方向

一 如何衡量作文质量

1 PEG 代理量度标准（注重表面结构而忽视文章语义）：作文长度反映流畅性、介词与关系代词等反映句子结构复杂度、词长情况反映文章措辞。

2 IEA 文本相似度分析：单词用法级别的统计模型，允许对文本段落的语义相似性进行分析

3 词性标注、句法分析

4 篇章结构评分

/////随机森林模型训练 浅层语言特征评分 /与专家系统对比得出较为接近的分数/确定不同特征的权重/多棵决策树 训练集：测试集 训练模型 测试 调优

篇章结构评分抽取的特征有两项：

(1) 浅层语言特征：

评价维度反映了文章的结构特征，例如用词丰富度，句式变化程度，文章长短等。这些维度包括：句子数/平均句长/单词数量/三级词汇占比/平均词长/未登录词占比/停用词占比/动词短语数量等。这些评价维度不涉及到文章的语义特征，更不涉及到文章的主题特征。

(2) 语义特征：

主要通过 LDA 主题模型/Word2Vec 方法构建文章的语义向量，反映文章的语义特征，主题特征等。

4.1 浅层语言特征

早年的 PEG 系统认为：计算机只能抽取文章的表层结构，而不能理解文章的语义。因此只提取文章的表层语言特征进行多元回归分析，而忽视了文章的内在，使得有心之人写出复杂结构的文章进而骗取计算机取得高分。

文章的浅层语言特征，即只抽取文章的单词，句子表层的特征，例如单词和句子的长度信息，单词和句子的数量信息，动词/介词/代词等的多样性使用。研究表明，某些特征项的组合可以稳定地反映作文的质量高低。 这些选出的特征项包括:句子数量、平均句长、三级词汇占比、词长大于 7 的单词数、词长大于 8 的单词数、词长大于 9 的单词数、单词数量、

类符形符比、平均词长、动词短语的数量、动词的类符形符比、副词的类符形符比、介词类数以及代词数量等。

///////todo 做成图表的形式

这些特征项从长度、词汇(比如词汇量、词汇多样性、复杂词汇和难词等)、搭配(比如动词短语等)、句子结构(比如复杂句式等)以及不同词性的词汇使用情况等不同角度反映作文的内在特点，每个特征项都有自己的用处:

(1)句子数、平均句长反映了句子的复杂程度。

(2)三级词汇占比反映了作文中难词的使用情况。单词数量反映了作文的长度，因为大学生英语写作的要求一般为 120 到 150 个单词，该特征能检查空作文、过短或者过长的作文。

(3)平均词长反映了作文整体的单词复杂度。

未登录词（OOV）占比、错误词（obvious typos）占比

停用词占比

句子连贯性特征，关键字重复率指数

(4)类符形符比反映了作文使用的词汇多样性，反映文本中词汇的丰富程度。

(5)单词的类符数量反映了写作者的词汇量。

(6)动词短语的数量，反映了写作者对动词掌握的情况，这个特征项已经经过前期相关学者的研究，证明　其在大学英语写作中对作文分数具有较高的预测力。

(7)动词的类符形符比、副词的类符形符比、介词类数以及代词数量反映了作文中各个词性词汇的掌握情况。

使用 N-gram 模型对其进行计算词向量，进行评分。

4.2 语义特征

Word2Vec 模型是一个深度学习的模型，是用来生成词向量的工具。

5 篇章分析

PEG 系统自动评分的生成因素只考虑了篇章结构的重要性，而未将篇章语义的重要性考虑进去。随后产生了一种基于内容分析的评分系统-IEA。IEA 是一种基于潜在语义分析的作

文自动评分系统。潜在语义分析，简称 LSA，是指文章中词之间存在着某种结构或语义上的关联。多义词之间有着不同的语义结构;同义词有着相同的语义结构。LSA 可以将训练语料库提炼为不同的概念，每个概念代表不同的含义，也与测量的主题吻合。每个概念的组成结构包括以下属性：各文档文本相关度，各此项相关度，该概念对主题评分的权重影响。

5.1 文本倾向性分析

文本倾向性分析即从文本中挖掘出用户对于某个事物的看法，判断该看法的情感倾向为褒义或贬义。基本的目标就是实现区分出正面，负面或者中性，即极性分析;按照程度进行划分，即星级评分。

