```
# 랭체인 설치
!pip install langchain langchain_openai langchain_experimental tiktoken langchain-text-splitters
```

```
Collecting langchain
  Downloading langchain-0.1.13-py3-none-any.whl (810 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 810.5/810.5 kB 6.7 MB/s eta 0:00:00
Collecting langchain_openai
  Downloading langchain_openai-0.1.1-py3-none-any.whl (32 kB)
Collecting langchain_experimental
  Downloading langchain_experimental-0.0.55-py3-none-any.whl (177 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 177.6/177.6 kB 6.8 MB/s eta 0:00:00
Collecting tiktoken
  Downloading tiktoken-0.6.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 11.0 MB/s eta 0:00:00
Collecting langchain-text-splitters
  Downloading langchain_text_splitters-0.0.1-py3-none-any.whl (21 kB)
Requirement already satisfied: PyYAML>=5.3 in /usr/local/lib/python3.10/dist-packages (from langchain) (6.0.1)
Requirement already satisfied: SQLAlchemy<3,>=1.4 in /usr/local/lib/python3.10/dist-packages (from langchain) (2.0.28)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in /usr/local/lib/python3.10/dist-packages (from langchain) (3.9.3)
Requirement already satisfied: async-timeout<5.0.0,>=4.0.0 in /usr/local/lib/python3.10/dist-packages (from langchain) (4.0.3)
Collecting dataclasses-json<0.7,>=0.5.7 (from langchain)
  Downloading dataclasses_json-0.6.4-py3-none-any.whl (28 kB)
Collecting jsonpatch<2.0,>=1.33 (from langchain)
  Downloading jsonpatch-1.33-py2.py3-none-any.whl (12 kB)
Collecting langchain-community<0.1,>=0.0.29 (from langchain)
  Downloading langchain_community-0.0.29-py3-none-any.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 20.4 MB/s eta 0:00:00
Collecting langchain-core<0.2.0,>=0.1.33 (from langchain)
  Downloading langchain_core-0.1.33-py3-none-any.whl (269 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 269.1/269.1 kB 17.7 MB/s eta 0:00:00
Collecting langsmith<0.2.0,>=0.1.17 (from langchain)
  Downloading langsmith-0.1.31-py3-none-any.whl (71 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 71.6/71.6 kB 5.5 MB/s eta 0:00:00
Requirement already satisfied: numpy<2,>=1 in /usr/local/lib/python3.10/dist-packages (from langchain) (1.25.2)
Requirement already satisfied: pydantic<3,>=1 in /usr/local/lib/python3.10/dist-packages (from langchain) (2.6.4)
Requirement already satisfied: requests<3,>=2 in /usr/local/lib/python3.10/dist-packages (from langchain) (2.31.0)
Requirement already satisfied: tenacity<9.0.0,>=8.1.0 in /usr/local/lib/python3.10/dist-packages (from langchain) (8.2.3)
Collecting openai<2.0.0,>=1.10.0 (from langchain_openai)
  Downloading openai-1.14.3-py3-none-any.whl (262 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 262.9/262.9 kB 18.2 MB/s eta 0:00:00
Requirement already satisfied: regex>=2022.1.18 in /usr/local/lib/python3.10/dist-packages (from tiktoken) (2023.12.25)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (1.9.4)
Collecting marshmallow<4.0.0,>=3.18.0 (from dataclasses-json<0.7,>=0.5.7->langchain)
  Downloading marshmallow-3.21.1-py3-none-any.whl (49 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 49.4/49.4 kB 2.7 MB/s eta 0:00:00
Collecting typing-inspect<1,>=0.4.0 (from dataclasses-json<0.7,>=0.5.7->langchain)
  Downloading typing_inspect-0.9.0-py3-none-any.whl (8.8 kB)
Collecting jsonpointer>=1.9 (from jsonpatch<2.0,>=1.33->langchain)
  Downloading jsonpointer-2.4-py2.py3-none-any.whl (7.8 kB)
Requirement already satisfied: anyio<5,>=3 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.2.0,>=0.1.33->langchain) (3.7.1)
Collecting packaging<24.0,>=23.2 (from langchain-core<0.2.0,>=0.1.33->langchain)
  Downloading packaging-23.2-py3-none-any.whl (53 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 53.0/53.0 kB 4.2 MB/s eta 0:00:00
Collecting orjson<4.0.0,>=3.9.14 (from langsmith<0.2.0,>=0.1.17->langchain)
  Downloading orjson-3.9.15-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (138 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 138.5/138.5 kB 2.2 MB/s eta 0:00:00
Requirement already satisfied: distro<2,>=1.7.0 in /usr/lib/python3/dist-packages (from openai<2.0.0,>=1.10.0->langchain_openai) (1.7.0)
```

더블클릭 또는 Enter 키를 눌러 수정

Fixed-size chunking

- 가장 일반적이고, 간단한 방식
- 단순히 청킹의 토큰수 or 문자길이 만큼 잘라내는 형식
- 간단함. NLP(자연어처리)를 위한 자원을 사용하지 않아도 됨.
- 문맥과 의미론적인 의미를 잃어버릴 가능성이 매우 높음. 비추천

```
from langchain_text_splitters import CharacterTextSplitter

text_splitter = CharacterTextSplitter(
    separator="", # 분리할 문자 지정가능 (문자단위로 분리할 때 사용)
    chunk_size=200,
    chunk_overlap=20,
    length_function=len,
    is_separator_regex=False,
)

doc = """혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI 전문기업으로 ㄱ
이와 함께, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다.
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이 진행
"""

texts = text_splitter.split_text(doc)

for t in texts:
  print(f"{t} : {len(t)}")
```

```
혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI 전문기업ㄷ
이와 함께, 파수는 데이터 보안 영역에서 국내 : 200
, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다.
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이
계 선두의 애플리케이션 보안까지, 파수는 디지털 혁신을 향해 진입장벽이 높은 고부가가치 기술 분야를 꾸준히 개척해 나가고 있습니다. : 73
```

