Wikivecs: A Fully Reproducible Vectorization of Multilingual Wikipedia

Brandon Duderstadt
Nomic AI
brandon@nomic.ai

Abstract

Dense vector representations have become foundational to modern natural language processing (NLP), powering diverse workflows from semantic search and retrieval augmented generation to content comparison across languages. Although Wikipedia is one of the most comprehensive and widely used datasets in modern NLP research, it lacks a fully reproducible and permissively licensed dense vectorization. In this paper, we present Wikivecs, a fully reproducible, permissively licensed dataset containing dense vector embeddings for every article in Multilingual Wikipedia. Our pipeline leverages a fully reproducible and permissively licensed multilingual text encoder to embed Wikipedia articles into a unified vector space, making it easy to compare and analyze content across languages. Alongside these vectors, we release a two-dimensional data map derived from the vectors, enabling visualization and exploration of Multilingual Wikipedia's content landscape. We demonstrate the utility of our dataset by identifying several content gaps between English and Russian Wikipedia.

1 Introduction

Dense vector representations have become foundational to modern natural language processing (NLP), powering diverse workflows from semantic search to retrieval augmented generation. As multilingual models have matured, these vectors have become particularly useful for bridging linguistic and cultural gaps, offering a shared representational space where texts from different languages can be meaningfully compared.

Simultaneously, Wikipedia has become one of the most important datasets in modern NLP research. Despite its importance, there exists no openly licensed, fully reproducible resource that provides dense vectors for the entirety of Multilingual Wikipedia. This gap limits the accessibility and transparency of multilingual Wikipedia research and hinders efforts to conduct scalable, comparative content analysis across languages.

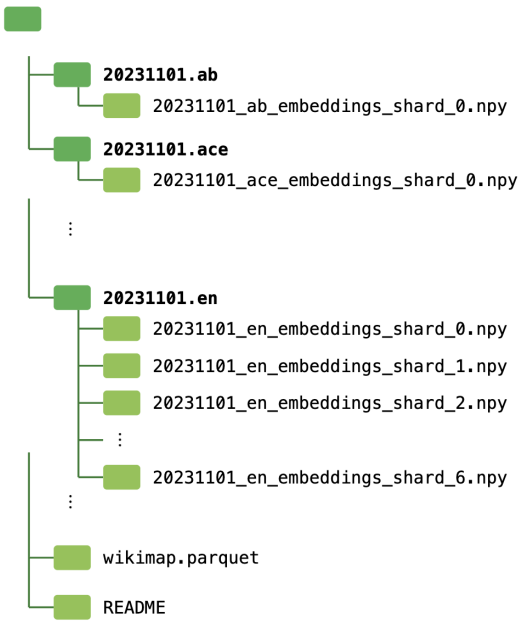


Figure 1: The directory structure of the Wikivecs dataset. Each folder corresponds to a language split and contains one or more sharded numpy arrays. When concatenated in ascending shard index order, these arrays correspond embeddings of the rows in the 2023-11-01 Wikidump dataset. For example, 20231101_en_embeddings_shard_0.npy contains embeddings for rows 0-999,999 of the Wikidump file 20231101.en, 20231101_en_embeddings_shard_1.npy contains embeddings for rows 1,000,0000-1,999,999 of the Wikidump file 20231101.en, etc... Also included in the dataset is wikimap.parquet, a file which contains 2d positions for every article in the Wikidump, enabling subsequent visualization.

In this work, we introduce Wikivecs, a fully reproducible and permissively licensed dataset of dense vector embeddings for every article in Multiulingual Wikipedia. Using a state-of-the-art multilingual encoder, Wikivecs captures the semantic content of articles in a vector space that can

Proceedings of the 2nd Workshop on Advancing Natural Language Processing for Wikipedia (WikiNLP 2025), pages 1–9

be compared meaningfully across languages. We also leverage recent research in scalable dimensionality reduction to construct a 2 dimensional map of the entirety of Multilingual Wikipedia. We demonstrate how these resources can be used to surface topics that lack cross-lingual coverage in Wikipedia, highlighting several content gaps between English and Russian Wikipedia.

