第五章 基于文档的问答

使用大语言模型构建一个能够回答关于给定文档和文档集合的问答系统是一种非常实用和有效的应用场景。**与仅依赖模型预训练知识不同，这种方法可以进一步整合用户自有数据，实现更加个性化和专业的问答服务。**例如,我们可以收集某公司的内部文档、产品说明书等文字资料，导入问答系统中。然后用户针对这些文档提出问题时，系统可以先在文档中检索相关信息，再提供给语言模型生成答案。

这样，语言模型不仅利用了自己的通用知识，还可以充分运用外部输入文档的专业信息来回答用户问题，显著提升答案的质量和适用性。构建这类基于外部文档的问答系统，可以让语言模型更好地服务于具体场景，而不是停留在通用层面。这种灵活应用语言模型的方法值得在实际使用中推广。

基于文档问答的这个过程，我们会涉及 LangChain 中的其他组件，比如：嵌入模型（Embedding Models)和向量储存(Vector Stores)，本章让我们一起来学习这部分的内容。

一、直接使用向量储存查询

1.1 导入数据

```
from langchain.chains import RetrievalQA  #检索QA链，在文档上进行检索
from langchain.chat_models import ChatOpenAI  #openai模型
from langchain.document_loaders import CSVLoader #文档加载器，采用csv格式存储
from langchain.vectorstores import DocArrayInMemorySearch  #向量存储
from IPython.display import display, Markdown #在jupyter显示信息的工具
import pandas as pd

file = '../data/OutdoorClothingCatalog_1000.csv'

# 使用langchain文档加载器对数据进行导入
loader = CSVLoader(file_path=file)

# 使用pandas导入数据，用以查看
data = pd.read_csv(file,usecols=[1, 2])
data.head()
```

	name	description
0	Women's Campside Oxfords	This ultracomfortable lace-to-toe Oxford boast...
1	Recycled Waterhog Dog Mat, Chevron Weave	Protect your floors from spills and splashing ...
2	Infant and Toddler Girls' Coastal Chill Swimsu...	She'll love the bright colors, ruffles and exc...
3	Refresh Swimwear, V-Neck Tankini Contrasts	Whether you're going for a swim or heading out...
4	EcoFlex 3L Storm Pants	Our new TEK O2 technology makes our four-seaso...

数据是字段为 `name` 和 `description` 的文本数据：

可以看到，导入的数据集为一个户外服装的 CSV 文件，接下来我们将在语言模型中使用它。

1.2 基本文档加载器创建向量存储

```
#导入向量存储索引创建器
from langchain.indexes import VectorstoreIndexCreator

# 创建指定向量存储类，创建完成后，从加载器中调用，通过文档加载器列表加载
index = VectorstoreIndexCreator(vectorstore_cls=DocArrayInMemorySearch).from_loaders([loader])
```

1.3 查询创建的向量存储

```
query ="请用markdown表格的方式列出所有具有防晒功能的衬衫，对每件衬衫描述进行总结"

#使用索引查询创建一个响应，并传入这个查询
response = index.query(query)

#查看查询返回的内容
display(Markdown(response))
```

Name	Description
Men's Tropical Plaid Short-Sleeve Shirt	UPF 50+ rated sun protection, 100% polyester fabric, wrinkle-resistant, front and back cape venting, two front bellows pockets
Men's Plaid Tropic Shirt, Short-Sleeve	UPF 50+ rated sun protection, 52% polyester and 48% nylon fabric, wrinkle-free, quickly evaporates perspiration, front and back cape venting, two front bellows pockets
Girls' Ocean Breeze Long-Sleeve Stripe Shirt	UPF 50+ rated sun protection, Nylon Lycra®-elastane blend fabric, quick-drying and fade-resistant, holds shape well, durable seawater-resistant fabric retains its color

在上面我们得到了一个 Markdown 表格，其中包含所有带有防晒衣的衬衫的 `名称(Name)` 和 `描述(Description)` ，其中描述是语言模型总结过的结果。

二、 结合表征模型和向量存储

由于语言模型的上下文长度限制，直接处理长文档具有困难。为实现对长文档的问答，我们可以引入向量嵌入(Embeddings)和向量存储(Vector Store)等技术：

