机器学习业务实践之路

课程7：文本分析-文本关键信息抽取、相似度分析

阿里云 李博（傲海）

目录

本节文本相关算法用途

1.关键词抽取：抽取能代表文本语义的关键词，常用来做文本内容的标签。

例：“中国队男子足球队，经过一场激烈的比赛战胜了巴西足球队。”

关键词：足球 中国队

2.文本摘要：提取可以概况文章内容的语句。

原文：就央视报道"70名大三学生本科学历变专科"事件，今天上午首师大召开新闻发布会，称此事为中介公司招生虚假宣传，首师大及美术学院不知情，将起诉该中介公司东方致远公司。另外通知书上美术学院公章不是原章。(记者杜丁)

摘要：70名大三学生本科学历变专科首师大将起诉中介公司

3.文本相似度分析：判断文章间的语义相似度

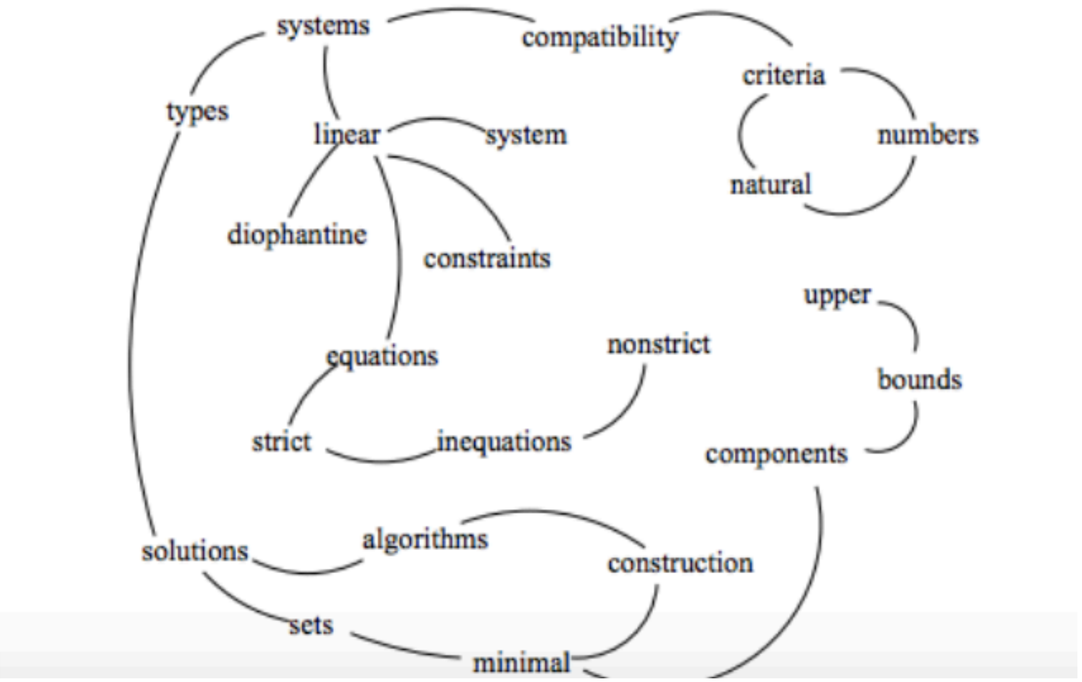
关键词抽取算法

关键词抽取：抽取目标样本中的关键词。

关键词抽取算法：

1.TF-IDF
2.LDA
3.基于Graph的关键词抽取
……

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.



