Do Sentence Transformers Learn Quasi-Geospatial Concepts from General Text?

Ilya Ilyankou[1,*], Aldo Lipani[1], Stefano Cavazzi[2], Xiaowei Gao[1] and James Haworth[1]

[1]*SpaceTimeLab, University College London, UK*

[2]*Ordnance Survey, Southampton, UK*

Abstract

Sentence transformers [1] are language models designed to perform semantic search. This study investigates the capacity of sentence transformers, fine-tuned on general question-answering datasets for asymmetric semantic search, to associate descriptions of human-generated routes across Great Britain with queries often used to describe hiking experiences. We find that sentence transformers have some zero-shot capabilities to understand quasi-geospatial concepts, such as route types and difficulty, suggesting their potential utility for routing recommendation systems.

Keywords

Semantic search, sentence transformers, language models, hiking

1. Introduction

Semantic search is different from traditional, keyword search in that it tries to capture the *intent* of the searcher [2]. For a good semantic search system, the sentence *'The fox sat on the mat'* should be similar to *'An animal rested upon the rug'*, but different from *'Fox News on Sat'*.

SentenceTransformers [1] is a Python library that contains a collection of predominantly BERT-family [3] transformer-based neural network models that are fine-tuned for semantic search. This study investigates the extent to which asymmetric semantic search models, designed for shorter queries and longer documents [4], fine-tuned on general (non-geospatial) question-answering datasets, can understand vague, subjective, and complex quasi-geospatial concepts. For example, do such models consistently associate a query *'a walk with a variety of landscapes'* with documents describing longer walks going through various terrains over the documents describing shorter, solely urban or rural walks?

This topic is important because when searching for (hiking) activities, people tend to describe their physical abilities or desired experiences [5, 6, 7] over using precise geospatial terms.

2. Methodology

In our experiment, we take existing user-generated routes across Great Britain, add geospatial context to generate their textual descriptions, use Sentence Transformers to create vector embeddings for each description, and compare these vectors with embeddings of user queries.

2.1. Data

We use user-generated hiking routes from the Ordnance Survey's OS Maps app—specifically, a subset of 501,294 routes classified by Ballatore et al. [8]. We remove very short (under 1 km) and very long (over 50 km) routes to focus on those that would typically be of interest to leisure hikers, and can be completed in a day. After further removing routes with obvious GPS signal issues, we have 496,723 routes whose average length is 11,289 metres. We then assign a set of attributes to each route using geopandas [9] and several Ordnance Survey datasets. A full list of attributes is shown in Table 1 in Appendix.

2.2. Generating descriptions

We use simple template language to generate 3-4 sentence descriptions for each route based on route attributes. Descriptions mention length, shape, start and end points, total elevation gain and steepness. We explicitly state the walk is *'predominantly uphill'* or *'predominantly downhill'* where the total elevation gain exceeds (or is less than) the total elevation loss by at least 100 metres; this categorisation applies to approximately 8% of all routes. The descriptions also include the types of areas the routes go near or through, including coast, surface water, woodland, green space, urban areas, and national parks. For each area type, we mention for how long in percentages, swapping numbers (e.g., '60 percent') for words (e.g., 'sixty percent' or 'most') roughly half the time. We do this because language models are known to under-perform when required to work with numbers [10].

The resulting descriptions are between 112 and 589 characters long, with mean and median of 299 and 296 characters respectively. The longest description reads:

> 'This is a twenty-five km walk that begins in Yelverton, West Devon, Devon and ends in Plymouth, City of Plymouth. Total elevation gain is seven hundred and thirty-nine metres, and elevation grade is 2.9. The walk is predominantly downhill. About 25 percent of the walk is within a national park, about thirty percent of the walk is in a wooded area, about thirty-three percent of the walk goes through an urban area, about 12 percent of the walk is within green space, about twenty percent of the walk is along the coast, about forty-three percent of the walk is alongside a body of water.'

Other examples of descriptions of various lengths are shown in Table 2 in Appendix.

2.3. Matching queries with descriptions

We calculate vector representations, or embeddings, of all textual descriptions and user queries using `msmarco-{MiniLM-L6|distilbert}-cos-v5` and `multi-qa-{MiniLM-`

`L6|distilbert|mpnet-base}-cos-v1` models. These are based on MiniLM [11], DistilBERT [12], and MPNet [13] architectures, and fine-tuned on MS MARCO [14] (about 500k records) and/or a compilation of question-answering datasets which we refer to here as Multi-QA [15] (about 215M records). Neither collection is specific to the geospatial domain. Models tuned on MS MARCO support input sequences of up to 384 tokens (word pieces), and those tuned on Multi-QA support up to 512 tokens; our inputs fit comfortably within both limits, with the longest description of 589 characters represented by 129 tokens.

