人工智能与NLP 5期 项目1 分析报告

周易华，李岩，李湘湘，吴林莉

2019.12.05

1. 简介

本项目使用了Word2vec的语言模型，结合SIF（Smooth Inverse Frequency），加上KNN平滑等技术，实现了一个简单的新闻摘要提取系统。

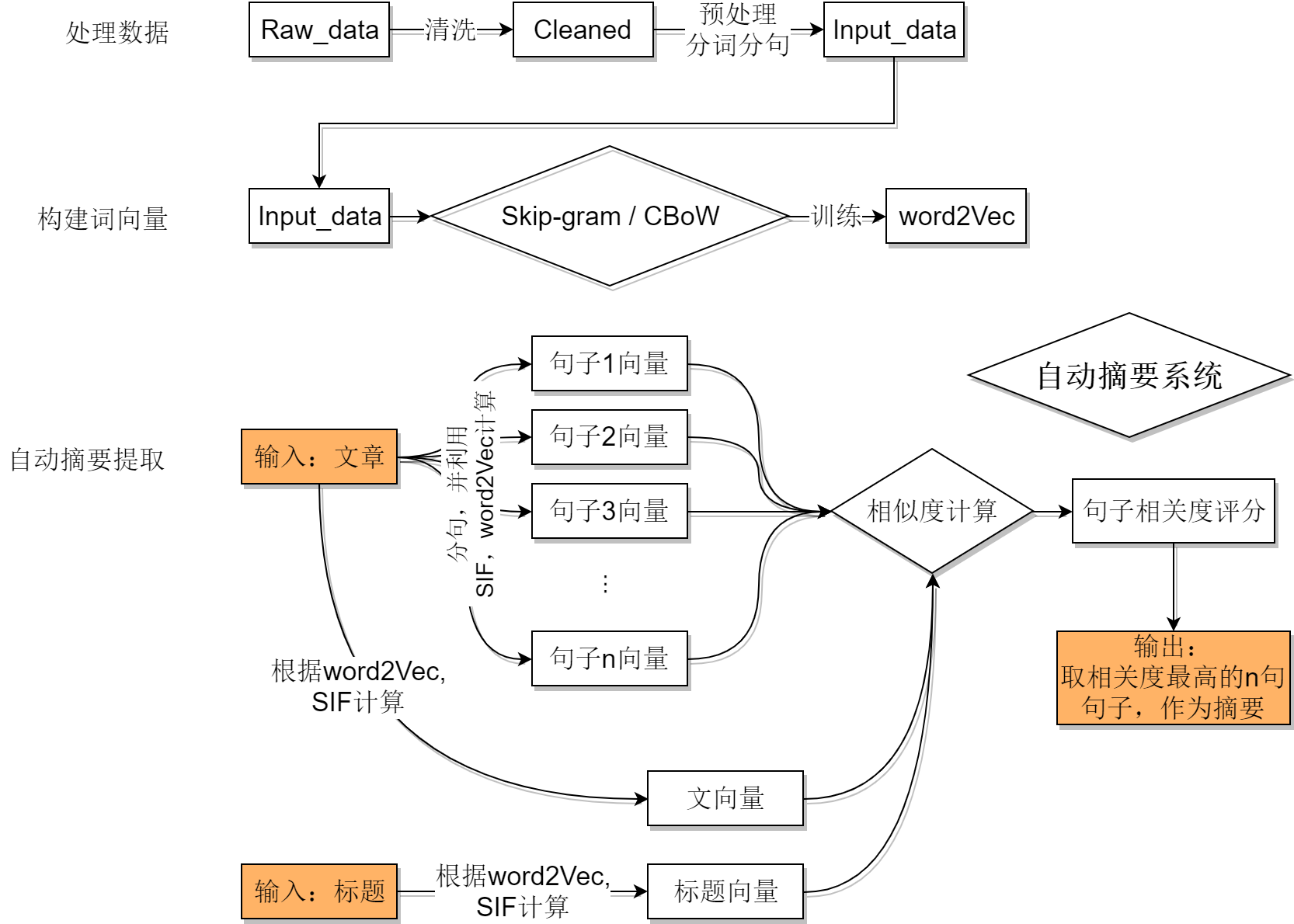


图1. 项目流程图

1. 项目的优缺点

项目的**优点**：

1. 计算速度快，占用资源少：word2Vec是预训练的，SIF句向量部分的计算是无监督的，并且计算也不复杂，因此模型简单快速效率高
2. 模型比较灵活，对输入的文本长度没有限制
3. 基于句子筛选的摘要，由于对句子本身不作改变，所以一般来讲不会出现逻辑错误，句子内容错误等重大问题，且具有较好的连贯性和通顺性

项目的**缺点**：

* 1. 模型简单，上限有限
  2. 采用的是无监督方法，且缺乏评价指标来评估sen2Vec，以及摘要的好坏
  3. 基于句子筛选的摘要，由于对句子本身不作改变，所以对于句子来讲就没有办法继续精简了，如果正文中的句子本身较长，那么摘要中的句子也会较为冗长
  4. 对于长文本新闻，模型的相似度分布较为均匀，没有考虑句子与摘要的注意力机制（例如文章的首尾段落应该获得较高的注意力，这点可以使用监督学习来改进）

1. 总结与未来工作

总的来说，这个项目采用了Word2vec，SIF模型，加上KNN平滑等技术，实现了一种简单快速的摘要提取功能，对于大部分的语料效果都还不错，但是对于特定类型的语料例如访谈，辩论等，效果一般。

针对不同方向，未来可以做出如下改进：

本模型调优方面：

* 1. 针对开头和起始段落可能包含更多主旨信息的事实，增大开头及首尾段落的比重

模型变更及改进方面：

1. 考虑使用seq2seq的有监督方法来训练sentenceVec生成器，通过有效降低损失来使生成的sentenceVec更加准确和可解释
2. 进一步，可以考虑使用Transformer模型例如BERT来进行训练sentenceVec生成器，从注意力分配的角度理解句子与摘要的关系