文档级别的情感识别，即为了准确的识别级性，可以考虑对文本的主客观语句分类，提取出 n 个最主观的句子概括评论的褒贬倾向。

基于特征的情感识别，需要从上下文提取出评价的对象，提取对象的特征，然后判断倾向性描述在每个特征上的极性。

//////////文本倾向性分析

5.1.1 褒贬倾向

(1)确定语言的褒贬倾向

(2)训练语料

5.1.2 情感分析

5.2 文本分类

每个文本都有其所属类别，文本分类即使文本集合按照一定的标准进行分类。文本分类主要包括训练阶段与预测阶段。分类模型即训练阶段得到的分类依据。预测阶段根据分类模型对新文本分类。

常见的分类方法由支持向量机(SVM)，K 临近算法和朴素贝叶斯。

(1)特征提取

特征提取指的是从所有词中，选取最有助于分类决策的词语。理想状态下所有词语都有助于分类决策，但现实情况是，如果将所有词语都纳入计算，则训练速度将非常慢，内存开销非常大且最终模型的体积非常大。

本系统采取的是卡方检测，通过卡方检测去掉卡方值低于一个阈值的特征，并且限定最终特征数不超过100万。

6 主题分析

作文是否符合题意也是评分的一个因素。现今的评分作文大多数的情景是，给定一段文本，考生自行阅读分析，根据给定题目制定文章主题。

主题分析即计算文章题目与文本内容的相关性。

6.1 文本相似度计算

6.1.1 基于词频的文本相似度计算

基于词频的文本相似度计算方式只考虑文本之间词语的重合程度。考虑到题目的文本一般较短，且由于中文表达的方式很多，不同的单词可表达相同的意义。如果简单的进行相同词的匹配会导致重合度过低。因此要对题目文本进行同义词库扩充。

(1)利用 TF-IDF 算法或 TextRank 算法计算题目文本的关键词集 W。

(2)对关键词集 W 进行同义词扩充。同义词库（///////todo）

(3)对文章进行分词处理，过滤掉停用词等操作。计算两个词语集的重叠度。

6.1.2 基于词向量的文本相似度计算

分析得知，部分作文词语丰富度低，滥用题目文本关键词，导致使用基于词频的文本相似度计算得分过高，与真实结果不符。因此引入一种新的计算方式：Word2Vec 来计算词之间的语义相似性。

6.1.3 基于篇章向量的文本相似度计算

计算文章与题目文本的相似度，可将文本表示为语义向量，计算两个向量之间的余弦相似度。采用 LDA 主题模型计算，LDA 主题模型包括词，主题和文档三层结构。该模型人唯一片文档的生成过程是：先挑选若干主题，在为主题挑选若干词语。最终，这些词语就生成了一篇文章。如果一个单词 w 对于主题 t 非常重要，而主题 t 对于文章 d 非常重要，那么单词 w 对于文章 d 就非常重要，并且在同类词中 w 的权重也会比较大。

6.1 关键字抽取

关键词可以代表一篇文章的核心内容，关键词抽取技术的任务是从待分析文本中自动抽取可反映文章核心内容的词语。抽取的方式一般有两种，一是通过训练语言模型实现，二是通过词语间关系从文本中抽取,，不需要事先对多篇文档进行学习训练。TextRank 算法是后者的典型实现方式，因无需训练，所以较为方便。

6.2.1 TF-IDF 算法

词项-文档矩阵，简称为 TF-IDF 算法。该矩阵是为了评价词项对文章的重要度出现的

文本相似度的应用及其广泛，除了匹配文字，还可以匹配音频，图片等。文本相似度计算的实质即计算范本之间的个体相似度。

TF，即词频，是词语在文本中出现的概率。计算文本相似度选用的词向量应是去除了停用词等对文章的中心思想无贡献的词语。计算的前提是先计算文本的关健词。因此，在词频 TF 的基础上又引入了反文档频率 IDF 的概念。不同的词应该有不同的权重，可以反映对文章的重要程度。对文章内容影响较小的词应赋予较小的权重，对文章内容影响较大的词应赋予较大的权重。TF-IDF 值反映了词语的权重，值越高权重越大。