2 Background

Early methods for computational text comparison relied on normalized word count statistics (Spärck Jones, 1972) The field of text representation shifted to a much more connectionist paradigm in 2013 with the publication of Word2Vec (Mikolov et al., 2013), which used a shallow neural network to learn a vector representation for each token in a sentence based on its context.

BERT (Devlin et al., 2019) utilized the transformer (Vaswani et al., 2023) architecture to train a deep neural network to produce contextualized vector representations of tokens in an input text. BERT could also be used to compare texts by running it in a cross-encoder configuration. In the cross-encoder configuration, two input texts were fed to the BERT model together to produce a document similarity score.

Computing the pairwise similarity of all texts in a large corpus is computationally burdensome, making it difficult to utilize BERT for document comparison in large corpora. To remedy this, Reimers and Gurevych introduced SBERT (Reimers and Gurevych, 2019), which utilized a bi-encoder architecture and a triplet training procedure to map sentences to dense vectors endowed with a semantic similarity structure. This enabled document similarity to be computed efficiently by measuring the angle between documents' dense vector representations.

The efficiency of this bi-encoder approach has led it to become the dominant technique for large scale text representation and comparison. As a result, a plethora of text bi-encoders have been subsequently developed and released. (Wang et al., 2024; Xiao et al., 2024; Günther et al., 2024; Nussbaum et al., 2024; Li et al., 2023; Chen et al., 2024).

Unfortunately, the almost all state-of-the-art bi-encoders are not suitable for open science and open knowledge use, since large parts of their training recipe (e.g. training data, training code, weights, etc...) remain proprietary or are restrictively licensed. In contrast, the recently released Nomic-Embed-Text-v2 (Nussbaum and Duderstadt, 2025) is the first state-of-the-art multilingual bi-encoder to release all the elements of its training recipe under a permissive Apache 2.0 license, making it an ideal candidate for open science and open knowledge work.

wikimap.parquet (61,614,907 rows)

Column	Type	Description
x	float	X-coordinate for visualization
y	float	Y-coordinate for visualization
title	string	Wikipedia article title
subset	string	Language subset identifier
url	string	Wikipedia article URL
wid	integer	Wikipedia article ID

Figure 2: The schema of wikimap.parquet

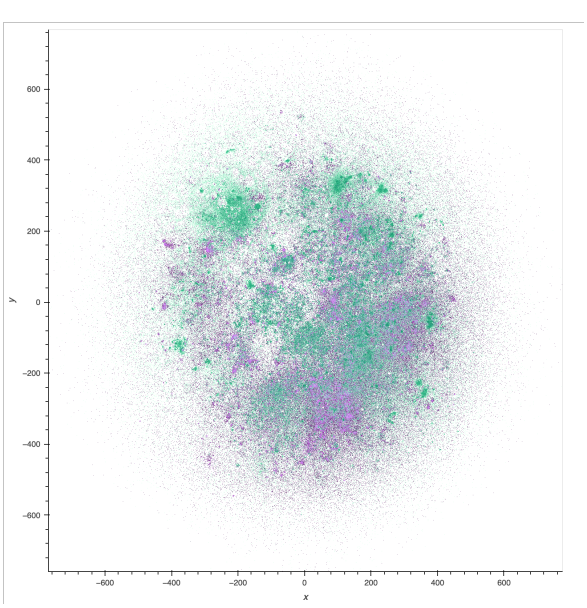
Table 1: NOMAD Projection Hyperparameters

Hyperparameter	Value
n noise	10,000
n neighbors	64
n cells	128
epochs	600
momentum	0.0
lr scale	0.1
learning rate decay start time	0.1
late exaggeration time	0.7
late exaggeration scale	1.5
batch size	70,000
cluster subset size	2,500,000
cluster chunk size	1,000

