Векторные базы данных

6 октября 2024 г.

Сравним следующие векторные базы данных, используя бенчмарки с сайта https://ann-benchmarks.com/

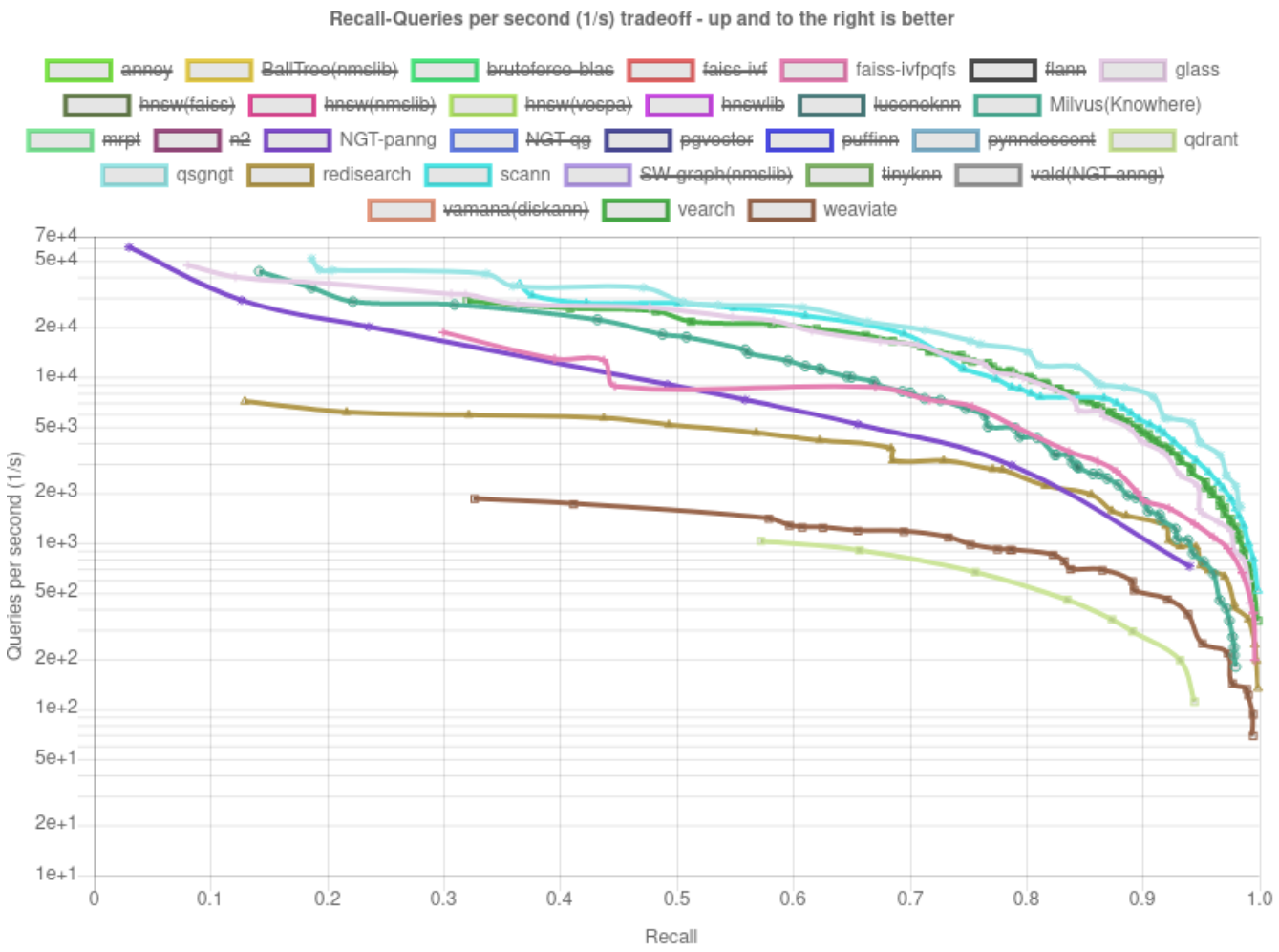
- redisearch
- weaviate
- qdrant
- scann
- milvus
- glass
- faiss-ivfpqfs
- vearch
- NGT-qg
- qsgngt

