Dashboard / My courses / [S22]IR / Part 5. Advanced search topics / Quiz #3

Started on	Monday, 11 April 2022, 9:12 AM
State	Finished
Completed on	Monday, 11 April 2022, 9:32 AM
Time taken	19 mins 55 secs
Grade	**10.00** out of 10.00 (**100**%)

Question **1**

Correct

Mark 1.00 out of 1.00

Mark all data structures which are suitable for non-vector representation of the data.

☑ a. Hierarchical clustering with k-Means ✓ k-Means indeed does not require vector representation!

□ b. annoy

☑ c. navigable small world ✓ this is a metric graph, no vectors needed

☑ d. vantage-point tree ✓ Uses only distance and variance

□ e. kd-tree

Your answer is correct.

The correct answers are:
vantage-point tree,

navigable small world,

Hierarchical clustering with k-Means

Question **2**

Correct

Mark 1.00 out of 1.00

Match the statements about clustering techniques in the scope of search (hierarchical clustering).

Does not guarantee convex cluster form, which can lead to search recall loss.	DBSCAN ✓
Is not a clustering technique.	k-NN ✓
Can be considered as clustering technique. Forms convex partitions, but these partitions don't form Voronoi diagram.	Binary space partitioning ✓
Even if guarantees convex cluster form, does not promise to make clusters balanced in items count.	k-Means ✓

Your answer is correct.

The correct answer is:
Does not guarantee convex cluster form, which can lead to search recall loss.

→ DBSCAN,

Is not a clustering technique.

→ k-NN, Can be considered as clustering technique. Forms convex partitions, but these partitions don't form Voronoi diagram. → Binary space partitioning, Even if guarantees convex cluster form, does not promise to make clusters balanced in items count. → k-Means

Question **3**

Correct

Mark 1.00 out of 1.00

Propose the best expansion to the query "**cookie recipe**". "Best" here refers to an idea, that *recall grows* (we get more relevant) *without precision reduction* (not together with garbage).

○ a. +Trump

○ b. +soup

● c. +biscuit ✓

○ d. +monster

Your answer is correct.

If should not significantly influence the meaning, but expand document set, e.g. with synonyms.

The correct answer is:
+biscuit

Question **4**

Correct

Mark 2.00 out of 2.00

You want to use KD-tree for nearest-neighbour and range search. You will store k=17-dimensional data in the index. Which dataset **sizes** N **are suitable** for this kind of index? *We mean, that it can potentially answer exact NN queries faster than full scan.* **Mark all**.

□ a. 100

□ b. 10 000

☑ c. 1 000 000 ✓

☑ d. 100 000 000 ✓

Your answer is correct.

kd-tree is a binary search tree.

It gives O(log(N)) guarantee on search speed and thus utilizes ~log2(N) first dimensions of the vector for tree levels.

if log2(N) < k, then remaining dimensions can unpredictably influence data distribution in leaf nodes.

Log2 values here are around 7, 14, 20, 27. Thus, you should choose those, where tree will be deeper than 17 levels.

The correct answers are:
1 000 000,

100 000 000

Question **5**

Correct

Mark 2.00 out of 2.00

Vector index of *normed* embeddings can be stored with a smaller memory footprint. For this you can apply some techniques, which can predictably influence (distort) the Euclidean distance metric. Mark all such techniques:

□ a. random walks

□ b. t-SNE

☑ c. random projections ✓ random projections due to Johnson–Lindenstrauss lemma preserve Euclidean metric in (1-eps...1+eps) range.

□ d. PCA

□ e. power iterations method

☑ f. scalar quantization ✓ SQ is method, which discretizes separate dimensions to store them in a smaller data type. E.g. float32 will be stored in int8. This, obviously, adds some error to distance metric, but we can estimate the error as sqrt(d * ((step)^2 + 2*step))

☑ g. product quantization ✓ PQ is a technique which discretizes subvectors and replaces them with indices. This method allows to make even better compression than SQ.

Your answer is correct.

The correct answers are:
random projections, scalar quantization, product quantization

Question **6**

Correct

Mark 1.00 out of 1.00

Mark all **necessary** conditions to build Navigable Small World index.

□ a. Nodes should be elements of vector space

☑ b. Graph should be connected ✓

□ c. New nodes can never be inserted

☑ d. Nodes should be elements of metric space ✓

Your answer is correct.

Classic algorithm is based only on **metric function**, which can be applied to any pair of nodes. Even if vectors have few well-known metrics (cosine, Lx, ...), elements doesn't have to vectors. Strings + editorial distance is also ok.

Graph construction is based on a sequence of trivial insertions, thus **adding new node can be assumed as a continuation of index construction**. It is totally legal. Graph should preserve the same properties.

Search algorithm **starts at a random point**, thus graph **should be connected** to end successfully for each query.

The correct answers are:
Nodes should be elements of metric space,

Graph should be connected

Question **7**

Correct

Mark 2.00 out of 2.00

In the middle of XXth century experimental data showed that human society is better described with small world ✓ graphs, rather than random ✓ . In 2013 a group of scientists proposed navigable small world ✓ graphs. They added a distance function which can be computed for any pair of nodes. We call such graphs metric ✓ . Moreover, they require this graph to have edges between nodes which are close in metric space. This additional requirement defines proximity ✓ graphs.

Couple of years later they improved practical properties of the data structure and proposed hierarchical navigable small world ✓ .

Your answer is correct.

The correct answer is:
In the middle of XXth century experimental data showed that human society is better described with [small world] graphs, rather than [random]. In 2013 a group of scientists proposed [navigable small world] graphs. They added a distance function which can be computed for any pair of nodes. We call such graphs [metric]. Moreover, they require this graph to have edges between nodes which are close in metric space. This additional requirement defines [proximity] graphs.

Couple of years later they improved practical properties of the data structure and proposed [hierarchical navigable small world].

◀ IR13. Image and video retrieval

Jump to...

Data retention summary
Get the mobile app