第一章_德塔自然语言图灵系统

测试速度：单机联想 Y7000 笔记本 win10 实测峰值每秒 中文分词 1630~1650 万+中文字， 词库 65000+，函数准确率 100%，缺失语法函数 0.3%-， 算法准确率 99.7%+， 100%完整开放源码，在 api 与书籍中。

测试效果：输入：如果从容易开始于是从容不迫天下等于是非常识时务必为俊杰沿海南方向逃跑他说的确实在理结婚的和尚未结婚的提高产品质量中外科学名著内科学是临床医学的基础内科学作为临床医学的基础学科重点论述人体各个系统各种疾病的病因发病机制临床表现诊断治疗与预防

输出结果：如果+从+容易+开始+于是+从容不迫+天下+等于+是非+常识+时务+必+为+俊杰+沿海+南+方向+逃跑+他+说+的+确实+在理+结婚+的+和+尚未+结婚+的+提高+产品质量+中外+科学+名著+内科学+是+临床+医学+的+基础+内科学+作为+临床+医学+的+基础+学科+重点+论述+人体+各个+系+统+各种+疾病+的+病因+发病+机制+临床+表现+诊断+治疗+与+预防+++++

定义：德塔分词是一种-- 基于神经网络索引字典切割-- 进行前序遍历词性组合匹配-- 按文学语法定义搭配 的切词引擎。

德塔分词的催化切词优化方式主要包含：

1 索引字典进行细化拆分加速。

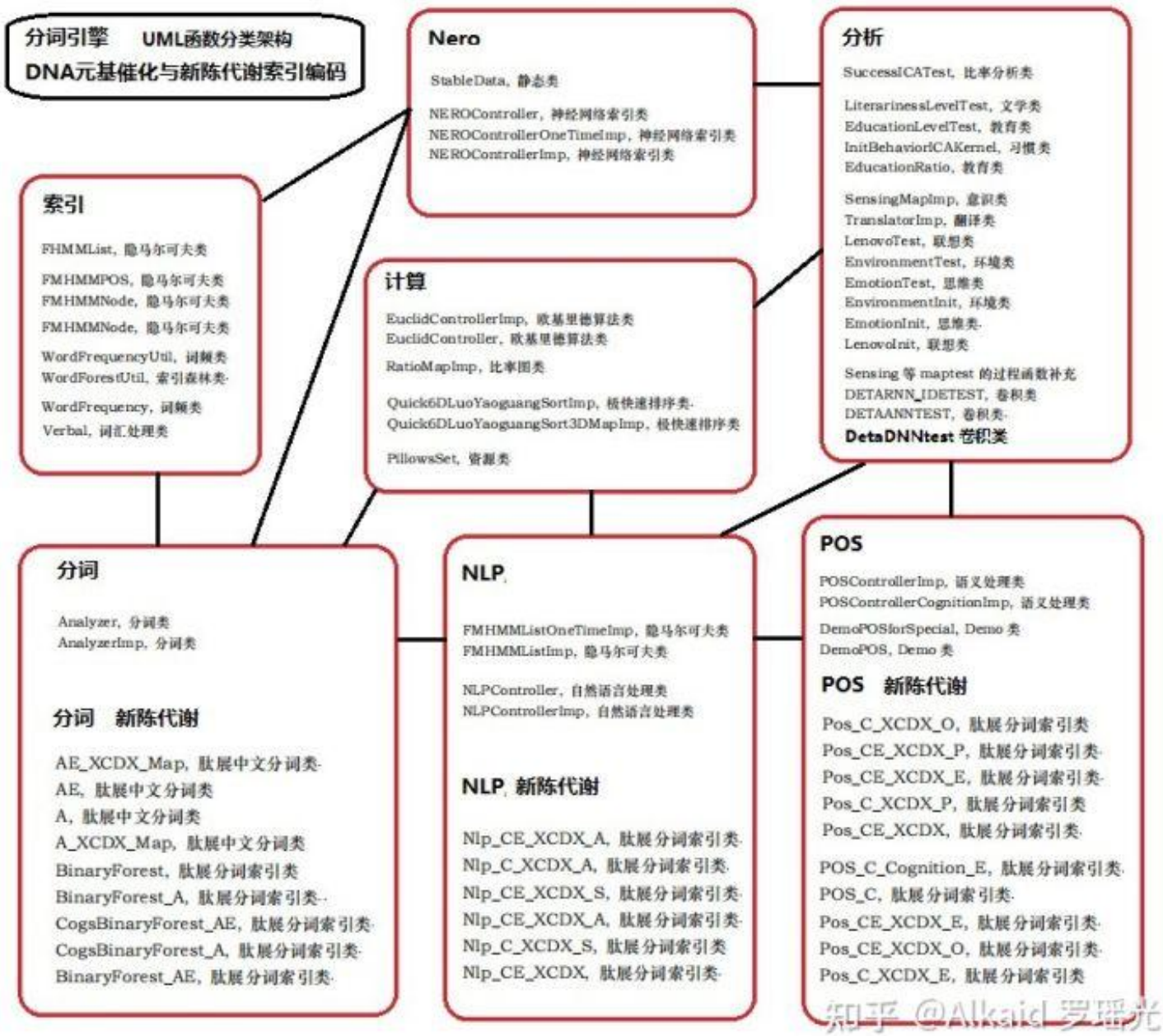
2 函数进行使用频率统计排列加速优化。

3 动态类卷积遍历内核的关键字优化。

4 函数文件和 函数文件名 进行新陈代谢，二次新陈代谢优化索引编码加速。

5 文学切词语法函数的细化优化加速。

定义者 罗瑶光



分词，

1 德塔的分词是一种前序《排队论》逐字遍历文字索引，通过索引中的词汇匹配 按长度进行提取，然后将提取的词汇串 进行词性切分的过程。refer page 12 ~

2 德塔的分词文字索引采用关联分类生成小文件 map 集（词性 map，词长 map，词类 map）， 进行整体加速，作为一个催化细化过程。refer page 44，54， 92，

3 德塔的词汇匹配目前有多个国家语言字符集，可统一，可拆分，目前最大划分处理长度为 4，划分切词采用动态类似 CNN 卷积（遍历 pos 函数语句的内核计算，非卷积的积分叠加计算） StringBuilder 核做 POS 识别。refer page 45，119,120，