Split by tokens

- LLM 토큰의 수로 chunk split
- tiktoken
- https://python.langchain.com/docs/modules/data_connection/document_transformers/split_by_token

```
from posixpath import splitext
from langchain_text_splitters import CharacterTextSplitter


doc = """혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI 전문기업으로 ㄱ
이와 함께, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다.
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이 진행
"""

text_splitter = CharacterTextSplitter.from_tiktoken_encoder(
    separator="",
    encoding_name="cl100k_base",
    chunk_size=100,
    chunk_overlap=10
)


texts = text_splitter.split_text(doc)

for t in texts:
  print(f"{t} : {len(t)}")
```

```
혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터 : 80
이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI : 83
용을 돕는 AI 전문기업으로 거듭나고 있습니다.
이와 함께, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다.
이 뿐만 아 : 82
.
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 : 73
루션, 업계 최고의 컨설턴트들이 진행하는 정보보호 컨설팅, 인공지능 기반 노트 앱, 블록체인 서비스, 그리고 자회사로 독립한 업계 선두의 : 76
계 선두의 애플리케이션 보안까지, 파수는 디지털 혁신을 향해 진입장벽이 높은 고부가가치 기술 분야를 꾸준히 개척해 나가고 있습니다. : 73
나가고 있습니다. : 9
```

더블클릭 또는 Enter 키를 눌러 수정

Recursive Chunking

- 고정길이 chunking 과 컨텐츠 인식 chunking 방식의 혼합
- 청크가 충분히 작아질 때까지 순서대로 분할하려고 시도.
  기본 목록은 ["\n\n", "\n", " ", ""] 임.

이는 모든 단락(그리고 문장, 단어)을 가능한 한 오랫동안 함께 유지하려는 효과가 있음. 즉 Context 유지 효과

- 문장을 구분하는 구분자로 문장을 추출한 다음, 문장이 원하는 고정 사이즈(fixed size) 보다 클 경우 다시 fiexed size 로 자른 후 나머지 문장을 문장 구분자로 재귀적으로 호출하는 방식

```
# 랭체인 이용

from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 50, # 분할된 청크(chunk)의 최대 길이
    chunk_overlap = 5, # 분할된 청크 사이의 중복 길이를 지정, 인접 chunk 사이의 n 개의 문자가 overlap
    length_function = len, # 길이를 계산하는 함수, 토큰수로 하고 싶으면 토큰 수를 계산하는 함수를 넘기면 됨
    is_separator_regex = False, #is_separator_regex=True로 설정하면 separators(텍스트를 분리하는 기준)에 정규식을 사용할 수 있음
)

doc = """혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI 전문기업으로 ㄱ
이와 함께, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다.
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이 진행
"""

texts = text_splitter.split_text(doc)

for t in texts:
  print(f"{t} : {len(t)}")
```

```
혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 : 50
생성형 AI 시대를 맞아 : 13
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ : 43
보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI : 46
AI 활용을 돕는 AI 전문기업으로 거듭나고 있습니다. : 30
이와 함께, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다. : 48
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 : 48
빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이 진행하는 정보보호 컨설팅, : 48
컨설팅, 인공지능 기반 노트 앱, 블록체인 서비스, 그리고 자회사로 독립한 업계 선두의 : 48
선두의 애플리케이션 보안까지, 파수는 디지털 혁신을 향해 진입장벽이 높은 고부가가치 기술 : 49
기술 분야를 꾸준히 개척해 나가고 있습니다. : 24
```

Semantic Chunking

```
from langchain_experimental.text_splitter import SemanticChunker
from langchain_openai.embeddings import OpenAIEmbeddings

semantic_text_splitter = SemanticChunker(OpenAIEmbeddings(api_key="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",model="text-embedding-3-small")
  breakpoint_threshold_type="percentile", # 분할시점 결정방식, 분할의 기본 방법은 백분위수를 기준으로 합니다. 이 방법에서는 문장 간의 모든 차이를 계산
  breakpoint_threshold_amount=10 # 분한 기준 수치
)
```

```
doc = """혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI 전문기업으로 거
이와 함께, 파수는 데이터 보안 영역에서 국내를 넘어 글로벌 시장을 선도하고 있습니다.
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이 진행하
파수는 개발1본부와 2본부가 있으며, 개발2본부에는 클라우드 개발팀이 속해있습니다.
"""

texts = semantic_text_splitter.split_text(doc)

for t in texts:
  print(f"{t} : {len(t)}")
```

```
혁신적인 솔루션과 서비스를 제공해 온 엔터프라이즈 소프트웨어 기업인 파수는 본격적인 생성형 AI 시대를 맞아
▲ AI-Ready 데이터 ▲ 엔터프라이즈 LLM ▲ AI-Ready 보안 ▲ AI-Powered 애플리케이션을 AI 비전으로 삼고, 고객의 생성형 AI 활용을 돕는 AI 전문기업으
이 뿐만 아니라, 문서가상화 기술을 활용한 기업용 문서관리 플랫폼, 압도적인 퍼포먼스의 빅데이터 개인정보 비식별화 솔루션, 업계 최고의 컨설턴트들이
파수는 개발1본부와 2본부가 있으며, 개발2본부에는 클라우드 개발팀이 속해있습니다. : 46
 : 0
```

코딩을 시작하거나 AI로 코드를 생성하세요.