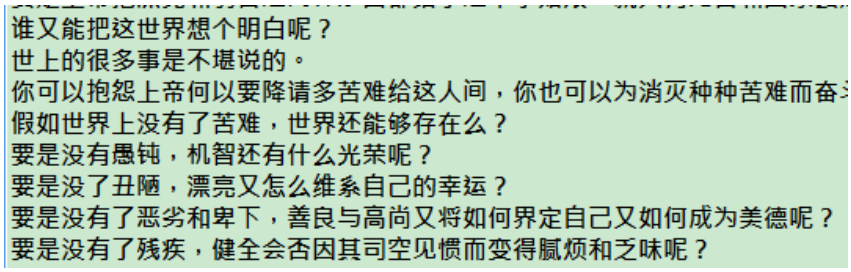
数据分析

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

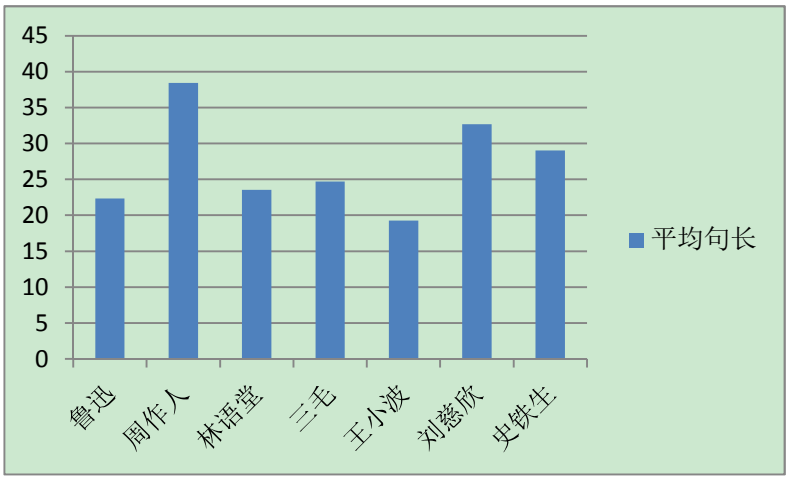


上图统计的是各个作家使用的疑问句和感叹句的比例。可以发现，史铁生使用疑问句的比例远比其他作家高。熟悉史铁生作品的人应该知道，他的散文无论是偏议论还是偏叙事，很多时候会用平静的问句来推动人物的对话或思考的逻辑。例如：


谁又能把这世界想个明白呢？
世上的很多事是不堪说的。
你可以抱怨上帝何以要降请多苦难给这人间，你也可以为消灭种种苦难而奋
假如世界上没有了苦难，世界还能够存在么？
要是没有愚钝，机智还有什么光荣呢？
要是没了丑陋，漂亮又怎么维系自己的幸运？
要是没有了恶劣和卑下，善良与高尚又将如何界定自己又如何成为美德呢？
要是没有了残疾，健全会否因其司空见惯而变得腻烦和乏味呢？

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



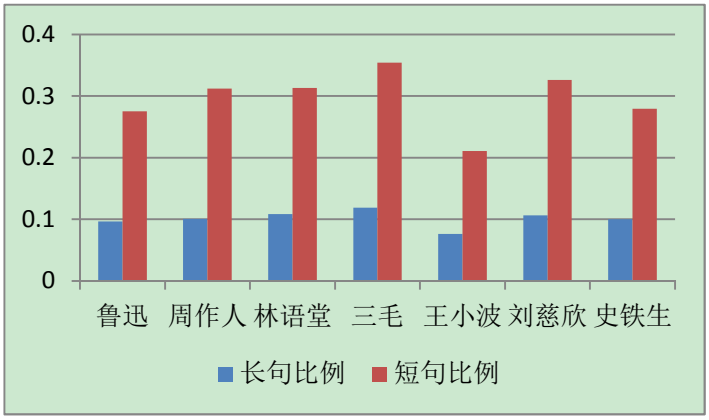
以上是各作家作品的平均句长。显然周作人的句子较长，王小波的句子较短。然而这只是一种很粗糙的统计，算法使用句号等标点符号显式切分句子，但是中文文学的句子会以逗号表示意义的承接，用句号表示逻辑的结束。所以这里的一个句子可能包含多个顺次承接的短句。这个特点在周作人的作品中表现更明显。

