1. Email spam filtering models often use a **bag-of-words** representation for emails. In a bag-of-words representation, the descriptive features that describe a document (in our case, an email) each represent how many times a particular word occurs in the document. One descriptive feature is included for each word in a predefined dictionary. The dictionary is typically defined as the complete set of words that occur in the training dataset. The table below lists the bag-of-words representation for the following five emails and a target feature, SPAM, whether they are spam emails or genuine emails:

- "*money, money, money*"
- "*free money for free gambling fun*"
- "*gambling for fun*"
- "*machine learning for fun, fun, fun*"
- "*free machine learning*"

| | Bag-of-Words | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ID | MONEY | FREE | FOR | GAMBLING | FUN | MACHINE | LEARNING | SPAM |
| 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | true |
| 2 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | true |
| 3 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | true |
| 4 | 0 | 0 | 1 | 0 | 3 | 1 | 1 | false |
| 5 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | false |

a. What target level would a nearest neighbor model using **Euclidean distance** return for the following email: "*machine learning for free*"?

b. What target level would a *k*-NN model with *k* = 3 and using **Euclidean distance** return for the same query?

c. What target level would a **weighted *k*-NN** model with *k* = 5 and using a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query, return for the query?

d. What target level would a *k*-NN model with *k* = 3 and using **Manhattan distance** return for the same query?

e. There are a lot of zero entries in the spam bag-of-words dataset. This is indicative of

**sparse data** and is typical for text analytics. **Cosine similarity** is often a good choice when dealing with sparse non-binary data. What target level would a 3-NN model using cosine similarity return for the query?