

데이터 수집 보고서

제출일자 : 2025.02.12. (수)

작성 팀원 : 김요은, 장정호, 김혜서

▶ 프로젝트 개요

- 프로젝트 주제 : 카메라 사용자 매뉴얼 검색 시스템

- 프로젝트 목표

- 1) 고객상담 업무 효율성을 극대화하기 위해 구매고객을 대상으로 서비스되고 있는 문서의 검색 시스템을 질의 응답 형식으로 구축
- 2) 카메라 사용자의 카메라에 대한 정보 접근성 확대 및 초기 카메라 문제 발생 시 간편 해결 방안 제시로 편의성 증대

▶ 데이터 수집 및 전처리 목적

- 수집 데이터 : 카메라 사용자 매뉴얼

- 1) 각 브랜드 별 홈페이지에 업로드 되어있는 카메라 사용 매뉴얼 데이터
- 2) 활용 브랜드 : 캐논, 소니, 후지 필름
- 3) 활용 모델
 - a) 캐논 : EOS 200DII, EOS M50MarkII, EOS R50MarkII, EOS R6, PowerShotG7XMarkIII
 - b) 소니 : ILCE-6400 α6400, ILCE-7M3 α7III, DSC-RX100M7, ZV-1, ZV-E10
 - c) 후지 : GFX100II, X-E4, X-S20, X-T5, X100V

- 데이터 수집 목적 : 카메라 사용자 매뉴얼 내에서 검색할 수 있는 모델을 구현하기 위한 수집

- 데이터 전처리 목적 : 이미지, 표, 텍스트가 혼합되어있는 PDF 파일을 모델이 이해하고, 비슷한 방식으로 답변을 생성할 수 있도록 하기 위해
이미지의 위치, 표의 텍스트 등까지 처리할 수 있게 전처리.

▶ 데이터 수집 및 전처리 과정

- 데이터 수집 과정 : 활용 브랜드 별 홈페이지에 접속하여 사용자 메뉴얼 PDF 다운로드

- 데이터 전처리 과정

1) 텍스트 파싱

- a) 사용 Parser : LLama Parser, Pymupdf
- b) 문제점 : 텍스트와 함께 혼용되고 있는 이모티콘의 제대로 된 파싱이 진행되지 않아
텍스트의 맥락을 잃어버리는 경우가 발생함.
- c) 해결 방안 : LLama Parser 에서 MultiModal 모드를 활용하여
LLM을 통한 파싱 데이터를 받아옴.

2) 이미지 파싱

- a) 사용 Parser: LLama Parser, Upstage Parser
- b) 텍스트로 설명하지 않고 이미지를 통해 설명되어있는 데이터가 다수 있어
이미지에 대한 설명 혹은 이미지를 LLM 모델에 함께 전달해야 함.
- c) 해결 방안 : LLama Parser 를 통해 파싱 데이터에 이미지 위치를 함께 전달,
Upstage Parser 를 통해 PDF 내에서 이미지를 추출하고 크롭하여 저장

3) 메타데이터 생성

- a) 각 브랜드 노드별 여러 모델의 PDF 를 사용하기 때문에, 데이터의 필터링을 위한 브랜드,
모델명 필요
- b) PDF 1개의 페이지 수가 적으면 300장, 많으면 1000장이 넘어가고, 이미지를 함께
처리해야 하기 위한 페이지 수 필요
- c) 메뉴얼 파일이기 때문에 존재하는 인덱스 페이지 제거 및 효율적인 인덱스 작업을 위한
메타데이터 추가 필요
- d) 진행 방안 : 파싱 진행 시 메타데이터 내 PDF 모델명, 브랜드명, 페이지 수 추가, 파싱된
인덱스 페이지를 바탕으로 챕터 혹은 메인 인덱스 형식으로 메타 데이터 추가