王小波的作品擅长以短句嬉笑怒骂。例如：

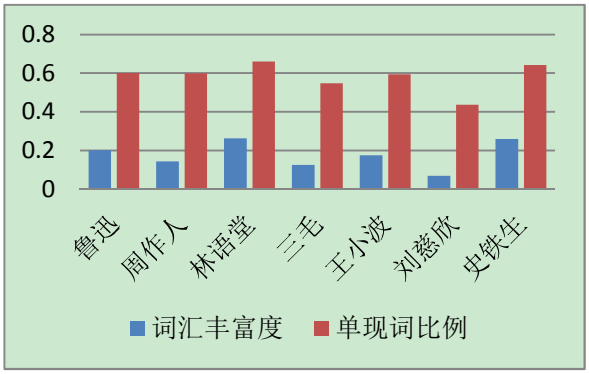
那天晚上我没去找她，倒进了医院。
这事原委是这样：
早上我到牛圈门前时，有一伙人等不及我，已经在开圈拉
大家都挑壮牛去犁田。
有个本地小伙子，叫三闷儿，正在拉一条大白牛。
我走过去，告诉他，这牛被毒蛇咬了，不能干活。
他似乎没听见。
我劈手把牛鼻绳夺了下来，他就朝我挥了一巴掌。
亏我当胸推了他一把，推了他一个屁股墩。
然后很多人拥了上来，把我们拥在中间要打架。
北京知青一伙，当地青年一伙，抄起了棍棒和皮带。
吵了一会儿，又说不打架，让我和三闷儿摔跤，三闷儿摔
我一脚把三闷儿踢进了圈前的粪坑，让他沾了一身牛屎。
三闷儿爬起来，抢了一把三齿要砍我，别人劝开了。

