

Имя функции	Описание
<code>extractFileText</code>	Чтение из PDF, Microsoft Word и обычного текста
<code>textscan</code>	Считывание отформатированных данных из текстового файла или строки
<code>readtable</code>	Создание таблицы из файла
<code>compose</code>	Преобразование данных в форматированный массив строк
<code>xlsread</code>	Чтение файла электронной таблицы Microsoft Excel
<code>webread</code>	Чтение содержимого из веб-службы RESTful
<code>TabularTextDatastore</code>	Хранилище данных для табличных текстовых файлов
<code>FileDatastore</code>	Хранилище данных с пользовательским устройством чтения файлов
<code>SpreadsheetDatastore</code>	Хранилище данных для файлов электронных таблиц



Импорт

Извлечение текста из файлов Microsoft® Word®, PDF-файлов, текстовых файлов и электронных таблиц.

“Performed preventive maintenance serviceing on a broken pump”

Предварительная обработка

Удалите менее полезные артефакты, такие как общие слова, знаки препинания и URL-адреса, и примените нормализацию текста к корням слов.

Имя функции	Описание
<code>tokenizedDocument</code>	Разбить документы на коллекции слов
<code>normalizeWords</code>	Удаление окончания из слов с помощью стеммера Портера
<code>bagOfWords</code>	Модель мешок слов
<code>stopWords</code>	Список стоп слов
<code>context</code>	Поиск документов для употребления слов в контексте
<code>removeWords</code>	Удаление выбранных слов из документа или мешка слов
<code>removeLongWords</code>	Удаление длинных слов из документа или мешка слов
<code>removeShortWords</code>	Удаление коротких слов из документа или мешка слов
<code>removeInfrequentWords</code>	Удаление редко встречающихся слов из модели мешка слов
<code>erasePunctuation</code>	Удаление пунктуации из текста и документов

Имя функции	Описание
<code>str = "Hello,world"</code>	Объявление строковой переменной
<code>str = ["Hello", "World"]</code>	Объявление массива строк
<code>str = string(C)</code>	Преобразование символьного вектора C в строку
<code>str2double</code>	Преобразование строки к типу double
<code>strlen</code>	Возвращает длину строки
<code>isstring</code>	Определяет является ли вход массивом строк
<code>join</code>	Объединение строк
<code>split</code>	Разделение строк в строковом массиве
<code>splitlines</code>	Разделяет строку на символы новой строки
<code>replace</code>	Поиск и замена подстрок в строковом массиве
<code>contains</code>	Определите, находится ли шаблон в строке
<code>erase</code>	Удаление подстрок в строках
<code>extractBetween</code>	Извлечение подстрок между индикаторами
<code>extractAfter</code>	Извлечь подстроку после указанной позиции
<code>extractBefore</code>	Извлечение подстроки перед указанной позицией
<code>strcmp</code>	Сравнение строк
<code>regex</code>	Соответствие регулярному выражению (чувствительно к регистру)

"Hello,world"

Строка

Эффективно комбинируйте, сравнивайте и храните текстовые данные.