▶ 데이터 수집 및 전처리 과정

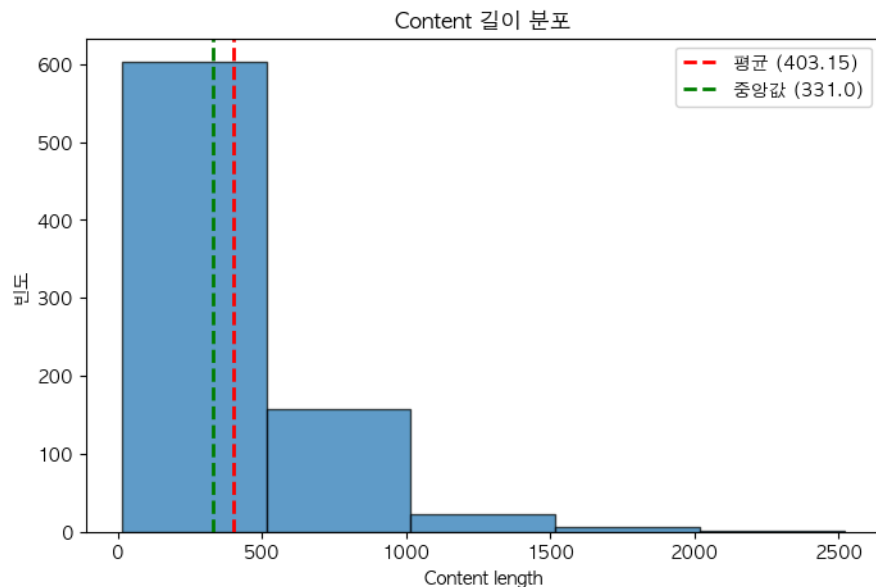
4) 청킹

a) 파싱된 데이터 분할

b) 목적 : LLM 모델 구현 시, 모델에 들어가는 토큰 개수를 효율적으로 활용하기 위함.

c) 진행 과정 : 각 브랜드 모델 별 다른 방식으로 구현

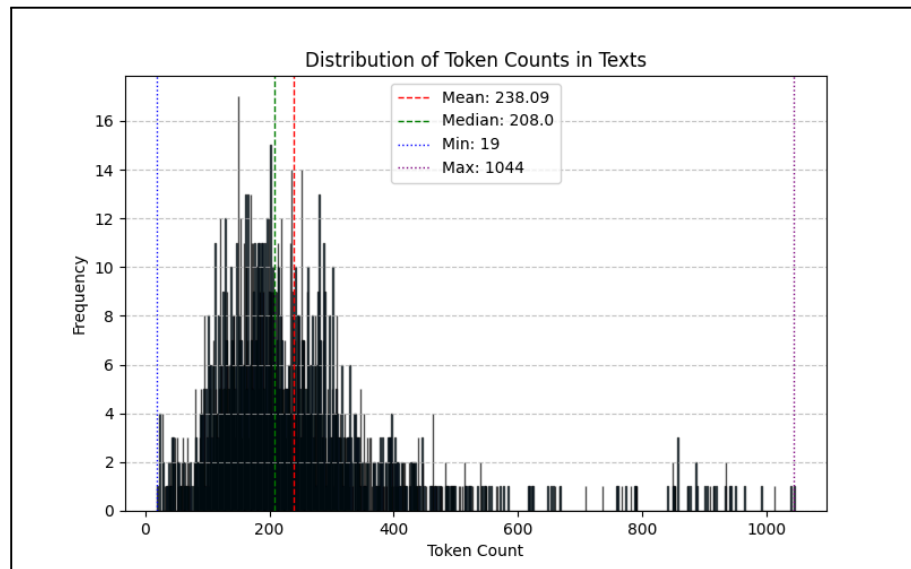
- 캐논 : RecursiveCharacterTextSplitter 진행(청크 사이즈 1000, 오버랩 500)
 - 페이지별 청크 개수 : 평균 403.15(최대 2519, 최소 15)
 - 주의사항과 같은 페이지는 청크가 큰 경우가 있어 효율적인 RAG 운영을 위한 split 진행
 - page 별 토큰 히스토그램



- 1000 토큰을 넘는 경우가 거의 없으므로, 1000청크를 기준으로 청킹 진행
- 소니 : 파싱된 json파일 데이터 문서를 기반으로 텍스트 청킹을 진행
 - page 단위로 데이터를 읽음
 - 청킹 함수(chunk_text)를 사용하여 max_length=512 최대 길이 지정
 - 문장 잘림을 방지하여 overlap=100 지정 (슬라이딩 윈도우 방식)

▶ 데이터 수집 및 전처리 과정

- 후지
 - GPT-4o의 Context Window 제한 : 128,000 tokens,
Output Token 제한 : 16,384 tokens
 - 후지 PDF 파일들의 Page 별 토큰 크기 비교

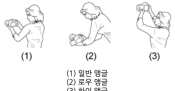
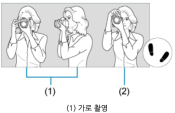




- **평균(Mean):** 238 tokens, **중앙값(Median):** 208 tokens
- 즉, 일반적인 토큰 크기는 약 200~250 tokens
- 입력 예상 토큰 수 (3장 입력 시 약 700 tokens) 는 GPT-4o의 Context Window(128,000 tokens) 대비 매우 적은 수준이기 때문에 chunking 없이도 token 제한에 문제가 없을 것으로 판단되어 페이지 별로 청킹된 데이터 사용