下图是词频的计算方法：

////词频图

词频=某个词的词频/总词数

词频=某个词的词频/出现最多的词频

词频标准化的目的是把所有的词频在同一维度上分析。方式一样本过大，结果数值过小，各单词词频的差值不大，不便于分析。方式二更适用。

下面是反文档频率的计算方法：

$$反文档频率(IDF) = \log(\frac{语料库的文章总数}{包含该词的文档数+1})$$

TF-IDF = 计算的词频(TF)*计算的反文档频率(IDF)。即 TF-IDF 值与在该文档中出现的次数成正比，与包含该词的文档数成反比。

可用余弦相似度来计算文本相似度。余弦相似度即计算向量空间中两个向量夹角的余弦值作为衡量两个个体之间差异的大小。相似度的范围为 0-1。余弦相似度的特点是余弦值越接近 1，即向量夹角愈趋近于 0,向量越相似。TF-IDF 算法是根据词频来计算，余弦相似度的计算只反映了语句之间重合词的概率，而没有具体考虑语句的语义。

计算公式如下：

向量 a=(x1, x2, x3, x4, x5)，

向量 b=(y1, y2, y3, y4, y5),

余弦相似度=

(x1*y1)+(x2*y2)+(x3*y3)+(x4*y4)+(x5*y5) /sqrt(x1*x1+x2*x2+x3*x3+x4*x4+x5*x5)

由上述公式可知，计算语句的文本相似度步骤如下：

1.中文分词器将语句分为词语集合

2.计算两个词语集合的并集

3.计算各自词集的词频并把词频向量化

4.带入向量计算模型就可以求出文本相似度

计算文章的相似度步骤如下所示：

1.找出各自文章的关键词并合成一个词集合

2.求出两个词集合的并集

3.计算各自词集的词频并把词频向量化

4.带入向量计算模型就可以求出文本相似

TF-IDF 算法的优点是简单快速，符合客观情况。缺点是考量维度太多单一，只有“词频”作为衡量标准，其他因素没有考虑到，不够全面具体，例如词性，词语的位置。实际情况而言，词频相同但出现位置不同的单词的权重应是不同的。

6.2.2 TextRank 算法

根据 Google 的 PageRank 算法衍生而来，PageRank 算法是为了评估网页的重要性，以此作为搜索结果排序的重要依据。TextRank 算法将语句之间的关系看成是一种推荐或投票的关系，由此构建 TextRank 网络图，利用投票的原理，每个词语可以对自己的相邻词语进行投

票，票的权重取决与自己得票的高低。利用矩阵迭代收敛的方式进行收敛，选取得票最高的几个词语作为关键词。

RageRank 算法的公式如下：

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

///////解释

TextRank 算法的公式如下：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

///////解释

6.2 生成摘要

文章自动生成摘要与文章自动抽取关键字类似，即生成文章的关键句。模拟人类对文章语句的理解，给各句进行权重子打分，迭代选取得分最高的几个句子。自动摘要生成算法使用 TextRank 算法，算法如下：

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

WS 表示句子的权重，WS 等于相邻句子对本句的不同贡献求和。每个入度的代表的////////////

二 中文分词器

中文文章结构解析必然存在着分词的问题，即将文章解析为最小粒度-词语。 中文文章结构分析必须要用到分词算法，分词不是最终的目的，而是为后续的文章语义分析、篇章分析提供底层支持，更是中文文章处理的基础技术。中文文章的特点是相邻的词之间无明显界限，且要进行处理的是非结构化的自然语言文本，因此首先要将词分开来。通过分词算法进行扫描将之改变形态－词的集合。

１分词的考量维度

1.1 准确率

准确率是最为重要的考量因素之一，决定了分词系统的可靠性。例如，待分析文本为“李晓明是一位勤恳且热爱工作的老师”，分词结果应该是以下情况:”李晓明/是/一位/勤恳/且/热爱/工作/的/老师”，若出现其他情况，必定会对篇章结构的准确性造成一定的影响。

1.2 分词速度

中文词语多达成千上万，在语料库如此丰富的情况下，分词速度对于系统友好型的影响十分之大。若想提升分词速度，要选择适合大规模搜索的数据结构。

1.3 可扩展性

在词典库不够丰富的情况下，是否允许用户自定义词典，并且可根据分词器之前的分词表现进行自学习。

1. 4 词性标注

是否可根据上下文环境进行词性判别，在不同的上下文环境中，相同的词汇可能会有不同的词性。例如，下列文本：“教授正在课堂教授高等数学课程。”两处“教授”分别是名词、动词。分词器应具有良好的词性识别能力，不可将二处相同的词语混为一谈，这就要求词典中词语不止一种词性。

1.5 歧义词识别

//////todo

1.6 未登录词识别

未登录词不仅包括日常中的词语而未被收录进词典中，命名实体更是占了未登录词的很大一部分比例，而未登录词又在文本信息中有着很重要的位置。因此，未登录词的识别应该在分词中进行优化。命名实体主要包括：中文人名、地名、机构名、音译人名和日本人名等。

