一、有关课程成绩的构成及比例、考核方式以及评定标准的详细说明

1. 课程成绩的构成及比例、考核方式

   课程成绩由论文成绩（占 60%）和平时成绩（占 40%）构成。

   课程论文要求以社交媒体数据的分析和挖掘为主题，不少于 2000 字，并必须要包含以下三部分内容：

   (1) 数据描述与预处理：论文数据可以为自行抓取的数据也可以是网上的公共数据集，要求在此部分要详细说明数据的来源、数据的规模和数据预处理的过程，如文本的去标签、分词和去停用词等。

   (2) 挖掘过程：根据所选数据，选择图挖掘和/或文本挖掘的方法。如果选择图挖掘的方法，要求给出基本的网络度量分析(包括中心性分析、凝聚性分析）、描述社区发现或信息传播建模的过程；如选择文本挖掘的方法，给出主题分析或情感分析的过程。此外，分析过程所使用的程序代码要作为论文的附录提交。

   (3) 结果分析：利用图表等工具对挖掘得到的结果进行分析，分析其理论和现实意义。

   平时成绩主要考察学生的出勤情况(占 10%),课后作业和课堂练习的完成情况(占 30%)

2. 评定标准

   论文成绩根据选题与课程内容的相符性、论文内容的完整性、数据分析的准确性和深度以及论文写作的规范性等来综合评定。具体评分标准如下：

   (1) 优秀(90 分以上)：选题与课程内容高度相关；论文内容完备并配有完整的程序代码作为附录，数据分析过程思路清晰，逻辑性强且有深度，能合理利用图表等工具对结果可视化，结论鲜明，有说服力；语言得体，表述清晰，论文按格式编排。

   (2) 良好(80-89 分)：选题与课程内容相关；论文内容较完备并配有完整的程序代码作为附录，数据分析过程清晰，逻辑性较强且有一定的深度，能合理利用图表等工具对结果可视化，结论前后无矛盾，有说服力；语言得体，表述清晰，论文按格式编排。

   (3) 中等(70-79 分)：选题与课程内容基本相关；论文内容基本完备并配有简单的程序代码作为附录，数据分析过程大体清晰，但分析不够深入，能运用简单图表工具对结果可视化，结论基本清楚；论文虽有少量的用词不规范及表述不清晰的地方，但基本符合论文写作规范和格式编排要求。

   (4) 及格(60-69 分)：选题与课程内容勉强相关；论文内容不够完备或缺少程序代码作为附录，数据分析过程简单，缺乏深度，结果分析和展示过于简单；论文存在语言不得体和表述不够清晰的问题，格式有个别地方未按要求编排。

   (5) 不及格(59 分及以下)：选题与课程内容不相关；论文存在明显的抄袭行为或缺少数据分析过程；论文写作随意，表述混乱，格式未按要求编排。

二、建议选题

1. 在线社会网络结构分析 :度分布、中心性、凝聚性、子群分析等 ,可使用的数据集: douban dataset, enron email dataset, epinion trust network dataset, flickr dataset, youtube dataset

注：对大规模网络可抽样来分析，但抽取的样本网络的主要网络度量应与全网络相近

2. 网络结构的多层次分析：

(1) flickr dataset: 在分析整体好友网络结构的基础上 ,可分析或比较一个或多个群组 group 内用户之间的好友关系。对于整体好友网络划分社区，分析社区与群组之间的关系

(2) youtube dataset: 可以根据用户在共享视频，共同订阅等网络中的位置和结构来划分社区，分析有共同兴趣的用户

(3) enron email dataset: 50 名高管及与其有邮件往来的员工组成的网络，可根据度中心性发现 50 名高管，并分析这 50 名高管之间的关系网络。

3. 文本分析与可视化：
(1) 可选用近几年/几届的政府工作报告/党的代表大会的报告等，分析高频词的变化，并以词云展示
(2) 可选用中美互加关税这一事件相关的微博数据，分析事件随时间变化的趋势以及文本内容随时间变化的趋势，以高频词和词云来展示
4. 文本聚类分析：以 20newsgroup 或搜狗体育新闻数据为例，对文本做聚类分析，算法不限，但需要对聚类结果进行解释和评估。结果的解释：选取每一类内的高频词来表示类的主题；结果的评估：组内距离，组间距离，F-measure 等
5. 文本分类分析：以 20newsgroup 数据或电影评论数据集为例，划分训练集和测试集对文本做分类，分类方法可以使用一种或多种，但要求使用不同的特征选取方法，并以准确率、召回率和 F-measure 来比较不同特征选取方法的结果。
6. 其他自行抓取或下载的数据