文本摘要算法

文本摘要：提取可以概况文章内容的语句。

实现形式：

1.返回文本中的关键句，从原文中抽取关键的语句返回。

2.解析原文的主、谓、宾等关键词，基于语义自动生成摘要概况原文语义。

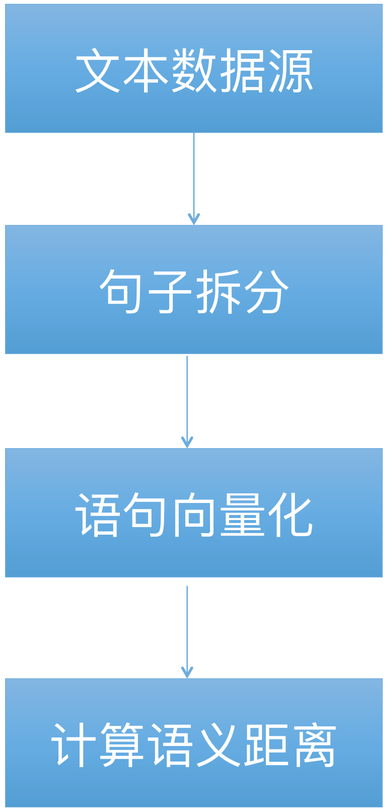
文本相似度分析

文本相似度分析：分析比较文章间的语义相似度。

实现方式：

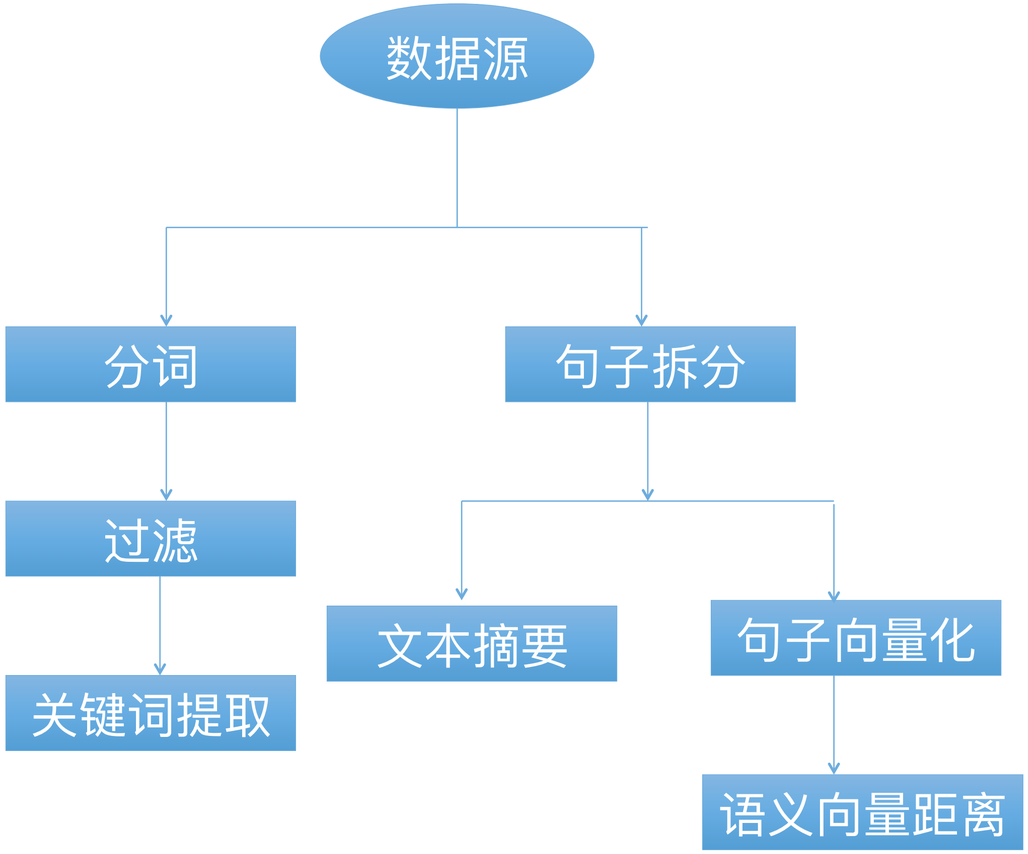
第一步：将文本按照语义向量化

第二步：通过向量距离判断文本的语义距离

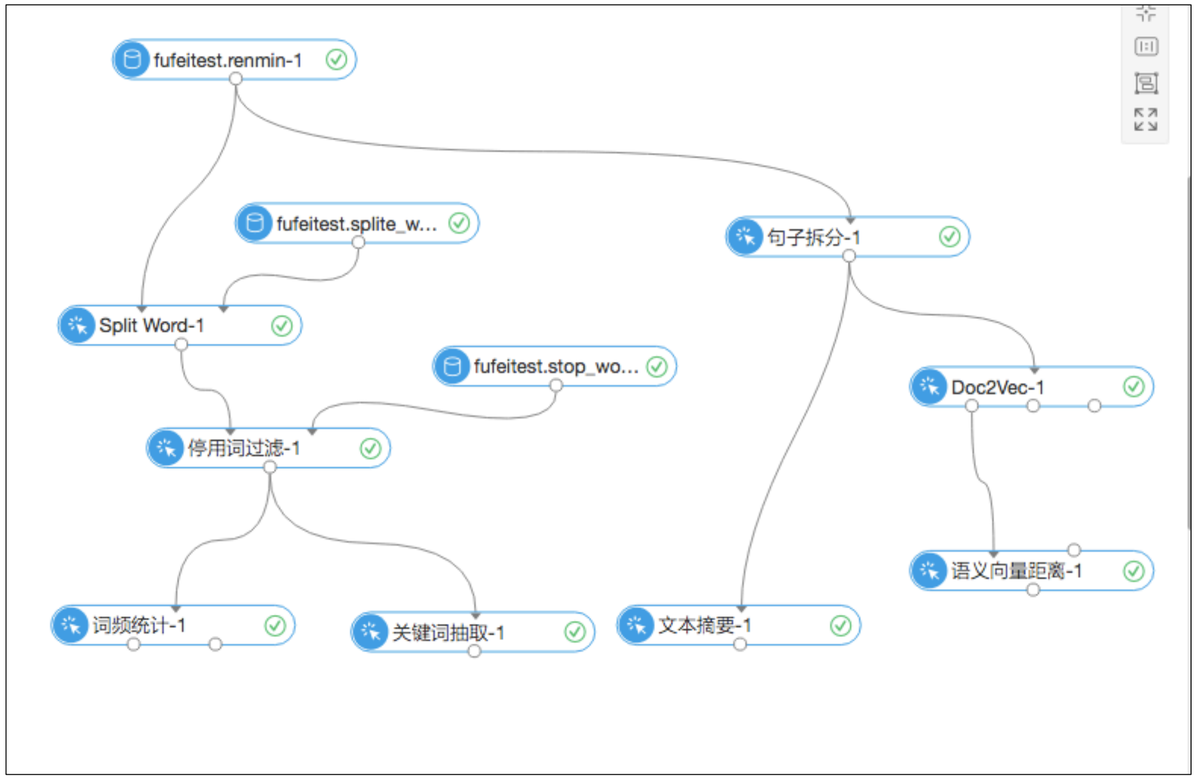


PAI平台实现文本算法

算法流程图



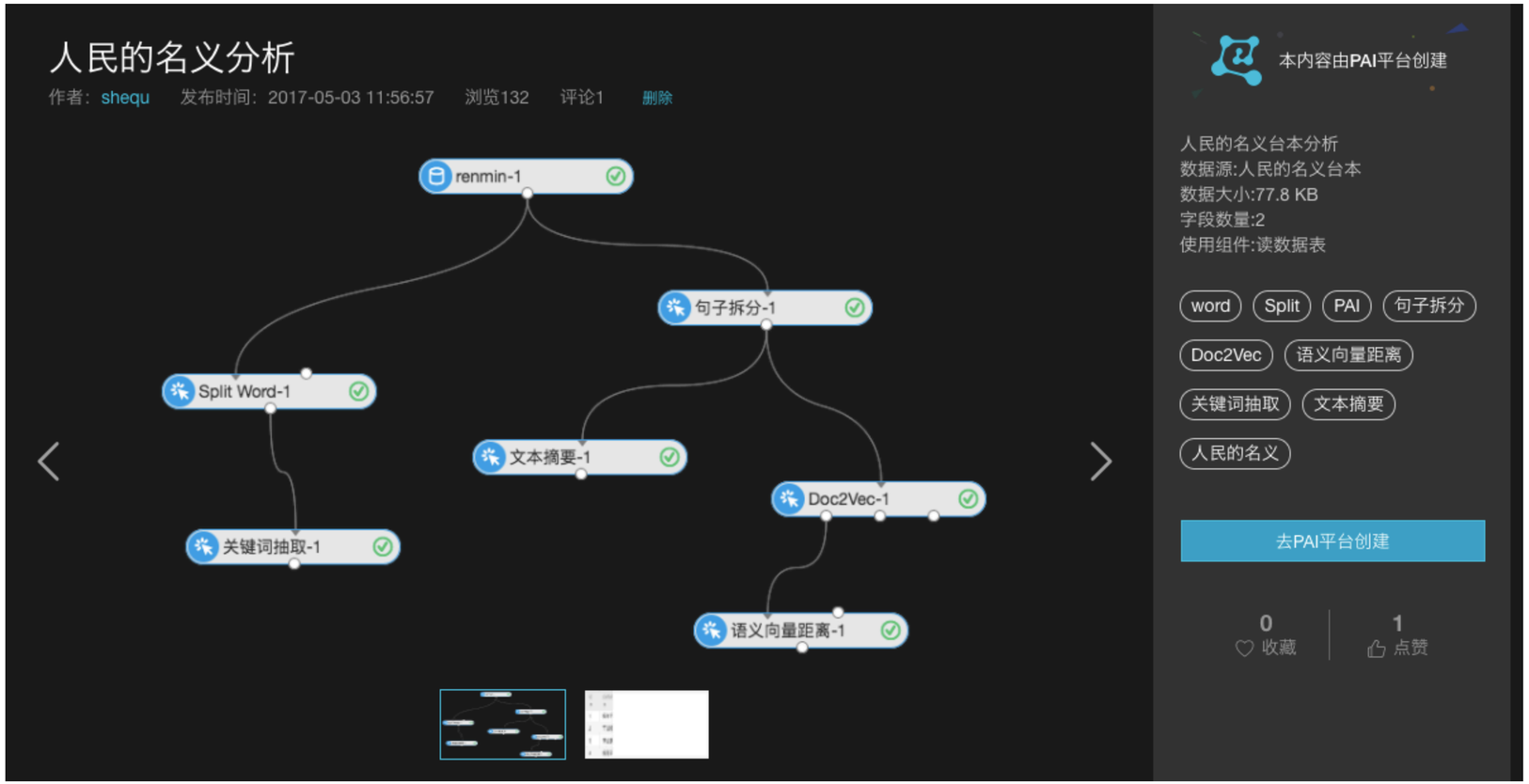
PAI平台架构图



基于主题算法的新闻分类

社区案例地址：
https://yq.aliyun.com/articles/75305

需要补充停用词过滤模块。



相关资料

推荐学习材料：

- 《机器学习实践》
- 《统计学习方法》
- 吴恩达的机器学习相关课程

推荐实验环境：机器学习PAI https://data.aliyun.com/product/learn
相关论文：
https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
https://arxiv.org/pdf/1509.00685.pdf

我的个人微信公众号（与我交流）：凡人机器学习

为了无法计算的价值 | 阿里云