首先，**使用文本嵌入(Embeddings)算法对文档进行向量化**，使语义相似的文本片段具有接近的向量表示。其次，**将向量化的文档切分为小块，存入向量数据库**，这个流程正是创建索引(index)的过程。向量数据库对各文档片段进行索引，支持快速检索。这样，当用户提出问题时，可以先将问题转换为向量，在数据库中快速找到语义最相关的文档片段。然后将这些文档片段与问题一起传递给语言模型，生成回答。

通过嵌入向量化和索引技术，我们实现了对长文档的切片检索和问答。这种流程克服了语言模型的上下文限制，可以构建处理大规模文档的问答系统。

2.1 导入数据

```
#创建一个文档加载器，通过csv格式加载
file = '../data/OutdoorClothingCatalog_1000.csv'
loader = CSVLoader(file_path=file)
docs = loader.load()

#查看单个文档，每个文档对应于CSV中的一行数据
docs[0]
```

```
Document(page_content=": 0\nname: Women's Campside Oxfords\ndescription: This
ultracomfortable lace-to-toe Oxford boasts a super-soft canvas, thick cushioning,
and quality construction for a broken-in feel from the first time you put them
on. \n\nSize & Fit: Order regular shoe size. For half sizes not offered, order up
to next whole size. \n\nSpecs: Approx. weight: 1 lb.1 oz. per pair.
\n\nConstruction: Soft canvas material for a broken-in feel and look. Comfortable
EVA innersole with Cleansport NXT® antimicrobial odor control. Vintage hunt, fish
and camping motif on innersole. Moderate arch contour of innersole. EVA foam
midsole for cushioning and support. Chain-tread-inspired molded rubber outsole
with modified chain-tread pattern. Imported. \n\nQuestions? Please contact us for
any inquiries.", metadata={'source': '../data/OutdoorClothingCatalog_1000.csv',
'row': 0})
```

2.2 文本向量表征模型

```
#使用OpenAIEmbedding类
from langchain.embeddings import OpenAIEmbeddings

embeddings = OpenAIEmbeddings()

#因为文档比较短了，所以这里不需要进行任何分块,可以直接进行向量表征
#使用初始化OpenAIEmbedding实例上的查询方法embed_query为文本创建向量表征
embed = embeddings.embed_query("你好呀，我的名字叫小可爱")

#查看得到向量表征的长度
print("\n\033[32m向量表征的长度： \033[0m \n", len(embed))

#每个元素都是不同的数字值，组合起来就是文本的向量表征
print("\n\033[32m向量表征前5个元素： \033[0m \n", embed[:5])
```

```
向量表征的长度：
 1536

向量表征前5个元素：
 [-0.019283676849006164, -0.006842594710511029, -0.007344046732916966,
-0.024501312942119265, -0.026608679897592472]
```

2.3 基于向量表征创建并查询向量存储

```
# 将刚才创建文本向量表征(embeddings)存储在向量存储(vector store)中
# 使用DocArrayInMemorySearch类的from_documents方法来实现
# 该方法接受文档列表以及向量表征模型作为输入
db = DocArrayInMemorySearch.from_documents(docs, embeddings)

query = "请推荐一件具有防晒功能的衬衫"
#使用上面的向量存储来查找与传入查询类似的文本，得到一个相似文档列表
docs = db.similarity_search(query)
print("\n\033[32m返回文档的个数: \033[0m \n", len(docs))
print("\n\033[32m第一个文档: \033[0m \n", docs[0])
```

```
返回文档的个数:
 4

第一个文档:
 page_content=": 535\nname: Men's TropicVibe Shirt, Short-Sleeve\ndescription:
This Men’s sun-protection shirt with built-in UPF 50+ has the lightweight feel
you want and the coverage you need when the air is hot and the UV rays are
strong. Size & Fit: Traditional Fit: Relaxed through the chest, sleeve and waist.
Fabric & Care: Shell: 71% Nylon, 29% Polyester. Lining: 100% Polyester knit mesh.
UPF 50+ rated – the highest rated sun protection possible. Machine wash and dry.
Additional Features: Wrinkle resistant. Front and back cape venting lets in cool
breezes. Two front bellows pockets. Imported.\n\nSun Protection That Won't Wear
Off: Our high-performance fabric provides SPF 50+ sun protection, blocking 98% of
the sun's harmful rays." metadata={'source':
'../data/OutdoorClothingCatalog_1000.csv', 'row': 535}
```