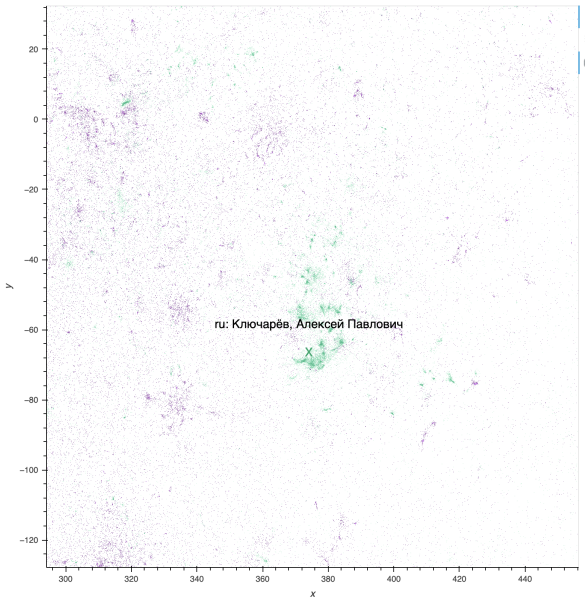
3 Dataset Description

Our dataset is an extension of the 2023-11-01 Wikidump dataset (Wikimedia), which we accessed via Hugging Face.

We produce a dense vector for each article in the Wikidump using Nomic-Embed-Text-v2, a fully open and permissively licensed multilingual text embedding model. Each article is truncated to the first 512 tokens so that it fits in the Nomic-Embed-Text-v2 context window, and no task specific prefixes are prepended to the article texts. Nomic-Embed-Text-v2 then converts each article to a 768 dimensional dense vector. These article vectors are saved in shards corresponding to each of the language splits in the Wikidump. Figure 1 details the directory structure of the resulting shards. We note that process of generating and storing these vectors



(a) A data map overlaying the English (purple) and Russian (green) Wikipedia corpora. Some areas contain dots of both colors, indicating conceptual overlap. Other areas contain dots of only one color, indicating a potential content gap.



(b) A zoomed view of the English-Russian Wikipedia data map. The location of an article on Ключарёв, Алексей Павлович is indicated. Further investigation reveals article is about a Russian Nuclear Physicist, and that it has no English translation. Analysis of the surrounding cluster reveals that it is a cluster of articles about Russian scientists, most of which lack English translations.

Figure 3: Comparison of English and Russian Wikipedia content coverage

is nontrivial, requiring several H100 days of compute, and resulting in approximately a terabyte of data.

Nomic-Embed-Text-v2 achieves state-of-the-art performance on the MIRACL benchmark (Zhang et al., 2022), as well as strong performance on the bitext mining task of the MMTEB benchmark (Enevoldsen et al., 2025). As a result, we can reasonably expect the vectors it produces to map semantically similar content in different languages to similar locations in vector space. This property enables downstream cross-lingual analysis using the produced vectors.

Once all the articles are vectorized, we apply NOMAD Projection (Duderstadt et al., 2025), a scalable nonlinear dimensionality reduction algorithm, to project them into a 2d space for subsequent visualization and inference. We run NOMAD Projection with the hyperparameters outlined in Table 1. The output of NOMAD Projection is used to generate wikimap.parquet, whose schema is detailed in Figure 2. A key benefit of the 2d Wikivecs is their size; working with the 2d vectors enables meaningful corpus analysis to be performed on a standard laptop, greatly increasing the accessibility of our dataset.

4 Access and Reproduction

All of the choices in our vectorization pipeline were made with accessibility and reproducibility in mind. Nomic-Embed-Text-v2 releases its training data, code, and weights under a permissive license, meaning the Wikivectors themselves have fully open provenance. Further, all code for generating the vectors, as well as the subsequent visualizations, is open sourced and permissively licensed. Finally, the 2d vectorization alleviates the large computational burden of working with the 768 dimensional vectors, enabling meaningful corpus analysis to be performed on a standard laptop.

The final dataset can be accessed at https://huggingface.co/datasets/nomic-ai/nomic-embed-v2-wikivecs, and the code for reproducing the dataset can be accessed at https://github.com/nomic-ai/wikivecs.

5 Example Application: Cross Lingual Content Gap Analysis

As an example of the utility of our dataset, we use the 2d positions derived from the Wikivecs to surface content gaps between English and Russian Wikipedia. We define a content gap as a collec-

Article Title	English	Russian
Ключарёв, Алексей Павлович	No	Yes
Глазков, Анатолий Александрович	No	Yes
Воробьёв, Леонид Евгеньевич	No	Yes
Загорец, Павел Авксентьевич	No	Yes
Харитонов, Анатолий Михайлович	No	Yes
Разборов, Александр Александрович	Yes	Yes
Рябинин, Валериан Николаевич	No	Yes
Аржаников, Николай Сергеевич	No	Yes
Золотов, Юрий Александрович (химик)	No	Yes
Левшин, Геннадий Егорович	No	Yes
Total	1	10

Table 2: Ten articles about Russian scientists from a homogeneously colored region in the interactive data exploration application. Manual investigation reveals that English Wikipedia lacks articles on almost all of these scientists, indicating a content gap.

tion of thematically related articles that exist in one language, but not another language. Our dataset enables both qualitative and quantitative surfacing of content gaps. We perform both the qualitative and quantitative analysis on a standard M2 Macbook Air with 24GB of RAM, which highlights the accessibility of the 2d vectors.

