Consider a vocabulary of size d. One hot representation of a word i is "1" at the location (index) corresponding to that word and zero else where. Given a document that contains P words, $\mathbf{w_1}, \ldots, \mathbf{w_P}$, we compute

$$\mathbf{x} = \sum_{i=1}^{P} \mathbf{w}_i$$

Then,

(A) **[Ans]** **x** is the histogram of the words, with x_i as the frequency of i th word.

(B) **x** is the probability distribution with x_i as the probability of words in that document.

(C) **[Ans]** **x** is in R^d independnet of the number of words in the document.

(D) **x** is in R^P independent of the vocabulary size.

(E) **[Ans]** $\sum_i x_i$ is P (x_i is the i th element of **x**)

Consider a document is represented by a histogram of the words in the document. **h**. i.e., h_i is the number of occurrence of the i th word in the document.

We define a linguistic operation: Paraphrasing (P1). P1 is defined as permuting sentences in a document and rewriting a sentence by permuting the words.

(A) **[Ans] h** is invariant to the P1

(B) **h** is not invariant to the P1

(C) **h** is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

(D) a Euclidean distance computed over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a".)

Consider a document is represented by a histogram of the words in the document **h** i.e., h_i is the number of occurrence of the i th word in the document.

We define a linguistic operation: Paraphrasing (P2). P2 is defined as replacing a set of words by their synonyms.

(A) **h** is invariant to the P2

(B) **[Ans] h** is not invariant to the P2

(C) **h** is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

(D) a Euclidean distance computed over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a")

A professor suspected that students while submiting home works are doing the paraphasing operations i.e., both P1 and P2. This resulted in failure of some similarity tests.

Professor designs a $d \times d$ word similarity matrix $\mathbf{S}$ such that $\mathbf{S}_{ij} = \mathbf{S}_{ji} = 1$ if words i and j are synonyms and zero else. (Note: d is the size of vocabulary).

Now to compare two documents, professor multiplies the histogram representations by $\mathbf{S}$.

$$\mathbf{h}'_i = \mathbf{S}\mathbf{h}_i$$

(Note: $\mathbf{h}'_i$ is the new representation. Also, note, after multiplying with the $\mathbf{S}$, the dimension does not change)

(A) **[Ans]** the new representation is invariant under the operation$P1$ and $P2$. (i.e., All the plagiarism now will be detected.)

(B) the new representation is not invariant for $P2$ and it does not help.

(C) the new representation helps for detecting people who have paraphrased with P2. But now it fails for the documents that were not paraphrased (like the original ones/sincere students!).

(D) the idea is worth, but then $\mathbf{S}$ should not have made symmetric. with only one of $\mathbf{S}_{ij}$ or $\mathbf{S}_{ji}$ as 1. The method could have worked as expected.

We want to compare two documents i and j which are represented as histogram (popular known as bag of words) of words h_i and h_j.

Here is what four students argued:

(A) **[Ans]** histograms should be normalized by dividing by the number of words in the document so that the comparison operation becomes "some what invariant" to another linguistic operation: "summarization".

(B) **[Ans]** Cosine distance is a popular distance to compare two documents using this representation.

(C) **[Ans]** Since these are probability distributions, KL-Divergence (may be a symmetric one) is an ideal candidate to compare.

(D) **[Ans]** we should remove the stop words (common words in the language) from the sentence so that the comparison will be more useful. Two documents have the same number of 'the' does not mean any useful similarity between them.