1. Сопоставьте терминам определения

1) Токенизация	а) перевод последовательности байт в последовательность символов
2) Нормализация	б) приведение грамматических форм слова и однокоренных слов к единой основе с помощью простых эвристических правил
3) Стемминг	в) приведение токенов к единому виду
4) Лемматизация	г) приведение грамматических форм слова и однокоренных слов к единой основе с использованием словарей и морфологического анализа
	д) удаление наиболее частых слов в языке, не содержащих информации о содержании текста
	е) разбиение последовательности символов на части

2. Закон Хипса:
а) описывает зависимость между количеством слов в корпусе и размером словаря
б) имеет теоретическое доказательство
в) справедлив не только для уникальных слов, но и для других понятий окружающего мира
г) означает, что при увеличении количества изучаемых текстов скорость заполнения словаря увеличивается

3. TF-IDF:
а) статистическая мера, используемая для оценки важности документа, являющегося частью коллекции документов или корпуса
б) выбор основания логарифма в формуле не имеет значения
в) большой вес получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах
г) может использоваться для построения векторной модели

4. Naive Bayes:
а) хорошо работает, потому что точно предсказывает вероятности классов
б) предполагает, что слова появляются в документе на разных позициях с разными вероятностями
в) не требует знания априорных вероятностей классов
г) требует квадратичного по размеру словаря количества памяти при обучении

1. Сопоставьте терминам определения

1) Токенизация	а) перевод последовательности байт в последовательность символов
2) Нормализация	б) приведение грамматических форм слова и однокоренных слов к единой основе с помощью простых эвристических правил
3) Стемминг	в) приведение токенов к единому виду
4) Лемматизация	г) приведение грамматических форм слова и однокоренных слов к единой основе с использованием словарей и морфологического анализа
	д) удаление наиболее частых слов в языке, не содержащих информации о содержании текста
	е) разбиение последовательности символов на части

2. Закон Хипса:
а) описывает зависимость между количеством слов в корпусе и размером словаря
б) имеет теоретическое доказательство
в) справедлив не только для уникальных слов, но и для других понятий окружающего мира
г) означает, что при увеличении количества изучаемых текстов скорость заполнения словаря увеличивается

3. TF-IDF:
а) статистическая мера, используемая для оценки важности документа, являющегося частью коллекции документов или корпуса
б) выбор основания логарифма в формуле не имеет значения
в) большой вес получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах
г) может использоваться для построения векторной модели

4. Naive Bayes:
а) хорошо работает, потому что точно предсказывает вероятности классов
б) предполагает, что слова появляются в документе на разных позициях с разными вероятностями
в) не требует знания априорных вероятностей классов
г) требует квадратичного по размеру словаря количества памяти при обучении