James Mortensen

1)

a)

When we define term clusters based on the top terms in a document cluster, we're focusing on frequency or importance of words within that specific context of a document cluster. These words may or may not have semantic coherence. They might be the most used terms in the documents but not necessarily related to each other in a meaningful way.

In contrast, when we use K-means to cluster terms, we're grouping words together based on their co-occurrence and distribution across all documents, not just a specific document cluster. The K-means algorithm tries to find inherent structure in the data and groups the words based on their similarities, which can lead to clusters of semantically related terms.

Top terms in a document cluster are more likely to represent the main topics or themes of that specific cluster of documents. On the other hand, term clusters from K-means can represent broader themes or topics across the entire corpus, not just specific to one cluster of documents.

b)

Each document can be represented by the clusters of terms it contains. For example, if we have term clusters for topics like 'sports', 'politics', and 'technology', a document containing terms from the 'sports' and 'technology' clusters might be represented as belonging to these two clusters.

By grouping documents that contain similar term clusters, we can cluster documents based on semantic similarity. For example, all documents containing most terms from the 'sports' cluster could form a 'sports' document cluster.

Term clustering can also be used to reduce noise and dimensionality in the document. Rarely used terms can add noise and make the document clustering process inefficient. By focusing on term clusters, we can potentially ignore these less informative terms and focus on the terms that contribute more to the understanding of the document.

Overall, the method chosen for clustering—whether it's using the top terms in a document cluster or employing an algorithm like K-means—can greatly influence the results of the clustering and the interpretation of the documents' contents.

2)

Probability refers to the measure of the likelihood of a specific outcome or event occurring, given a certain set of known parameters or conditions. It's the measure of how likely a specific event is to occur out of all possible events. Probabilities range from 0 to 1, with 0 indicating the event is impossible and 1 indicating it is certain. Probability is typically used in a forward-thinking perspective; given a population or model and wanting to predict an outcome.

In contrast, likelihood is a measure of the plausibility of a particular statistical model or set of model parameters given some observed data. It's not a measure of certainty, but rather a measure of how well the assumed model explains the observed data. Likelihood doesn't have the same strict 0-1 bounds as probability does. Likelihood is typically used in a reverse-thinking perspective; given some observed data and wanting to evaluate how likely the data could have come from different models or model parameters.

An example to contrast the two further is of a coin flip problem, where one would expect a fair coin to have a probability of landing face up equal to .5. However, if the coin were flipped 100 times and the coin lands face up only 10 times, then one might assume the coin’s true probability is not actually .5. This would make someone doubt whether the model that was used to arrive at a probability of .5 for that coin was not accurate in predicting the real outcomes.