4 德塔的词性切分按照 4 字词 3 字词 2 字词 单字 进行逐级按词汇的 POS 搭配语法模式进行归纳，按文本的 POS 出现频率进行流水阀门方式优化。refer page 97,116，

（德塔分词逻辑， 已经纠正红色字 ‘卷积’改为‘内核’，因为第四修订版本已经在申请中，ppt 所有书中的原图纠正内容统一更新在第 5 版，罗瑶光）

排序，

1 德塔分词排序思想原型采用 Sir Charles Antony Richard Hoare 的 快速排序思想。

refer page 版权原因无文字收录 已经 refer 快速排序算法_百度百科

2 德塔分词排序源码原型采用 Introduction to Algorithms 的 快速排序 4 代源码。

refer page 版权原因无源码收录 已经
refer https://github.com/yaoguangluo/Data_Processor/blob/master/DP/sortProcessor/Quick_4D_Sort.java

3 基于 1 和 2 原型，德塔分词排序 采用 Theory on YAOGUANG's Array Split Peak Defect 的微分催化算子优化思想 2013 年开始优化。refer page 247,248,250,529,620，

4 优化过程为 小高峰左右比对法， 波动算子过滤思想，离散条件归纳微分思想(如狄摩根计算，流水阀门计算等)，目前为 TopSort5D。refer page 658，下册 134

5 德塔分词的函数优化方式和算法优化方式，包括分词引擎，读心术，NLP 分析等核心组件均采用 微分催化系统。refer page 661，

神经网络索引，

1 德塔分词的词汇字典用 map 进行索引，因为 jdk8+的 map 对象的 key 支持 2 分搜索，搜索速度到了峰值。refer page,129,131

2 德塔分词的索引不断的将大 map 进行细化分类，如词长 map，词类 map，词性 map，让搜索再次加速。refer page 55，

3 德塔分词的索引 map 支持 2 次组合计算，支持分布式服务器进行索引 cache。关于 2 次组合计算作者不建议单机使用。refer page 92，

4 德塔分词 map 的 key 用 string 的 char 对应 ASCII int 进行标识来执行 find key，方便二分搜索存储和 StringBuilder 高速计算，实现底层核统一。refer page 92

