基于计量的文本文采识别与可视化

- 问题出发点：文本文采量化，能够通过指标判断一段文本是否有文采且能将其文采的表现与构成展示出来。
  - 语料选择：由于不同语体下，文采的具体表现形式不同，故而需要分语体进行分析。
    - 文艺语体
      - 选择散文、小说为代表语料，散文选取/散文阅读网/中的原创作品，小说选取小说阅读网中的原创作品。
        - 统计分析与指标选取——筛选出在语言学上具有较好解释性的指标

          通过显著性（$p<0.05$，指在不同类别文本中分布具有差异显著性的指标）与相关性分析（$r > 0.3$，$p<0.05$指指标与文本类别具有显著相关性）进行指标筛选——此时只做简单的描述分析

          - 支线任务：对每个筛选出的指标进行深入分析，如词汇进行具体类别分析，挖掘具有文采属性的词语。
        - 将指标运用于文字片段的文采判断（有无）

          验证以上指标在机器学习上是否有效，能否达到较好的分类效果
        - **片段文采表现的可视化**
          - 从色彩美、形象美、情感美、声音美、装饰美、哲理美6个方面合并词、句、篇章三个维度的指标，每个指标根据相关性高低进行加权，每个方面的大指标以该类型下所有文本的平均数A为60分，文本在该方面所得分数=60/A*（指标a*权重+指标b*权重+……+指标z*权重）
    - 科技语体
    - 政论语体

- 公文语体

- 语料处理
  - 以500字为单位，删掉句末最后一个句号后不完整的内容，然后再筛选一遍，确保每个片段语义完整。散文500个片段，小说500个片段
- 语料标注

  工具：excel。方法：0.选择朱自清、张爱玲、冰心、林语堂、沈从文、张爱玲等作家散文作品中的优秀片段；鲁迅、巴金、萧红、莫言等作家小说作品中的优秀片段，作为文采标注的参考语料，标注者需要先对这些语料进行阅读，以获得基本的文采语感。1.以20段为基本单位，标注者首先阅读完20个语段，之后再对每段话进行有无文采的判断。
  2.为保证标注的结果尽可能反映出人们的普遍认知，每段话由4人判断，若每一判断结果 > 2人一致则收录该语料。 = 2人则舍弃该语料。3.对语料进行辞格标注4.句类标注
- 构建文采特征体系——

  音——声音美
  形——装饰美
  义——色彩美、情感美、形象美、哲理美
  - 词汇
    - 词汇语音——声音美
      - 1. 2次反复词数 2. 2次反复词比例 3. 3次反复词数 4. 3次反复词比例 5. 4次及以上反复词数 6. 4次及以上反复词比例 7. 拟声词形符数 8. 拟声词比例 9.拟声词类符数 10. 类符比例11. 叹词数 12. 叹词比例13.连绵词数14.连绵词比例15.非双声叠韵连绵词数 16.非双声叠韵连绵词比例 17.双声词数 18.双声词比例 19.叠韵词数 20.双声叠韵词数 21.双声叠韵词比例 22.叠字连绵词数 23.叠字连绵词比例24. AAB类叠音词数 25. AAB类叠音词比例 26. ABB类叠音词数 27. ABB类叠音词比例 28. AABB类叠音词数 29. AABB类叠音词比例 30.AA类叠音词数 31.AA类叠音词比例

