1. **What does it mean to perform sentiment analysis on text, and give an example?**

Sentiment analysis is the process of understanding the subjective emotions that are associated with a text, as well as realizing its affective states. For example, given three texts:

- "I love using the new Google Pixel!"
- "The pixel battery life is awfully short"
- "Out of all my phones, the Pixel is average"

using sentiment analysis, a program could map the first text to positive emotion, the second text to negative emotion, and the third text to neutral emotion.

2. **Explain how one can use a bag-of-words model to featurize a text document. Name a pro and con of bag-of-words**

Bag-of-words is used to featurize a document by counting the number of times a word appears in a document. It maps each unique word to a frequency, and gives features to the document with this map. A pro of bag-of-words is that it is fairly low computation, easy to implement, and it has multiple adjustments that can be made to it to make it more complex. A con of bag-of-words is that it can lead to a very large vocabulary size, as well as it not taking into account the order of words.

3. **In bag-of-visual-words for image classification, what makes up the vocabulary in this model, how is this similar to text?**

In bag-of-visual-words for image classification, the vocabulary is made from different portions of the image. This is similar to the text model because just like words are the building blocks of a document, an image is built from its unique portions.

4. **What is the purpose of interest point detection, and how is it different from a regular grid model?**

The purpose of interest point detection is to featurize an image and break it up into smaller parts. Interest point detection picks out the well defined and visible parts of an image and makes these parts of the image the features. This is different from a regular grid because a regular grid splits the whole image into a grid regardless of what is in that grid so the whole image is taken into account. In interest point detection, however, there are certain parts that are deliberately picked out to use because of the qualities of that portion of the image.

5. **Why is it sometimes important to reduce the size of a vocabulary, and what are some ways to do this?**

It is important to reduce the size of a vocabulary because the document vocabulary should be stripped of unnecessary and misleading words. We only want to keep the words that truly define the document. In addition, we want to prevent from too large of a vocabulary vector

that will lead to large amounts of computation. A few ways to do this is by ignoring capitalization and punctuation, using stemming, and removing stop words.

6. **What are stop-words and why are they relevant to bag-of-words?**

Stop words are words such as "a", "the", etc. that tend to appear very frequently in text, but add no actual meaning to the text. Thus, they are usually removed prior to performing bag-of-words.

7. **What are n-grams and how can they be used to help model a document?**

N-grams are contiguous sequences of n words within a document. They are useful in analyzing text because they allow for features to store the ordering of words and also can give context for words, compared to bag of words(n=1) which is just an unordered collection of each word and its count. In certain situations, bi-grams(n=2) have been shown to drastically improve model performance compared to single words.

8. **What is the main goal of a SVM and how does it relate to sentiment analysis/image classification?**

The goal of the SVM is to maximize the distance between the two classes' data points. By doing this, we ensure that we are picking a decision boundary that separates the data points by the furthest margin, and therefore most optimally. Depending on the dimension N of the data, the svm will make a decision boundary in an N - dimensional plane to separate the data. After training on data and arriving at an optimal model, text or image features can be extracted and passed through the SVM to predict the class of unknown data.

9. **What is the main purpose of bag-of-words?**

    A. Feature Generation
    B. Regularization
    C. Classification
    D. Clustering

A. Feature Generation

The purpose of bag-of-words is to turn documents of text into features usable by a model. Bag of words itself is just the process of generating these features, but those features can then be used in a variety of ways.

10. **Which of the following could be used in conjunction with bag-of-words?**
    A. PCA
    B. K-means
    C. SVM

D. All of the above

D. All of the above

Bag of words is simply responsible for creating features, afterwards it does not make any assumptions about what type of model or technique the features are used on.

11. Which of the following makes sense to use instead of the raw count of a word?

A. Frequency (n_occurances / n_words_in_document)
B. log (count)
C. Normalized feature vectors (divide each bag by its l2 norm)
D. All of the above

D. All of the above

A and C can both be used to account for differences in lengths of documents, since just using raw count will be proportional to the length of a document. Using log can be useful in "dampening" the effect of high frequency words so they don't dominate over medium frequency words as much.

12. How does tf-idf scoring differ from traditional bag-of-words?

A. Puts more weight on words that appear in many documents and also puts more weight on words that occur a lot in a single document
B. Takes away weight from words that appear in many documents and puts less weight on words that occur a lot in a single document
C. Takes away weight from words that appear in many documents and puts more weight on words that occur a lot in a single document
D. Puts more weight on words that appear in many documents and puts less weight on words that are occur a lot in a single document

C. Takes away weight from words that appear in many documents and puts more weight on words that occur a lot in a single document

Tf-idf score is trying to weight words accordingly based on how often they appear in a single document as well as how many documents they appear in. If a word occurs at least once in many different documents, then we should take away weight from this word because it is common amongst most documents, and therefore not an important word to discriminate between documents. If a word appears many times in a single document, we want to give it more weight because it is a very important part of the composition of that single document.