**基于网络搜索数据研究艾滋病发病数**

**一、为什么选这个题，阐述下有趣的点、有研究价值的点**

1.HIV是一个敏感而又吸引人的话题，还可以谈谈男同

2.契合大数据时代主题，契合金融智能课程，涉及到机器学习，做的是相关性分析

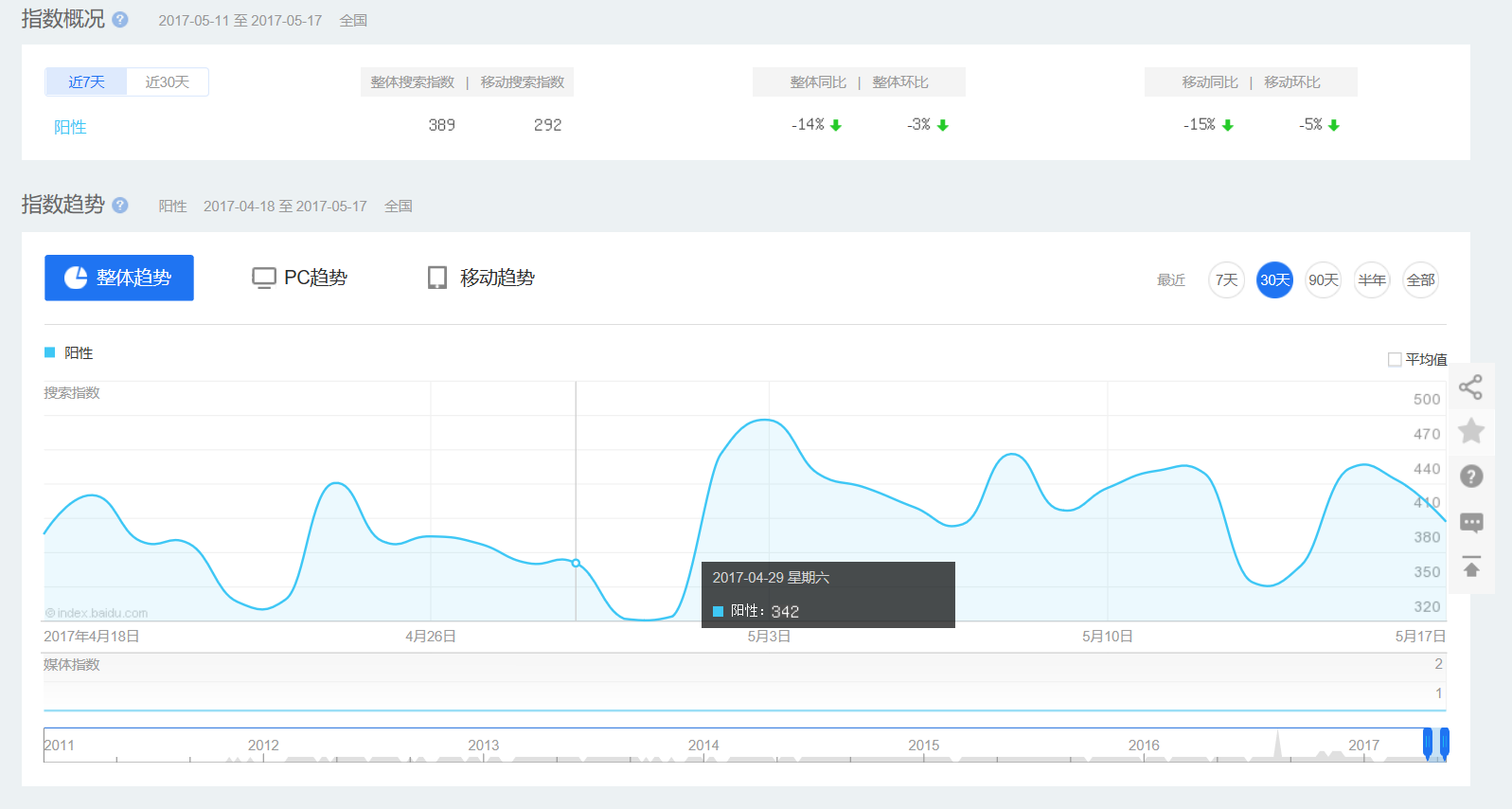
3.将互联网技术应用到医学领域，无疑是有巨大价值的

**二、是否能直接找到数据，还是需要编写爬虫去找，数据质量如何**

1.关键词

结合患者搜索心理，采用文献综述法、百度相关推荐词推荐法、手动查找补充法来选取一定数量关键词

2.百度指数数据



百度提供自2006年6月至今任意时间段的PC端搜索指数，2011年1月至今的移动端无线搜索指数；百度指数大众版暂不支持数据下载，如有强需求，可沟通后付费定制。

3.艾滋病发病数月度数据

（1）公共卫生科学数据中心2004.01-2014.12，但好像要花钱 用户名：panda 密码：panda001

（2）联合国艾滋病规划署数据

（3）百度搜索

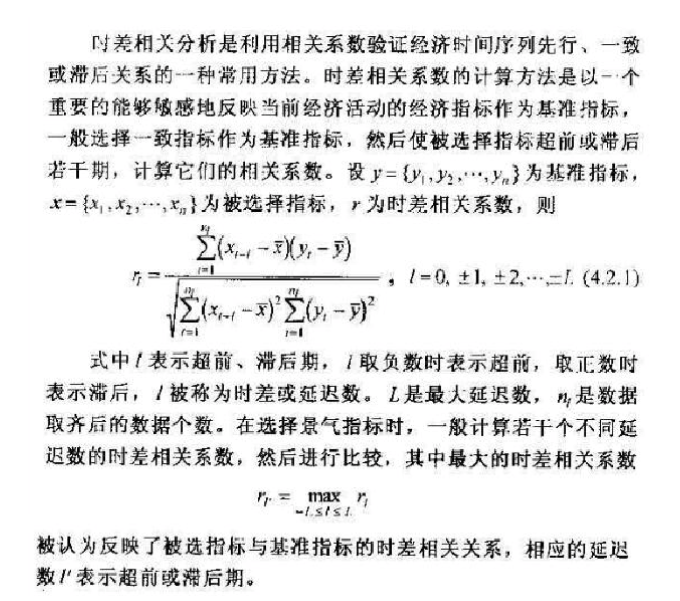
综上所述，我们的数据需求如下：

几十个搜索关键词；每个关键词的百度搜索月数据（可由日数据加总获得）；艾滋病发病数月数据。数据的时间跨度大概应在2011年1月至2014年12月，或者从2011年1月至2016年12月。

**三、模型大致是怎样的，用怎样的研究方法**

在我们的模型开始前，可以先做传统预测模型：灰度预测模型、时间序列ARIMA模型(自回归移动平均模型)等，用以突出我们模型的优越性。下面是我们的做法：