- 词汇外形
  - 装饰美
    - 1.同义词类数；2.同义词数；3.同义词比例4.反义词类数；5.反义词数；6.反义词比例7.同音词类数；8.同音词数；9.同音词比例10.缩略词数；11.缩略词比例12.平均词长；13.单字词比例；14.双字词比例；15.三字词比例；16.四字词比例；17.四字以上词比例；18.字词比19.词形符数；20.词类符数；21.词类符形符比；22.单次词数；23.单次词比例；24.成语数；25.成语比例；26.方言词数；27.方言词比例；28.口语词数；29.口语词比例；30.外来词数；31.外来词比例；32.古语词数；33.古语词比例；34.文言词数；35.文言词比例；36.惯用语数；37.惯用语比例38.名词比例；39.动词比例；40.形容词比例；41.副词比例42.实词数 43.实词比例44.常用词、非常用词使用指标—R2； 45. 非常用词词型数比例-R3； 46.词汇使用均衡度R4
  - 词汇语义
    - 色彩美、形象美（在此将色彩美与形象美的词汇特征进行合并，统称为形象色彩词汇。因为色彩美在大框架下仍然属于视觉类形象色彩词。）
      - 1.空间觉形容词比例 2.味觉形容词比例 3.视觉形容词比例 4.触觉形容词比例 5.听觉形容词比例6.嗅觉形容词比例
    - 情感美
      - 1.褒义词数 2.褒义词比例 3.贬义词数 4.贬义词比例 5.兼类词比例
    - 哲理美
      - 1.原型意象类型数 2.原型意象词数3.典故词数；4.哲理性成语数
- 句
  - 音

- 声音美
  - 1.等长句比例2.句内虚词比例
  - 3. 隔行韵词比例 4. 排韵词比例 5.双行韵比例
  - 6.句子破碎度7.句子离散度
  - 8.句内对偶次数9. 句内对偶比例10.句内顶真次数11. 句内顶真比例
- 形
  - 装饰美
    - 语句修辞：1.反复次数2.反复比例3.对偶次数4.对偶比例5.顶真次数6.顶真比例
    - 句形变换：1.句子总数2.平均句长（基于字）3.平均句长（词）4.最大句长（基于字）5.最大句长（基于词）
    - 句法结构：
      - 短语句法结构：1.单句数；2.单句比例；3.复句数；4.复句比例；5.平均句法树高度；6.最大句法树高度；7.高度>16 的句法树数量；8.高度>16 的句法树比例；9.平均名词短语数；10.平均动词短语数；11.平均形容词短语数；12.平均副词短语数；13.平均介词短语数；14.名词短语平均长度；15.动词短语平均长度
      - 依存句法结构16.平均主语长度；17.最大主语长度；18.平均修饰语个数；19.平均修饰语长度；20.平均主要动词前的词数；21.主要动词前的最大词数；22.平均句子依存距离；23.最大依存距离
- 义
  - 形象美
    - 语句修辞（具象化：1.比喻次数2.比喻比例3.比拟次数4.比拟比例5.列锦比例6.通感比例
  - 色彩美

- 情感美
  - 1.反复比例 **2.夸张比例**
- 哲理美
  - 1.用典比例2.引用比例3.谚语比例4.歇后语比例5.格言比例

- 篇章
  - 音
    - 声音美
      - 修辞1.排比比例
  - 形
    - 装饰美
      - 修辞1.排比比例
      - 篇章复杂性 1.篇章段落数；2.平均段落长度（字）；3.平均段落长度（词）4.最长段落长度（基于字）；5.最长段落长度（基于词）；
      - 篇章衔接性1.代词比例2.人称代词比例；3.第一人称代词比例；4.第三人称代词比例；5.平均句间重叠词数；6.平均句间重叠词占比；7.平均句间重叠实词数；8.平均句间重叠名词数；9.平均段落间重叠词数；10.平均段落间重叠词占比；11.平均段落间重叠实词数；12.平均段落间重叠名词数；13.介词比例；14.连词比例；15.并列连词比例；16.选择连词比例；17.承接连词比例；18.递进连词比例；19.转折连词比例；20.因果连词比例；21.假设连词比例；22.比较连词比例；23.让步连词比例；24.目的连词比例
  - 义
    - 情感美
      - 1.设问比例2.反问比例
      - 1.感叹句比例2.陈述句比例3.祈使句比例4.疑问句比例

- 1.标点符号数