BBC News Article Classification using Custom Bag of Words and TF-IDF Implementations

**Objective:** Implement Bag of Words and TF-IDF from scratch, then use these implementations to classify BBC news articles into their respective categories.

**Dataset:** You will use the provided BBC news corpus, where each article contains:

1. The full text of the news article
2. The category of the article (e.g., politics, technology, sports, business, entertainment)

**Tasks:**

1. **Data Preprocessing:**
   1. Clean the text data by removing punctuation, converting to lowercase, and removing stop words.
   2. Tokenize the text into individual words.
2. **Implement Bag of Words:**
   1. Create a function to build a vocabulary from the training set.
   2. Implement a function that converts a document into a BoW vector.
   3. Use your implementation to represent each document in both training and testing sets as a BoW vector.
3. **Implement TF-IDF:**
   1. Create a function to calculate term frequency (TF) for each term in a document.
   2. Implement a function to calculate inverse document frequency (IDF) for each term in the corpus.
   3. Combine TF and IDF to create TF-IDF vectors for each document.
4. **Analysis:**
   1. For a given category, find the top 10 words with the highest average TF-IDF scores.
   2. Identify words that have high TF scores but low IDF scores, and vice versa.