1.7 关键词提取

识别文本关键词可以分析出中文文本信息的主题，进而分析文章内容与标题的契合度，以此作为评分的一个维度。

2 数据结构

分词的前提是需要一个词语数量足够多且丰富的词库。理想的词库被认为是包含着现有的所有词，只靠字符串匹配即可在词库中查到所需的词。考虑到词库的容量大小，词语的存储结构和检索速度必然是重要的，存储结构以占用内存少为标准。检索速度以快为标准。根据这两个特性，引入一种 Trie 树以及双 Trie 树结构。

1.1 Trie 树

树形结构，又称单词查找树，一种哈希树的变种。典型应用是用于统计，排序和保存大量的字符串（但不仅限于字符串），所以经常被搜索引擎系统用于文本词频统计。它的优点是：利用字符串的公共前缀来减少查询时间，最大限度地减少无谓的字符串比较，查询效率比哈希表高。对于给定的字符串 b1,b2,b3,b4...bn,最多需要 n 次匹配即可完成查找，缺点是空间开销大且浪费率太高。

Trie 要解决的问题：词频统计与前缀匹配。

Trie 树的特性：

第一：根节点不包含字符，除根节点外的每一个子节点都包含一个字符。

第二：从根节点到某一节点，路径上经过的字符连接起来，就是该节点对应的字符串。

第三：每个单词的公共前缀作为一个字符节点保存。每个节点的所有子节点包含的字符都不相同

1.2 双数组 Trie

双数组 Trie 是基于 Trie 树的变形的出现是为了解决空间利用率的问题。

2　中文分词算法

现有的分词算法可以分为三类：

基于词典的分词，即事先准备一个可供分词的词典，根据词典对文本进行切分分词。按照扫描方向的不同，可分为正向最大匹配法、逆向最大匹配法、双向最大匹配法。

基于统计的分词，即相邻字，同时出现的频率高于与其他字组合，就可以认为这几个字组成了一个词。公式描述为：C(a,b)=P(ab)/P(a)P(b)，a、b 为不同的字，P(a)表示 a 在文本出现的概率，P(ｂ)表示ｂ在文本出现的概率。P(ab)即 ab 同时出现的概率。

基于理解的分词，即进行句法分析与语法分析消除歧义。消除歧义不是只能靠一部词典就可以解决的，进行句法以及语法、篇章分析的目的是通过自然语言处理模拟人类对语言的理解。

2.1 正向最大匹配算法 MM

基本思想如下所示：

(1) 通过标点符号的切割将文章分割为句子。

(2) 循环读入每一个句子。

(3) 已知词典中长度最长的词条的长度为 N，搜索长度为 M，处理待分析文本时从左向右选取前 N 个字符，起始位置 Begin=1，结束位置 End=N。

(4)文本开始搜索词典中有无该词条。

(4) 若词典中有该词条，Begin=Begin＋M，End=End+M，M=N，重复步骤(4)

(5) 若词典无该词条，N=N-1,End=End-1，重复步骤(4)

(6)若 Begin=End-1，且词典中无该词条，即该词条为单字

正向最大匹配算法的优点是代码简单，分词效率高，但错误率较高，经常出现歧义词切分，例如待分析发文本为：“学习机器的使用方式”，使用正向最大匹配算法会被切分为如下字符串：“学习机、器、的、使用、方式”。

2.2 逆向最大匹配算法 RMM

由于中文句子的重心一般在句子的后半部分，即前半部分大多是句子的主语等，后半部分是句子真正表达的意思，因此出现了逆向最大匹配算法。逆向最大匹配算法是正向最大匹配算法的逆向思维。基本思想如下：

(1) 通过标点符号的切割将文章分割为句子。

(2) 循环读入每一个句子，假设句子长度为 L。

(3) 已知词典中长度最长的词条的长度为 N，搜索长度为 M=N，处理待分析文本时从右向左选取后 N 个字符，起始位置 Begin=L-N，结束位置 End=N

(4) 文本开始搜索词典中有无该词条。

(5) 若词典中有该词条，Begin=Begin-M，End=End-M，M=N，重复步骤(4)

(6) 若词典无该词条，M=M-1,Begin=Begin+1，重复步骤(4)

(7) 若 Begin=End-1，且词典中无该词条，即该词条为单字

逆向最大匹配算法弥补了正向最大匹配算法不能识别歧义词的缺点。上文待分析字符串：“学习机器的使用方式”，若使用逆向最大匹配算法切分，则切分结果为：“学习、机器、的、使用、方式”，结果没有出现歧义词。

