Приложение 1

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

17 марта 2016 г. в 13.15 в ауд. 327

по адресу: Санкт-Петербург, Петергоф, Университетский просп., 35

(указывается дата, время и место заседания Ученого совета)

состоится заседание Ученого совета
факультета прикладной математики - процессов управления СПбГУ

(указывается название института/факультета)

ПОВЕСТКА ДНЯ:

1. *Научный доклад «Тематическое моделирование корпуса текстовых документов». Докладчик: Добрынин В. Ю., доцент кафедры технологии программирования СПбГУ.*

2.Проведение конкурса на замещение должностей НПР. Докладчик: Петросян Л. А., профессор, декан ф-та ПМ-ПУ СПбГУ, председатель Учёного совета ф-та ПМ-ПУ СПбГУ.

Председатель Ученого совета
факультета  Петросян Л. А.

Аннотация научного доклада доцента Кафедры технологии программирования СПбГУ Добрынина Владимира Юрьевича

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КОРПУСА ТЕКСТОВЫХ ДОКУМЕНТОВ

Задача тематического моделирования корпуса текстовых документов заключается обычно в построении взвешенных списков слов, характеризующих раз личные наиболее важные темы, неявно представленные в текстах большой коллекции документов. Знание этих тем позволяет решать множество задач, включая семантический поиск, навигацию по корпусу документов и т.п.

В докладе будут представлены три различных модели, включая латентное семантическое индексирование (Latent Semantic Indexing, LSI), латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) и контекстная кластеризация документов (Contextual Document Clustering, CDC).

Первая модель (LSI) основана на построении сингулярного разложения для матрицы частот всех слов во всех документах коллекции. На основе этого разложения находится малоранговая аппроксимация указанной матрицы (ранга 100 - 200), на основе которой и формируются представления как слов, так и документов в виде векторов в вещественном пространстве (семантическое пространство) относительно небольшой размерности.

Модель LDA является вероятностной порождающей моделью, параметры которой оцениваются в процессе максимизации функции правдоподобия. Результатом являются модели для тем, в которых каждая тема представляется распределением вероятностей на множестве слов, и модели для документов, в которых каждый документ представляется распределением вероятностей на множестве тем. В обоих случаях используются распределения Дирихле.

Модель CDC основана на интерпретации различных идей из области семиотики методами теории информации. Результатом является семейство тем (вероятностные распределения на множестве слов) и жесткая кластеризация (группирование) документов вокруг найденных тем.

Будут приведены следующие примеры: кластеризация результатов поиска (LSI), тематический анализ Википедии (LDA) и тематический анализ публикаций в газете New York Times за 20 лет (CDC).