1、互联网上每天都会产生数以亿计的博文,博文实时性强、内容丰富,其蕴含知识的发现、分析和处理越发重要和困难。如何及时获取并有效分析和挖掘其中的信息,成为一个重要的研究课题

* 1. 互联网上数量多不容易爬取
  2. 互联网数据实时性高，一格较短的时间内，会产生大量的数据。

作为新兴网络应用的新浪微博,从2009 年8月起开始内部测试。截止到2015  
年3月,微博的日活跃用户数量达到了8900万人次,月活跃用户数量更是达到了1.98亿人次。大量的政府机构官员、金业和个人都是微博的活跃用户。新浪微博借助这种开放的传播机制使其成为了中国的“公共议事厅”。目前微博在公益类、文体类、人物类、社会事件类等方面都获得了人们的重点关注,每一个话题都有数以万计用户参与其中,还有些话题甚至引起了国际上的广泛关注。微博用户对自己感兴趣的话题参与评论并持续关注,因此微博上民众对某些敏感话题的舆论有时甚至会改变整个事态的发展。

由于微博平台上的信息量巨大,传播速度非常迅速,如果仅仅依靠人工对信息进行搜集、浏览、评论以及观点的分析,是完全不可行的。微博信息采集的及时性、数据的结构化都有一定难难度,通过人工方式进行浏览和分析效率极低。针对中文交本分析的机器学习算法相对比较成熟,但对于微博的特性,如何对采集的微博数据信息进行预处理并选择适合的算法完成微博舆情的分析有待进一步研究。