分词在线性文本搜索中应用，

1 德塔分词的搜索建立在 map 类的权重计算方法上，不同的权重叠加产生的打分进行排序输出。refer page 下册 64

2 权重的计算方法按词性的主谓宾如代 名动形 ，和 POS 如 动名形谓介分类。refer page 下册 66

3 权重与词长，词频进行耦合 bit 叠加计算(bit 位计算比乘法要快一个数量级)，生成最终输出结果。 refer page 下册 68

4 权重与词长的 比值可以精度调节，确定搜索的精确性和记录个人搜索偏好。refer page 下册 68

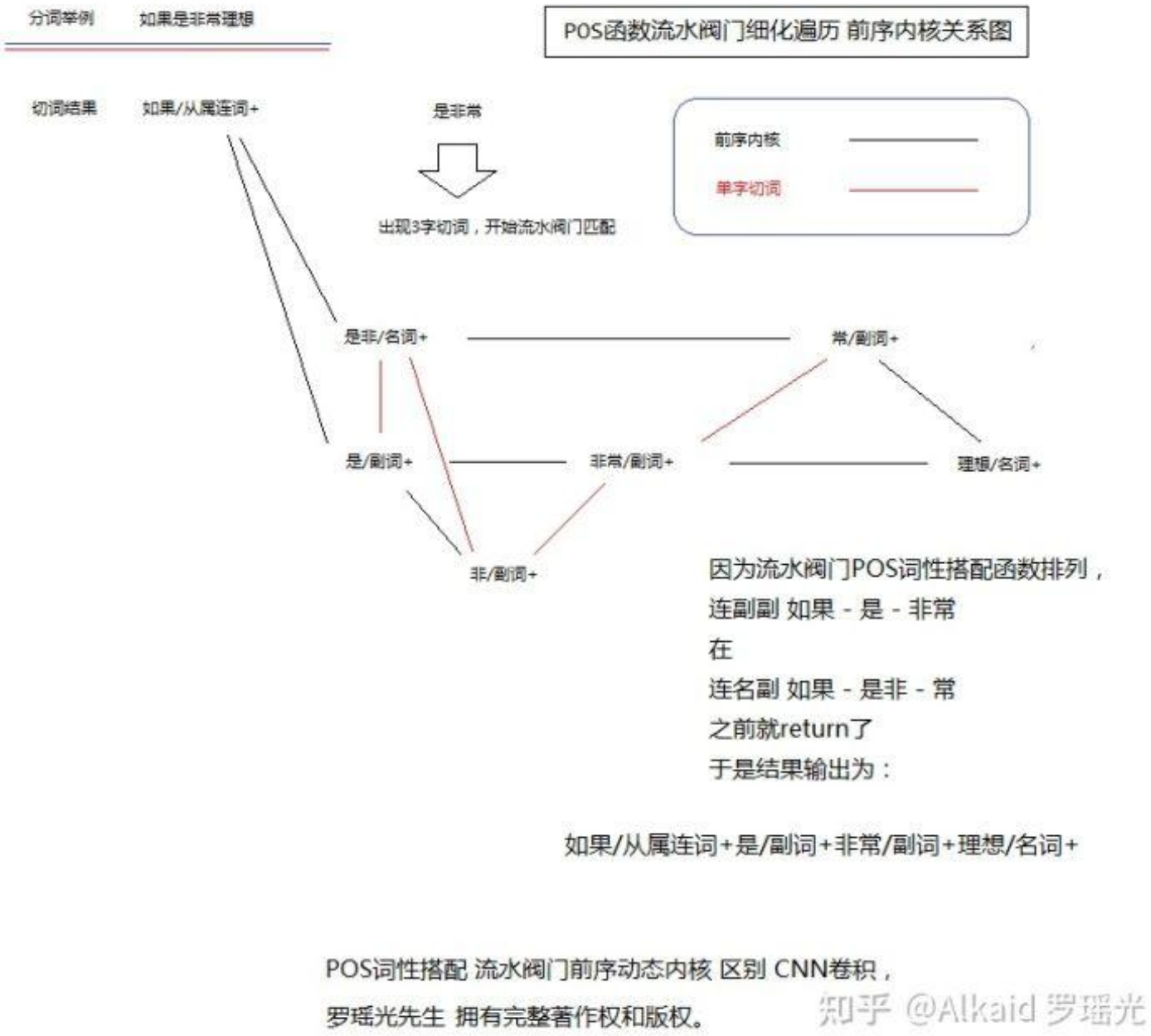
动态 POS 函数流水阀门细化遍历 内核匹配，

1 动态的核分为前序核和后序核两种。根据词汇分析的位置进行实时变动更新。refer page 97

2 前序核主要缓存存储词汇的位置和词性，用于 POS 词性搭配的 POS 函数流水阀门细化遍历 计算。refer page 97

3 后序核主要缓存词汇的切词链 后面准备 跟进的词语。用于 POS 语法的修正计算，如连词匹配。refer page 97