We use 20 queries (see Table 3 in Appendix) that resemble questions (e.g., *'what is a walk for an expert hiker'*), and calculate cosine similarity between all queries and route descriptions to rank the relevance of each description for each query. The queries are inspired by various research papers that studied hiking experiences [5, 7, 6, 16, 8, 17].

2.4. Visualising results

Our experiment is difficult to assess using standard information retrieval quality metrics such as mean reciprocal rank (MRR) or mean average precision (MAP) [18] given we cannot easily label an individual route description (document) as relevant to the query or not. Instead, we are interested in the overall patterns of how the documents are ranked.

We decided to plot **cumulative means** of relevant route attributes for ranked documents, sorted from best to worst match (one can think of it as plotting `average@k` of relevant route attribute means for all `k` between 1 and 496,723). As such, the y-value of the left-most point of each line chart represents the attribute value of the top-matching document, while the y-value of the right-most point represents the mean attribute value for the whole dataset. An increasing cumulative mean (i.e., a line going up) signifies higher-ranking documents (on the left of the x-axis) typically having lower values than lower-ranking documents (to the right), and vice versa. We utilise a logarithmic scale for the x-axis to highlight the cumulative means of top-ranking documents, while also presenting the overall trend for all documents.

3. Results

The results are mixed. Unsurprisingly, sentence transformers do better when user queries have similar terminology to the documents. All five models are able to associate a 'seaside walk' with routes described as having longer stretches along the coast. Four models clearly associate 'a walk for someone who enjoys town walks' with routes going through urban areas.

Figure 1 shows cumulative mean length, grade, and elevation gain for more complex queries targeting easier walks produced by `multi-qa-mpnet-base-cos-v1`. The queries mentioning a 'beginner hiker' and a 'person with limited mobility' are indeed associated with shorter and flatter routes; the results are less clear for the query mentioning an 'elderly person'.

Conversely, Figure 2 shows results produced by the same model for queries aimed at more challenging hiking experiences. While the top-10 or so results for an 'expert hiker' are indeed longer walks, both the slope and total elevation gain patterns are not convincing. A 'sporty person' will receive similarly disappointing suggestions. But 'someone who likes climbing uphill' will be pleasantly surprised, given both the grade and total elevation gain are much higher for best matches.
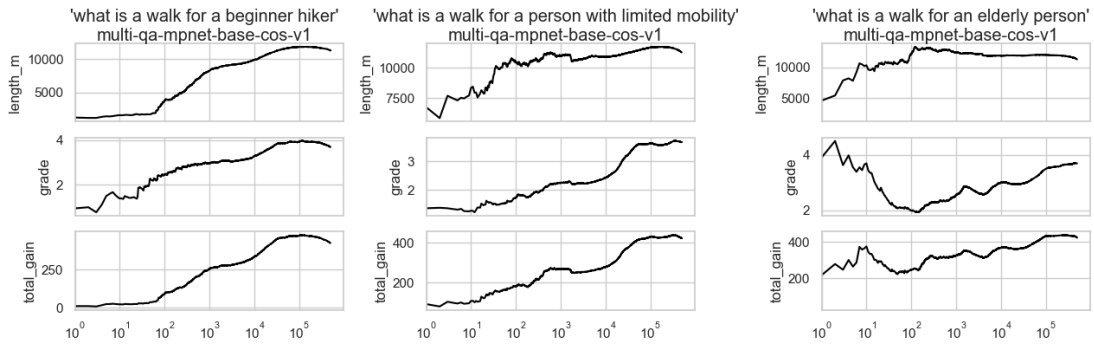
Figure 1: Cumulative mean length, grade, and elevation gain for what should be *easier* routes
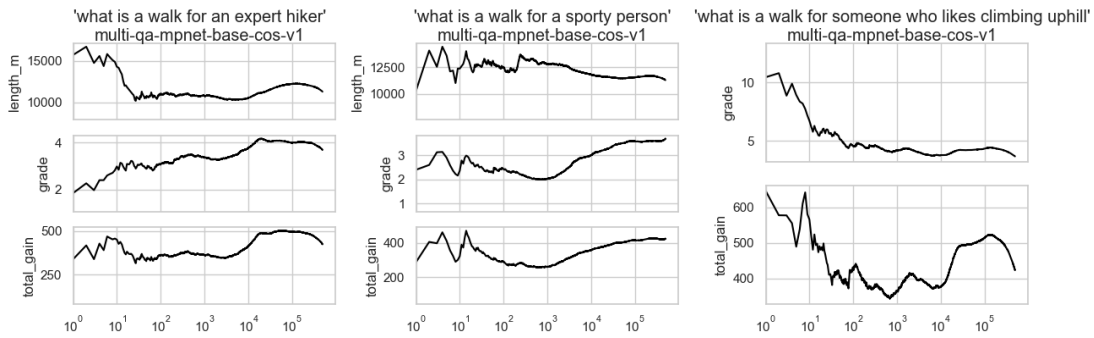
Figure 2: Cumulative mean length, grade, and elevation gain for what should be *harder* routes