2.3 双向最大匹配算法

正向最大匹配算法与逆向最大匹配算法都遵循着“长词优先”的原则，即认为文本切分的结果中词语的数量越少越有可能是正确的。双向最大匹配法是将正向最大匹配法得到的分词结果和逆向最大匹配法的到的结果进行比较，从而决定正确的分词方法。研究表明，中文中 90.0%左右的句子，正向最大匹配法和逆向最大匹配法完全重合且正确，只有大概 9.0%的句子两种切分方法得到的结果不一样，但其中必有一个是正确的（歧义检测成功），只有不到 1.0%的句子，或者正向最大匹配法和逆向最大匹配法的切分虽重合却是错的，或者正向最大匹配法和逆向最大匹配法切分不同但两个都不对（歧义检测失败）。这正是双向最大匹配法在实用中文信息处理系统中得以广泛使用的原因所在。

2.4 最短路分词

基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)。采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

待分析文本切分成句，生成有向无环图，即对于待分析文本句子, 根据词典进行查询 , 生成几种可能的句子切分。

DAG 即有向无环图，记录的是句子中某个词的开始位置, 从 0 到 n-1(n 为句子的长度), 每个开始位置作为字典的键, value 是个 list, 其中保存了可能的词语的结束位置(通过查字典得到词, 开始位置+词语的长度得到结束位置)。

例如:{0:[1,2,3]} 这样一个简单的 DAG, 就是表示 0 位置开始, 在 1,2,3 位置都是词, 就是说 0~1, 0~2,0~3 这三个起始位置之间的字符。

利用动态规划算法查找最大概率路径，即从后向前扫描生成的有向无环图，根据句子的前半部分多是形容词与主语等对语义影响不大的部分，句子的重心一般落在后半部分，因此从后向前扫描的正确率要高于从前向后扫描。算法描述如下：

(1) 遍历 DAG 图，查找切好的每个词条 List 在词典中出现的频率 f(List)

(2)若词典中无该词条，则 f(List)=f(Min(词典))，即该词条的频率暂定为词典中出现频率最小的词条频率。

(3)从后向前计算最大概率，设某个节点为 N，则动态规划的递归公式为：P(N-1)=f(N-1)*Max()

///////todo 知网

3 未登录词识别

3.1 命名实体简介

命名实体，即人名/地名/机构名/音译名与日本人名等在未登录词中占了很大的比例。命名实体识别是自然语言处理的基本流程，在词性标注和分词过程中的重要环节，也在机器翻译/句法分析/信息检索/文章自动评分等领域有着重要的作用。中文文章的特殊篇章结构决定了中文分词的复杂性，因中文文本各词语之间无显著的分隔符等信息，命名实体与已登录词更是相互关联，你中我有，我中有你的关系。未登录词在文章中占了很大的比例，不考虑未登录词的解析，对文章的句法分析/篇章分析的质量也有一定的影响。但未登录词没有一定的规则，例如，人名或许与某些词重合，机构名或长或短，没有统一的模式，因此也较难分析。其中人名/地名/机构名的分析识别更是最为复杂的环节。

3.2 命名实体特点

命名实体的特点是数量众多且构成规律复杂，且各命名实体之间存在着相互嵌套的现象。这就表明了命名实体的识别不是独立的，而是各组成成分相互关联。例如，人名细分又可分为音译名与日本人名，中国人名;因为机构种类多样，机构名称的组成方式更是复杂，只有开头或结尾的词比较通俗相同。

3.3 已有的工作

未登录词的识别已有的方式分别是基于统计的方式与基于词典的方式。

基于规则的方式较为简单，一般是根据命名实体的特点来完成的。人名识别中主要是依据对现有人名姓氏进行登记，根据姓氏来识别是否为人名。机构名识别主要是依据对现有机构的结尾词进行识别，例如：XX 局，XX 中心，结尾的“局”与“中心”二字代表了本词是一个特定的机构。地名的识别主要是根据现阶段地名的特点来决定的，例如 XX 市 XX 县等。基于规则识别的优势是简单可行，缺点是识别的准确率不高，可扩展性较低。

基于统计的方式主要是通过对训练语料库和上下文进行分析统计，构建语言模型来进行命名实体的识别。可采用的模型有基于 Agent 的方法/隐马尔可夫模型/最大熵模型和基于类的三元语言模型等。

