**Лабораторная работа №3**

Классификация текстовых сообщений в системе R

В файле SMSSpamCollection.txt (SMSSpamCollection.csv) содержится набор текстовых сообщений, часть из которых является спамом (размечены как spam), остальные - обычные сообщения (размечены как ham).

В системе статистических вычислений R требуется:

1. Построить классификатор, который на основе обучающей выборки размеченных смс-сообщений сможет относить новые сообщения к одному из двух классов: «спам» и «не спам».

2. Для построения классификатора использовать любую математическую модель: модель логистической регрессии, Байесовский классификатор, метод опорных векторов (SVM), нейронные сети и т.д.

3. На начальном этапе осуществить предобработку сообщений: все символы привести к нижнему регистру, удалить пунктуационные знаки, удалить «стоп-слова» и т.п.

4. Используемую выборку разбить на обучающую (70%) и тестовую (30%).

5. Для построения классификатора использовать подход «bag of word», а также «tf-idf»-взвешивание элементов матрицы документов.

6. Подготовить отчет о проделанной работе и полученных результатах.

Дополнительные полезные материалы:

**Статьи с объяснениями**

<https://www.r-bloggers.com/document-classification-using-r/>

<https://www.svm-tutorial.com/2014/11/svm-classify-text-r/>

<https://www.springboard.com/blog/text-mining-in-r/>

<https://blog.paperspace.com/intro-to-datascience/>

**Код для классификации текстовых сообщений в R**

<https://github.com/sureshgorakala/Text-Mining/blob/master/R%20Text%20classfication%20using%20CSV%20files>

<https://gist.github.com/primaryobjects/094d24084d1045c011b7>

https://github.com/alexandrekow/svmtutorial/tree/master/ClassifyTextWithR