Также стоит упомянуть chroma и pinecone, у которых отсутствуют бенчмарки на этом сайте, но которые, тем не менее, могут быть довольно эффективными

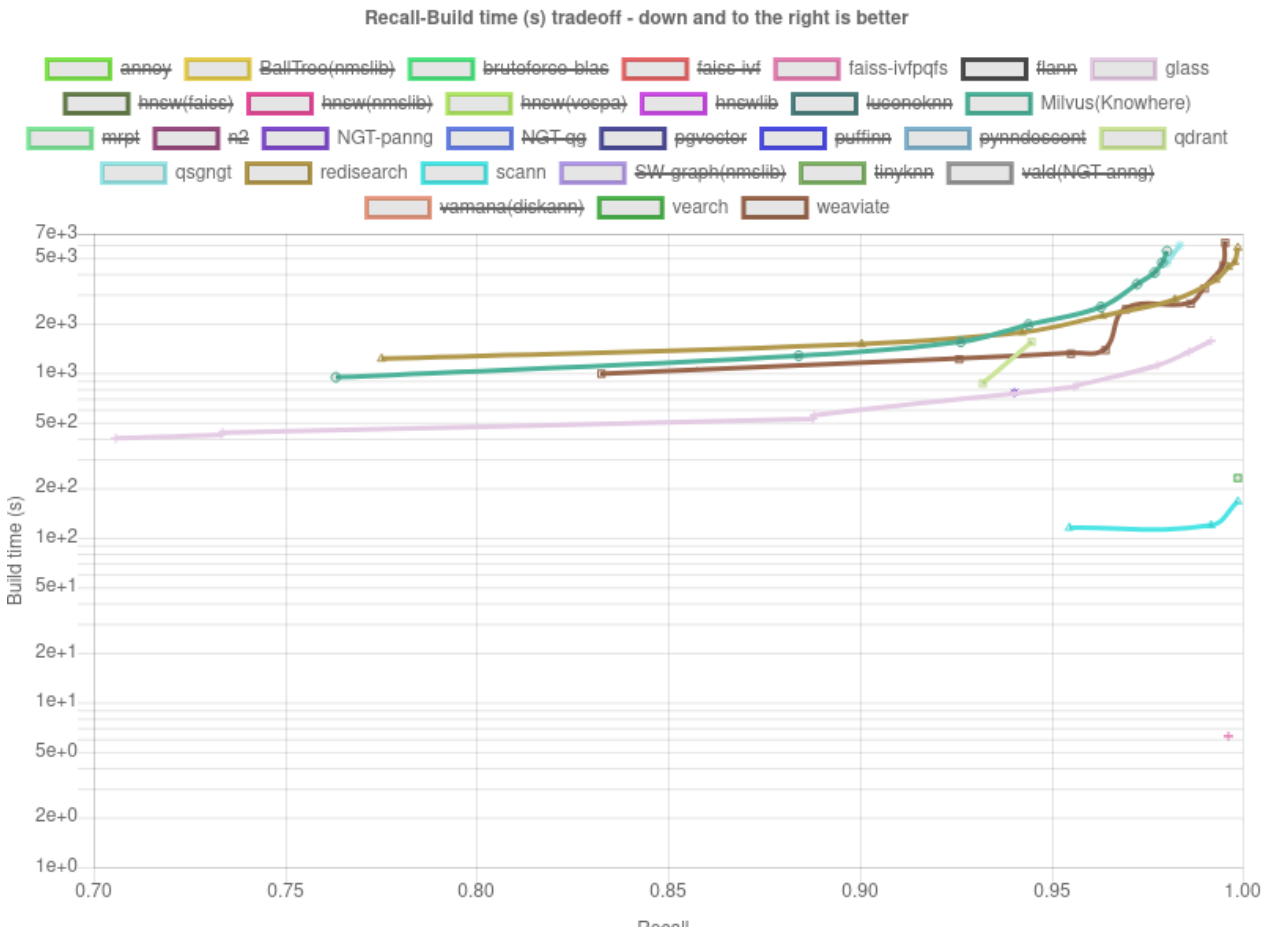
TODO: найти бенчмарки на chroma и pinecone