Peculiarly, even when fine-tuned on the same Multi-QA dataset, MiniLM, DistilBERT, and MPNet models are in total disagreement over the walks that can be completed in under an hour (Figure 3). We generally expect these to be routes of under 5 km [19]. While MiniLM shows a logical pattern of top-matching routes being shorter, DistilBERT ranks the results in a near reverse order; cumulative mean length of routes ranked by MPNet seems to hover around the dataset mean.

Figure 3: Three models, even though fine-tuned on the same dataset, disagree on which walks can be completed in under 1 hour

None of the models are good at associating 'long' and 'very long' walks with higher kilometre values. Regrettably, most models associate 'someone seeking greater challenges' with descriptions of shorter and flatter walks, and queries for people 'preferring wilderness to man-made' barely relate to walks going through national parks. Full results are available in the Appendix.

4. Conclusion

Sentence transformers, fine-tuned on a corpus of general question-answer pairs for asymmetric semantic search, demonstrate some zero-shot ability to associate short and subjective queries looking for particular hiking experiences with synthetically composed route descriptions. One tested model was able to relate 'beginner hikers' and those with 'limited mobility' to shorter and flatter walks; another was able to associate 'walks that can be completed in under an hour' with shorter walks. Models fine-tuned on the same dataset sometimes showed very different results, signalling that architectures and pre-training matter.

In future work, a more systematic approach to evaluate sentence transformers and other language models for geospatial understanding should be introduced. We suggest focusing on four aspects: model architecture, datasets for fine-tuning, geospatial descriptions, and evaluation. Firstly, a wider array of sentence transformers should be tested to identify which architectures, BERT-based and beyond, achieve better results. Secondly, existing general question-answering datasets used for fine-tuning should be evaluated both independently and in combination to see how dataset size, theme, and quality affect learning. Although the user queries we tested are not 'traditionally' geospatial, will using *smaller* datasets of *primarily geographic* questions make sentence transformers better understand hiking and other active living experiences? Thirdly, we recognise that using a generic template to describe routes is just one of many ways to represent geospatial data as text. As such, we suggest exploring more sophisticated approaches of generating descriptions for routes and other geospatial objects (e.g., specific locations represented as points, and general areas represented as polygons), given that various sources (people, websites) describe such objects differently. And lastly, a more formal way of evaluating ranked results should be explored, accounting for the fact that user queries tested here are more subjective and incomplete than those used in typical information retrieval tasks.

Acknowledgments

References

[1] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL: http://arxiv.org/abs/1908.10084, arXiv:1908.10084 [cs].

[2] W. Wei, P. M. Barnaghi, A. Bargiela, Search with Meanings: An Overview of Semantic Search Systems (2008).

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805 [cs].

[4] K. Zhang, W. Wu, H. Wu, Z. Li, M. Zhou, Question Retrieval with High Quality Answers in Community Question Answering, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, Shanghai China, 2014, pp. 371–380. URL: https://dl.acm.org/doi/10.1145/2661829.2661908. doi:10.1145/2661829.2661908.

[5] N. Davies, Who walks, where and why? Practitioners' observations and perspectives on recreational walkers at UK tourist destinations, Annals of Leisure Research 21 (2018) 553–574. URL: https://doi.org/10.1080/11745398.2016.1250648. doi:10.1080/11745398.2016.1250648, publisher: Routledge _eprint: https://doi.org/10.1080/11745398.2016.1250648.

