# 稿件配图模型说明

本文目的是根据稿件内容从图库中推荐图片配给稿件。想法是计算图片文本和稿件文本之间的相似度，选取文本相似度高的图片推荐给稿件。具体方法是将文本都映射到topic的维度，然后计算相似度。所有模型方法以及计算方法来自gensim。

topic模型采用[LSI](http://en.wikipedia.org/wiki/Latent_semantic_indexing)(Latent semantic indexing），中文叫做浅层语义索引。通常我们考察两个词语间的关系常常考虑它们在一个窗口长度的共现情况。比如A 和B共现，B和C共现，那么LSI所做的就是挖掘A和B的隐含关系。

1. 输入文档

输入文档为处理好的38,922条图片文本和20,000条补充新闻类文本。

1. 预处理

Jieba是python 中文分词组件，本次主要采用了分词和词性标注的功能，由于目的是将文本映射到topic维度，一般来讲topic都是名词，所以根据词性去除了除名词之外的所有词组。

预处理也包括去停用词和标点符号。

1. 文档向量

我们可以通过这些文档抽取一个词袋，同时过滤掉低频的词。这样就可以将字符串表示的文档转换为用id表示的文档向量，此时每个单词有一个id以及其词频。然后基于这些文档计算一个TF-IDF模型，将由词频表示的文档向量转为用一个tf-idf值表示的文档向量。

1. LSI模型

有了tf-idf表示的文档向量，可以用它来训练一个LSI模型，一般设置topic在200-500之间，黄金topic数目为300。这只是一个经验数据，还是要根据样本数量来调整。

LSI最核心的意义是将文档向量组成的矩阵做秩为topic数目的SVD分解，我们可以得到所需要数目的topic以及每个topic的单词组成。通过这个模型可以将文档（tfidf文档向量）映射到topic空间中。

1. 建立索引index

建立索引的目的是可以给定一个query，直接找到最相关的文本。使用gensims中similarities.MatrixSimilarity来建立索引。由于我们只有图片文本才是和图片url直接联系的，所以索引这一部分我只采用了图片文本的文档向量映射到topic空间作为参数。

1. 测试

索引建立好之后就可以进行测试了。将给定文本做文档向量处理后通过训练好的LSI模型映射到topic空间，就可以与index中的文本计算余弦相似度了，排序之后取前20个最相思的文本并找到其对应的图片url，就可以获得所需要的图片了。  
7. 可视化

此次web演示是基于Flask框架。