Посмотрим на общие графики зависимости скорости выполнения запросов, времени построения и размера индекса от recall (на датасете glove-100-angular (k = 10))

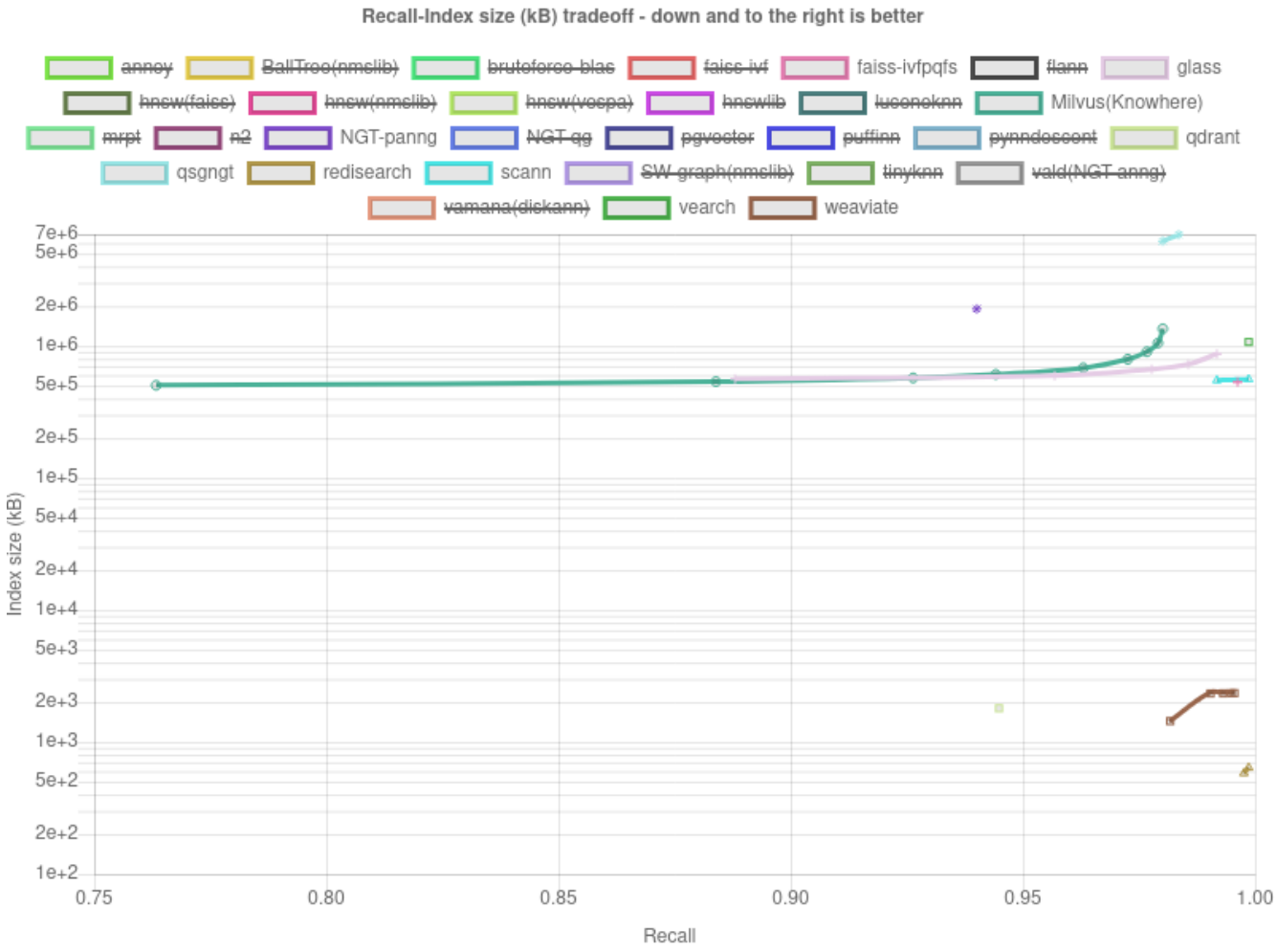
Recall/Queries per second (1/s)



Recall/Build time (s)