[6] M. Molokáč, J. Hlaváčová, D. Tometzová, E. Liptáková, The Preference Analysis for Hikers' Choice of Hiking Trail, Sustainability 14 (2022) 6795. URL: https://www.mdpi.com/2071-1050/14/11/6795. doi:10.3390/su14116795, number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[7] R. B. Hull, W. P. Stewart, The Landscape Encountered and Experienced While Hiking, Environment and Behavior 27 (1995) 404–426. URL: https://doi.org/10.1177/0013916595273007. doi:10.1177/0013916595273007, publisher: SAGE Publications Inc.

[8] A. Ballatore, S. Cavazzi, J. Morley, The context of outdoor walking: A classification of user-generated routes, The Geographical Journal 189 (2023) 485–500. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/geoj.12511. doi:10.1111/geoj.12511, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geoj.12511.

[9] GeoPandas 0.dev+untagged — GeoPandas 0+untagged.50.g5558c35.dirty documentation, 2024. URL: https://geopandas.org/en/stable/index.html.

[10] D. Petrak, N. S. Moosavi, I. Gurevych, Arithmetic-Based Pretraining – Improving Numeracy of Pretrained Language Models, 2023. URL: http://arxiv.org/abs/2205.06733, arXiv:2205.06733 [cs].

[11] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, 2020. URL: http://arxiv.org/abs/2002.10957. doi:10.48550/arXiv.2002.10957, arXiv:2002.10957 [cs].

[12] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL: http://arxiv.org/abs/1910.01108. doi:10.48550/arXiv.1910.01108, arXiv:1910.01108 [cs].

[13] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: Masked and Permuted Pre-training for Language Understanding (2020).

[14] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, 2018. URL: http://arxiv.org/abs/1611.09268. doi:10.48550/arXiv.1611.09268, arXiv:1611.09268 [cs].

[15] sentence-transformers/multi-qa-MiniLM-L6-cos-v1 · Hugging Face, 2024. URL: https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1.

[16] L. T. Sarjakoski, P. Kettunen, H.-M. Flink, M. Laakso, M. Rönneberg, T. Sarjakoski, Analysis of verbal route descriptions and landmarks for hiking, Personal and Ubiquitous Computing 16 (2012) 1001–1011. URL: https://doi.org/10.1007/s00779-011-0460-7. doi:10.1007/s00779-011-0460-7.

[17] J.-P. Calbimonte, S. Martin, D. Calvaresi, N. Zappelaz, A. Cotting, Semantic Data Models for Hiking Trail Difficulty Assessment, in: J. Neidhardt, W. Wörndl (Eds.), Information and Communication Technologies in Tourism 2020, Springer International Publishing, Cham, 2020, pp. 295–306. doi:10.1007/978-3-030-36737-4_24.

[18] A. Bellogín, P. Castells, I. Cantador, Statistical biases in Information Retrieval metrics for recommender systems, Information Retrieval Journal 20 (2017) 606–634. URL: https://doi.org/10.1007/s10791-017-9312-z. doi:10.1007/s10791-017-9312-z.

[19] S. M. Club, Scottish Mountaineering Club Journal, Scottish Mountaineering Club., 1893.

A. Tables

Table 1
Route attributes

Attribute	Description	Dataset
`length_m`	Route length in metres	-
`total_gain`	Total elevation gain (ascent)	OS Terrain 5
`total_loss`	Total elevation loss (descent)	OS Terrain 5
`grade`	Hiking grade, calculated as ($\texttt{total\_gain} \div \texttt{length\_m} \times 100$)	-
`is_circular`	*True* if route start and end points are within 500m of each other	-
`is_out_and_back`	*True* if route is circular, and its 25m buffer has a high overlap with itself	-
`start_place`	Name of the nearest place to the route's start point, limited to 1km	OS Open Names
`end_place`	Name of the nearest place to the route's end point, limited to 1 km	OS Open Names
`along_surfacewater`	Percent of route length that falls within 50m of a body of water	OS Zoomstack
`along_coast`	Percent of route length that lies within 150m of GB boundary	OS Zoomstack
`is_coastal`	*True* if at least 50% of the walk length is along the coast	-
`in_national_parks`	Percent of route length that falls within a national park	OS Zoomstack
`in_greenspace`	Percent of route length that falls within a green space boundary	OS Zoomstack
`in_woodland`	Percent of route length that falls within a woodland boundary	OS Zoomstack
`in_urban`	Percent of route length that falls within an urban area boundary	OS Zoomstack

