# Математические методы анализа текстов

## Лабораторая работа №4

### Векторные представления слов, тематическое моделирование. Анализ тональности.

**Дедлайн: 08.05.2017**

Результатом выполнения задания является отчёт в формате pdf, написанный в LaTeX, а также скрипты, реализующие все проделанные эксперименты. Отчёт должен содержать формулировку задания, описание всех использованных моделей, инструментов, подробный рассказ о проведённых экспериментах и внятные чёткие выводы. Вся проделанная работа должна быть понятна из этого текста. Претензии к отчёту радикальным образом скажутся на итоговой оценке за лабораторную работу. При этом особенно качественно и опрятно написанный отчёт будет поводом для получения бонусных баллов.

Отчёты, написанные не в LaTeX, не принимаются.

**Часть первая (5 баллов)** В этом пункте задания требуется решить задачу классификации текстовой коллекции EUR-lex, пробуя различные признаковые пространства.

Формулировка задания:

1. **(1 балл)** Предобработка данных. Требуется привести данные в подходящий для работы вид, применить средства предобработки (лемматизация, фильтрация словаря и т.п). Все проделанные операции подробно описать в отчёте.

2. **(3 балла)** На основе имеющихся данных подготовить с помощью изученных в курсе инструментов **минимум четыре** признаковых пространства, описывающих датасет. Обязательно должны быть признаки, полученные с помощью word2vec, тематического моделирования и doc2vec. Выбор реализаций инструментов свободный. Отсутствие любого из обязательных признаковых пространств означает невыполнение данного пункта (следующий пункт при этом можно попробовать выполнить).
   **Примечание:** Реализацию doc2vec лучше всего использовать из пакета gensim, а именно модель DBOW с параметром `dbow_words=0`.

3. **(1 балл)** Выбрать один или несколько алгоритмов классификации (на свой вкус) и применить их к полученным признакам, оценивая качество классификации. Сделать исчерпывающие выводы о результатах.
   **Примечание:** алгоритм классификации стоит выбирать так, чтобы он позволял посчитать указанные ниже метрики качества, которые предлагается оптимизировать.

Датасет EUR-lex предлагается в виде двух файлов, `eurlex_data.txt` и `eurlex_labels.txt`. Файл с данными содержит два столбца, первый — идентификатор документа, второй — сам документ. Файл с метками состоит из трёх столбцов,

первый — имя метки класса, второй — документ, к которому эта метка относится (у каждого документа может быть несколько меток классов), третий — константа 1, которую можно игнорировать.

Метриками качества будут ROC AUC и Precision-Recall AUC (PR AUC), посчитанные на тестовой выборке. Будем считать эти величины следующим образом: для каждого документа будем рассматривать вектор, длиной в число меток классов. В векторе ответов будут 1 на позициях верных классов, и 0 - на остальных позициях. В векторе предсказаний на каждой позиции будет вероятность данной метки в данном документе. Будем рассчитывать AUC для указанной пары векторов, после чего просуммируем и усредним получившиеся значения по всем тестовым документам. Для подсчёта обоих типов AUC для одного документа рекомендуется использовать реализации метрик из `sklearn`.

В обязательном порядке требуется написать код, демонстрирующий истинные и предсказанные метки классов и включить в отчёт примеры его работы.

**Часть вторая (3 балла)** Важно не только извлечь данные, но и грамотно их визуализировать.

Формулировка задания:

В этом пункте требуется визуализировать векторы, полученные с помощью word2vec, и матрицу «слова-темы», полученную с помощью тематического моделирования. Выбор инструментов построения модели и визуализации свободный. Требуется, чтобы визуализация была наглядной, давала какую-то информацию о данных и была, по возможности, красивой. Рекомендуется смотреть в сторону t-SNE и библиотек визуализации графов и тематических моделей (самый простой вариант — LDAvis). Пример простейшей визуализации word2vec давался на лекции, пример визуализации матрицы «слова-темы» можно найти здесь. Визуализации должны быть представлены в отчёте вместе с подробным описанием того, как они были получены и какие выводы были сделаны на их основании.

Особенно хорошая визуализация будет поводом для получения бонусных баллов.

**Часть третья (2 балла)** В этом пункте задания требуется решить бинарную задачу анализа тональности текстов постов Twitter. Данные можно скачать по этой ссылке.

Формулировка задания:

1. **(1 балл)** Предобработка данных. Требуется привести данные в подходящий для работы вид, применить средства предобработки (лемматизация, фильтрация словаря и т.п). Все проделанные операции подробно описать в отчёте.
   **Примечание:** данные сырые, поэтому без грамотной предобработки получить нормальные признаки для сантимент анализа не получится. Если итоговое качество будет низким из-за недостаточной предобработки данных, это приведёт к штрафу.

2. **(1 балла)** Собственно классификация. Допускается модель Naive Bayes или любая другая, подходящая для решения данной задачи. Выборку следует разбить случайным образом на две части в соотношении 9:1. Большая часть должна

быть использована для обучения, меньшая — для тестирования.

**Примечание**: подозрительно низкое качество классификации говорит о том, что на каком-то этапе эксперимента была сделана ошибка, стоит проверить всё ещё раз.

Выполненное задание следует присылать на почту курса с подписью «Лабораторная 4 - ФИО». На этот же адрес можно писать свои вопросы по заданию.