我们可以看到一个返回了四个结果。输出的第一结果是一件关于防晒的衬衫，满足我们查询的要求：`请推荐一件具有防晒功能的衬衫`

2.4 使用查询结果构造提示来回答问题

```
#导入大语言模型, 这里使用默认模型gpt-3.5-turbo会出现504服务器超时，
#因此使用gpt-3.5-turbo-0301
llm = ChatOpenAI(model_name="gpt-3.5-turbo-0301",temperature = 0.0)

#合并获得的相似文档内容
qdocs = "".join([docs[i].page_content for i in range(len(docs))])


#将合并的相似文档内容后加上问题（question）输入到 `llm.call_as_llm`中
#这里问题是：以Markdown表格的方式列出所有具有防晒功能的衬衫并总结
response = llm.call_as_llm(f"{qdocs}问题：请用markdown表格的方式列出所有具有防晒功能的衬
衫，对每件衬衫描述进行总结")

display(Markdown(response))
```

衣服名称	描述总结
Men's TropicVibe Shirt, Short-Sleeve	男士短袖衬衫，内置UPF 50+防晒功能，轻盈舒适，前后通风口，两个前口袋，防皱，最高级别的防晒保护。
Men's Tropical Plaid Short-Sleeve Shirt	男士短袖衬衫，UPF 50+防晒，100%聚酯纤维，防皱，前后通风口，两个前口袋，最高级别的防晒保护。
Men's Plaid Tropic Shirt, Short-Sleeve	男士短袖衬衫，UPF 50+防晒，52%聚酯纤维和48%尼龙，防皱，前后通风口，两个前口袋，最高级别的防晒保护。
Girls' Ocean Breeze Long-Sleeve Stripe Shirt	女孩长袖衬衫，UPF 50+防晒，尼龙Lycra®-弹性纤维混纺，快干，耐褪色，防水，最高级别的防晒保护，适合与我们的泳衣系列搭配。

2.5 使用检索问答链来回答问题

通过LangChain创建一个检索问答链，对检索到的文档进行问题回答。检索问答链的输入包含以下

- `llm`: 语言模型，进行文本生成
- `chain_type`: 传入链类型，这里使用stuff，将所有查询得到的文档组合成一个文档传入下一步。其他的方式包括：
    - Map Reduce： 将所有块与问题一起传递给语言模型，获取回复，使用另一个语言模型调用将所有单独的回复总结成最终答案，它可以在任意数量的文档上运行。可以并行处理单个问题，同时也需要更多的调用。它将所有文档视为独立的
    - Refine： 用于循环许多文档，际上是迭代的，建立在先前文档的答案之上，非常适合前后因果信息并随时间逐步构建答案，依赖于先前调用的结果。它通常需要更长的时间，并且基本上需要与Map Reduce一样多的调用
    - Map Re-rank： 对每个文档进行单个语言模型调用，要求它返回一个分数，选择最高分，这依赖于语言模型知道分数应该是什么，需要告诉它，如果它与文档相关，则应该是高分，并在那里精细调整说明，可以批量处理它们相对较快，但是更加昂贵