Table 2
Examples of generated descriptions of various character lengths

Description	Length
This is a circular, ten km walk that begins and ends in Priddy, Somerset. Total elevation gain is 130 metres, and elevation grade is 1.3.	138
This is a circular, 12 km walk that begins and ends in Tarrant Gunville, Dorset. Total elevation gain is one hundred and ninety-one metres, and elevation grade is 1.6.	168
This is a 2 km coastal walk that begins in Glenuig Bay, Highland and ends in Smirisary, Highland. Total elevation gain is 142 metres, and elevation grade is 4.9. About seven percent of the walk is in a wooded area, about 62 percent of the walk is along the coast.	263
This is a circular, twenty-five km walk that begins and ends in Milton, Stirling. Total elevation gain is 1846 metres, and elevation grade is 7.4. The walk is entirely within a national park, about twenty percent of the walk is in a wooded area, about ten percent of the walk is alongside a body of water.	305
This is a 6 km walk that begins in Pebbly Hill, Cotswold, Gloucestershire and ends in Stow-on-the-Wold, Cotswold, Gloucestershire. Total elevation gain is 215 metres, and elevation grade is 3.5. The walk is predominantly uphill. About seven percent of the walk is in a wooded area, about 8 percent of the walk goes through an urban area.	337
This is a 22 km walk that begins in Rampart Head, Cumberland and ends in Little Caldew, Cumberland. Total elevation gain is 222 metres, and elevation grade is 1.0. About 6 percent of the walk is in a wooded area, about 9 percent of the walk goes through an urban area, about 7 percent of the walk is within green space, about 18 percent of the walk is along the coast, about twenty-seven percent of the walk is alongside a body of water.	437
This is a nineteen km walk that begins in Millbrook, Caerffili - Caerphilly and ends in Ynysfro Reservoirs, Casnewydd - Newport. Total elevation gain is seven hundred and twenty-four metres, and elevation grade is 3.7. The walk is predominantly downhill. About seventeen percent of the walk is in a wooded area, about forty-five percent of the walk goes through an urban area, about eight percent of the walk is within green space, about 17 percent of the walk is alongside a body of water.	490
This is a 7 km coastal walk that begins in Rhoose Cardiff International Airport, Rhoose / Y Rhws, Bro Morgannwg - the Vale of Glamorgan and ends in Storehouse Point, Bro Morgannwg - the Vale of Glamorgan. Total elevation gain is 194 metres, and elevation grade is 2.7. About thirteen percent of the walk is in a wooded area, about 19 percent of the walk goes through an urban area, about 17 percent of the walk is within green space, 77 percent of the walk is along the coast, about twelve percent of the walk is alongside a body of water.	539

Table 3
User queries

Query	Relevant attributes
what is a short walk	`length_m`
what is a very short walk	`length_m`
what is a long walk	`length_m`
what is a very long walk	`length_m`
what is a walk by the seaside	`is_coastal`
what is a walk through the woods	`in_woodland`
what is an urban walk	`in_urban`
what is a country walk	`in_urban, in_greenspace`
what is a walk for a beginner hiker	`length_m, grade, total_gain`
what is a walk for an expert hiker	`length_m, grade, total_gain`
what is a walk for a sporty person	`length_m, grade, total_gain`
what is a walk for a person with limited mobility	`length_m, grade, total_gain`
what is a walk for an elderly person	`length_m, grade, total_gain`
what is a walk that can be completed in an hour	`length_m`
what is a walk for someone who likes climbing uphill	`grade, total_gain`
what is a walk with a variety of landscapes	`in_urban, in_woodland, in_national_parks, along_surfacewater, along_coast`
what is a walk for someone seeking greater challenges	`length_m, grade, total_gain`
what is a walk for someone who enjoys town walks	`in_urban`
what is a walk for someone who is interested in nature	`in_urban, in_woodland, in_national_parks, along_surfacewater, along_coast`
what is a walk for someone who prefers wilderness to man-made	`in_urban, in_woodland, in_national_parks, along_surfacewater, along_coast`