▶ 데이터 전처리 결과

- 각 RAG 모델 별 파싱 및 전처리 결과

- 캐논

원본	Llama Parser	Upstage Parser
<p>촬영 자세</p> <p>● 스크린을 보면서 촬영하기 촬영하면서 스크린을 잘라거나 조정할 수 있습니다. 자세한 내용은 스크린 사용법기를 참조하십시오.</p>  <p>(1) 일반 앵글 (2) 로우 앵글 (3) 하이 앵글</p> <p>● 뷰파인더를 보면서 촬영하기 선명한 이미지를 얻으려면 카메라를 안정되게 잡아 카메라 흔들림을 최소화해야 합니다.</p>  <p>(1) 가로 촬영 (2) 세로 촬영</p> <p>1. 오른손으로 카메라의 그림자를 단단히 잡으십시오. 2. 왼손으로 렌즈를 아래에서부터 감싸듯이 받치십시오. 3. 오른손의 엄지를 세로 바깥 방향으로 뻗어주세요. 4. 양팔과 양손을 상체의 편에 가깝게 밀착하십시오. 5. 팔꿈치와 손목은 팔과 어깨에 90도 안팎의 각도를 유지할 수 있도록 하십시오. 6. 카메라를 얼굴 가까이 대고 뷰파인더를 들여다보십시오.</p>	<p>## 셔터 버튼</p> <p>셔터 버튼은 2단계로 되어 있습니다. 셔터 버튼을 반누름한 다음 추가적으로 완전히 누를 수 있습니다.</p> <p>### 반누름</p> <p>! [그림 자리 (셔터 버튼 반누름 설명)]</p> <p>자동 초점 및 셔터 스피드와 조리개 값을 설정하는 자동 노출 시스템이 실행됩니다. 노출값 (셔터 스피드와 조리개 값) 이 스크린이나 뷰파인더에 약 8초간 표시됩니다 (측광 타이머) .</p> <p>### 완전 누름</p> <p>! [그림 자리 (셔터 버튼 완전 누름 설명)]</p> <p>셔터를 개방시켜 사진을 촬영합니다.</p> <p>### 카메라 흔들림 방지하기</p> <p>카메라를 손에 들고 있을 때 노출 순간에 일어나는 카메라의 움직임을 카메라 흔들림이라고 합니다. 카메라 흔들림이 발생하면 이미지가 흐릿해질 수 있습니다. 카메라 흔들림을 방지하려면 아래의 사항에 유의하십시오:</p> <ul style="list-style-type: none"> - **촬영 자세**의 설명을 따라 카메라를 안정적으로 잡으십시오. - 먼저 셔터 버튼을 반누름하여 자동으로 초점을 맞춘 다음 버튼을 천천히 끝까지 누르십시오. 	 <p>(1) 일반 앵글 (2) 로우 앵글 (3) 하이 앵글</p>  <p>(1) 가로 촬영 (2) 세로 촬영</p>

데이터 수집 및 저장

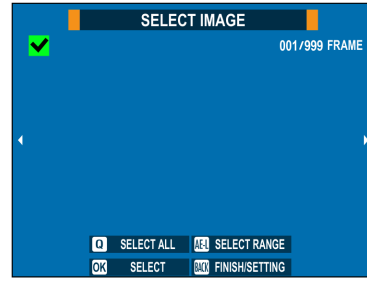
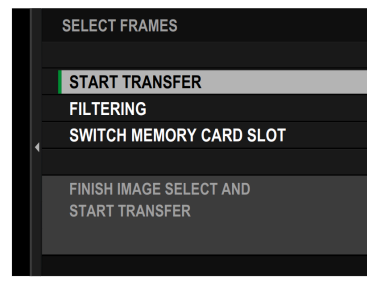
▶ 데이터 전처리 결과