4 内核采用 StringBuilder 做核载体进行计算加速。refer page 97



2019 年 3 月 18 日之前作者 Github 的 该算法函数编码框架已经出现

https://github.com/yaoguangluo/Deta_Parser/commit/25b90c9847d15df85c5c991448f2c271e0ad8106

注意：链接的 **CNN 关键词**的 历史记录 属于作者用词错误，作者当年基础学术累积不够，关于卷积的知识仅仅学了**计算机视觉**的理论课，以为带内核计算的都叫 CNN 卷积，

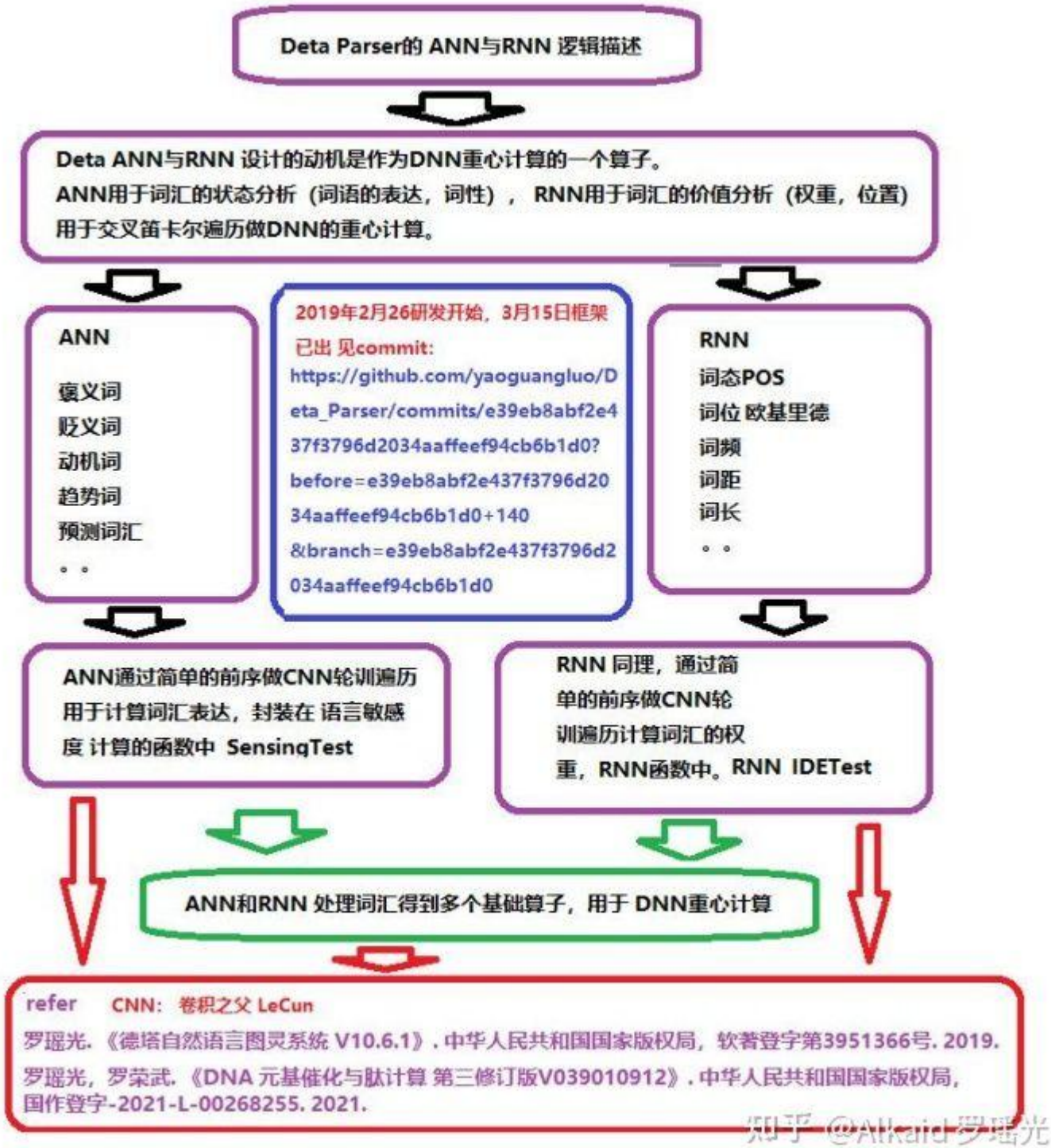
另外作者发现自己还有一个错误， 就是以为序列链表方式计算就叫隐马科夫链计算。所以 **CNN+隐马可夫**这两个技术词汇，伴随作者 10 年之久。今天进行 ppt 严谨定义，翻阅大量定义文献资料，才发现这些错误。予以纠正。作者的 ANN 和 RNN 出现的文本分析内核计算才是真正的 CNN 卷积计算。

