# РАБОТА С МЕТОДОМ ОПОЛНЫХ ВЕРКОРОВ (SVM)

## АНАЛИЗ ТЕКСТОВ

Задание основано на применении пакета Python Scikit-learn для изучения метода опорных векторов для анализа текстовой информации (определения того, к какой из тематик относится новость: атеизм или космос).

Цели практики:

- освоить основы работать с методом опорных векторов (SVM)
- овладеть подходами нахождения оптимальных параметров для метода опорных векторов
- начитаться работать с текстовыми данными

Начало работы

Для начала вам потребуется загрузить данные. Для этого нужно воспользоваться модулем datasets:

```
from sklearn import datasets

newsgroups = datasets.fetch_20newsgroups(
                    subset='all',
                    categories=['alt.atheism', 'sci.space']
                )
```

После выполнения этого кода массив с текстами будет находиться в поле **newsgroups.data**, номер класса - в поле **newsgroups.target**.

Одна из сложностей работы с текстовыми данными состоит в том, что для них нужно построить числовое представление (в данной работе применяется вычисление TF-IDF. В Scikit-Learn это реализовано в классе **sklearn.feature_extraction.text.TfidfVectorizer.**

Преобразование обучающей выборки нужно делать с помощью функции **fit_transform**, тестовой - с помощью **transform**.

Реализация SVM-классификатора находится в классе **sklearn.svm.SVC**.

Веса каждого признака у обученного классификатора хранятся в поле **coef**.

Подбор параметров удобно делать с помощью класса **sklearn.grid_search.GridSearchCV** (**sklearn.model_selection.GridSearchCV**).

Пример:

```
cv = KFold(n_splits=5, shuffle=True, random_state=241)
```

Первым аргументом в **GridSearchCV** передается классификатор, для которого будут подбираться значения параметров, вторым **–** словарь **(dict),** задающий сетку параметров для перебора.

После того, как перебор окончен, можно проанализировать значения качества для всех значений параметров и выбрать наилучший вариант:

```
for a in gs.grid_scores_:
# a.mean_validation_score
# a.parameters
```

**Задание**

1. Загрузите объекты из новостного датасета «20 newsgroups», относящиеся к категориям «космос» и «атеизм».

2. Вычислите характеристики TF-IDF для всех текстов. В данной задаче требуется рассчитать TF-IDF по всем данным. Такой подход подразумевает, что признаки в обучающем наборе данных используют информацию из тестовой выборки, соответственно, значения целевой переменной из тестового набора не используются. На практике не редко возникают ситуации, когда признаки объектов тестовой выборки известны на момент обучения, и, следовательно, их можно использовать при обучении алгоритма.

3. Подберите минимальный лучший параметр C из множества $[10^{-5}, 10^{-4} ... 10^{4}, 10^{5}]$ для SVM с линейным ядром **(kernel='linear')** при помощи кросс-валидации по 5 блокам.

Укажите параметр **random_state=241** и для SVM, и для **KFold**. В качестве меры качества используйте долю верных ответов (accuracy).

4. Обучите SVM по всей выборке с лучшим параметром **C**, найденным на предыдущем шаге.

5. Найдите 10 слов с наибольшим по модулю весом. Они являются ответом на это задание. Укажите их через запятую, в нижнем лексикографическом порядке.

**Ответ на каждое задание** — текстовый файл, содержащий ответ в первой строчке (#.txt). Обратите внимание, что отправляемые файлы не должны содержать перевод строки в конце.

**Утонения по выполнению задания:**

Набор данных 20 групп новостей содержит около 18000 сообщений в группах новостей по 20 темам, разделенных на два подмножества: одно для обучения (или разработки), а другое для тестирования (или оценки производительности).
Разделение между обучающим и тестовым наборами основано на сообщениях, опубликованных до и после определенной даты.
https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html
.

## ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

Метод опорных векторов (**Support Vector Machine, SVM**) представляет собой разновидность линейных классификаторов. Его оптимизируемый функционал направлен на максимизацию ширины разделяющей полосы между классами. Согласно теории статистического обучения, данная ширина тесно связана с обобщающей способностью алгоритма, и максимизация её позволяет противостоять переобучению.

Привлекательность линейных методов, в том числе SVM, обусловлена их эффективностью на разреженных данных. Такими данными называют выборки с большим количеством признаков, где у большинства объектов большинство признаков равно нулю.

Разреженные данные часто возникают при анализе текстов. Тексты удобно кодировать с использованием «мешка слов», при этом формируется столько признаков, сколько уникальных слов встречается в текстах.

Значение каждого признака равно числу вхождений соответствующего слова в документ. Поскольку общее число уникальных слов может быть значительным, а конкретное слово часто встречается только в небольшом подмножестве текстов, возникает необходимость более сложных методов кодирования.

Одним из таких методов является использование **TF-IDF** для кодирования текстов. Этот показатель представляет собой произведение двух компонент: **TF** (частота термина) и **IDF** (инверсия частоты документа). Первая компонента отражает отношение числа вхождений слова в документ к общей длине документа, в то время как вторая зависит от того, в скольких документах выборки встречается это слово. TF-IDF высок для слов, часто встречающихся в данном документе и редко встречающихся в других.