Q1. (5 *points* in total)

- Bag-of-Words: Representing a documents as a bag of words.
- Vector Space: A high dimensional space defined on ”basic concepts”. Documents can be embedded as vectors in this space.

With Bag-of-Words, we can construct the vector space for a set of documents, such that the correlation among documents can be modeled and computed.

Q2. (6 *points* in total)

(a) $TF(apple) = \frac{4}{200} = 0.02$ (2 *points*)

(b) $IDF(apple) = 1 + \log \frac{10^6}{10^2} = 5$ (2 *points*)

(c) $TF - IDF(apple) = 0.02 * 5 = 0.1$ (2 *points*)

Note: Intentionally question (a) asks for the normalized TF. So here we show the results of normalized TF. No point is deducted if you answer is the raw TF.

Q3. (5 *points* in total)

(a) $Prob(\text{“apple”}|D) = \frac{c(\text{“apple”},D)}{|D|} = \frac{1}{5} = 0.2$ (2 *points*)

(b) If Laplace smoothing is applied, $Prob(\text{“apple”}|D) = \frac{c(\text{“apple”},D)+1}{|D|+|V|} = \frac{1+1}{5+12} = 0.1176$ (3 *points*)

Q4. (4 *points* in total)

To assign non-zero probability to unseen words (or n-grams).