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。



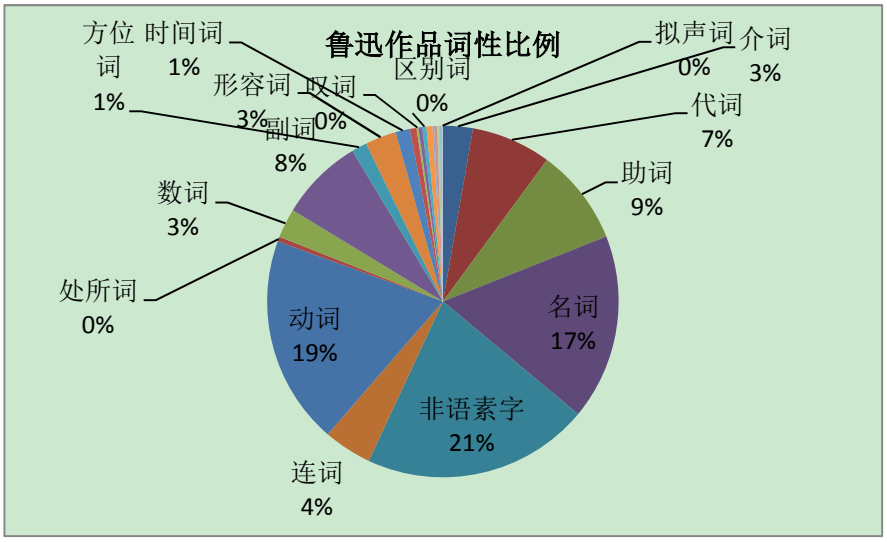
四、 词汇丰富度与单现词比例

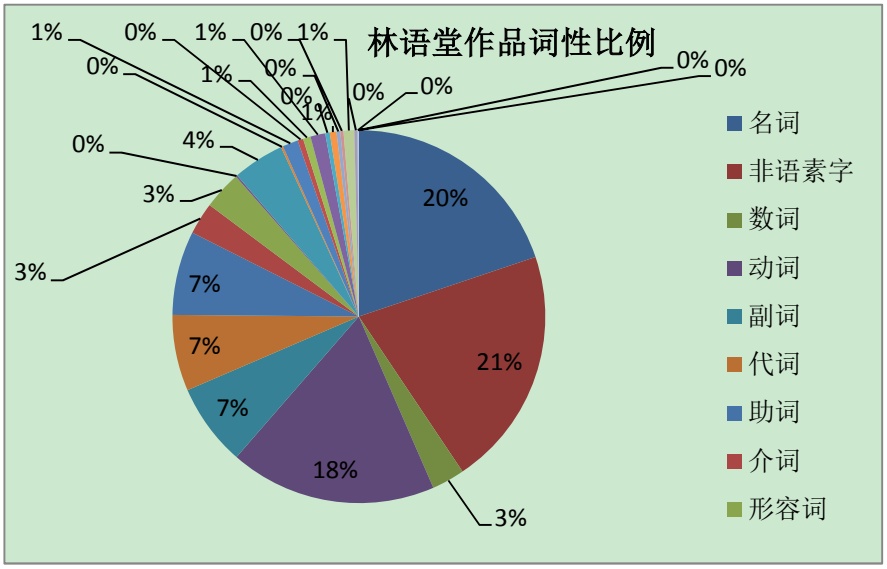


词汇丰富度和单现词比例信息如图。刘慈欣的作品在这两个指标上都比较低，这也是“硬科幻”作家被诟病的原因之一——过于注重技术细节和人文伦理的讨论，而淡化人物塑造和情景描写。因为人物面谱化严重，对人物的描写上用词重复就会导致新鲜的词汇较少。与此相对，林语堂的文学语言中词汇较丰富，艺术味道较浓。

五、词性比例

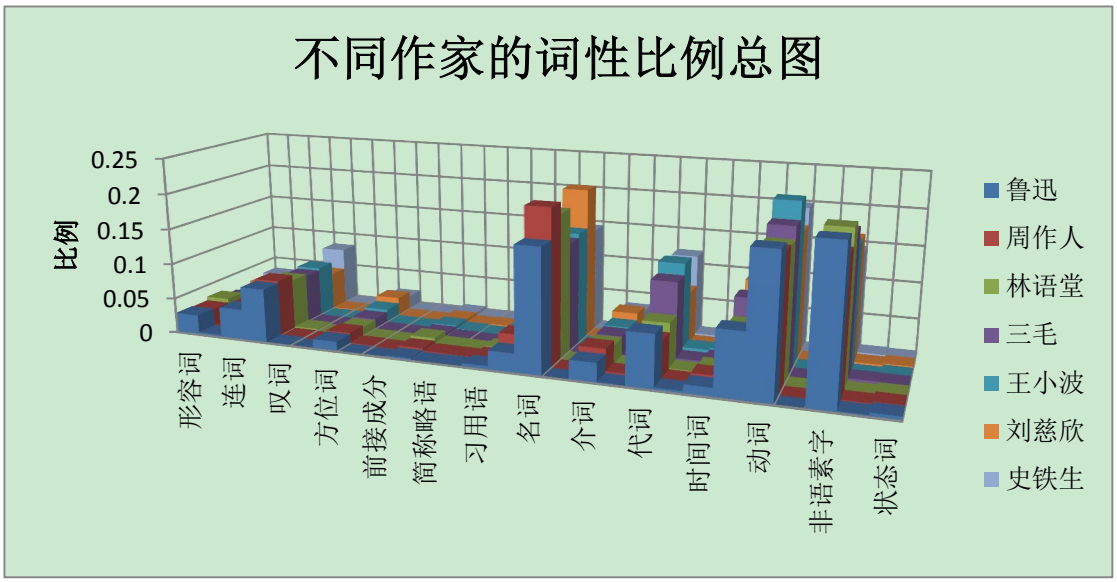
以下饼状图显示的是鲁迅和林语堂作品的词性比例。实际上中国现代作家的词性比例是相差不大的，原因是白话文普及之后，汉语的句法基本没有太大的变化。名词、动词和非语素字分别占约五分之一，助词、副词和代词分别占略低于 10%，剩下各种词性比例少许。

