## Introduction

输入query

输出相近的documents

**老方法**：布尔，包含多的好 会有负数

**新方法**：向量空间

Document 是vector， query 也是vector

用 distance角度，找document 和 query 的相关度

One-hot的向量表示方法，dimentsion会比较大

Tf-idf

文本

描述已自动生成

文本

描述已自动生成

文本

描述已自动生成

相乘



计算 document的tf-idf 也直接用在query vector上

然后计算 cosine 相似度

图片包含 图示

描述已自动生成

一个输入query ，和所有 document 的vector计算角度

排列10个最小的角度，也就是最大的cosine值

文本

描述已自动生成

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

流程================================

获取数据集

Parser解析出来要的信息（文件如果比较大的话估计要用sax解析？不知道develop集多大）

创建idf 和 if 数据，对应相乘得到所有ducument的idf-if数据

输入query

得到query vector的idf-if

与所有document 的idf-if计算cosine较多

排列

打印