5.1 Qualitative Analysis: Interactive Data Explorer

To facilitate the qualitative discovery of content gaps between different languages on Wikipedia, we created an interactive data exploration application for the 2d positions in the Wikivecs corpus. The application plots the 2d positions of articles in two or more language splits of Wikivecs as differently colored dots in a scatter plot. The explorer enables users to pan and zoom the plot, hover their mouse over dots to reveal article titles, and click on dots to open their associated articles in a new tab. The application workflow facilitates the investigation and discovery of content gaps across Multilingual Wikipedia.

Figure 3a shows the application comparing English (purple) and Russian (green) Wikipedia. Some areas of the plot contain dots of both colors, indicating conceptual overlap. Other areas contain dots of only one color, indicating a potential content gap.

Figure 3b shows a zoomed in view of an area in the English-Russian Wikipedia data map containing a concentration of green dots. This area of the map corresponds to articles about Russian scientists on Russian Wikipedia. The location of an article on Ключарёв, Алексей Павлович is indicated by the green X. Google translate reveals that this article is about Alexey Pavlovich Klyucharyov, the head of the Nuclear Physics Department at Kharkov University from 1943-1944. We manually verify that Alexey Pavlovich Klyucharyov has no corresponding article in English Wikipedia.

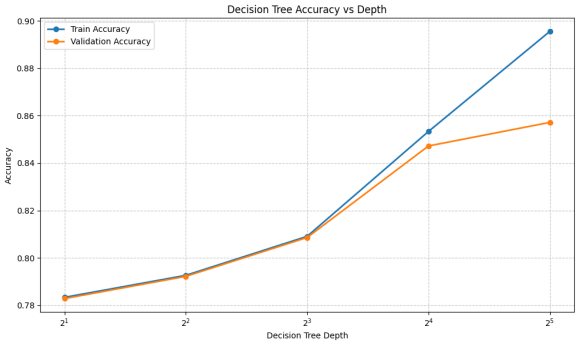
We further investigate this potential content gap by manually inspecting 9 other articles that are proximal to Alexey Pavlovich Klyucharyov in the data explorer. The results of our manual inspection are presented in Table 2. Overall, we find that 9 out of the 10 Russian Wikipedia articles we selected did not have counterparts on English Wikipedia, indicating a content gap.

The discovery of this cluster of Russian scientists who have no matching articles in English Wikipedia demonstrates our dataset's utility for powering qualitative applications that surface content gaps between language splits on Wikipedia.

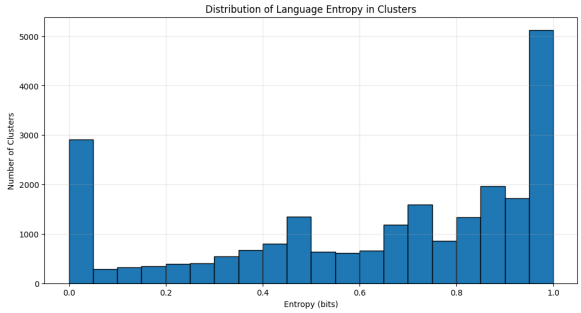
5.2 Quantitative Analysis: Cluster Entropy

To further investigate content gaps between English and Russian Wikipedia, we cluster the 2d positions in the English-Russian Wikipedia map, and investigate clusters with low inter-cluster language distribution entropy.