3.4 层叠隐马尔可夫模型

层叠隐马尔可夫模型旨在将各类命名实体的识别融合到一个相对统一的理论模型中。首先是在词语粗且分的结果集上，采用底层隐马尔可夫模型识别出普通无嵌套的人名/地名和机构名等，然后依次采取高层隐马尔可夫模型识别出嵌套了人名/地名的复杂地名和机构名。层叠隐马尔可夫模型试图在统一的隐马尔可夫模型中识别各类命名实体，并在这些模型中建立起一定的联系。此模型由三层互相联系的隐马尔可夫模型构成，自底向上分别为人名识别隐马尔可夫模型，地名识别隐马尔可夫模型和机构名识别隐马尔可夫模型。各层隐马尔可夫模型的以如下方式联结：

(1)每一层的隐马尔可夫模型都采用 N-Best 策略，将生成结构中 N 条最大概率路径作为参数输入到上一层的隐马尔可夫模型中。

(2)采用维特比算法(Viterbi)实现 N-Best 策略。

(2)第一层的人名识别是粗切分的分词识别，第三层的机构名识别将在第一层人名识别和第二层地名识别的基础上进行机构名识别。

3.4.1 基于角色标注的命名实体识别

角色标注的基本思想是：根据各类命名实体的组成结构与用词特点制定一套角色标注集

//////////////////todo

4 关键词识别

关键词可以代表一篇文章的核心内容，关键词抽取技术的任务是从待分析文本中自动抽取可反映文章核心内容的词语。抽取的方式一般有两种，一是通过训练语言模型实现，二是通过词语间关系从文本中抽取,，不需要事先对多篇文档进行学习训练。TextRank 算法是后者的典型实现方式，因无需训练，所以较为方便。

TextRank 算法是根据 Google 的 PageRank 算法衍生而来，PageRank 算法是为了评估网页的重要性，以此作为搜索结果排序的重要依据。TextRank 算法利用投票的原理，每个词语可以对自己的相邻词语进行投票，票的权重取决与自己得票的高低。利用矩阵迭代收敛的方式进行收敛，选取得票最高的几个词语作为关键词。

RageRank 算法的公式如下：

////// todo

TextRank 算法的公式如下：

///////

5 词性标注

词性标注即对于一个给定的词，根据文本上下文信息给予正确的词性标注，也即确定每个词是名词/动词/形容词或其他词性的过程。词性标注在文本信息处理中有着重要的意义，也在许多应用领域作出了巨大贡献，例如文本分类 文本索引、语料库加工 语言合成中，词性标注都是 必不可少的环节。汉语中词性标注较为简单，因汉语的每个词语大多只有一种词性，词性多变的情况较为少见。即便词语有多种词性，出现频次最高的词性的频率远远高

于其他词性的频率。现阶段研究词性标注所使用的语言模型有基于规则的方法与基于统计的方法两大类。

基于规则的方式即设计者在预先准备好的词典中对每个词条进行词性标注，这种方式解析简单，缺点是若词条有多种词性，分词器不得不面临着多种选择。例如，待分析文本：“教授在课堂教授学生们线性代数”，两处“教授”若不根据上下文进行处理，极有可能被解析为同一种词性，这是不正确的。根据最高频选取词性的方法，可高达 80%的准确率。

基于统计的分词方式弥补了基于规则的分词方式的缺点，可根据上下文解析来选择正确的词性。基于统计的分词方式采用一阶隐马模型。隐马尔可夫模型在计算待分析文本的最大概率词性标注序列时，不仅考虑上下文，还要依靠初始概率/转移概率，即依靠人工标注的训练语料来完成。且必须依靠大量的训练语料，基于隐马尔可夫模型的词性标注才能达到很高的正确率。

