**计算机新闻作业**

**小组成员：卢家希、王亦茗、刘妮**

**项目主题/题目**

网络论坛中的国家大事

——以天涯论坛网民对“GDP”相关话题的讨论为例

**项目引言**

改革开放以来，“GDP”（国内生产总值）三个字母牵动着上至中南海，下至普通百姓的心。“GDP”总量反映的是国家实力，人均反映的是国民的富裕程度，而其增速又反映了国家的经济发展状况。

我国“GDP”增速从九十年代到二十一世纪初的狂飙猛进，数值一度达到10%以上，致使在2010年“GDP”总量超越日本，跃居世界第二。2012年党的“十八大”召开，提出的经济结构调整的方针，使得“GDP”增速保持在6.5%，并开始注重经济发展的质量。

天下兴亡，匹夫有责。网络论坛中，关于“GDP”这个话题网友不乏讨论。网友的讨论反映的是普通人民对国家经济的看法，了解关于此话题的表达有助于了解民众对于国家大事的关注程度与关注面向。

**研究思路**

我们选择国内最大的论坛“天涯论坛”作为案例展开研究。关于此话题，研究主要解决以下几个问题：

1、“天涯论坛”关于“GDP”话题帖子的总体情况如何？

2、“天涯论坛”关于“GDP”话题的帖子中，网友们在讨论什么？

3、“天涯论坛”关于“GDP”话题的帖子的相似性如何？

4、参与“天涯论坛”关于“GDP”话题讨论的网友，其关系网络是怎么样的？

基于此，研究的思路如下：

1、抓取“天涯论坛”经济版中前20页标题中包含“GDP”的帖子的帖子列表、作者信息、所有网友的回复。

2、对帖子列表中点击与回复数进行分析。

3、对所有帖子中的网友回复进行文本挖掘并建立主题模型，分析在“GDP”这个主题下，网友的关注面向与看法。

4、计算帖子间的相似性，构建给网友推荐帖子的推荐系统。

5、构建“GDP”相关话题的帖子中，网友间的回帖网络，并对网络进行描述与分析。

6、总结以上5个部分的发现，得出天涯论坛网民对“GDP”相关话题讨论的相关结论。

**研究发现**

**从数据的整体表现上发现：**

1、点击量，回复量等各数据的数值的整体样本并不十分符合正态分布，尤其是回复量偏离正态分布更加严重，主要原因是少部分的帖子热度非常大，其点击量和回帖量等都超过平均水平很多。

2、各数据之间除了点击量和回复量有很明显的线性关系

3、天涯关于GDP的讨论热度在13年增幅很大，在14年继续增长达到高峰，但是之后就一直减退，15年跌幅近50%。

**从文本内容中发现：**

1、关于“GDP”，网友在2012-2015年的关注点大致相同，均为经济增长以及与发达国家如美国、日本的比较。但在2016年上半年，网友除了关心经济增长外，话题还延伸到了城镇化相关的社会问题与工业化相关的问题上，这与2016年经济形势是符合的。

2、在2012-2016年网友关于“gdp”回复的词语大致可分为两类，一类是较为宏观的词语，另一类是对于上述宏观词语的细化，体现出网友们“既能谈论天下大事，又能说得有理有据”。

**从建立的主题模型中发现：**

1、天涯关于GDP的帖子可以根据title分成美国、增速、才能、居安思危、蝉联冠军、增长、中国、三季度、数据、关注、CPI、黄金、超过、黄金、广东、城市、房价、造假、增速、美元、注水等一些主题。

2、从建立主题模型的结果上来看，关于中国、增长和地方的主题比较明显。还可以发现，一些文章在很多主题下都有更高的权重，因为他们比较长并且包含的关键词出现在不同主题里，其他一些文章在大多主题下权重都很低。

**从回帖网络可以发现：**

1、关于"GDP"这个话题，1040个帖子中共有1046个网友发表了留言。由于帖子抓取自“天涯论坛经济版”，且帖子数量与发表评论网友数量相差不大，说明有部分网友经常在此版发表评论。

2、网络回帖所构成的网络较为分散、疏松，连接度也比较低。

3、网友的回帖网络符合BA网络特性，此网络由少数核心节点串联，大部分节点连接度低于网络平均连接度。说明在“GDP”这个话题下，大多数网友只是单独发表自己意见，相互间所形成的讨论较少。

**结论**

从研究发现中可见，网友对GDP的讨论在13-14年达到高峰，但是2015年的讨论有所褪温。有些帖子点击和回帖量远远超出平均值。从网友讨论的话题情况来看，2012-2016的关注集中在对日本、美国的讨论以及国内重庆等热门城市的讨论。另外从回帖网络中可见网友更倾向于单独发表意见较少一起讨论的回帖习惯。另外，我们组构建的推荐模型可以帮助网友推荐最关注的帖子。