**《计算新闻传播学课程》期末小组报告**

孙倩 周洋帆 徐蓓清 葛晨捷

**一、项目主题：**

天涯论坛关于“EXO”话题数据分析

**二、项目摘要：**

通过爬取在天涯论坛上关于“EXO”话题的相关帖子作为数据集，在此基础上运用主题模型分析、社会网络分析与文本分析，希望用理性的方法以洞察中国网民对于韩国男子团体EXO的看法与情感。

**三、引言**

EXO是韩国SM娱乐有限公司于2012年4月8日正式推出的12人男子流行演唱团体。现以9名成员的形式展开活动，分别为金珉锡、金俊勉、张艺兴、边伯贤、金钟大、朴灿烈、都暻秀、金钟仁、吴世勋，其中张艺兴为中国籍成员。EXO这个名字从意为太阳系以外的外部行星“EXO PLANET”一词中得来，蕴含了他们是从未知的世界来的新星之意。

近年来，SM公司将EXO的一部分市场转移至中国，这个团体在中国掀起了一阵潮流之风，无数粉丝为之疯狂。即使有部分中国成员例如鹿晗、黄子韬先后退出该团，EXO的热度依旧不减，而那些退出的成员也未因单飞而人气下滑，相反知名度更高了。

而作为网民发表自己观点的的网络论坛中，“EXO”也成为一个颇具人气的热门话题。在论坛上，网民可以发表支持偶像的言论，同样也能够表达自己的负面情绪与评价。相比只在官方后援团的网站上爬取到清一色“迷妹”们支持的意见，使用天涯上的相关数据则更具有现实意义与价值。

因此，我们选取在天涯论坛上“EXO”的相关帖子进行具体舆情分析，可以了解到EXO组合在内地发展的现状，以及网民的支持程度。

**四、研究思路：对于研究项目的系统的描述**

在整个项目过程中，我们将对抓取的数据进行文本分析、建立主题模型、建构社区网络，希望了解并解决如下问题：

* 在天涯论坛中，对于“EXO”话题的讨论热度如何？
* 天涯用户关于“EXO”的讨论有哪些具体内容？分内容的主题有哪些？
* 讨论者们对于“EXO”组合的态度是怎样的？
* 对“EXO”感兴趣的天涯用户形成了怎样的网络模型？又有怎样的特点？

我们的数据分析项目，通过如下步骤展开：

* 数据挖掘及清洗：

数据挖掘是之后所有流程的第一步。我们抓取了在天涯论坛上共计150页含有“EXO”话题的帖子，抓取内容包含帖子列表、作者信息、网友的回复的所有文本信息。而在数据前期准备中，还需要对已挖掘数据文本进行清洗。

* 文本分析：

针对已经进行分词操作的文本内容进行词频统计与分析，并且做出词云图。

* 建立主题模型：

在这一步骤中，我们将基于LDA模型进行主题建模，采用了词袋的方法，对系列已抓取的内容进行主题建模，以分析不同主题在所有帖子以及回帖话语中出现的权重分布，来寻找网民对于“EXO”的关注点与兴趣点所在。

* 生成推荐系统：

将根据已有用户对于现有主题的兴趣程度来，根据帖子的主题分类，对用户进行阅读更多内容的推送。

* 构建回帖网络：

根据对主贴人与回帖人所在的节点进行连接，绘制关系图，以观察在天涯论坛中对于“EXO”话题的社会网络。

* 得出结论：

对上述步骤所得结果进行综合分析。

以上为对于天涯论然“EXO”话题数据挖掘部分的具体研究思路。

众所周知，微博是公众意见表达的重要场所。所以，除了上述数据烦恶习，我们还爬取了在微博上关于已退团成员“鹿晗”的部分相关微博，同样地进行包括文本处理、词云等方面的深入分析，以研究公众对于“EXO”与“鹿晗”两个话题的讨论时的关联与差异。

**五、研究发现**

1、数据描述：

从对文本的基本信息描述中，我们可以发现：所有有关“EXO”话题帖子的点击量总量接近3261万，每个帖子的平均点击量是大约在4349次，四分之三位点的回复量是1208次，从这可以看出这个数据分布的非常不平衡，有些贴子的点击量非常大，拉高了整体平均的情况很严重。另同理可以见，3/4位点的回帖数只有22，但是由于极少部分的帖子的回帖量很大，左右拉高了整体的回帖数，回帖量仍然分布不平均。因此，点击量和回帖数热度极高的帖子只是极少数。并且，点击量和回贴数有比较强的线性关系，用户发贴数量和评论数量有较强的线性关系。

