****

# Задание для демонстрационного экзамена по комплекту оценочной документации № 1.1 по компетенции № F5 «Машинное обучение и большие данные»

Вариант 1

Модули с описанием работ

Модуль 1: Предобработка данных и визуализация зависимостей

**Исходные файлы:**

1. Выгрузка 9 апреля.xls (Набор данных)
2. Модуль 1.pdf (Инструкция к 1 модулю)
3. Алгоритм.pdf (Правила работы с документами)

**Результаты работы:**

1. Data.zip (Предобработанные данные)
2. Report.html + Report.ipynb (Отчет о проделанной работе)
3. Readme.txt (Дополнительные комментарии)

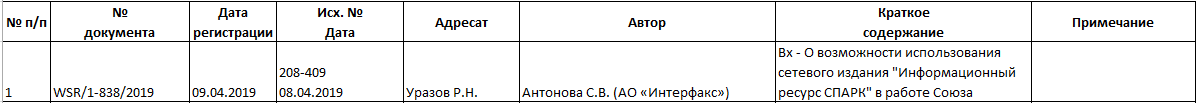
**ВВЕДЕНИЕ**

На этом демонстрационном экзамене вам предстоит разработать программный модуль (чат-бот) для системы автоматической сортировки входящей корреспонденции Союза “Агентство развития профессиональных сообществ и рабочих кадров “Молодые профессионалы (Ворлдскиллс Россия)” (далее – Союз).

В настоящее время Союз использует систему электронного документа оборота «Практика» (далее – СЭД). Возможности этой системы представлены на сейте <http://sd-praktika.ru/>

В качестве исходного набора данных представлена выгрузка из СЭД на момент 09.04.2019. Данные представлены в таблице формата .xls и имеют следующие атрибуты: порядковый номер, номер документа, дата регистрации, исходящий номер документа, дата, адресат, автор, краткое содержание и примечание.

Поля формируются следующим образом:



1. Номер строки
2. Номер входящего документа, присваиваемый СЭД
3. Дата получения документа
4. Номер исходящего документа
5. Адресат получения документа в Союзе
6. Автор документа, его должность и организация
7. Содержание документа
8. Любые комментарии или переадресации документа

Вся входящая корреспонденция направляется сотрудникам Союза в соответствии с Приказом №01.06.2018-2 от 01.06.2018 Об усовершенствовании системы документооборота по основной деятельности. В соответствии с этим приказом адресат определяется на основании алгоритма, приведенного в Приложении 1. Текст приказа и приложения находятся в файле Алгоритм.pdf

В этом модуле необходимо произвести предобработку данных для дальнейшего исследования и построения модели обучения.

**ЗАДАНИЕ**

| 1.1 Предобработка данных и выделение значимых атрибутов |
| --- |
| Задача автоматической сортировки корреспонденции заключается в определении адресата для входящих документов. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на классификацию входящих документов и оставить только их для последующего обучения. Также необходимо обосновать выбор атрибутов |

| 1.2 Разбиение сложных атрибутов |
| --- |
| В исходных данных есть поля, представляющие собой конкатенацию нескольких, иногда разнотипных, значений. Необходимо выделить и разбить такие поля на несколько других |

| 1.3 Дополнение недостающими данными |
| --- |
| Согласно алгоритму направления входящей документации, письма от конкретных лиц и организаций направляются разным сотрудникам Союза. Так, например, от первых лиц министерств – непосредственно Генеральному директору, независимо от указанного адресата. Однако в исходных данных отсутствует информация о том, кем является отправитель. Такая же ситуация с адресатом, система, определяющая направление документа по фамилии будет являться частным случаем решения, и в случае кадровых перестановок не сможет функционировать. В связи с этим, необходимо проанализировать исходный набор данных и дополнить его |

| 1.4 Визуализация зависимостей данных |
| --- |
| Используя программные средства, визуализируйте зависимости атрибутов в наборе данных. Визуализация должна отражать влияние атрибутов на определение класса (адресата). Произведите расчеты зависимостей по выбранным алгоритмам. Приведите интерпретацию полученным результатам |

| 1.5 Подготовка отчета |
| --- |
| Подготовьте отчет о проделанной работе по итогам модуля, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip |

Модуль 2: Построение модели машинного обучения

**Исходные файлы:**

1. Data.zip (Предобработанные данные)
2. Модуль 2.pdf (Инструкция ко 2 модулю)

**Результаты работы:**

1. Data.zip (Результаты модуля)
2. Report.html + Report.ipynb (Отчет о проделанной работе)
3. Readme.txt (Дополнительные комментарии)

**ВВЕДЕНИЕ**

В этом модуле вы продолжаете работать с данными, подготовленными в предыдущем модуле. На основе предобработанных данных необходимо построить и обучить модель классификации, а также добиться наиболее точной работы построенной модели. Какая-либо работа, обусловленная задачами предыдущего модуля, выполненная в ходе текущего, оцениваться не будет, поэтому проделывайте её только в случае необходимости.

**ЗАДАНИЕ**

| 2.1 Классификация документов |
| --- |
| Выберите модель классификации входящих документов по адресату, приведите обоснование выбора модели. Разделите исходный набор данных на обучающую и тестирующую выборки оптимальным образом. |

| 2.2 Обучение |
| --- |
| Проведите обучение выбранной модели на обучающей выборке, сформированной в предыдущем модуле. Протестируйте работу обученной модели на тестовой выборке. Определите показатели точности работы выбранной модели, сравните с остальными рассматриваемыми моделями. |

| 2.3 Feature Engineering |
| --- |
| Путём преобразования набора данных, добейтесь более точной работы выбранной модели. Опишите приемы генерации новых данных и результаты, к которым они привели, рассматривая все ранее определенные показатели точности |

| 2.4 Подготовка отчета |
| --- |
| Подготовьте отчет о проделанной работе по итогам модуля, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip |

1. **Необходимые приложения**

КОД 1.1 Вариант 1 Приложение 1.

Содержит:

1. Выгрузка 9 апреля.xls - набор данных;
2. Алгоритм.pdf - правила работы с документами.