- 소니

원본	Llama Parser	Upstage Parser
 <p>SONY 도움말 안내</p> <p>디지털 카메라 ZV-1</p> <p>장면 선택</p> <p>장면에 따른 사진 설정을 사용해서 촬영할 수 있습니다.</p> <p>① MODE (모드) 버튼 → [장면 선택] → 조작 휠의 가운데를 누릅니다. * MENU →  (카메라 설정) → [촬영 모드]를 선택해서 촬영 모드를 설정할 수도 있습니다.</p> <p>② 조작 휠을 돌려서 원하는 설정을 선택하여 주십시오.</p> <p>특히 항목 세부 내용</p> <p>! 인물: 배경을 흐리게 해서 피사체가 선명하게 나오도록 합니다. 피부 색을 부드럽게 강조합니다.</p> <p>! 스포츠 액션: 움직이는 피사체가 정지된 것처럼 보이도록 고속 셔터로 촬영합니다. 셔터 버튼을 누르고 있는 동안 제품이 연속으로 이미지를 촬영합니다.</p> <p>! 일물: 일물의 붉은 색을 아름답게 촬영합니다.</p>	<p># 도움말 안내</p> <p>디지털 카메라 ZV-1</p> <p>## 장면 선택</p> <p>장면에 따른 사진 설정을 사용해서 촬영할 수 있습니다.</p> <p>1. **MODE (모드) 버튼 → [장면 선택] → 조작 휠의 가운데를 누릅니다. ** - MENU → !1 → [촬영 모드]를 선택해서 촬영 모드를 설정할 수도 있습니다.</p> <p>2. **조작 휠을 돌려서 원하는 설정을 선택하여 주십시오. **</p> <p>## 메뉴 항목 세부 내용</p> <p>### 인물: 배경을 흐리게 해서 피사체가 선명하게 나오도록 합니다. 피부 색을 부드럽게 강조합니다.</p> <p>! 인물 사진</p> <p>### 스포츠 액션: 움직이는 피사체가 정지된 것처럼 보이도록 고속 셔터로 촬영합니다. 셔터 버튼을 누르고 있는 동안 제품이 연속으로 이미지를 촬영합니다.</p> <p>! 스포츠 액션 사진</p> <p>... ### 일물: 일물의 붉은 색을 아름답게 촬영합니다.</p> <p>! 일물 사진</p>	        

▶ 데이터 전처리 결과

- 후지

원본	LLama Parser	Upstage Parser
<p>The Playback Menu</p> <p>TRANSFER IMAGE TO SMARTPHONE Select photos for upload to a paired smartphone or tablet via Bluetooth (see page 249).</p> <p>1 Select TRANSFER IMAGE TO SMARTPHONE > SELECT FRAMES. Select RESET to remove "upload to smartphone" marking from all pictures before proceeding.</p> <p>2 Mark pictures for upload. Highlight pictures and press MENU/OK to mark them for upload. • To switch to the memory card in the other slot or display only pictures that meet selected criteria, press DISP/BACK before marking begins. • To select all pictures, press the Q button. • Selecting any two pictures with the AEL button also selects all pictures between them.</p> <p>3 Press DISP/BACK and select START TRANSFER. The selected pictures will be uploaded.</p> <p>239</p>	<p># Transfer Image to Smartphone</p> <p>Select photos for upload to a paired smartphone or tablet via Bluetooth (see page 249).</p> <p>1. **Select TRANSFER IMAGE > SELECT FRAMES.**</p> <p>!Select Image Screen</p> <p>*Select RESET to remove "upload to smartphone" marking from all pictures before proceeding.*</p> <p>2. **Mark pictures for upload.** Highlight pictures and press **MENU/OK** to mark them for upload. - To switch to the memory card in the other slot or display only pictures that meet selected criteria, press **DISP/BACK** before marking begins. - To select all pictures, press the **Q** button. - Selecting any two pictures with the **AEL** button also selects all pictures between them.</p> <p>3. **Press DISP/BACK and select START TRANSFER.** The selected pictures will be uploaded.</p> <p>!Start Transfer Screen</p> <pre>{ 'page': 239, 'model': 'x-t5', 'chapter': 'Playback and the Playback Menu', 'section': 'The Playback Menu', 'subsection': ['TRANSFER IMAGE TO SMARTPHONE']} </pre>	 

▶ 데이터 저장 및 관리

- 데이터 저장 장소 : Pinecone Index, BM25 Retriever 객체
- 각 데이터를 임베딩 모델을 활용하여 저장
- 사용하는 각 브랜드별로 DB를 따로 저장하여 관리함.
- 단, Vectorstore 의 경우 API 기반 서버리스 벡터스토어에 저장함.
- 캐논
 - 1) VectorStore : OpenAi embedding 활용하여 저장(사용 모델 : text-embedding-3-small)
 - 2) BM25 Retriever : BM25 tokenizer kiwi 사용하여 형태소 분할 및 retriever 객체로 저장
- 소니
 - 1) VectorStore : OpenAi embedding 활용하여 저장(사용 모델 : text-embedding-3-small)
 - 2) BM25 Retriever : BM25 tokenizer kiwi 사용하여 형태소 분할(토큰화) 및 retriever 객체로 저장
- 후지
 - 1) VectorStore : OpenAi embedding 활용하여 저장(사용 모델 : text-embedding-3-small)
 - 2) BM25 Retriever : porter_stemmer 사용하여 어간 추출(Stemming)및 retriever 객체로 저장