从时间趋势的维度上看，五、六、十、十一月这四个月是讨论EXO的高峰期。尽管9月帖子的数量并不算高，但从总点击量和平均点击量来看，却是最高的。由此可以看出9月也同样是论坛用户参与EXO主帖讨论参与度较高的月份，并且原创的文章所获的点击量和回复量远高于转载的文章。

2、文本分析：

通过使用TF-IDF与Text Rank对于文本词频的分析后，我们选用第一组词语（炸死、弯弯、飞弹、表示、台湾、导弹）来比较两个算法分别对这些词语词频的统计。从上图可以发现，在TF-IDF算法中，“炸死”、“弯弯”、“飞弹”、“表示”、“台湾”、“导弹”的权重分别是0.75、0.82、0.8、0.3、0.48、0.49左右；而在Text Rank算法中，其权重分别为1.00、0.99、0.68、0.67、0.67、0.67左右。不难发现，使用Text Rank得出的权重普遍高于TF-IDF算法，并且高出的值十分明显。

而在基于TF-IDF算法绘制出的词云图中可以明显发现，“EXO”、“吴亦凡”、“鹿晗”、“粉丝”、“转载”、“张艺兴”、“大势”等词为热门词汇；除此之外，别列入的明星名单还有吴京、李宇春、普京、SNH48等。

3、主题建模：

通过对文本进行主题建模，共产生十组50个主题。其中包括与内地男子团体TFboys等组合的比较、对于新专辑的讨论、以及对于“EXO”同人的幻想。

在最后，也对主题模型的准确度进行验证，发现验证成绩一般，不够理想。

4、推荐系统：

根据针对我们所做的验证测试，从给“淼焱森晶鑫磊犇众”用户推荐打分前20的帖子的排名及成绩中，我们可以发现：该推荐算法的准确度可以接受，但是还有较大的提升空间。排名前两位的打分很高，为3455783.0和8850.0分，说明主要的推荐帖子精准度是较高的；但是三名以外的内容成绩一下子大幅跳水，与第一名产生了量级的差别。

因此，该推荐系统需要提高其末位推荐的精准度。

5、回帖网络：

通过分析运算得出的两张可视化网络图，不难发现，在天涯对于“EXO”话题的回帖网络中，最主体的结构是有一个聚集度十分明显的中心点，以此节点为中心向外辐射，密度较小、分布稀疏，并且其他子分支很少。而在主体结构旁的辐射圈外还有一个额外延伸的复杂网络，枝节交错，难名形状。除此之外，还有少部分的孤单、零散的单向关系线。

由此可见，这个回帖网络并不符合小世界模型，更偏向于BA无标度模型。

6、与微博内容的比较分析：

首先，爬取微博中关于“鹿晗”的相关内容。

接着，运用python相关算法进行数据清洗与文本分析。接着，运用python相关算法进行数据清洗与文本分析。可以发现，“鹿哥”、“鹿晗”、“糊鹿娃"、“葫芦娃”、“紫娃”等词为最热词汇，其他相关形容词有“最帅”、“可爱”、“开心”、“快乐”等。

将以上词云图与之前天涯论坛关于“EXO”的帖子内容所生成的词云图相比较，其实鹿晗在“EXO”词云图中也是最热门的词汇之一，可见其在中国粉丝圈中的人气；而针对他个人的搜索生成的词云图呈现偏向更加针对他个人属性以及粉丝给他的标签

**六、结论：**

通过数据挖掘、文本分析、主题模型构建、推荐系统与回帖网络的探索后发现，天涯论坛上对于“EXO”话题的热度与组合当时的活动与推出新曲有关，简而言之就是与炒作营销的程度有关。并且，网民对于已退团成员的关注度拉动了对于“EXO”的讨论，也就是说天涯论坛上用户对于“EXO”的实际关注度实际不如我们预期。因此，根据以上科学分析，“EXO”在中国市场的进一步发展走向不太清晰，他们应该抓牢仅剩不多的中国成员，以保持“EXO”在国内的人气。