**Безопасность больших данных**

**1 блок**

1. Опиши область обработки Больших данных и основные характеристики данной области.
2. Охарактеризуйте, какие технологические вызовы связаны с обработкой больших данных и как эти вызовы влияют на выбор технологий для их обработки.
3. Опишите процесс очистки Больших данных для последующего анализа.
4. Опишите операции стемминга и лемматизации данных.
5. Дайте характеристику основным этапам обработки текстов в NLP, нарисуйте схему этапов обработки текстов.
6. Приведите характеристику различных методов векторизации, таких как Bag of words, Tf-idf и другие.
7. Опишите векторизацию данных с помощью word embeddings: Word2Vec, FastText и т.д.
8. Опиши работу с Большими данными интернет-угроз, такими как DDoS, Man in the Middle, Malware, Phishing, SQL injection.
9. Опиши основные типы интернет-угроз могут быть обнаружены с помощью методов машинного обучения.
10. Опишите этапы классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз. Охарактеризуйте алгоритмы классификации Support Vector Machine и Метод k-ближайших соседей.
11. Дайте характеристику методам классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз. Опишите алгоритмы классификации Decision Tree и Random Forest.
12. Опишите этапы классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз. Охарактеризуйте алгоритмы классификации XgBoost и CatBoost.
13. Опишите, какие типы нейронных сетей обычно используются для классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
14. Дайте характеристику нейронной сети прямого распространения (Feedforward neural network) для классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
15. Опишите применением сверточных нейронных сетей (CNN) для классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
16. Объясните, как рекуррентные нейронные сети (RNN) могут быть использованы для классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
17. Дайте характеристику нейронной сети LSTM (Long Short-Term Memory) и почему они эффективны в задачах классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
18. Охарактеризуйте модель BERT (Bidirectional Encoder Representations from Transformers) для классификации DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
19. Объясните, какие методы можно использовать для оценки качества модели, обученной на задаче обнаружения DDoS, Man in the Middle, Malware, Phishing, SQL injection угроз.
20. Опишите технологические решения необходимы для обеспечения анализа больших данных интернет-угроз

**2 блок**

1. Опишите, какие методы обработки больших данных необходимо использовать для классификации нейронными сетями.
2. Опишите стратегии, которые используются для борьбы с переобучением нейронных сетей в задачах классификации больших данных.
3. Приведите методы балансировки в задачах классификации больших данных.
4. Опишите основные принципы работы моделей GPT и ChatGPT. Какие технологии лежат в их основе.
5. **Опиши использование** GPT для анализа текстовых данных. Приведите примеры конкретных задач.
6. **Объясните,** как размер тренировочного датасета влияет на производительность моделей GPT. Объясните, почему это важно при работе с большими данными.
7. **Опишите**, как можно адаптировать предварительно обученную модель GPT для специфических задач анализа больших данных.
8. Охарактеризуй этические проблемы, которые могут возникнуть при использовании GPT для анализа данных. Обсудите потенциальные риски и способы их минимизации.
9. Приведите примеры использования ChatGPT для поддержки принятия решений на основе анализа больших данных.
10. Опишите основные отличия и преимущества использования GPT по сравнению с традиционными моделями машинного обучения для анализа текстовых данных.
11. Опишите тенденции и развитие, которые можно ожидать в области использования GPT и подобных моделей для анализа больших данных в ближайшие годы
12. **Опишите инструмент** Power BI и какие основные компоненты входят в состав этого инструмента.
13. Охарактеризуйте, какие источники данных можно подключать к Power BI? Приведите примеры и описания процесса импорта данных из одного из этих источников.
14. **Опишите связи данных** в Power BI и как они управляются.
15. **Охарактеризуйте** DAX и для чего он используется в Power BI. Приведите пример простой функции DAX и ее использования в реальном сценарии.
16. **Опишите** типы визуализаций данных в Power BI? Опишите процесс создания интерактивной визуализации.
17. **Опишите Web interface Power BI, какие он имеет особенности.**
18. Опишите методы можно использовать для оптимизации производительности отчетов в Power BI.
19. Охарактеризуйте уровень защищенности больших данных в веб-приложениях.
20. Опишите разработку веб-приложения с использованием защиты данных методами машинного обучения

