



2.th Chunking, Embedding, OpenSearch: Store&Search

www.youtube.com/@HONGJOONG-SHIN

theprismdata@gmail.com

1. Chunking

목적 Large Language Model이 받아 들일 수 있는 크기의 문자(토큰)로 긴 문장을 조각
작은 토큰으로 잘라내면서, 상용 LLM의 비용 절약 효과

오는 25일부터 30일까지 설 연휴 기간 데이터 통신 ...데이터 도매대가를 낮출
예정이다. 또 중고폰의 안전한 거래 환경 조성을 위한 중고폰 안심거래 사업
자 인증제도를 추진한다.

.....

오는 25일부터 30일까지 설 연휴 기간.

데이터 통신 요금 부담 없이 영상통
화가
무료로 제공된다



1. Chunking

종류 RecursiveCharacterTextSplitter

separators (["\n\n", "\n", " ", ""])를 기준으로 문장을 분할

chunk_size : 글자를 설정된 길이를 넘지 않도록 분할

chunk_overlap : 분할 된 문장에서 이전 문장에 포함할 글자 수

chunk overlap을 사용할 경우 검색시 문맥 단절 현상 방지

온 가족이 함께 과학문화를 즐길 수 있도록 설 연휴기간 동안 중앙(대전), 과천, 광주, 부산 등 소관 국립과학관 5곳의 상설전시관을 무료로 개방한다. 각 과학관에서 을사년 뱀관련 별자리 해설, 민속놀이 등

질문

답변 마지막

온 가족이 함께 과학문화를 즐길 수 있도록 설 연휴기간 동안 중앙(대전), 과천, 광주, 부산 등 소관 국립과학관 5곳의 상설전시관을 무료로 개방한다. 각 과학관에서 을사년

을사년 뱀관련 별자리 해설, 민속놀이 등 다양한 과학문화 전시와 체험행사를 즐길 수 있다.

추가 내용

1. Chunking

종류 NLTKTextSplitter

NLTK의 문장 토크 나이저를 사용하여 문장 단위로 텍스트 분할

chunk_size : 최대 길이를 넘지 않도록 함

chunk_overlap : 분할 된 문장에서 이전 문장에 포함할 글자 수

chunk overlap을 사용할 경우 검색시 문맥 단절 현상 방지

온 가족이 함께 과학문화를 즐길 수 있도록 설 연휴기간 동안 중앙(대전), 과천, 광주, 부산 등 소관 국립과학관 5곳의 상설전시관을 무료로 개방한다. 각 과학관에서 을사년 뱀관련 별자리 해설, 민속놀이 등

고물가 시대 국민의 가계통신비 부담을 완화하고 소비자 후생을 향상하기 위해 발표된 알뜰폰 경쟁력 강화방안에 따라 데이터 도매대가를 낮출 예정이다. **또 중고폰의 안전한 거래 환경 조성을 위한 중고폰 안심거래 사업자 인증제도를 추진한다.**

또 중고폰의 안전한 거래 환경 조성을 위한 중고폰 안심거래 사업자 인증제도를 추진한다. 설 장바구니 물가 부담을 낮출 수 있도록 오는 30일까지 '2025년 우체국쇼핑 설 선물대전'을 진행한다.

2. Embedding

목적 사람이 쓰는 자연어를 컴퓨터가 이해할 수 있는 숫자의 나열인 Vector(행렬)로 바꿈

2025.01.20. 오후 12:01

https://n.news.naver.com/article/092/0002360502?cde=news_media_pc&type=editn

[오는 25일부터 30일까지 설 연휴 기간 데이터 통신 요금 부담 없이 영상통화가 무료로 제공된다.]



[-3.04232180e-01 -6.26503825e-01 -4.61999416e-01 -4.66390222e-01
-1.56113058e-01 2.99530268e-01 -6.98186040e-01 -1.18409050e+00
1.00436628e+00 4.64943588e-01 -9.49845433e-01 4.15850103e-01

....

2.33040616e-01 -3.78922403e-01 1.32098496e-01 -9.12056491e-02
-2.00412273e-01 6.72386527e-01 1.11323878e-01 -8.46709847e-01
-1.10680744e-01 -1.24335192e-01 -8.31869721e-01 -1.72620559e+00
3.48826170e-01 -3.57889205e-01 3.85624588e-01 9.57825303e-01]

임베딩 모델 KR-SBERT-V40K-klueNLI-augSTS

3. Store&Search

목적 문자열, 단어 검색이 아닌 유사한 Vector를 검색하기 위해 사용

Stage 1 OpenSearch Docker 설치

windows cmd에서 초기 암호 설정 (admin의 암호가 복잡해야 설치됩니다.)
set OPENSEARCH_INITIAL_ADMIN_PASSWORD=juntPass123!