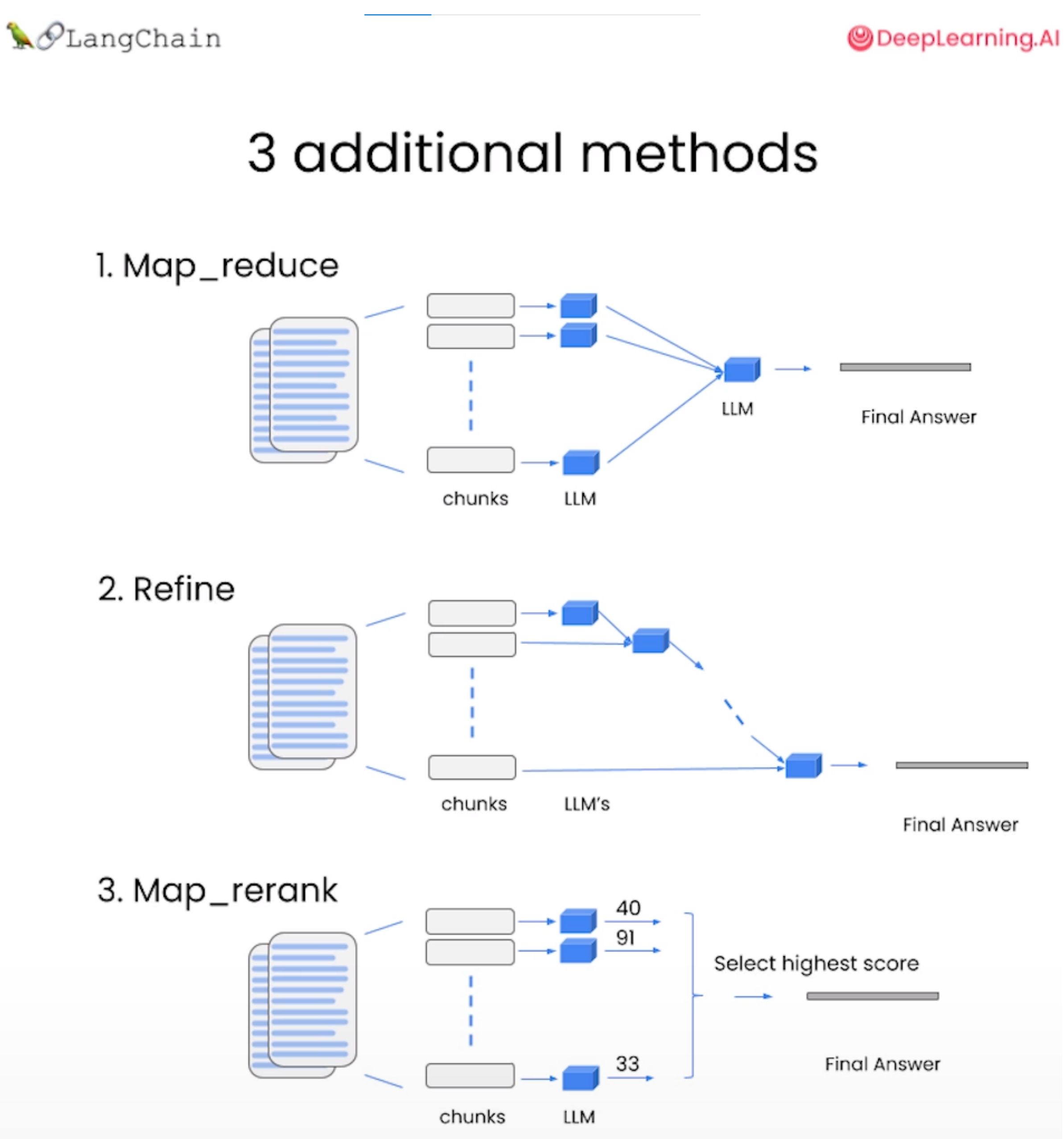

图 3.5 检索问答链

- `retriever`:检索器

```
#基于向量储存，创建检索器
retriever = db.as_retriever()

qa_stuff = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=retriever,
    verbose=True
)

#创建一个查询并在此查询上运行链
query =  "请用markdown表格的方式列出所有具有防晒功能的衬衫，对每件衬衫描述进行总结"

response = qa_stuff.run(query)

display(Markdown(response))
```

```
> Entering new RetrievalQA chain...

> Finished chain.
```

编号	名称	描述
618	Men's Tropical Plaid Short-Sleeve Shirt	100%聚酯纤维制成，轻便，防皱，前后背部有通风口，两个前面的褶皱口袋，UPF 50+防晒等级，可阻挡98%的紫外线
374	Men's Plaid Tropic Shirt, Short-Sleeve	52%聚酯纤维和48%尼龙制成，轻便，防皱，前后背部有通风口，两个前面的褶皱口袋，UPF 50+防晒等级，可阻挡98%的紫外线
535	Men's TropicVibe Shirt, Short-Sleeve	71%尼龙和29%聚酯纤维制成，轻便，防皱，前后背部有通风口，两个前面的褶皱口袋，UPF 50+防晒等级，可阻挡98%的紫外线
293	Girls' Ocean Breeze Long-Sleeve Stripe Shirt	尼龙Lycra®-弹性纤维混纺，长袖，UPF 50+防晒等级，可阻挡98%的紫外线，快干，耐褪色，可与我们的泳衣系列轻松搭配

总结：这些衬衫都具有防晒功能，防晒等级为UPF 50+，可阻挡98%的紫外线。它们都是轻便的，防皱的，有前后背部通风口和前面的褶皱口袋。其中女孩的长袖条纹衬衫是由尼龙Lycra®-弹性纤维混纺制成，快干，耐褪色，可与泳衣系列轻松搭配。

