
DocFlow Content Analyzer — поиск и классификация документов по содержимому

Datacult:

Денис Власюк - сбор и анализ информации

Яна Володина – занималась сайтом и frontend разработкой

Вячеслав Камлин – backend, логика кода



Цель:

Создать приложение
которое оптимизирует
работу с
многотысячными базами
данных

Задачи:




















- 1) преобразовать pdf файлы в текст
- 2) построить иерархическую модель хранения файлов
- 3) классифицировать при помощи nlp файлы на категории и подкатегории
- 4) апробировать данное приложение

План работы:

- определение цели и задач;
- работа с источниками;
- создание логики кода;
- проработка отцифровки pdf файлов;
- кластеризация данных;
- написание frontend части проекта;
- создание иерархической системы для отображения файлов;
- создание поисковика.

-
- Почему tesseract?

All models 

Model Name	Performance Sentence Embeddings (14 Datasets) 	Performance Semantic Search (6 Datasets) 	 Avg. Performance 	Speed 	Model Size 
all-mpnet-base-v2 	69.57	57.02	63.30	2800	420 MB
multi-qa-mpnet-base-dot-v1 	66.76	57.60	62.18	2800	420 MB
all-distilroberta-v1 	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v2 	68.70	50.82	59.76	7500	120 MB
multi-qa-distilbert-cos-v1 	65.98	52.83	59.41	4000	250 MB
all-MiniLM-L6-v2 	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-cos-v1 	64.33	51.83	58.08	14200	80 MB
paraphrase-multilingual-mpnet-base-v2 	65.83	41.68	53.75	2500	970 MB
paraphrase-albert-small-v2 	64.46	40.04	52.25	5000	43 MB
<u>paraphrase-multilingual-MiniLM-L12-v2 </u>	64.25	39.19	<u>51.72</u>	<u>7500</u>	<u>420 MB</u>
paraphrase-MiniLM-L3-v2 	62.29	39.19	50.74	19000	61 MB
distiluse-base-multilingual-cased-v1 	61.30	29.87	45.59	4000	480 MB
distiluse-base-multilingual-cased-v2 	60.18	27.35	43.77	4000	480 MB