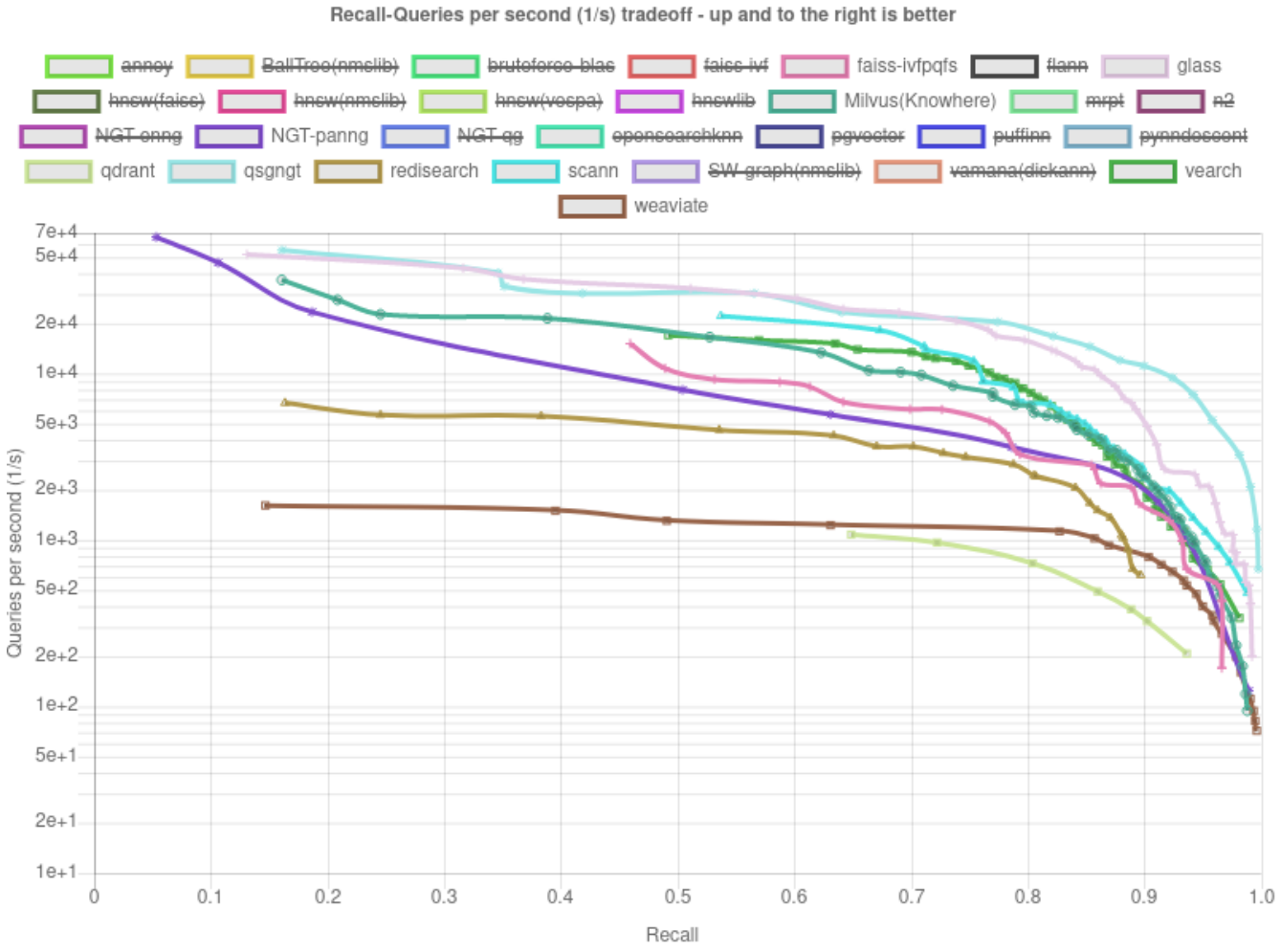


Recall/Index size (kB)