可以看到 2.5 和 2.6 部分的这两个方式返回相同的结果。

英文版提示

1. 直接使用向量储存查询

```python
from langchain.document_loaders import CSVLoader
from langchain.indexes import VectorstoreIndexCreator

file = '../data/OutdoorClothingCatalog_1000.csv'
loader = CSVLoader(file_path=file)

index = 
VectorstoreIndexCreator(vectorstore_cls=DocArrayInMemorySearch).from_loaders([loader])

query ="Please list all your shirts with sun protection \
in a table in markdown and summarize each one."

response = index.query(query)

display(Markdown(response))
```

Name	Description
Men's Tropical Plaid Short-Sleeve Shirt	UPF 50+ rated, 100% polyester, wrinkle-resistant, front and back cape venting, two front bellows pockets
Men's Plaid Tropic Shirt, Short-Sleeve	UPF 50+ rated, 52% polyester and 48% nylon, machine washable and dryable, front and back cape venting, two front bellows pockets
Men's TropicVibe Shirt, Short-Sleeve	UPF 50+ rated, 71% Nylon, 29% Polyester, 100% Polyester knit mesh, machine wash and dry, front and back cape venting, two front bellows pockets
Sun Shield Shirt by	UPF 50+ rated, 78% nylon, 22% Lycra Xtra Life fiber, handwash, line dry, wicks moisture, fits comfortably over swimsuit, abrasion resistant

All four shirts provide UPF 50+ sun protection, blocking 98% of the sun's harmful rays. The Men's Tropical Plaid Short-Sleeve Shirt is made of 100% polyester and is wrinkle-resistant

2. 结合表征模型和向量存储

```
from langchain.document_loaders import CSVLoader
from langchain.embeddings import OpenAIEmbeddings
from langchain.vectorstores import DocArrayInMemorySearch


embeddings = OpenAIEmbeddings()
embed = embeddings.embed_query("Hi my name is Harrison")

print("\n\033[32m向量表征的长度: \033[0m \n", len(embed))
print("\n\033[32m向量表征前5个元素: \033[0m \n", embed[:5])

file = '../data/OutdoorClothingCatalog_1000.csv'
loader = CSVLoader(file_path=file)
docs = loader.load()
embeddings = OpenAIEmbeddings()
db = DocArrayInMemorySearch.from_documents(docs, embeddings)

query = "Please suggest a shirt with sunblocking"
docs = db.similarity_search(query)
print("\n\033[32m返回文档的个数: \033[0m \n", len(docs))
print("\n\033[32m第一个文档: \033[0m \n", docs[0])


# 使用查询结果构造提示来回答问题
llm = ChatOpenAI(model_name="gpt-3.5-turbo-0301",temperature = 0.0)

qdocs = "".join([docs[i].page_content for i in range(len(docs))])

response = llm.call_as_llm(f"{qdocs} Question: Please list all your \
shirts with sun protection in a table in markdown and summarize each one.")

print("\n\033[32m使用查询结果构造提示来回答问题: \033[0m \n", docs[0])
```

```
display(Markdown(response))


# 使用检索问答链来回答问题
retriever = db.as_retriever()

qa_stuff = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=retriever,
    verbose=True
)


query =  "Please list all your shirts with sun protection in a table \
in markdown and summarize each one."

response = qa_stuff.run(query)

print("\n\033[32m 使用检索问答链来回答问题： \033[0m \n")
display(Markdown(response))
```

```
向量表征的长度:
 1536

向量表征前5个元素:
 [-0.021913960932078383, 0.006774206755842609, -0.018190348816400977,
-0.039148249368104494, -0.014089343366938917]

返回文档的个数:
 4

第一个文档:
 page_content=': 255\nname: Sun Shield Shirt by\ndescription: "Block the sun, not
the fun – our high-performance sun shirt is guaranteed to protect from harmful UV
rays. \n\nSize & Fit: Slightly Fitted: Softly shapes the body. Falls at
hip.\n\nFabric & Care: 78% nylon, 22% Lycra Xtra Life fiber. UPF 50+ rated – the
highest rated sun protection possible. Handwash, line dry.\n\nAdditional
Features: Wicks moisture for quick-drying comfort. Fits comfortably over your
favorite swimsuit. Abrasion resistant for season after season of wear.
Imported.\n\nSun Protection That Won\'t Wear Off\nOur high-performance fabric
provides SPF 50+ sun protection, blocking 98% of the sun\'s harmful rays. This
fabric is recommended by The Skin Cancer Foundation as an effective UV
protectant.' metadata={'source': '../data/OutdoorClothingCatalog_1000.csv',
'row': 255}

使用查询结果构造提示来回答问题:
 page_content=': 255\nname: Sun Shield Shirt by\ndescription: "Block the sun, not
the fun – our high-performance sun shirt is guaranteed to protect from harmful UV
rays. \n\nSize & Fit: Slightly Fitted: Softly shapes the body. Falls at
hip.\n\nFabric & Care: 78% nylon, 22% Lycra Xtra Life fiber. UPF 50+ rated – the
highest rated sun protection possible. Handwash, line dry.\n\nAdditional
Features: Wicks moisture for quick-drying comfort. Fits comfortably over your
favorite swimsuit. Abrasion resistant for season after season of wear.
Imported.\n\nSun Protection That Won\'t Wear Off\nOur high-performance fabric
provides SPF 50+ sun protection, blocking 98% of the sun\'s harmful rays. This
fabric is recommended by The Skin Cancer Foundation as an effective UV
protectant.' metadata={'source': '../data/OutdoorClothingCatalog_1000.csv',
'row': 255}
```

