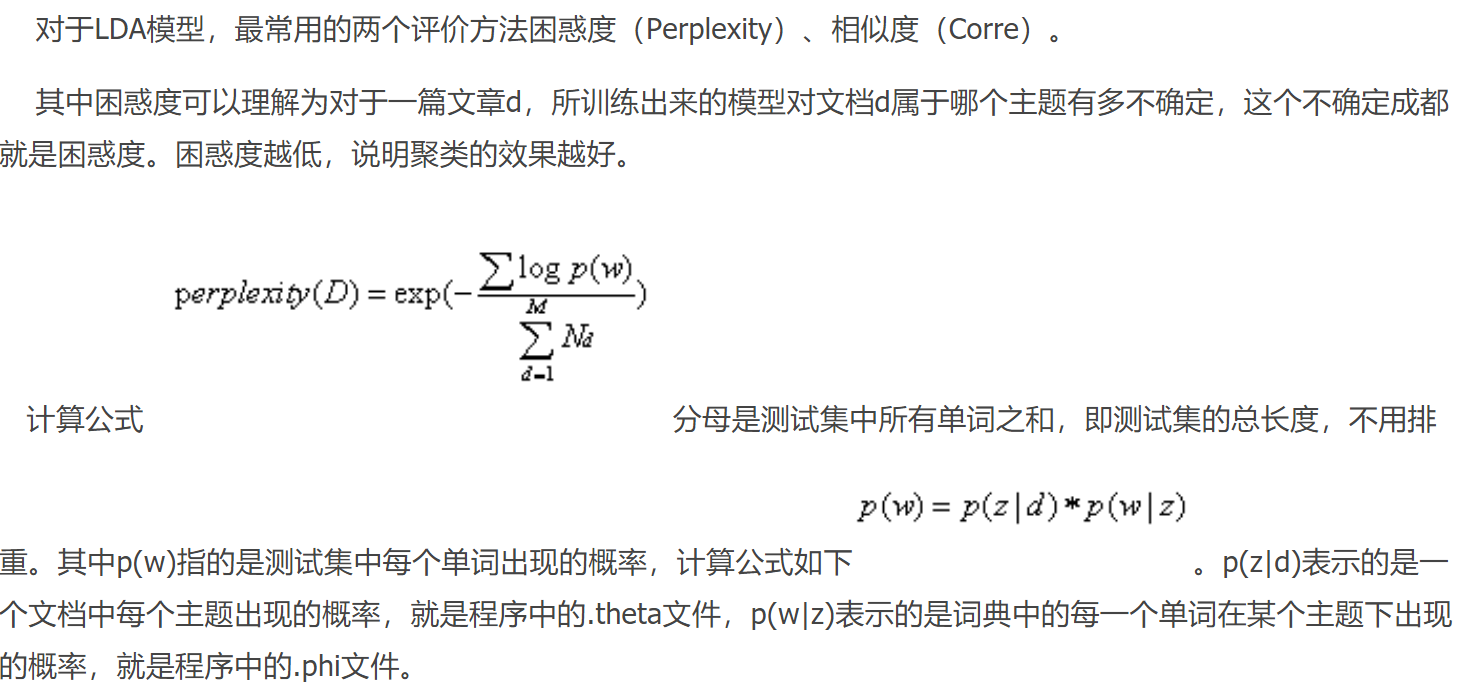
# 计算LDA困惑度perplexity确定最优主题个数

****LDA主题模型困惑度：****



****简单来讲我们分别计算主题个数在一定范围内的困惑度的变化情况，当困惑度值在对应某个主题数量值最小时，选取这个主题数量就是最优数量。结合我们数据的规模大小，最大主题数为十个主题。****

****计算困惑度函数的代码：****

**def** perplexity(ldamodel, testset, dictionary, size\_dictionary, num\_topics):  
 *"""calculate the perplexity of a lda-model"""  
 # dictionary : {7822:'deferment', 1841:'circuitry',19202:'fabianism'...]* print (**'the info of this ldamodel: \n'**)  
 print (**'num of testset: %s; size\_dictionary: %s; num of topics: %s'**%(len(testset), size\_dictionary, num\_topics))  
 prep = 0.0  
 prob\_doc\_sum = 0.0  
 topic\_word\_list = [] *# store the probablity of topic-word:[(u'business', 0.010020942661849608),(u'family', 0.0088027946271537413)...]* **for** topic\_id **in** range(num\_topics):  
 topic\_word = ldamodel.show\_topic(topic\_id, size\_dictionary)  
 dic = {}  
 **for** word, probability **in** topic\_word:  
 dic[word] = probability  
 topic\_word\_list.append(dic)  
 doc\_topics\_ist = [] *#store the doc-topic tuples:[(0, 0.0006211180124223594),(1, 0.0006211180124223594),...]* **for** doc **in** testset:  
 doc\_topics\_ist.append(ldamodel.get\_document\_topics(doc, minimum\_probability=0))  
 testset\_word\_num = 0  
 **for** i **in** range(len(testset)):  
 prob\_doc = 0.0 *# the probablity of the doc* doc = testset[i]  
 doc\_word\_num = 0 *# the num of words in the doc* **for** word\_id, num **in** doc.items():  
 prob\_word = 0.0 *# the probablity of the word* doc\_word\_num += num  
 word = dictionary[word\_id]  
 **for** topic\_id **in** range(num\_topics):  
 *# cal p(w) : p(w) = sumz(p(z)\*p(w|z))* prob\_topic = doc\_topics\_ist[i][topic\_id][1]  
 prob\_topic\_word = topic\_word\_list[topic\_id][word]  
 prob\_word += prob\_topic\*prob\_topic\_word  
 prob\_doc += math.log(prob\_word) *# p(d) = sum(log(p(w)))* prob\_doc\_sum += prob\_doc  
 testset\_word\_num += doc\_word\_num  
 prep = math.exp(-prob\_doc\_sum/testset\_word\_num) *# perplexity = exp(-sum(p(d)/sum(Nd))* print (**"the perplexity of this ldamodel is : %s"**%prep)  
 **return** prep

****作图来展示主题数与perplexity的关系****



以我们data2为例进行计算绘图结果：

最低困惑度值在5处，data2最优主题数为5。

所有结果如下：

data1 3

data2 5

data3 4

data4 5

data5 5

data6 5

data7 4

data8 4

data9 1

data10 10

总结：基于perplexity选择的主题，语义上与人工的判别还是有一定差距。每篇数据的主题数大致与其文章规模大小相关，同时还与文章的词句复杂程度等因素相关，具体因各种情况的文章而不同。