数据仓库与数据挖掘 作业 1

1. 作业概述

该作业要求同学对数据挖掘中的常见主题进行论文阅读，每组选择一个主题，并在选定的主题中选择至少 1 篇指定的论文和至少 3 篇该主题的相关论文（所提供论文列表之外）进行仔细阅读，并做课堂展示报告。报告内容应包含**问题背景**、**涉及方法**、**总结分析**与**实验结果**，以及结合数据挖掘课程所学知识，自己对于该方向的**理解与感想**。

2. 作业要求

1) 报告主题通过在网络学堂“课程讨论”中对应话题后留言进行选择，每个主题限至多 3 组选，先到先得。留言后请刷新该话题以确保之前选择该主题组数未多过 2 组。如已有多于 2 组选择了该话题，该选择为不成功，请重新选择。如成功选定主题，则不可重选。
2) PPT 分为三部分
   a) 首页：包括**组员信息**、**组员分数调整**（可选）。本次作业按照总分 100 分计算，每个组员可将自己不超过 5 分的分数转移给组内其他组员，请按照“张三：-3、李四：-1、王五：+4”这样的格式注明在组员姓名后。若该组本次作业的评分为 85 分，则上面的示例中最终三人的得分分别为：张三 82，李四 84，王五 89。注意某个组员减掉的分数不能超过 5 分，但是某个组员加上的分数可以超过 5 分的，只要最终保证所有组员的正负值和为零即可。如未指明分数调整办法，则组内各成员得分一致。
   b) 正文：5-10 页
   c) 附录：0-5 页。本部分不需进行课堂展示，内容可包含但不限于正文部分未展示的论文细节、更多的扩展阅读或理解分析等。此部分会作为评分依据之一。请在 ppt 中明确标识该部分的起止位置。
3) 组内的每个同学都需要进行汇报（如某位同学未参加则视为放弃本作业得分）。
4) 每组报告时间为 5 分钟，请各组合理安排时间（助教会进行计时，超出时间将会被打断）。

3. 作业提交

1) 通过阅读论文总结出来的 PPT 于 **2017 年 12 月 18 日 23:59（含）**之前提交到网络学堂。超过时间提交将视为放弃本次作业，请各位同学务必提前提交作业，以免最终因为网络等问题提交失败。
2) 报告文件只接受 **ppt** 或 **pptx** 文件，提交文件命名为：姓名 1_学号 1_姓名 2_学号 2_姓名 3_学号 3.ppt(x)。如“向昌盛_2015213505_何涛_2015214231.ppt”
3) 助教将于 **2017 年 12 月 18 日 23:59** 前公布报告顺序，请各组同学提前查看，组内报告顺序各组自行决定。

4. 重要时间

2017 年 10 月 23 日 19:00-24:00	主题选择
2017 年 12 月 18 日 23:59 前	报告 ppt 提交
	公布每组报告时间、顺序
2017 年 12 月 26 日、2018 年 1 月 2 日	展示报告

5. 论文列表

Topic 1: Temporal Analysis

1: TrioVecEvent: EmbeddingBased Online Local Event Detection in GeoTagged Tweet Streams

2: Effective and Realtime InApp Activity Analysis in Encrypted Internet Traffic Streams

3: Toeplitz Inverse CovarianceBased Clustering of Multivariate Time Series Data

Topic 2: Novel Applications 1

1: Not All Passes Are Created Equal:" Objectively Measuring The Risk and Reward of Passes in Soccer from Tracking Data"

2: Luck is Hard to Beat: The Difficulty of Sports Prediction

3: PNP: Fast Path Ensemble Method for Movie Design

Topic 3: Network and Graph 1

1: On Finding Socially Tenuous Groups for Online Social Networks

2: Improved Degree Bounds and Full Spectrum Power Laws in Preferential Attachment Networks

3: Graph Edge Partitioning via Neighborhood Heuristic

Topic 4: Classification 1

1: Similarity Forests

2: Learning Certifiably Optimal Rule Lists

3: Structural Neighborhood based Classification of Nodes in a Network

Topic 5: Medical Data

1: Predicting Clinical Outcomes Across Changing Electronic Health Record Systems

2: Prognosis and Diagnosis of Parkinson's Disease Using MultiTask Learning

3: GELL: Automatic Extraction of Epidemiological Line Lists from Open Sources

Topic 6: Classification 2

1: KATE: KCompetitive Autoencoder for Text

2: Large-Scale Item Categorization in e-Commerce Using Multiple Recurrent Neural Networks

3: PPDsparse: A Parallel PrimalDual Sparse Method for Extreme Classification

Topic 7: Frequent Pattern and Association Rule

1: Frequent Pattern Mining with Uncertain Data

2: Discovering Frequent Patterns in Sensitive Data

3: Detecting Privacy Leaks Using Corpus-based Association Rules

Topic 8: Networks and Graphs 2

1: WeisfeilerLehman Neural Machine for Link Prediction

2: Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity

3: FORA: Simple and Effective Approximate SingleSource Personalized PageRank

Topic 9: Methodology

1: Constructivism Learning: A Learning Paradigm for Transparent Predictive Analytics

2: Discovering Reliable Approximate Functional Dependencies

3: The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables

Topic 10: Novel Application 2

1: A Data Mining Framework for Valuing Large Portfolios of Variable Annuities

2: Backpage and Bitcoin: Uncovering Human Traffickers

3: Quick Access: Building a Smart Experience for Google Drive

Topic 11: Representations

1: Collaboratively Improving Topic Discovery and Word Embeddings by Coordinating Global and Local Contexts

2: Struc2vec : Learning Node Representations from Structural Identity

3: Efficient Correlated Topic Modeling with Topic Embedding

Topic 12: Urban Planning

1: No Longer Sleeping with a Bomb: A Duet System for Protecting Urban Safety from Dangerous Goods

2: Planning Bike Lanes based on SharingBikes' Trajectories

3: The Simpler the Better: A Unified Approach to Predicting Original Taxi Demands based on LargeScale Online Platforms

Topic 13: Clustering

1: Towards an Optimal Subspace for KMeans

2: EgoSplitting Framework: from NonOverlapping to Overlapping Clusters

3: A Hierarchical Algorithm for Extreme Clustering

Topic 14: Web Applications

1: A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments

2: Cascade Ranking for Operational Ecommerce Search

3: A Practical Exploration System for Search Advertising

Topic 15: Classification 3

1: Interpretable Decision Sets: A Joint Framework for Description and Prediction

2: Learning Cumulatively to Become More Knowledgeable

3: “Why Should I Trust you?” Explaining the Predictions of Any Classifier

Topic 16: Humans and Crowds

1: Robust Top k Multiclass SVM for Visual Category Recognition

2: Accelerating Innovation Through Analogy Mining

3: Human Mobility Synchronization and Trip Purpose Detection with Mixture of Hawkes Processes

Topic 17: Recommendation

1: Online Ranking with Constraints: A PrimalDual Algorithm and Applications to Web TrafficShaping

2: Unsupervised P2P Rental Recommendations via Integer Programming

3: Largescale Collaborative Ranking in NearLinear Time