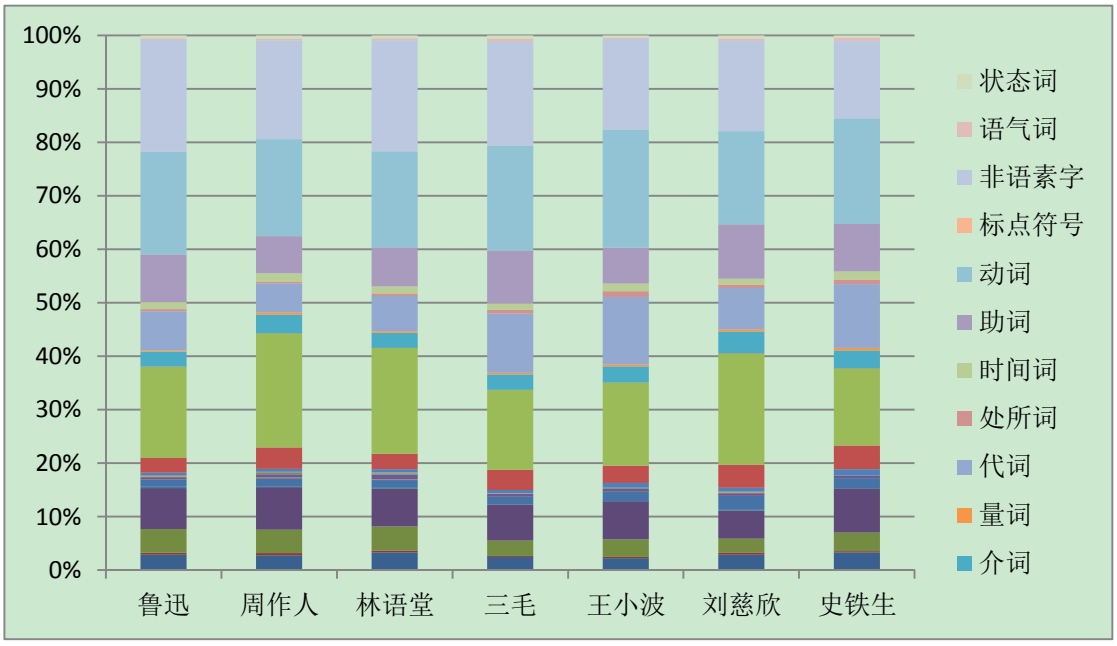




至于作家之间的词性比例差异可以参考下面的三维图。可以看出，基本上不同作家的用词比例差异不大，除了几个比较明显的峰值：刘慈欣、周作人和林语堂使用名词的比例略高，王小波、三毛、史铁生使用动词的

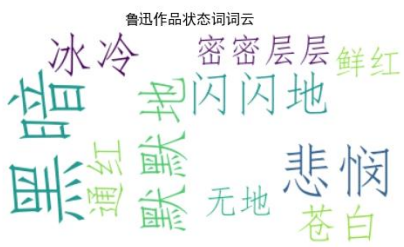
比例略高。这种差异比较细微，对于风格的解释意义不大。





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：

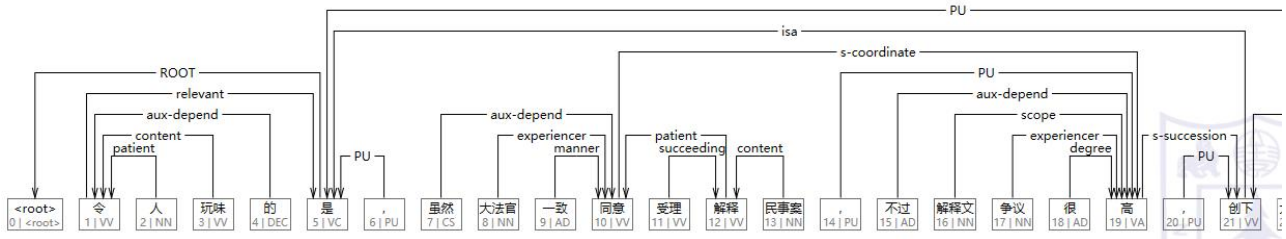


词云中字的大小代表频率。字越大出现频率越高。观察副词和助词的词云，可以发现被词云选中的高频词基本上是固定的几个，但是它们所构成的频率分布有显著的差异。例如，对于副词，鲁迅用“不”、“也”、“就”比较多，而刘慈欣用“都”、“也”、“就”比较多，周作人用“也”、“不”、“又”、“都”、“很”比较多。对于助词，鲁迅使用“等”的频率相对比周作人多，而使用“的话”的比例明显比刘慈欣少。

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：

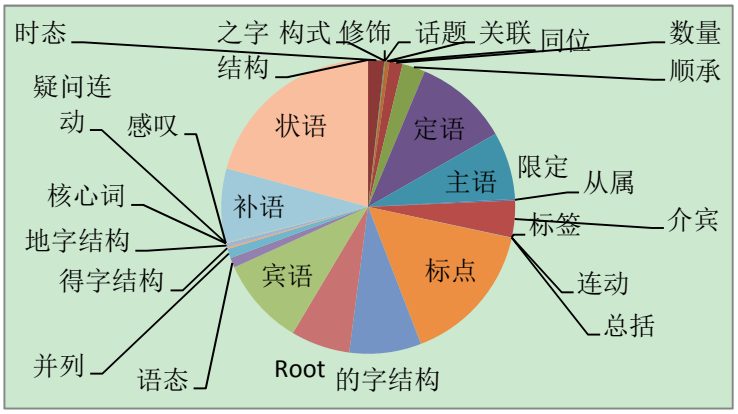


在上图中，每一个词以唯一的依存关系指向它的父节点。即使把所有位置上的词都更换为相同语法功能的另一个词，也不会改变这棵树的结构。所以句法依存树可以编码一个句子的逻辑架构。为了方便计算机处理，我们可以将句子编码为依存关系的序列。例如图中的句子，可以表示为

