大数据分析第二次作业

本次作业共分为几个部分：

1. pdf文件处理
2. 关于分词
3. 关于关键词提取的算法
4. 关于结果文件的写入
5. pdf文件处理

使用了pdfminer这一python库进行了pdf文件的读取

1. 分词

因为所给论文均为英文文本，所以使用了nltk这一较为成熟的分词以及词汇处理的库来进行stopwords的剔除以及无关词语（数字等）的处理以及词性的还原。

1. 关于关键词提取的算法

共使用了七种算法，但最终只有4种算法得出了结果，因为textrank和rake基本只适用于短文本的处理，而word2vec由于词库问题无法写出最后的关键词。

1. tf-idf:

tf即词频，idf即反向文本的词频，tf\*idf即可较客观的产生对应的权重。

1. lda模型

使用lda模型进行权重计算

1. lsa(i)模型

使用lsa模型进行权重计算

1. hdp模型

使用hdp模型进行权重计算

1. textrank

类似于pagerank，根据上下文构建对应的向量矩阵来进行权重的计算。

1. rake:

使用对应的词在短语和句子中出现的次数来进行打分，来实现word权重的表示，最终找出关键词

1. word2vec

使用语料库以及对应的模型进行对应的词向量的权重计算，最终再计算得分进行排名

1. 关于结果文件的写入

在代码中直接创建文件以及将结果写入。