Name	Description
Sun Shield Shirt	High-performance sun shirt with UPF 50+ sun protection, moisture-wicking, and abrasion-resistant fabric. Recommended by The Skin Cancer Foundation.
Men's Plaid Tropic Shirt	Ultracomfortable shirt with UPF 50+ sun protection, wrinkle-free fabric, and front/back cape venting. Made with 52% polyester and 48% nylon.
Men's TropicVibe Shirt	Men's sun-protection shirt with built-in UPF 50+ and front/back cape venting. Made with 71% nylon and 29% polyester.
Men's Tropical Plaid Short-Sleeve Shirt	Lightest hot-weather shirt with UPF 50+ sun protection, front/back cape venting, and two front bellows pockets. Made with 100% polyester.

All of these shirts provide UPF 50+ sun protection, blocking 98% of the sun's harmful rays. They also have additional features such as moisture-wicking, wrinkle-free fabric, and front/back cape venting for added comfort.

```
> Entering new RetrievalQA chain...

> Finished chain.

 使用检索问答链来回答问题：
```

Shirt Number	Name	Description
618	Men's Tropical Plaid Short-Sleeve Shirt	Rated UPF 50+ for superior protection from the sun's UV rays. Made of 100% polyester and is wrinkle-resistant. With front and back cape venting that lets in cool breezes and two front bellows pockets.
374	Men's Plaid Tropic Shirt, Short-Sleeve	Rated to UPF 50+ and offers sun protection. Made with 52% polyester and 48% nylon, this shirt is machine washable and dryable. Additional features include front and back cape venting, two front bellows pockets.

Shirt Number	Name	Description
535	Men's TropicVibe Shirt, Short-Sleeve	Built-in UPF 50+ has the lightweight feel you want and the coverage you need when the air is hot and the UV rays are strong. Made with 71% Nylon, 29% Polyester. Wrinkle resistant. Front and back cape venting lets in cool breezes. Two front bellows pockets.
255	Sun Shield Shirt	High-performance sun shirt is guaranteed to protect from harmful UV rays. Made with 78% nylon, 22% Lycra Xtra Life fiber. Wicks moisture for quick-drying comfort. Fits comfortably over your favorite swimsuit. Abrasion-resistant.

All of the shirts listed above provide sun protection with a UPF rating of 50+ and block 98% of the sun's harmful rays. The Men's Tropical Plaid Short-Sleeve Shirt is made of 100% polyester and has front and back cape venting and two front bellows pockets. The Men's Plaid Tropic Shirt, Short-Sleeve is made with 52% polyester and 48% nylon and has front and back cape venting and two front bellows pockets. The Men's TropicVibe Shirt, Short-Sleeve is made with 71% Nylon, 29% Polyester and has front and back cape venting and two front bellows pockets. The Sun Shield Shirt is made with 78% nylon, 22% Lycra Xtra Life fiber and is abrasion-resistant. It fits comfortably over your favorite swimsuit.