We start by training a decision tree to partition the 2d map positions into regions of low linguistic entropy. We use a held-out validation set consisting of a random sample of 20% of the English and Russian Wikivec data to determine the maximum depth hyperparameter of our decision tree, and we use a minimum leaf size of 10 to prevent small clusters from forming in our tree. The results of our hyperparameter sweep are presented in Figure



(a) The accuracy of decision tree classifiers trained to predict article languages given their 2d positions in the Wikivecs dataset. The train and validation accuracy curves deviate after a maximum tree depth of 2^4, indicating overfitting. As a result, a maximum tree depth of 2^4 is selected for the final decision tree.



(b) A histogram of the entropy of the language distribution in clusters derived from 2d positions of English and Russian articles in the Wikivecs dataset. There are nearly 3000 clusters with 0 bits of entropy, meaning that they consist entirely of articles in a single language.

Figure 4: Decision tree classifier analysis: (a) Accuracy vs. tree depth showing overfitting beyond depth 2^4, and (b) Entropy distribution of language clusters showing many pure single-language clusters.

4a.

We train a final decision tree with a maximum depth of 2^4 and a minimum leaf size of 10 on all the 2d positions corresponding to English and Russian articles in the Wikivecs dataset. We then interpret the leaf nodes of this decision tree as clusters, and compute the entropy of the language distribution in each cluster. Clusters with a low entropy correspond to spatially localized regions in the map consisting almost entirely of articles in a single language, making them strong candidates to investigate for content gaps.

Figure 4b shows a histogram of the clusters' language distribution entropies. There are a large number of clusters with 0 bits of entropy, meaning that they consist entirely of articles in a single language.

To get a sense of whether these zero entropy clusters actually correspond to content gaps, we manually inspect 10 random articles from three random zero entropy clusters. The results of our manual inspection are presented in Tables 3, 4, and 5. We find that all three clusters contain articles with coherent themes; namely, Beetles, Bibilical Codex, and Hungarian Villages. This is a positive indication that our vectors effectively group articles according to their semantics.

In the Beetles cluster, we find that only 1 of the 10 articles has a Russian translation. Similarly, in the Hungarian Villages cluster, we find that none of the articles have Russian translations. As a result, we conclude that both of these clusters represent thematic content gaps between English and Russian Wikipedia.

In the Biblical Codex cluster, we find that 6 of the 10 articles have Russian translations. This indicates that Russian Wikipedia has coverage of some of the content in this theme. We conclude that this cluster may not represent a content gap between English and Russian Wikipedia.

Overall, the manual verification confirms that our pipeline is able to surface strong candidates for thematic content gaps between English and Russian Wikipedia.

Article Title	English	Russian
Amara exarata	Yes	No
Amara fusca	Yes	No
Amara familiaris	Yes	No
Amara alpina	Yes	Yes
Amara praetermissa	Yes	No
Amara confusa	Yes	No
Amara quenseli	Yes	No
Amara pomona	Yes	No
Amara latior	Yes	No
Amara rubrica	Yes	No
Total	10	1

Table 3: Cluster: 34853 - Beetles

6 Conclusion

In this paper we introduced Wikivecs, the first open source, fully reproducible, and permissively licensed dense vectorization of Multilingual Wikipedia. We generate a 768 dimensional dense vector representation of each article on Multilingual Wikipedia using a fully reproducible and per-

Article Title	English	Russian
Hencse	Yes	No
Kálmáncsa	Yes	No
Fonyód District	Yes	No
Nagyberény	Yes	No
Szentbalázs	Yes	No
Gamás	Yes	No
Öreglak	Yes	No
Bélavár	Yes	No
Marcali	Yes	No
Nagykorpád	Yes	No
Total	10	0

Table 4: Cluster: 29172 - Hungarian Villages

Article Title	English	Russian
Codex Bezae	Yes	Yes
Codex Vaticanus 2061	Yes	Yes
Codex Marchalianus	Yes	Yes
Codex Speculum	Yes	No
Codex Brixianus	Yes	No
Codex Toletanus	Yes	Yes
Codex Sangallensis 1395	Yes	No
Codex Vaticanus 1829	Yes	No
Codex Agobardinus	Yes	No
Codex Boernerianus	Yes	Yes
Total	10	6

Table 5: Cluster: 40931 - Biblical Codex

missively licensed multilingual text embedder. We then perform large scale nonlinear dimensionality reduction on the 768 dimensional vectors to assign every article in Multilingual Wikipedia a position in a 2d semantic map.