컨테이너 Restart시 데이터 유지를 위해 볼륨 설정 (docker/docker-compose.yml 에서 찾을)

~~

volumes:

- **/c/Temp/opensearch**:/usr/share/opensearch/data

c:\Temp\opensearch 폴더가 생성됩니다.

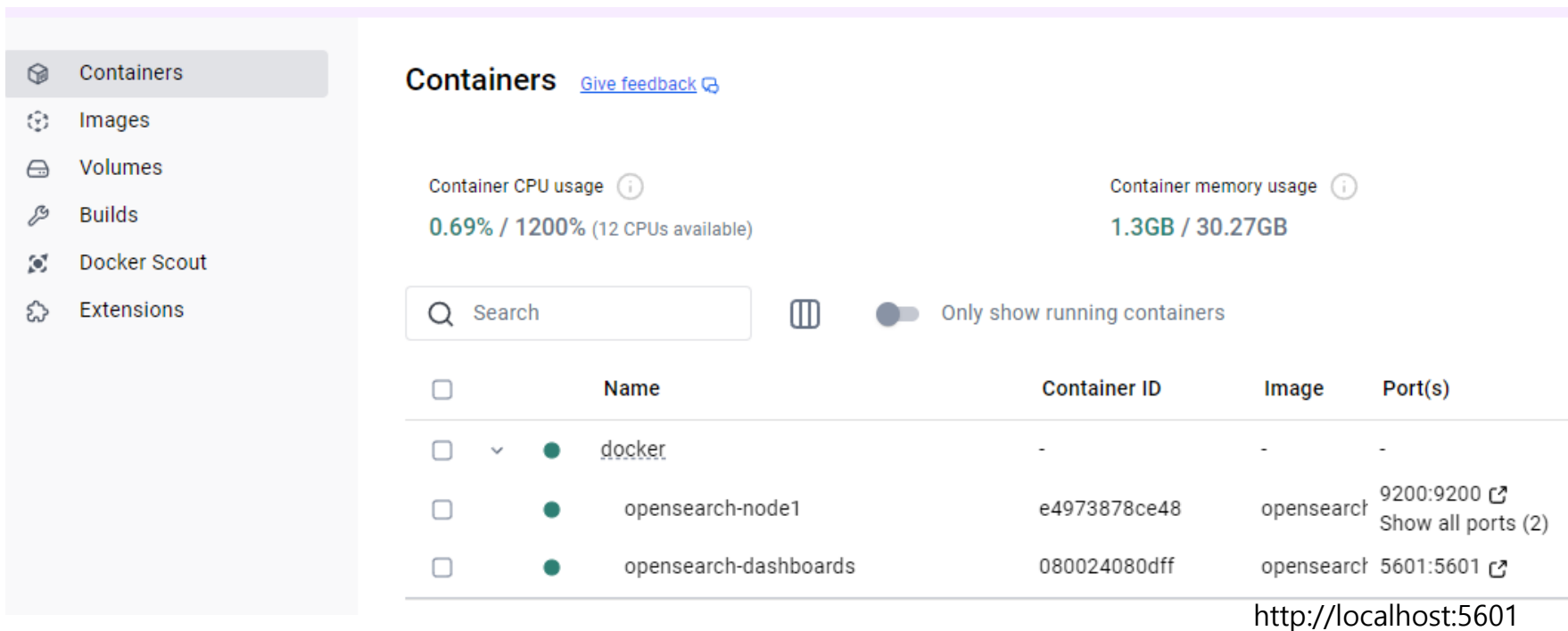
docker 컨테이너 실행

docker-compose -f docker\docker-compose.yml up

3. Store&Search

목적 문자열, 단어 검색이 아닌 유사한 Vector를 검색하기 위해 사용

Stage 1 OpenSearch Dashboard 컨테이너 접속



The screenshot shows the Docker Desktop interface. On the left is a sidebar with navigation options: Containers (selected), Images, Volumes, Builds, Docker Scout, and Extensions. The main area is titled 'Containers' and shows system metrics: 'Container CPU usage' at 0.69% / 1200% (12 CPUs available) and 'Container memory usage' at 1.3GB / 30.27GB. Below the metrics is a search bar and a toggle switch for 'Only show running containers'. A table lists the running containers:

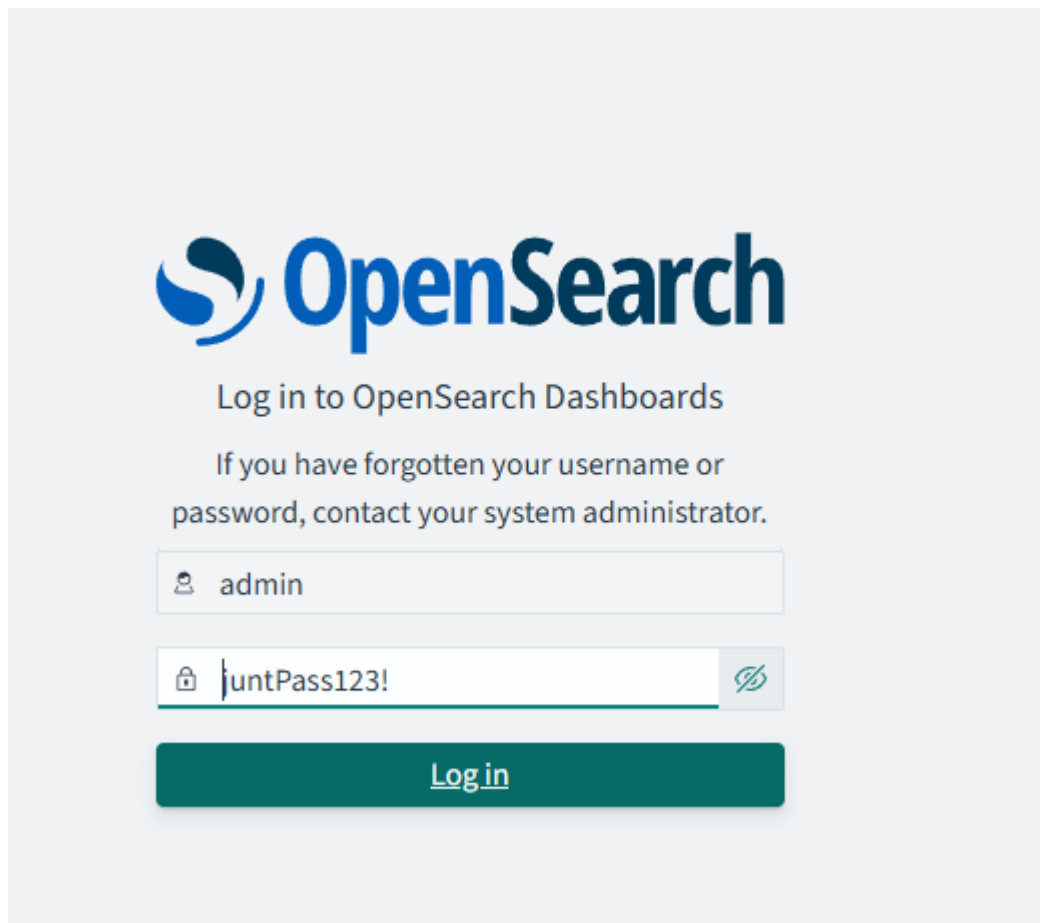
<input type="checkbox"/>	Name	Container ID	Image	Port(s)
<input type="checkbox"/>	docker	-	-	-
<input type="checkbox"/>	opensearch-node1	e4973878ce48	opensearch	9200:9200 Show all ports (2)
<input type="checkbox"/>	opensearch-dashboards	080024080dff	opensearch	5601:5601 Show all ports (2)

Below the table, the URL <http://localhost:5601> is displayed.

3. Store&Search

목적 문자열, 단어 검색이 아닌 유사한 Vector를 검색하기 위해 사용

Stage 1 OpenSearch Dashboard 로그인



The image shows the OpenSearch Dashboard login interface. At the top is the OpenSearch logo, which consists of a blue circular icon with a white swoosh and the text "OpenSearch" in a bold, blue, sans-serif font. Below the logo, the text "Log in to OpenSearch Dashboards" is displayed in a smaller, grey font. Underneath this, a message in grey text reads: "If you have forgotten your username or password, contact your system administrator." There are two input fields: the first is for the username, with a user icon on the left and the text "admin" entered; the second is for the password, with a lock icon on the left, the text "untPass123!" entered, and a toggle icon on the right. At the bottom is a large, dark teal button with the text "Log in" in white, underlined.

3. Store&Search

목적 문자열, 단어 검색이 아닌 유사한 Vector를 검색하기 위해 사용

Stage 2 OpenSearch 생성된 Index 확인

The screenshot shows the OpenSearch Dashboards interface. On the left, the 'Management' menu is expanded, and 'Index Management' is selected (labeled (2)). In the center, the 'Index Management' sidebar is open, and 'Indexes' is selected (labeled (3)). On the right, the 'State management policies' page is visible. Below it, the 'Indexes (8)' page is shown, listing various indices. The 'embedding_test' index is highlighted with a red box (labeled (4)).