1.时差相关分析



计算得到每个关键词与艾滋病发病数的（最大时差）相关系数

2.对关键词进行甄选，选出真正影响发病数的关键词

（1）方法一

首先可以通过弱相关标准剔除最大时差相关系数较小的关键词

然后可以查看入选关键词的显著性水平，要求系数在0.05显著性水平下不为0

（2）方法二

选取相关系数较大的关键词，认为它们是主要因素，再进行主成分分析，选取关键词

（3）方法三

先进行聚类分析，然后看分成几类，在每一类中选“主成分”关键词

3.构建综合搜索指数

不同方法构建综合搜索指数的方法应该不一样，可以看看别人的文献选取我们的方法

4.平稳性检验：排除非平稳可能，防止虚假回归，即两个无因果性的因素表现出相关性

5.以发病数为因变量，综合搜索指数为自变量，建立回归模型，注意要剔除季节变动、不规则变动的影响，还要注意长期趋势（流行病的特点）

6.求解模型并进行检验

7.模型预测

8.比较传统方法与我们的模型，从结果或建模角度阐释我们的优越性

9.进一步讨论，可以发现我们用科学的方法挑选出的关键词是真的有意义的，此外可以对男同与艾滋发病数进行回归分析巴拉巴拉

**四、结论大概是如何的**

要么我们的结果好，要么我们的建模思维好，要么都好

反正我们的结果对相关部门提前预估艾滋病的传染疫情是否变化并作出相应的防控措施有重要意义

**五、能否对结果进行评测**

可以呀

**六、相关文献丰富度、可参考度**

多得是