**3 блок**

1. Создайте небольшой скрипт по загрузке данных с помощью библиотеки Pandas и очистке данных от лишних символов с помощью регулярных выражений.
2. Создайте небольшой скрипт на Python с добавлением стемминга слов.
3. Создайте скрипт на Python с применением лемматизации слов.
4. Создайте скрипт использования one-hot encoding для категориальных переменных с помощью Pandas или Scikit-learn.
5. Создайте скрипт по работе с регулярными выражениями в Python для очистки текстовых данных.
6. Напишите скрипт векторизации данных с помощью метрики *Tf-idf* на Python.
7. Создайте скрипт на Pandas для загрузки нескольких данных и их слияния в единый DataFrame.
8. Создайте скрипт по предобработке данных и их разделения на train и test части.
9. Создайте скрипт масштабирования данных в библиотеке Scikit-learn.
10. Создай небольшой скрипт классификации текстов данных с помощью библиотеки Scikit-learn.
11. Напишите небольшой скрипт векторизации с помощью Bag of words на Python
12. Дан небольшой скрипт. Дополни его команда для классификации с помощью Decision Tree.

import tensorflow as tf

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense, Embedding, GlobalAveragePooling1D

# Пример данных

reviews = [

"Great movie! Must watch.",

"Worst movie ever.",

"It was a fantastic performance!",

"Not a good movie, boring plot.",

"The cinematography is pretty great.",

"I hated the storyline."

]

labels = [1, 0, 1, 0, 1, 0] # 1 - положительный, 0 - отрицательный

1. Дан небольшой скрипт. Дополни его команда для классификации с помощью Random Forest.

import tensorflow as tf

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense, Embedding, GlobalAveragePooling1D

# Пример данных

reviews = [

"Great movie! Must watch.",

"Worst movie ever.",

"It was a fantastic performance!",

"Not a good movie, boring plot.",

"The cinematography is pretty great.",

"I hated the storyline."

]

labels = [1, 0, 1, 0, 1, 0] # 1 - положительный, 0 - отрицательный

1. Дан небольшой скрипт. Дополни его команда для классификации с помощью Dense neural network

import tensorflow as tf

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense, Embedding, GlobalAveragePooling1D

# Пример данных

reviews = [

"Great movie! Must watch.",

"Worst movie ever.",

"It was a fantastic performance!",

"Not a good movie, boring plot.",

"The cinematography is pretty great.",

"I hated the storyline."

]

labels = [1, 0, 1, 0, 1, 0] # 1 - положительный, 0 - отрицательный

1. Дан небольшой скрипт. Дополни его команда для классификации с помощью Long short-term memory neural network

import tensorflow as tf

from tensorflow.keras.preprocessing.text import Tokenizer

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense, Embedding, GlobalAveragePooling1D

# Пример данных

reviews = [

"Great movie! Must watch.",

"Worst movie ever.",

"It was a fantastic performance!",

"Not a good movie, boring plot.",

"The cinematography is pretty great.",

"I hated the storyline."

]

labels = [1, 0, 1, 0, 1, 0] # 1 - положительный, 0 - отрицательный

1. Создать новую меру в Power BI, которая рассчитывает общий объем продаж по всем категориям продуктов.

**Данные:** Используйте таблицу Sales, содержащую колонки ProductID, UnitPrice, и Quantity.

**Формула:** Создайте меру Total Sales с использованием формулы DAX

1. Написать формулу DAX, которая вычисляет среднее количество проданных единиц продукции по дням.

**Данные:** Используйте таблицу Sales, содержащую колонки Date и Quantity.

**Формула:** Создайте меру Average Daily Sales

1. Опиши методы, которые можно применить для улучшения производительности загрузки и визуализации данных в большом отчете Power BI.
2. Опиши создание гистограммы, отображающей общий объем продаж по месяцам. Пошагово опишите выбор типа графика и добавление данных на график.
3. Опишите процесс трансформации данных в Power Query, например, как изменить формат данных в столбце с датами из американского формата (MM/DD/YYYY) в европейский (DD/MM/YYYY).