Index	Health	Manag...	Status	Total size	Size of ...	Total d...	Delete...	Primaries	Replicas
security-auditlog-2025.01.20	Yellow	No	Open	101.6kb	101.6kb	7	0	1	1
embedding_test	Yellow	No	Open	263kb	263kb	17	0	1	1
.ql-datasources	Green	No	Open	208b	208b	0	0	1	0
.plugins-ml-config	Green	No	Open	4kb	4kb	1	0	1	0
.opensearch-observability	Green	No	Open	208b	208b	0	0	1	0
.opendistro_security	Green	No	Open	82.5kb	82.5kb	10	0	1	0
.kibana_92668751_admin_1	Green	No	Open	5.3kb	5.3kb	1	0	1	0
.kibana_1	Green	No	Open	208b	208b	0	0	1	0

3. Store&Search

목적 문자열, 단어 검색이 아닌 유사한 Vector를 검색하기 위해 사용

Stage 3 OpenSearch Index pattern (embedding_test) 등록

Recently viewed

No recently viewed items

OpenSearch Dashboards

Overview

Discover

Dashboards

Visualize

Dashboards Management

Index patterns

Data sources

Saved objects

Advanced settings

Create index pattern

An index pattern can match a single source, for example, `filebeat-4-3-22`, or **multiple** data sources, `filebeat-*`.

[Read documentation](#)

Step 1 of 2: Define an index pattern

index pattern name

embedding_test

Next step >

Use an asterisk (*) to match multiple indices. Spaces and the characters `\, /, ?, *, <, >, |` are not allowed.

☐ Include system and hidden indices

✓ Your index pattern matches 1 source.

embedding_test Index

Rows per page: 10

embedding_test

This page lists every field in the **embedding_test** index and the field's associated core type as recorded by OpenSearch. To change a field type, use the OpenSearch [Mapping API](#).

Fields (7) Scripted fields (0) Source filters (0)

Search

All field types

Name	Type	Format	Searchable	Aggregatable	Excluded
_id	string		•	•	✎
_index	string		•	•	✎
_score	number				✎
_source	_source				✎
_type	string				✎
embedding_vector	unknown			•	✎
text	string		•		✎

Rows per page: 10

3. Store&Search

목적 문자열, 단어 검색이 아닌 유사한 Vector를 검색하기 위해 사용

Stage 4 OpenSearch Index (embedding_test)의 vector와 저장된 데이터 보기

The screenshot displays the OpenSearch Dashboards interface. On the left, the 'Recently viewed' sidebar shows 'OpenSearch Dashboards' with options for Overview, Discover, Dashboards, and Visualize. The main panel is titled 'Discover' and shows the 'embedding_test' index selected. The 'Selected fields' list includes '_source', and the 'Available fields' list includes '_id', '_index', '_score', '_type', 'embedding_vector', and 'text'. The search bar at the top contains the text 'embedding_test'. Below the search bar, the 'Filter by type' dropdown is set to '0'. The search results section shows '17 hits' and displays the '_source' field for each hit. The first three hits are visible, each containing a 'text' field and an 'embedding_vector' field. The text fields contain Korean sentences, and the embedding vectors are arrays of floating-point numbers.

Recently viewed

No recently viewed items

OpenSearch Dashboards

Overview

Discover

Dashboards

Visualize

embedding_test

Search

DQL

Refresh

Filter by type 0

17 hits

_source

text: 오는 25일부터 30일까지 설 연휴 기간 데이터 통신 요금 부담 없이 영상통화가 무료로 제공됩니다. embedding_vector: [-0.3042321801185608, -0.6265038251876831, -0.46199941635131836, -0.4663902223110199, -0.15611305832862854, 0.2995302677154541, -0.6981860399246216, -1.184090495109558, 1.0043662786483765, 0.4649435877799988, -0.9498454332351685, 0.41585010290145874, -1.0104620456695557, -0.34390145540237427, -0.21564026176929474, -0.013105439953505993, -0.48314887285232544, -0.0026497840881347656, 0.40082311630249023, -0.4428459107875824, -0.12366093695163727, 0.46892282366752625, -1.1840496063232422, 0.3373248279094696, 0.0004513902240432799, 0.18030832707881927, -0.6666765213012695, -0.00507548451423645, -1.2134196758270264, -0.11449635028839111, -0.657720685005188, 0.3660542070865631, -0.07126934826374054, -0.6809401512145996, 0.1826457530260086, -0.808596134185791, -0.2288435697555542, -0.010753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772, 0.38141903281211853, -0.4157448709011078, -0.22954432666301727, 0.989873468875885, -0.17050017416477203, 0.01567869633436203, -0.36870837211608887, -0.39498916268348694, -0.28632256388664246, 0.10320013016462326, 0.01953931525349617, -0.10753853246569633, 0.9966694712638855, 1.0294933319091797, 0.917744631575317383, 1.006557822227478, -0.40982767939567566, -0.1463317573070526, -0.5699802041053772