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

后面的实验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一般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。

对于未知文本的分类问题，仅凭借相似度是不可靠的。因为在上述相似度度量中每个特征都有相同的地位，对距离的贡献是平等的；但实际上有些特征比较重要（如功能词），另外一些特征比较次要（如标点），它们明显不能被赋予同样的分类权重。在这种情况下，我们希望借助文本本身的内在性质，自动选择合适的特征构建分类器，这就是需要借助机器学习的方法。

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。在这约 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。

数据划分如下：

样本数	鲁迅	周作人	林语堂	三毛	王小波	刘慈欣	江南
训练集	714	703	695	673	720	672	630
验证集	102	100	99	96	102	96	90
测试集	205	202	200	193	207	193	181
总计	1021	1005	994	962	1029	961	901

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。
以下方法都使用开源的 scikit-learn 工具包实现。

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
SVM 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

支持向量机的测试效果在准确率上比决策树要好。几乎所有作家的测试准确率都有或大或小的提升。其中最明显的是王小波作品的识别率从 88.6%升至 93.9%。

随机森林

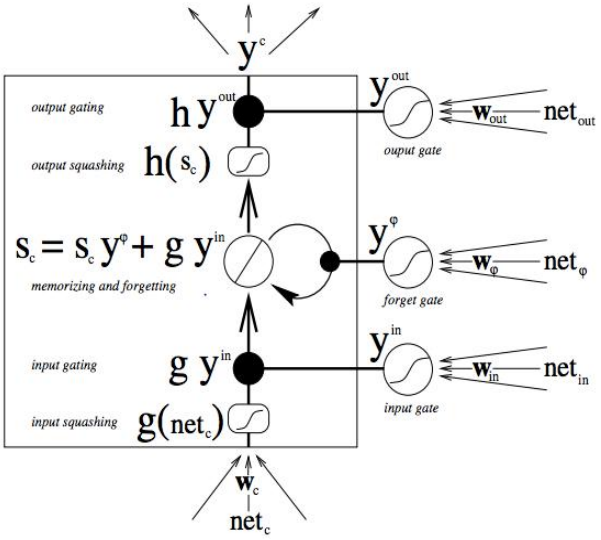
作为决策树的集成版本，使用随即森林一般会有更好的效果。这里设定随机森林的 base estimator 个数为 10，每个 base estimator 都是深度不超过 3 的决策树，决策树的构建方法跟第一个模型一样。

	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	0.919	0.932	**0.942**	0.876	0.935	0.855	0.938	0.914
召回率	0.799	0.809	0.824	0.928	0.691	0.816	0.837	0.815
F 值	0.854	0.866	0.879	0.902	0.795	0.835	0.885	0.859
正确率	0.906	0.867	0.931	0.936	**0.961**	0.861	0.841	0.900
错误率	0.104	0.100	0.051	0.090	0.106	0.215	0.124	0.113

可以发现，随机森林能够把最高准确率提高到 96.1%，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。

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，也未必能达到深度学习所要求的巨大数据量（千万级）。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。
本次试验采用的模型架构如下图。



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

使用框架	TensorFlow1.4.0
输入维数	300
时间戳个数	30
隐藏层单元个数	256
激活函数	softsign $f(x)=\frac{x}{1+\lvert x\rvert}$
正则化	L1
输出维数	10
损失函数	交叉熵
优化算法	AdaGrad

句法依存关系是一个 30 维 one-hot 向量。

{"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"}

LSTM 的输入设定为长度为 10 的序列，因此实际输入是 300 维向量。
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}*

时间戳个数表示每个作家的一个样本（已经切分好的依存关系序列）以长度为 10 的窗口通过 LSTM 需要的输入次数。TensorFlow 在输入的张量上增加额外一个维度表示时间戳。
选用的激活函数是 softsign，根据调参经验可以缓和梯度饱和问题。
输出的 10 维向量可以通过 softmax 分类器归一化，得到不同作家的分类概率。取最大的概率为预测输出。