POS,

1 德塔分词的核心类，包含了词性的搭配切分所有函数。refer page 97,116

NLP,

1 德塔分词的核心类，包含了词性的词长切分所有函数。refer page 119，120



ANN,

德塔词性的卷积计算 ANN，主要包含意识比率算子,环境比率算子,动机比率算子,情绪比率算子。这个四个算子的组合计算产生了一些高级决策,如 情感比重,动机比重,词权比重,持续度,趋势比重,预测比重,猜想比重,意识综合。这些决策在文本分析的领域可以拥有实际评估和决策的价值。同时意识综合 summing 也是德塔 DNN 计算的一个输入参数组件，用于文本中心思想词汇标识计算。

1 词性卷积计算 refer page 182

2 用于确定文本的中心

2.1 算子组成

2.1.1 S SENSING 意识比率

2.1.2 E ENVIRONMENT 环境比率

2.1.3 M MOTIVATION 动机比率

2.1.4 E EMOTION 情绪比率

refer page 18

RNN,

德塔的词位卷积计算 RNN, 主要包含词性比率, 词距比率算子和欧基里德熵算子。这三个算子主要用于求解 POS 距离, COVEX 距离, EUCLID 距离.这些权距 在一篇文章中能够很清楚的计算每一个词汇的使用度，出现的价值，和应用频率以及分布规律。用于文本的主要描述语句的重心所在位置计算。

1 词位卷积计算 refer page 178

2 用于确定文本的重心

2.1 算子组成

2.1.1 P POS 词性比率

2.1.2 C CORRELATION 词距比率

2.1.3 E E-DISTANCE 欧基里德熵

refer page 18

DNN,

德塔的词汇深度计算 可以理解为 德塔词性的卷积计算 ANN 与 德塔的词位卷积计算 RNN 的前序笛卡尔卷积计算。因为参数 由 文章中心思想 和 文章的重心词位 两类组成，因此适用于分析和计算 文章的 核心思想词汇的价值