As an example of the utility of our dataset, we use both qualitative and quantitative methods to surface content gaps between English and Russian Wikipedia. Qualitatively, we contribute an interactive data exploration application that enables users to visually compare the coverage of different Wikipedia language splits. We use this application to surface a content gap related to Russian scientists who lack articles in English Wikipedia. Quantitatively, we train a decision tree to surface spatially localized regions of low language entropy in the 2d semantic map. Manual investigation of a random sample of these low-entropy regions surfaces content gaps relating to beetles and Hungarian villages on Russian Wikipedia. We run both our qualitative and quantitative analysis on a standard M2 Macbook Air with 24GB of RAM, highlighting the accessibility of our dataset.

Overall, we believe that our dataset will significantly lower the barrier to performing modern NLP application development and analysis on Multilingual Wikipedia.

Limitations

There are several important limitations to consider regarding our dataset.

The Wikivecs corpus is nearly a terabyte of data resulting from several H100 days of processing time. Its scale undoubtedly affects scientists' ability to reproduce and analyze it in its entirety. We attempted to address this by contributing the much more wieldy 2d vectorization in conjuncton with the 768 dimensional vectorization, but the generalization of findings from the 2d vectors to the 768 dimensional vectors represents a limitation in and of itself.

Additionally, the quantitative and qualitative analyses we provided regarding content gaps could be expanded significantly. In particular, we were hindered by the highly manual process of verifying whether or not the clusters we surfaced indeed represented true content gaps. As a result, our evaluations were limited to one language comparison (English-Russian), and were quantified using relatively low evaluation set sample sizes. It is our hope that the release of this dataset will accelerate a much wider community effort to understand and validate content gaps across Multilingual Wikipedia.

Benefits and Risks

How does this work benefit the Wikipedia community? Our work stands to significantly benefit the Wikipedia community. Understanding and rectifying content gaps across different Wikipedia languages is an issue of central importance in the community, and our work demonstrably accelerates work in this area. More broadly, we believe that our work will enable researchers to much more easily build NLP applications and analyses with Multilingual Wikipedia.

What license are you using for your data, code, models? Are they available for community re-use? We made decisions in the interest of accessibility and reproducibility throughout the entirety of our project. We specifically selected Nomic-Embed-Text-v2 as our embedder due to its permissive Apache 2.0 license and end-to-end re-

producibility. Further, we provided a 2d vectorization in addition to the much more unwieldy 768d vectorization to enable corpus analysis using much more limited compute resources. Moreover, we selected a simple and efficient quantitative content gap surfacing method specifically to enable reproducibility on a standard laptop. The entire analysis in quantitative analysis section can be run on an M2 Macbook Air with 24GB of RAM in under 10 minutes.

Finally, all artifacts, data, and models used in this paper are released to the community under a permissive MIT license.

Did you provide clear descriptions and rationale for any filtering that you applied to your data? For example, did you filter to just one language (e.g., English Wikipedia) or many? Did you filter to any specific geographies or topics? The place where we most obviously filtered our data was in our analysis, where we rather arbitrarily selected English and Russian as the languages we would investigate for content gaps. The main rationale for filtering down to two languages is because an exhaustive pairwise comparison of the content between all languages on Wikipedia is a massive undertaking, and is beyond the scope of this paper.

If there are risks from your work, do any of them apply specifically to Wikimedia editors or the projects? We do not forsee any immediate risks to Wikimedia editors or their projects as a result of our work.

Did you name any Wikimedia editors (including username) or provide information exposing an editor's identity? We neither named any Wikimedia editors nor provided information exposing any identities.

Could your research be used to infer sensitive data about individual editors? If so, please explain further. We do not believe our work meaningfully accelerates the ability of any actors to deanonymize any individual editors.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Brandon Duderstadt, Zach Nussbaum, and Laurens van der Maaten. 2025. Nomad projection. *Preprint*, arXiv:2505.15511.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *Preprint*, arXiv:2502.13595.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *Preprint*, arXiv:2310.19923.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Zach Nussbaum and Brandon Duderstadt. 2025. Training sparse mixture of experts text embedding models. *Preprint*, arXiv:2502.07972.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

