设计说明书

本项目主要包含两个模块，分别是爬取中文网页的爬虫和对所爬取的网页进行聚类的K-means聚类模块。

# 爬虫设计

## 中文网页源选择

选取了 10 个具有明显特征的新闻频道，分别为：

* <https://www.myzaker.com/channel/9> 娱乐频道
* <https://www.myzaker.com/channel/7> 汽车频道
* <https://www.myzaker.com/channel/8> 体育频道
* <https://www.myzaker.com/channel/13> 科技频道
* <https://www.myzaker.com/channel/3> 军事频道
* <https://www.myzaker.com/channel/10386> 美食频道
* <https://www.myzaker.com/channel/12> 时尚频道
* <https://www.myzaker.com/channel/981> 旅游频道
* <https://www.myzaker.com/channel/10376> 游戏频道
* <https://www.myzaker.com/channel/10530> 电影频道

这样，我们就可以得到真实的簇为 10 个，确定 K = 10。

## 基于生产者-消费者实现爬虫

爬虫基于多线程，使用生产者-消费者实现。其中，一个生产者负责从网页中提取新闻链接，加入到阻塞队列中。数个消费者从阻塞队列中提取新闻链接，下载并提取网页中文。

# K-means聚类算法设计

## 提取网页中文部分、分词和去除停用词、统计词频

* + 1. **Map函数:**

Input: 网页编号，网页内容

Output: 单词|网页编号，1

其中的中文分词使用Jieba分词进行分词，然后过滤掉停用词，再进行统计词频。

* + 1. **Reduce函数:**

Input: 单词|网页编号，1

Output: 单词|网页编号，单词在网页编号出现的次数

## 统计每个网页中的单词在网页总单词数的占比

* + 1. **Map函数:**

Input: 单词|网页编号，单词在网页编号出现的次数

Output: 网页编号，单词|单词在网页编号出现的次数

* + 1. **Reduce函数:**

Input: 网页编号，单词|单词在网页编号出现的次数

Output: 单词|网页编号， 单词在网页编号出现次数/网页的总单词数

## 统计单词出现的网页个数、TFIDF、建立词表

* + 1. **Map函数:**

Input: 单词|网页编号， 单词在网页编号出现次数/网页的总单词数

Output: 单词，网页编号|单词在网页编号出现次数/网页的总单词

* + 1. **Reduce函数:**

Input: 单词，网页编号|单词在网页编号出现次数/网页的总单词数

Output: 单词|网页编号，TFIDF值

为每一个单词赋予ID，然后将词表存入temp文件夹中的dict文件中。

在计算TF-IDF的值时使用如下公式：

## 构造网页向量、随机选取初始中心点

* + 1. **Map函数:**

Input: 单词|网页编号，TFIDF值

Output: 网页编号，单词|TFIDF值

* + 1. **Reduce函数:**

Input: 网页编号，单词|TFIDF值

Output: 网页编号，单词编号1|TFIDF值1&单词编号2|TFIDF值2..

使用单词的TF-IDF值来构造网页向量。

## 实现K-means、

* + 1. **Map函数:**

Input: 网页编号，单词编号1|TFIDF值1&单词编号2|TFIDF值2..

Output: 网页对应中心点编号，单词编号1|TFIDF值1&单词编号2|TFIDF值2...

在计算网页间的距离时使用余弦距离，以此来代表网页间的相似度。

* + 1. **Reduce函数:**

Input: 网页对应中心点编号，单词编号1|TFIDF值1&单词编号2|TFIDF值2...

Output: 网页对应中心点编号，新的中心点向量

## 输出每个网页对应的编号

* + 1. **Map函数:**

Input: 网页编号，单词编号1|TFIDF值1&单词编号2|TFIDF值2..

Output:　网页编号，中心编号

* + 1. **Reduce函数:**

Input:　网页编号，中心编号

Output:　网页编号，中心编号

最后将网页的聚类结果输出到output文件夹

## 控制K-means迭代直至满足条件

* + 1. 指定k的值
    2. K-means算法的迭代停止条件：

当以下两个条件其中之一满足时，K-means停止迭代

1. 迭代次数达到参数maxIteration所设定的次数时
2. 当簇心向量不再变化时