德塔 DNN 词汇花展示

Alkaid 罗瑶光的视频

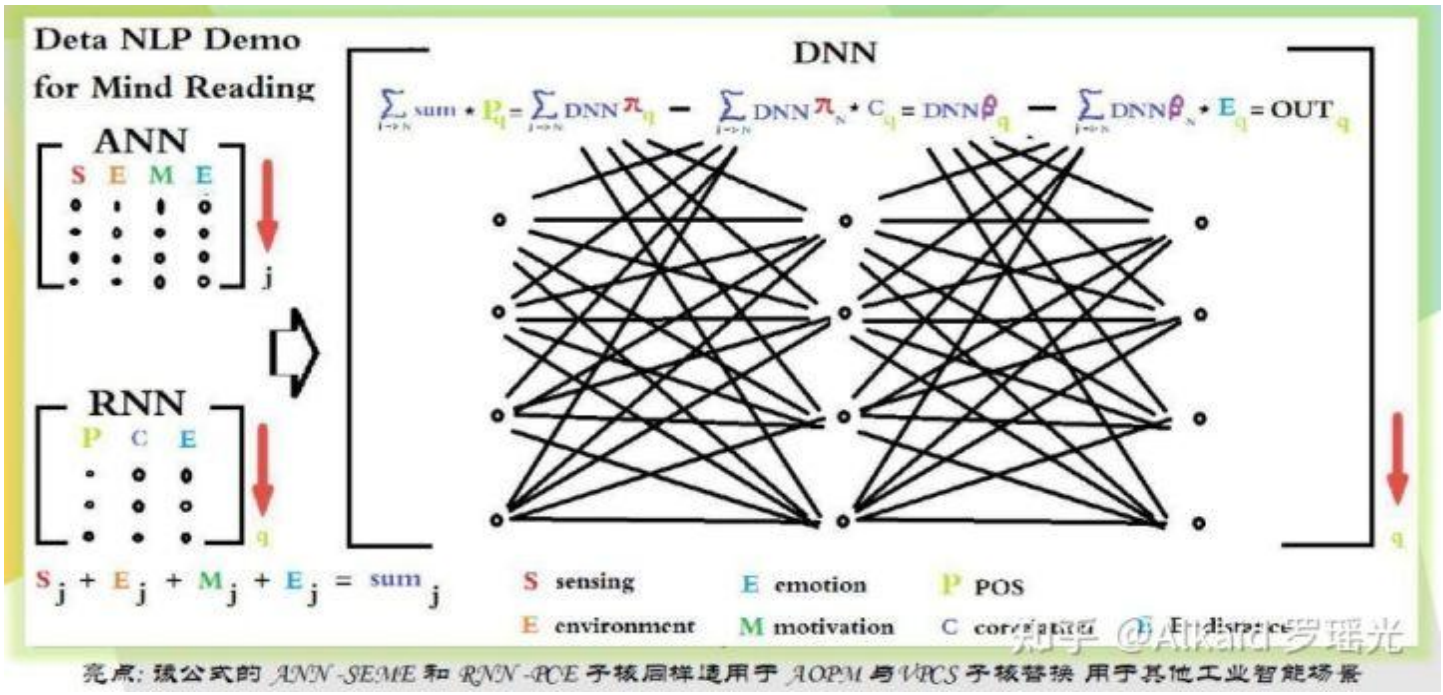
- 124 播放

1 词汇深度计算 refer page 183

2 用于确定文本的核心

2.1 深度计算 (ANN sum 核 -> RNN PCE)

refer page 18



图灵机，

1 文学分析 refer page 168

德塔文学分析主要用于文章的思想分析和挖掘，如确定多语意识的场景，当时的环境，动机，意识形态倾向和决策思维表达等。（**多语意识 ：通过人物的对话方式，语言特征，模式场景等因素 来 分析当时的人文情感，大众思想，从而了解所处时代的民族风情，社会建筑，时代背景。 教授人：作者导师白育芳，2007 年，总参解放军炮兵学院南京分院。**）

2 作品评估 refer page 167

德塔作品评估 可理解为教育程度评估，如语法，词汇的词性统计，专业词汇的统计，成语，三字词的词长词汇的统计，等等。如一个句子中含有的高级词汇的比率，4 字名词的比率，形容词的比率。（作者最早意识出现在 **2009 年** 在上海章鑫杰那 处理**法国 ESIEE 亚眠大学**的法语邮件项目， **Pascal 教授**曾传授作者关于 FLECH 法语