基于隐马尔可夫模型的词性标注过程：

(1) 训练语料，统计各单词词性出现频率，形成核心词典。

(2) 统计各标签的转移频次，得到转移矩阵。

(3) 利用转移矩阵和核心词典词频计算出隐马尔可夫模型的初始概率，转移概率，发射概率。

(4) 利用维特比(Viterbi)算法求得最大概率路径。

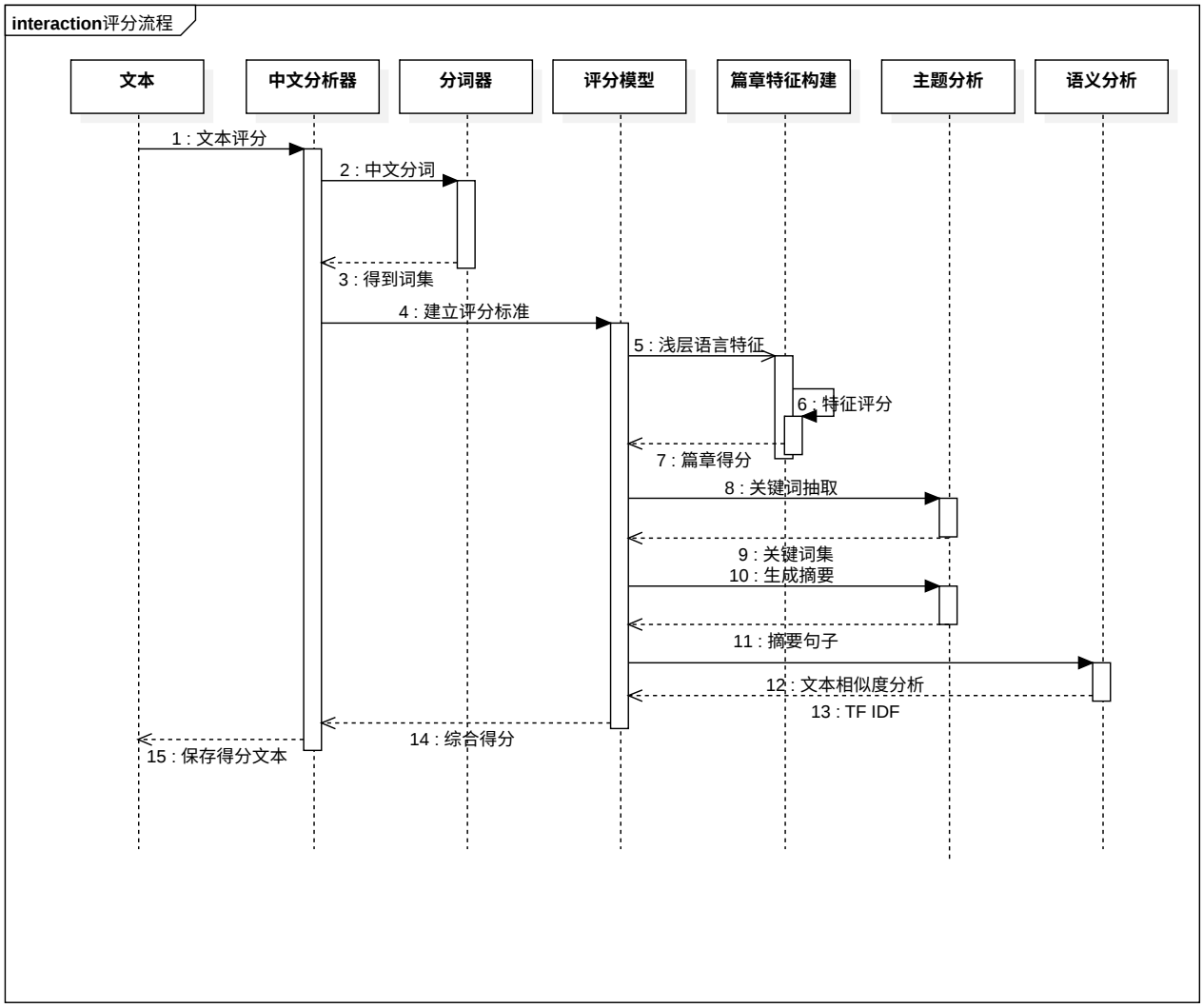
6 文本推荐

文本推荐即根据语义距离推荐出与本语句最相似的语句。

7 语义距离

四 系统设计与实现

1 概要设计

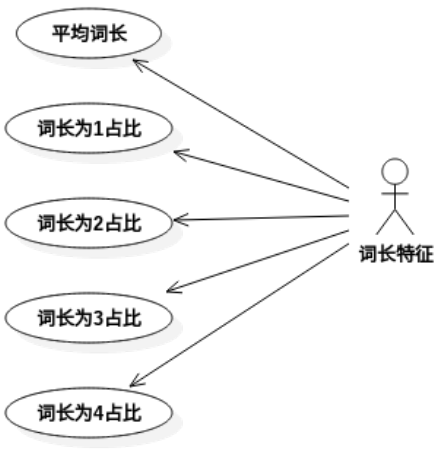


S

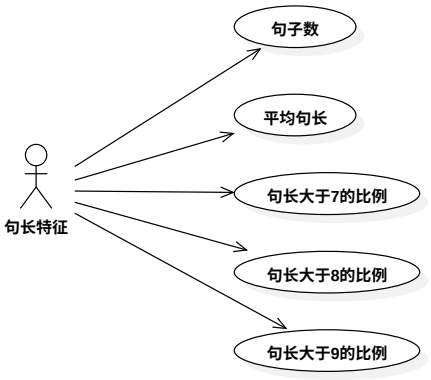
2 评分准则

2.1 特征分析

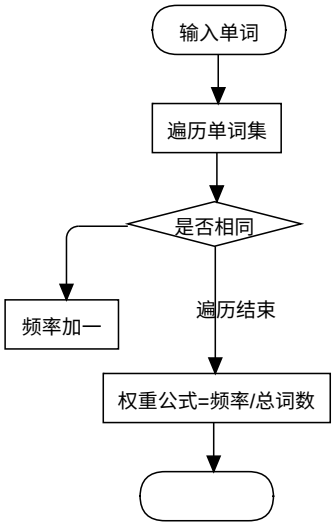
2.1.1 词长特征



2.1.2 句长特征



2.1.3 权重特征



4 类图

5 序列图

五 系统结果与测试分析

1 浅层语言特征

对三篇法律文章进行测试，得出以下结果。

1.1 句长特征

句长特征	文章一	文章二	文章三
平均句长	5.96	6.6	6.75
句长>7 的句数	11.0	6.0	9.0
句长>8 的句数	11.0	6.0	7.0
句长>9 的句数	8.0	5.0	4.0
句长>7 的比例	22.0%	30.0%	45.0%
句长>8 的比例	22.0%	30.0%	35.0%
句长>9 的比例	16.0%	25.0%	20.0%

1.2 词长特征

词长特征	文章一	文章二	文章三
平均词长	1.7	1.59	1.45
词长为 1 的单词数	131	64	83
词长为 1 的单词数	131	59	46
词长为 3 的单词数	19	8	4
词长为 3 的单词数	9	1	1
词长为 1 的词数占比	45.0%	48.0%	61.0%
词长为 1 的词数占比	45.0%	45.0%	34.0%
词长为 1 的词数占比	7.00%	6.0%	3.0%
