1. Спецкурс программы аспирантуры, полугодовой: **Аналитика больших данных: дополнительные главы**.

2. Преподаватели: доц. С.Т.Главацкий, н.с. И.Г.Бурыкин, с.н.с. Р.Р.Айдагулов.

3. Аннотация курса: методы поддержки систем принятия решений (recommendation systems); алгоритм Google определения значимости веб-страниц PageRank; современные методы понижения размерности массивов данных; методы графовой кластеризации; основные модели машинного обучения.

4. Тематическое содержание курса:

|  |  |
| --- | --- |
| **Тема 1.** | Анализ ссылок в Интернет. |
| **Тема 2.** | Вычисление PageRank. |
| **Тема 3.** | Модифицированные алгоритмы вычисления PageRank. |
| **Тема 4**. | Концентраторы и авторитеты, спам. |
| **Тема 5.** | Рекомендательные системы, вызов NetFlix. |
| **Тема 6.**  | Алгоритмы классификации пользователей и товаров. |
| **Тема 7.** | UV-разложение. |
| **Тема 8.** | Снижение размерности пространств данных. |
| **Тема 9.** | Линейные алгоритмы снижения размерности, метод главных компонент. |
| **Тема 10.** | Сингулярное разложение, CUR-декомпозиция. |
| **Тема 11.** | Анализ социальных сетей: графовая модель. |
| **Тема 12.** | Методы графовой кластеризации в социальных сетях. |
| **Тема 13.** | Концепция машинного обучения в анализе больших данных. |
| **Тема 14.** | Основные модели машинного обучения: линейные модели, деревья решений, нейронные сети. |

5. Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения, характеризующих этапы формирования компетенций.

PageRank.

Эффективное вычисление PageRank.

Эффективное представление матрицы перехода.

PageRank, чувствительный к теме

TrustRank.

Концентраторы и авторитеты.

Рекомендательные системы.

Матрица предпочтений.

Рекомендации на основе содержания.

Обнаружение особенностей документов.

Алгоритмы классификации.

Кластеризация пользователей и товаров.

Снижение размерности.

UV-разложение.

Оптимизация произвольного элемента.

Метод главных компонент.

Матрица расстояний.

Сингулярное разложение.

Запросы с использованием концептов.

Вычисление сингулярного разложения матрицы.

CUR-декомпозиция.

Методы графовой кластеризации в социальных сетях.

Дерево принятия решений, понятие энтропии.

Кластерный анализ в  интеллектуальном анализе процессов.

Оценка результатов для дерева решений.

Сети Петри как инструмент анализа процессов.

WF-сети и понятие бездефектности (soundness).

Альфа-алгоритм.

Обнаружение процессов (рrocess discovery): 4 критерия качества.

Модель бизнес-процессов (BPMN).

Анализ графа зависимостей.

Теорема CAP.

Основные модели машинного обучения: линейные модели, деревья решений, нейронные сети.

Примеры задач для самостоятельного решения.

1. Насколько разреженной должна быть матрица (то есть, какова в ней должна быть доля единичных элементов) для того, чтобы разреженное представление было более экономичным?

2. Используя компактный метод представления (источник, степень, преемники) представьте матрицы перехода определенных графов (в задании).

3. Матрица предпочтений представляет рейтинги по шкале 1-5 для восьми товаров, от *a* до *h*, для трех пользователей *A*, *B* и *C*. Вычислить следующие величины по данным из этой матрицы.

(а) Преобразуйте матрицу предпочтений в булев вид (есть оценка или ее нет) и вычислите расстояние по Жаккару между каждой парой пользователей.

(б) Повторите п. (а) для расстояния по косинусу.

(с) Преобразуйте рейтинги 3, 4 и 5 в 1; а 1, 2 и пустое значение – в 0. Вычислите расстояние по Жаккару между каждой парой пользователей.

(d) Повторите п. (с) для расстояния по косинусу.

(е) Нормализуйте матрицу вычитанием из каждого непустого значения среднего значения для данного пользователя.

(f) Используя нормализованную матрицу из п. (d), вычислить расстояние по косинусу между каждой парой пользователей.

4. Рассмотрим вершину в дереве решений с 80 экземплярами типа A и 70 экземплярами типа B. Вычислить энтропию данной вершины.

5. Рассмотрим следующий журнал событий L. С помощью альфа-алгоритма нарисуйте сеть Петри, соответствующую журналу L.

6. Рассмотрим следующий журнал событий L. Постройте матрицу следа (Footprint of L) и найдите коэффициент подтверждения на основе матрицы следа (Footprint-based conformance).

7. Рассмотрим следующую модель процесса РN и след < a,b,c,h,i >. Найти коэффициент подтверждения на основе выравнивания (correct alignment-based conformance) для этого следа, предполагая, что стоимость переходов для всех активностей равна 1.

6. Перечень основной и дополнительной учебной литературы:

А. Основная литература

1. Ю. Лесковец, А. Раджарамаран, Дж. Ульман. Анализ больших наборов данных. ДМК, Москва, 2016.

2. Wil M. P., van der Aalst. Process Mining Discovery, Conformance and Enhancement of Business Processes. ISBN: 978-3-642-19344-6 (Print) 978-3-642-19345-3 (Online).

3. Hasso Plattner. A Course in In-Memory Data Management. The Inner Mechanics of In-Memory Databases. ISBN:  978-3-642-55269-4 (Print)  978-3-642-55270-0 (Online).

MongoDB Data Modeling. Focus on data usage and better design schemas with the help of MongoDB. Wilson da Rocha França. 2015, Packt Publishing. ISBN 978-1-78217-534-6.

Б. Дополнительная литература

# 1. [Jure Leskovec](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople), [Anand Rajaraman](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB" \l "bookPeople), [Jeffrey D. Ullman](http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition?format=HB#bookPeople). Mining of Massive Datasets. 2nd Edition. Stanford University, California, 2014.

# 7. Перечень ресурсов информационно-коммуникационной сети «Интернет»:

1. http://www.cambridge.org/ru/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/mining-massive-datasets-2nd-edition#c3bEPha9lWJHRUhW.99

2. http://i.stanford.edu/~ullman/mmdsn.html

**Программа утверждена на заседании кафедры теоретической информатики**

**Протокол №**