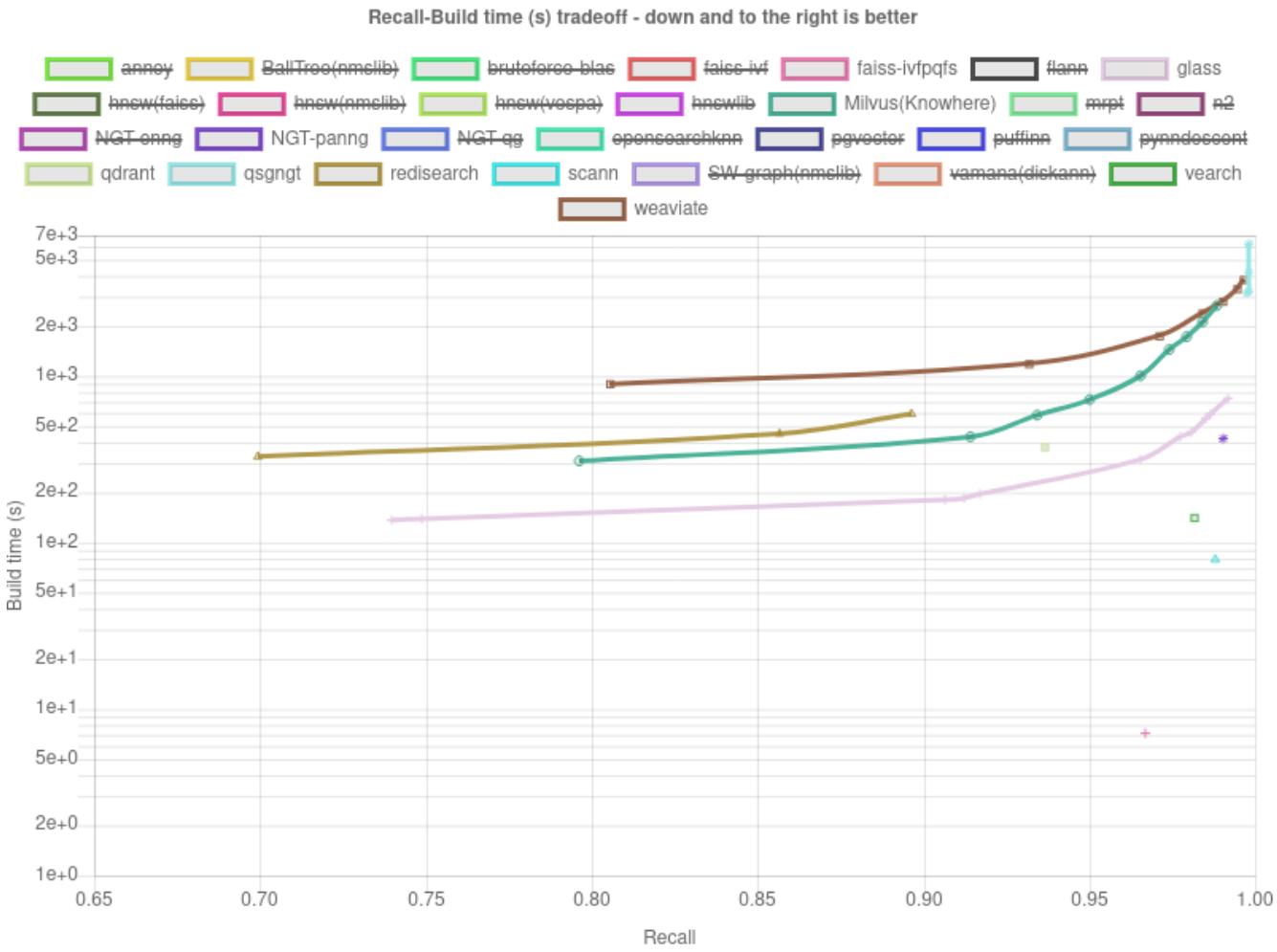


Посмотрим на общие графики зависимости скорости выполнения запросов, времени построения и размера индекса от recall (на датасете nytimes-256-angular (k = 10))