词长为 1 的词数占比	3.0%	1.0%	1.0%

1.3 词性及分类特征

词性特征	文章一	文章二	文章三
单词权重	79.00%	70.63%	71.75%
单词重复率	5.13%	3.03%	28.88%
停用词比例	24.0%	31.06%	31.85%
未登录词	17.0%	14.39%	17.04%
明显错误词占比	17.0%	14.39%	17.04%
类符形符比	0.15%	0.07%	0.07%
动词的类符形符比	0.0268%	0.0115%	0.0127%
副词的类符形符比	0.0031%	6.3925%	0.0019%

介词数	12.0	8.0	7.0
代词数	3.0	0.0	0.0
总字数	496.0	210	196.0
句子数	49.0	20.0	20.0

2 主题特征

2.1 关键字

关键字	10	5
文章一	案件 监察 机关 查处 违纪 立案 季度 问题 处分 违法 案件 监察 机关 查处 违纪	案件 监察 机关 查处 违纪
文章二	成立 水 研究会 水法 研究 中国 建设 法制 北京 促进	成立 水 研究会 水法 研究
文章三	总统 法律 苏联 最高 达 课 卢布 诽谤 污辱 荣誉	总统 法律 苏联 最高 达

2.1 摘要

摘要	文章一	文章二	文章三
3	监察机关一季度立案查处行政违纪案件一万余件; 监察机关查处的违法违纪案件线索绝大部分来自; 监察对象违法违纪问题或案件线索５８０８９件;	的中国水法研究会今天在北京成立; 中国水法研究会在京成立; 即水法规体系、水管理体系和水执法体系	可课以最高额达２５０００卢布的罚; 苏通过维护总统荣誉的法律; 苏联最高苏维埃两院通过了一项新的法律——
1	监察机关一季度立案查处行政违纪案件一万余件;	的中国水法研究会今天在北京成立;	可课以最高额达２５０００卢布的罚;

3 分词器速度比较

/////

篇章分析：文本分类/情感分析

验证方法:交叉验证来验证分类器性能。将原始数组集进行分组，一部分作为训练集，一部分作为验证集。用训练集对分类器进行训练，用验证集来测试训练得到的模型。来作为评价分类器的性能指标。