元音比重单词分析的表述。设计这个项目，进行了灵感发散。**德塔图灵分词全文没有任何单词分析和 非中文的语言分析，不涉及 flech 任何思想和逻辑，因此一直没有 refer。** 作者拥有完整著作权和版权）

3 动机分析 refer page 169

德塔动机分析 基于动机词典的 map key 匹配 进行决策表达。比较简单。因为词典定义 带有作者个人主观思维特征。所以没有太多描述。

适用 3，4，5

4 情感分析 refer page 159

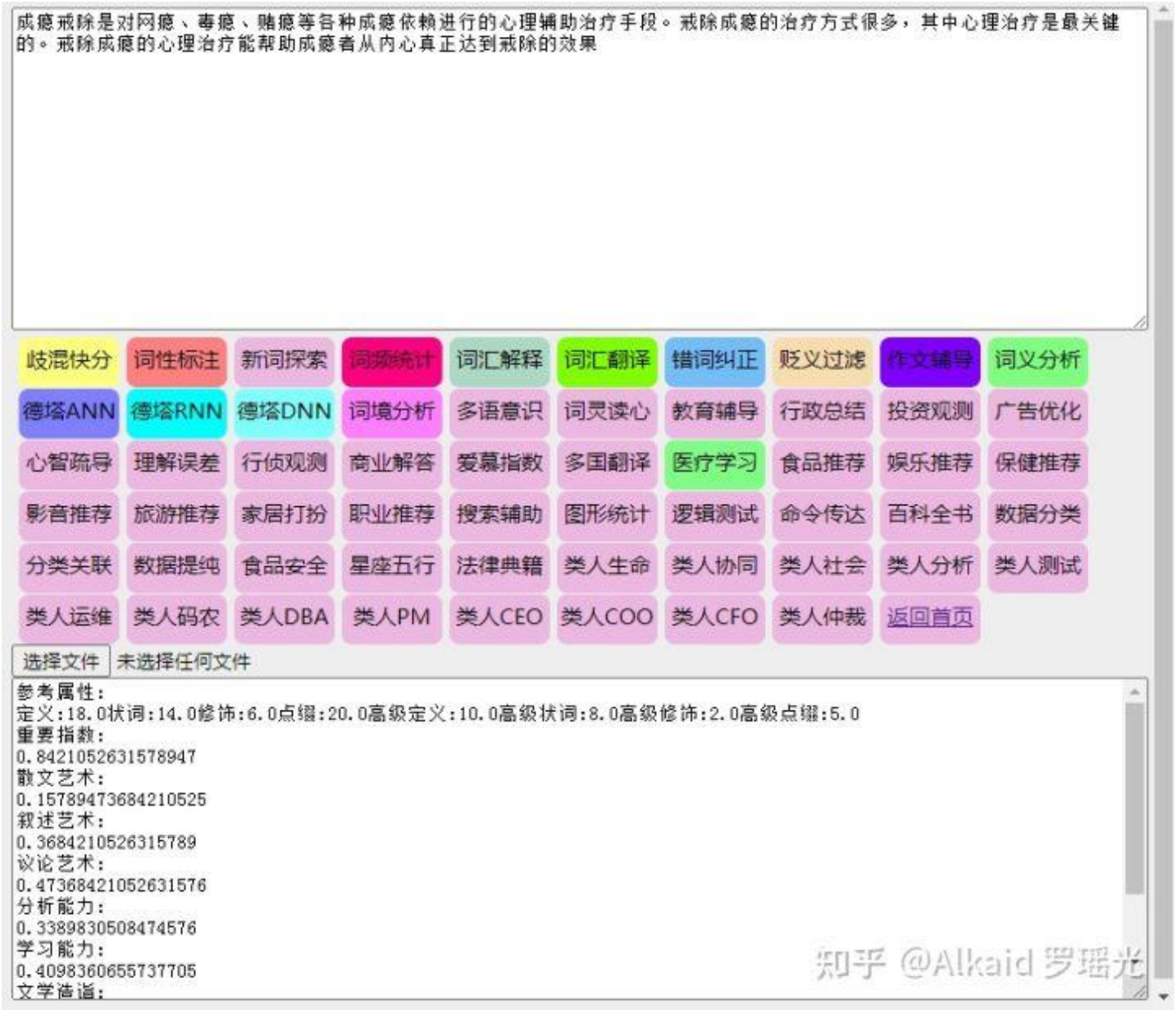
德塔情感分析 基于 褒义词 贬义词 和中性词 的 map key 匹配 进行决策表达。比较简单。因为词典定义 带有作者个人主观思维特征。所以没有太多描述。

