文档匹配算法：

目前世界上所用到的文档匹配算法大都基于文档向量空间模型(Vector Space Model, VSM)，向量空间模型的基本原理是通过分词等手段将文章分成一个个词或词语组成的特征，进而将文本以这些特征按照不同计算方法得到不同特征项的权重，最后以这些权重为分量组成一个向量，这样一篇文本就转化成了一个向量，然后再利用一些距离算法来求得不同向量之间的距离，大多数都采用余弦距离。

20世纪80年代，一种新的计算文本相似度的算法被提出，那就是基于隐性语义索引(Latent Semantic Indexing, LSI)的方法，它利用了矩阵理论中常用的“奇异值分解”。LSI可以利用VSM，将表示文档的一个个向量组成一个矩阵，然后对矩阵进行奇异值分解，较小的奇异值将被去除，然后将奇异值分解得到的矩阵以及新来的映射到一个同一空间中，在此空间下计算相似度。21世纪初提出的一种对离散数据集（如文档集）建模的概率主题模型，即LDA（Latent Dirichlet Allocation）模型。LDA模型在主题层和词层都引入了Dirichlet先验参数，解决了LSI模型的主题参数个数随文档数增加而增加的问题。

还有一些效果不错的相似度算法，比如基于汉明距离来计算文本相似度，该方法利用模二加等运算，与其他基于VSM模型利用乘法来计算相似性的方法相比速度有了比较大的提升。而基于压缩稀疏矩阵矢量相乘的文本相似度计算方法，能够减少计算和存储空间的开销。该方法仅对非零元素存储和表示,然后用压缩稀疏矩阵矢量相乘的方法计算文本和查询的相似度,可通过给定相似度阈值来判定一个文本是否和查询相似。还有基于词汇语义计算文本相似度，采用了基于知网的词汇语义网络计算方法来计算两篇文章向量的相关性，再利用最大匹配算法来计算两篇文章的相似度。

目前本项目综合了几种算法，以余弦距离为基础，计算了LSI下的相似度，LDA下的相似度，由于LSI和LDA模型都是基于概率的模型，结果可能每次不同，本项目利用基础的余弦距离来排除一下比较差的结果。由于相似度最终比较的结果是阈值，本项目根据用户对匹配的评价同时需求和服务文档的特殊性利用遗传算法来自动调整算法阈值，可以得到更优的结果。后期本项目会利用需求和服务文档的特殊性对算法进行进一步的优化，提高用户的满意度。