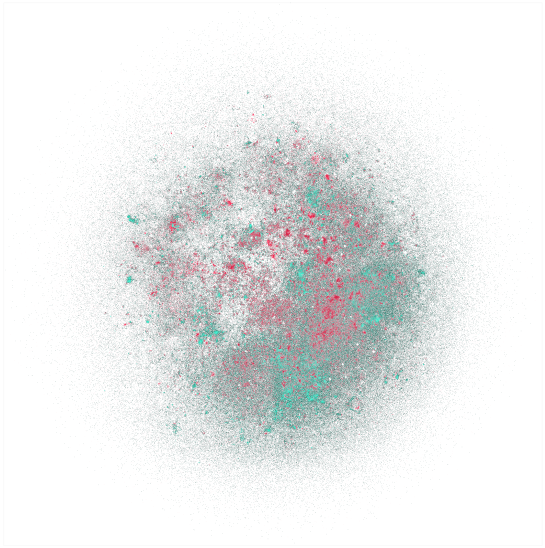
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

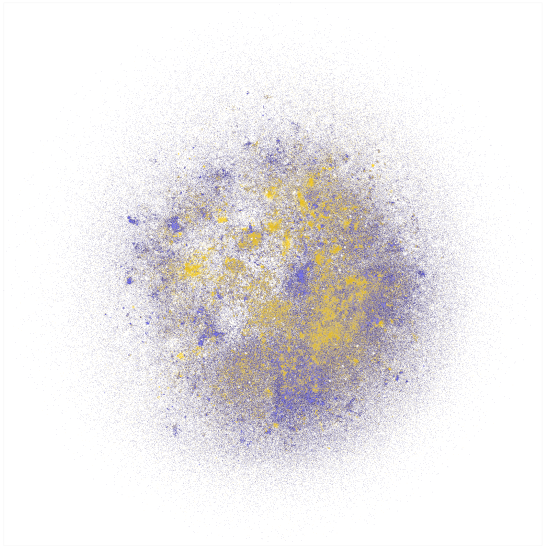
Wikimedia. Wikimedia downloads.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

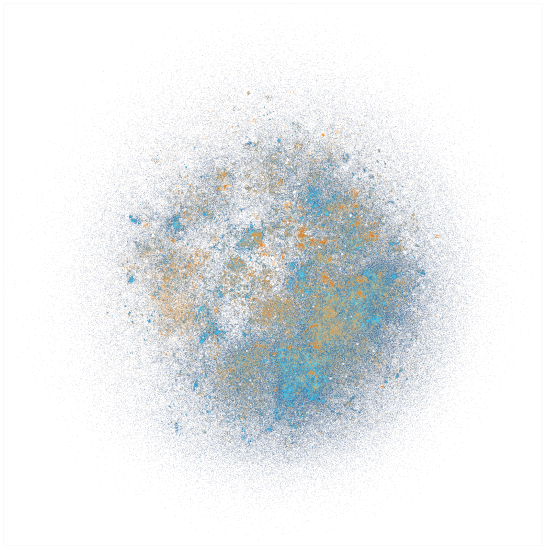
Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *Preprint*, arXiv:2210.09984.



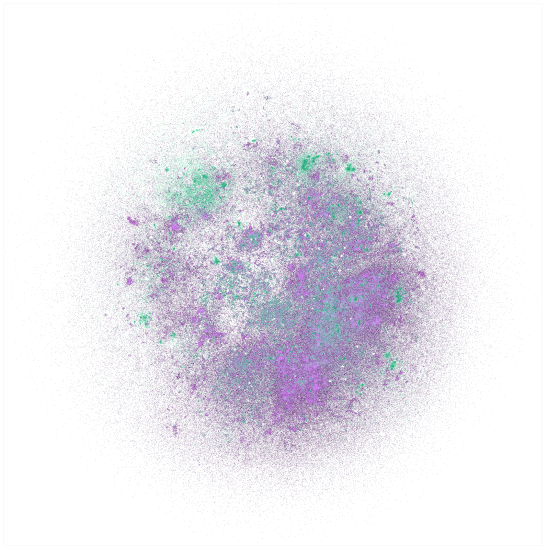
(a) English (teal) vs French (red)



(b) English (purple) vs German (yellow)



(c) English (blue) vs Spanish (orange)



(d) English (purple) vs Russian (green)

Figure 5: Additional visualizations of semantic overlap in Multilingual Wikipedia across several languages.