5 习惯分析 refer page 169

德塔习惯分析 基于 褒义词 贬义词 和中性词，动机词， 文学分析数据，作品评估比率，教育程度等数据 的全文比重，来确定一个人写作特征，和写作习惯。写作风格。因为词典定义 带有作者个人主观思维特征。所以没有太多描述。

6 教育程度评估 refer page 168

德塔教育程度评估体现在文章中的（**有效词汇**如词长超过 2 位）的 （**有价值词汇**如名动形谓状）的全文，全句，其它 POS 词性的比率来确定文章的句法特征。举个简单的例子，一个句子中有效有价值的形容词比重大的文章通常代表作者的分析表达和散文修饰能力比较强势。，思维来自作者初中语文学习。



应用

涉及著作权文件：

1.罗瑶光. 《德塔自然语言图灵系统 V10.6.1》. 中华人民共和国国家版权局，软著登字第 3951366 号. 2019.

2.罗瑶光. 《Java 数据分析算法引擎系统 V1.0.0》. 中华人民共和国国家版权局，软著登字第 4584594 号. 2014.

3.罗瑶光，罗荣武. 《类人 DNA 与 神经元基于催化算子映射编码方式 V_1.2.2》. 中华人民共和国国家版权局，国作登字-2021-A-00097017. 2021.

4.罗瑶光，罗荣武. 《DNA 元基催化与肽计算第二卷养疗经应用研究 20210305》. 中华人民共和国国家版权局，国作登字-2021-L-00103660. 2021.

5.罗瑶光，罗荣武. 《DNA 元基催化与肽计算 第三修订版 V039010912》. 中华人民共和国国家版权局，国作登字-2021-L-00268255. 2021.

6.罗瑶光. 《DNA 元基索引 ETL 中文脚本编译机 V0.0.2》. 中华人民共和国国家版权局，SD-2021R11L2844054. 2021. （登记号:2022SR0011067）软著登字第 8965266 号

7.类人数据生命的 DNA 计算思想 Github［引用日期 2020-03-05］ https://github.com/yaoguangluo/Deta_Resource

8.罗瑶光，罗荣武. 《DNA 元基催化与肽计算 第四修订版 V00919》. 中华人民共和国国家版权局，SD-2022Z11L0025809. 2022.

文件资源

1 Jar： https://github.com/yaoguangluo/ChromosomeDNA/blob/main/BloomChromosome_V19001_20220108.jar

2 UML： DNA 元基催化与肽计算 第四修订版 V00919

3 PPT: https://github.com/yaoguangluo/ChromosomeDNA/tree/main/ppt

4 Book:《DNA 元基催化与肽计算 第四修订版 V00919》上下册

https://github.com/yaoguangluo/ChromosomeDNA/tree/main/元基催化与肽计算第四修订版本整理

5 函数在 Git 的存储地址：Demos

Github：https://github.com/yaoguangluo/ChromosomeDNA/

Coding：公开仓库

Bitbucket：Bitbucket

Gitee：浏阳德塔软件开发有限公司 GPL2.0 开源大数据项目 (DetaChina) - Gitee.com

6 其它资源链接:

ZHIHU DNA 元基催化与肽计算第四修订版

CSDN DNA 元基催化与肽计算 UML 集_罗瑶光 19850525 的博客-CSDN 博客

CSDN DNA 元基催化与肽计算 第四修订版 V00919