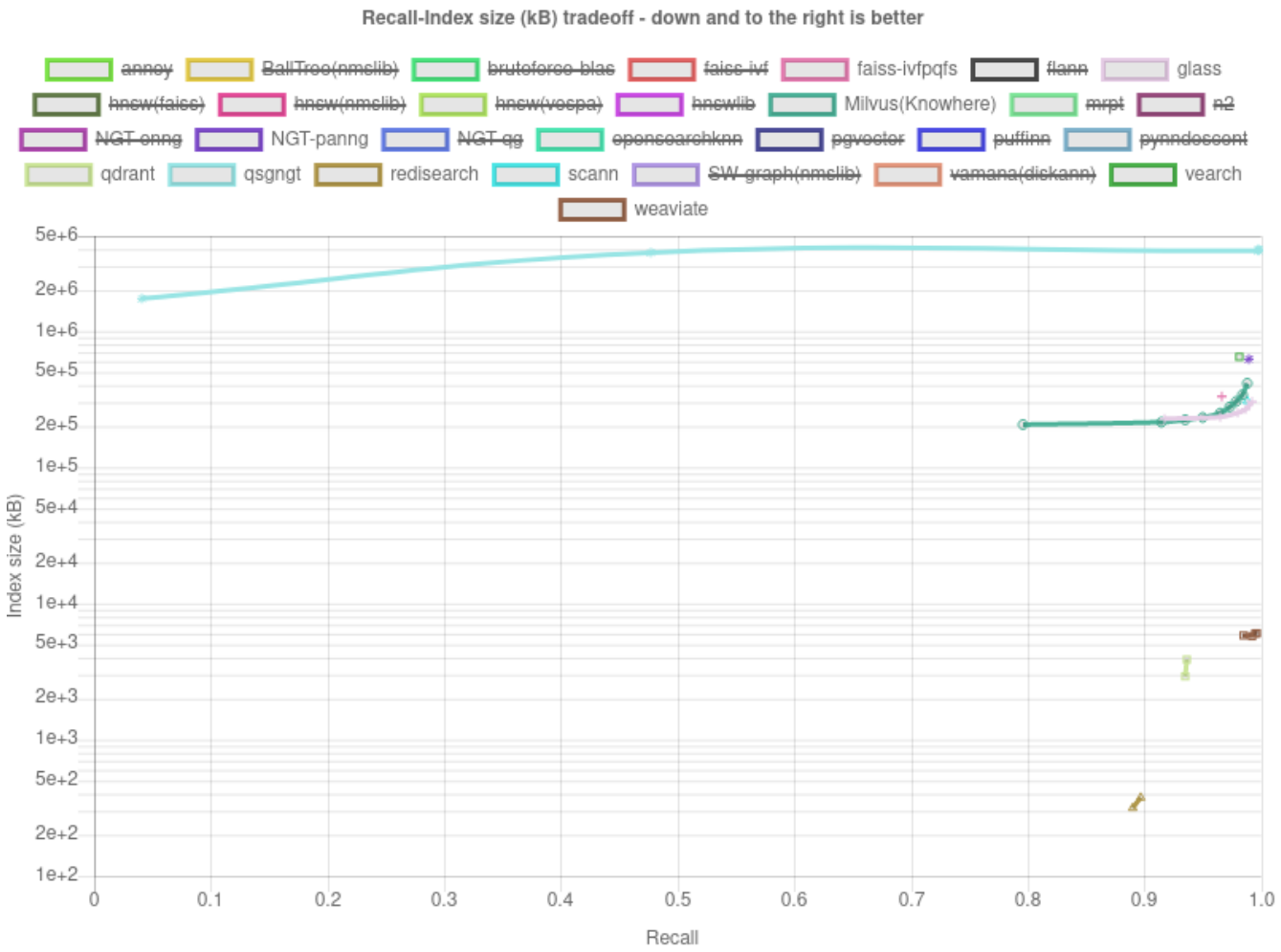
Recall/Queries per second (1/s)



Recall/Build time (s)



Recall/Index size (kB)



qdrant, weaviate и redisearch используют мало памяти, однако требуют довольно много времени на запрос. Из этих трёх redisearch на упомянутых датасетах быстрее и экономичнее по памяти.

Наиболее быстрая обработка запросов у scann, qsgngt, glass, vearch и milvus. Из них scann, glass и milvus довольно хороши по памяти.

faiss-invfpqfs и NGT-qg проигрывают 5 моделям выше и не особо выигрывают по памяти.