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	190	6	2	0	2	10	6
周作人	8	187	18	0	0	0	4
林语堂	0	2	172	4	1	0	0
三毛	0	3	0	170	1	0	0
王小波	12	0	0	2	202	12	1
刘慈欣	0	0	5	3	2	178	2
江南	5	4	3	14	4	3	168

	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	0.880	0.862	0.961	**0.977**	0.882	0.937	0.836	0.905
召回率	0.884	0.926	0.860	0.881	0.953	0.877	0.928	0.901
F 值	0.882	0.893	0.908	0.927	0.916	0.906	0.880	0.902

效果略有提升，在王小波作品识别率上可以达到 97.7%准确率。
利用深度学习的方法作家作品风格识别的问题，囿于数据量和时间限制，没有做更深入的探究。从目前的效果来看，使用深度学习和传统机器学习方法在监督分类上没有非常大的区别。

应用

中文作家写作风格识别最大的应用是在文学研究领域。通过分析作家的写作风格，可以为文学流派分析、师承关系分析、文学史研究提供参考，还能用于匿名文本考究，作者身份建模等领域。例如：历史上很多作家都会使用多个笔名发表作品，还有不怀好意之人冒用他人笔名发表文章，怎样判断一篇文章的确属于某个作者，除了用考据的手段，还能用作家风格识别的方式。

鲁迅是中外文学历史上笔名数量最多的作家之一 。《春末杂谈》是鲁迅于 1925 年发表在《莽原》上的一篇文章，用的是另外一个笔名“冥昭”。如果不是有其他证据的话，我们其实没有办法知道“冥昭”到底是不是鲁迅本人。这时唯一的证据就是文章《春末杂谈》本身。使用随机森林分类器对该文章进行分类，得到的结果是鲁迅。

北京正是春末，也许我过于性急之故罢，觉着夏意了，于是突然记起故乡的细腰蜂。那时候大约是盛夏，青蝇密集在凉棚索子上，铁黑色的细腰蜂就在桑树间或墙角的蛛网左近往来飞行，有时衔一支小青虫去了，有时拉一个蜘蛛。青虫或蜘蛛先是抵抗着不肯去，但终于乏力，被衔着腾空而去了，坐了飞机似的。……（《春末杂谈》全文约 2400 字）

除了作家自己的作品以外，对于高度模仿的作品，分类器也能很好地识别。
以下是网友模仿鲁迅风格的一段文字。除了内容不同外，句式结构和用词跟鲁迅的风格极似。
用 LSTM 进行分类，输出分类概率。

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。……（全文约 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

分类器可以识别出鲁迅风格的文字。由于模型是内容无关的，即使主题完全不一样，也能根据以上所阐述的 10 个方面的特征，计算出该文章与鲁迅风格最相近。值得注意的是，识别对象必须字数足够多，最好一千字以上，效果才比较明显。

更多的分类例子请参考项目完整代码。

结论

对于中文作家写作风格识别问题，本文在经典研究上做了进一步探索。从作品文本当中，选取 10 个方面的 48 维特征，使用基于树的方法构建分类器，达到较好的分类效果。使用 LSTM 为代表的深度网络，对作品句法依存序列进行建模，也能达到一流的识别效果。两种方法在模型复杂性、算法效率上各有优劣，但最终结果差别不大。根据这些方法构造的分类器，在有限作家集合的条件下，能够有效解决中文作家写作风格识别的问题。对于新的文本，分类器能够以较高的可信度识别出可能的作者。这项技术还能用于作者身份建模和匿名文本识别，对于中文文学研究有极大的应用价值。

作家风格识别在中文 NLP 领域是一个冷门话题，因为既没有统一评价标准，也没有巨大的商用价值。但作为课程的实验项目，能够充分利用已学的知识来构建分类系统的各个流程都有很大的实践意义。纵观整个风格识别问题，理论上的难点是选取合适的特征和合适的分类器，而工程上的难点是支持中文处理的工具包较少，效果不稳定。如果有更好的中文工具包可以使用，效果应该更好。

完整项目代码： https://github.com/FengZiYjun/Writing-Style-Recognition

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.

Martín AbadiAgarwal, Paul Barham, Eugene BrevdoAshish. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

PedregosaG.,Gramfort, A. Michel, V.F.,Varoquaux,. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 页 2825--2830.

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.