# 06. Clustering – Homework

## Overview

* Новостные статьи, разбитые на известных 20 групп, которые можно использовать для оценки качества кластеризации  
  <https://www.kaggle.com/crawford/20-newsgroups/data>
* Или датасет со спамом  
  <https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection>

## Evaluation Criteria

Максимум: 10

- Предварительный exploratory analysis (различные гистограммы длин тексов, распределения топовых по популярности слов, попробовать придумать визуализацию с помощью методов понижения размерности и т.д.) -1

- Попробовать различные векторные представления текстов (не только tf-idf. посмотреть в сторону LDA, word2vec, doc2vec, lda2vec) -2

- Построить k-Means с известным (двадцатью) количеством кластеров -1

- Определить количество кластеров для k-Means, предполагая его неизвестным -1

- Придумать визуализацию полученных кластеров -1

- Попробовать другие алгоритмы кластеризации (хотя бы еще два помимо стандартного kMeans) и потюнить их гиперпараметры -2

- Оценить все алгоритмы кластеризации используя метрики, использующие и неиспользующие известные лейблы -2