B. Full results

'what is a short walk'
multi-qa-mpnet-base-cos-v1

'what is a very short walk'
multi-qa-mpnet-base-cos-v1

'what is a long walk'
multi-qa-mpnet-base-cos-v1

'what is a very long walk'
multi-qa-mpnet-base-cos-v1

'what is a walk by the seaside'
multi-qa-mpnet-base-cos-v1

'what is a walk through the woods'
multi-qa-mpnet-base-cos-v1

'what is an urban walk'
multi-qa-mpnet-base-cos-v1

'what is a country walk'
multi-qa-mpnet-base-cos-v1

'what is a walk for a beginner hiker'
multi-qa-mpnet-base-cos-v1

'what is a walk for an expert hiker'
multi-qa-mpnet-base-cos-v1

'what is a walk for a sporty person'
multi-qa-mpnet-base-cos-v1

'what is a walk for a person with limited mobility'
multi-qa-mpnet-base-cos-v1

'what is a walk for an elderly person'
multi-qa-mpnet-base-cos-v1

'what is a walk that can be completed in an hour'
multi-qa-mpnet-base-cos-v1

'what is a walk for someone who likes climbing uphill'
multi-qa-mpnet-base-cos-v1

'what is a walk with a variety of landscapes'
multi-qa-mpnet-base-cos-v1

'what is a walk for someone seeking greater challenges'
multi-qa-mpnet-base-cos-v1

'what is a walk for someone who enjoys town walks'
multi-qa-mpnet-base-cos-v1

'what is a walk for someone who is interested in nature'
multi-qa-mpnet-base-cos-v1

'what is a walk for someone who prefers wilderness to man-made'
multi-qa-mpnet-base-cos-v1

'what is a short walk'
msmarco-MiniLM-L6-cos-v5

'what is a very short walk'
msmarco-MiniLM-L6-cos-v5

'what is a long walk'
msmarco-MiniLM-L6-cos-v5

'what is a very long walk'
msmarco-MiniLM-L6-cos-v5

'what is a walk by the seaside'
msmarco-MiniLM-L6-cos-v5

'what is a walk through the woods'
msmarco-MiniLM-L6-cos-v5

'what is an urban walk'
msmarco-MiniLM-L6-cos-v5

'what is a country walk'
msmarco-MiniLM-L6-cos-v5

'what is a walk for a beginner hiker'
msmarco-MiniLM-L6-cos-v5

'what is a walk for an expert hiker'
msmarco-MiniLM-L6-cos-v5

'what is a walk for a sporty person'
msmarco-MiniLM-L6-cos-v5

'what is a walk for a person with limited mobility'
msmarco-MiniLM-L6-cos-v5

'what is a walk for an elderly person'
msmarco-MiniLM-L6-cos-v5

'what is a walk that can be completed in an hour'
msmarco-MiniLM-L6-cos-v5

'what is a walk for someone who likes climbing uphill'
msmarco-MiniLM-L6-cos-v5

'what is a walk with a variety of landscapes'
msmarco-MiniLM-L6-cos-v5

'what is a walk for someone seeking greater challenges'
msmarco-MiniLM-L6-cos-v5

'what is a walk for someone who enjoys town walks'
msmarco-MiniLM-L6-cos-v5

'what is a walk for someone who is interested in nature'
msmarco-MiniLM-L6-cos-v5

'what is a walk for someone who prefers wilderness to man-made'
msmarco-MiniLM-L6-cos-v5

'what is a short walk'
msmarco-distilbert-cos-v5

'what is a very short walk'
msmarco-distilbert-cos-v5

'what is a long walk'
msmarco-distilbert-cos-v5

'what is a very long walk'
msmarco-distilbert-cos-v5

'what is a walk by the seaside'
msmarco-distilbert-cos-v5

'what is a walk through the woods'
msmarco-distilbert-cos-v5

'what is an urban walk'
msmarco-distilbert-cos-v5

'what is a country walk'
msmarco-distilbert-cos-v5

'what is a walk for a beginner hiker'
msmarco-distilbert-cos-v5

'what is a walk for an expert hiker'
msmarco-distilbert-cos-v5

'what is a walk for a sporty person'
msmarco-distilbert-cos-v5

'what is a walk for a person with limited mobility'
msmarco-distilbert-cos-v5

'what is a walk for an elderly person'
msmarco-distilbert-cos-v5

'what is a walk that can be completed in an hour'
msmarco-distilbert-cos-v5

'what is a walk for someone who likes climbing uphill'
msmarco-distilbert-cos-v5

'what is a walk with a variety of landscapes'
msmarco-distilbert-cos-v5

'what is a walk for someone seeking greater challenges'
msmarco-distilbert-cos-v5

'what is a walk for someone who enjoys town walks'
msmarco-distilbert-cos-v5

'what is a walk for someone who is interested in nature'
msmarco-distilbert-cos-v5

'what is a walk for someone who prefers wilderness to man-made'
msmarco-distilbert-cos-v5