*Теоретическое домашнее задание не сдается и не проверяется. Вместо проверки **в начале** каждого семинара будет устраиваться самостоятельная работа, которая будет включать в себя набор задач из домашнего задания. Время, отведенное на выполнение самостоятельной работы, будет невелико (около 10 мин). Использование чего-либо, кроме ручки и выданных листов бумаги, запрещено. В случае обнаружения списывания, подлога и т.д. выставляется оценка "0".*

Теоретическое домашнее задание №4

Препроцессинг наше всё

Задание 1.

Возраст	Зарплата (в тысячах долларов)	Есть ли водительские права	Марка автомобиля	Пол
45	2.2	Yes	Audi	M
	1.5		Renault	M
28	1.3	Yes	Citroen	F
35		No		F
30	2.5	No		M
32	1.8	No		F
45		Yes	Citroen	F
54	3.1			M
24	1.8	Yes	Kia	M
26	2.9	No		M
	3.5			M

1. Заполните пропуски в данных. Выберите способ заполнения самостоятельно. Объясните, почему вы выбрали ваш способ заполнения данных.

2. Сделайте One Hot Encoding для столбца "Есть ли водительские права".

3. Закодируйте столбец "Марка автомобиля" любым способом, объясните ваш выбор.

Задание 2.

Пусть переменная $mydatetime$ является признаком, содержащим дату и время (например, она может быть равна $pd.tslib.Timestamp$('2016-03-03 04:22:07.000')). Какой способ извлечения признаков будет наименее полезным?

a) $mydatetime.weekday() * 24 + mydatetime.hour$
b) $str(mydatetime)$ с последующим one hot encoding
c) $mydatetime.weekday()$
d) $mydatetime.hour$

Задание 3.

а) В некотором документе D, содержащем 100 слов, есть слово «экономика», которое встречается 5 раз. Всего у нас есть 1000 документов (включая документ D), и слово «экономика» встречается в 10 из них. Вычислите TF-IDF слова «экономика» в документе D.

б) Проведем стемминг всех документов из пункта а), после него следующие слова: «экономика», «экономист», «экономить»,«экономичный» превратятся в слово «эконом». До стемминга все эти слова встретились по 5 раз каждое в документе D. Также каждое из этих слов до стемминга упоминалось в 10 документах из 1000, причем в каждом из этих 10 документов встречались все описанные слова. Вычислите TF-IDF слова «эконом» в документе D после стемминга.

Убедитесь, что вы знаете ответы на следующие вопросы:

1) Как выглядят формулы масштабирования в MinMaxScaler и в StandardScaler?

2) Пусть данные имеют категориальные признаки с большим количеством (больше миллиона) различных значений. В чем заключается проблема метода One Hot Encoding?

3) Что делает LabelEncoder с категориальными признаками? Почему для кодирования категориальных признаков недостаточно применения LabelEncodera?

4) Какую особенность распределения слов в документах и документов в коллекции мы не учтем, если вместо tf-idf будем вычислять просто tf?

5) На каком основном предположении относительно контекстов слов базируется принцип работы Word2Vec?