MapReduce Programming Model

1) Title of Problem

Implement a MapReduce-style program to count word frequencies from multiple text files while applying the following conditions:

- Ignore common stopwords such as `the`, `and`, `of`, etc.
- Display the Top 5 least frequent words.
- Count only words with more than 5 characters.
- Extend the program to process multiple text files.

2) Code for Problem

The following Python code implements the above requirements:

```
from collections import defaultdict
import glob

# Mapper function
def mapper(file_path, stopwords):
    mapped = []
    with open(file_path, 'r', encoding="utf-8") as f:
        for line in f:
            words = line.strip().split()
            for word in words:
                w = word.lower().strip(",.?!:;\"'()[]{}")
                if w and w not in stopwords and len(w) > 5:
                    mapped.append((w, 1))
    return mapped

# Reducer function
def reducer(mapped_data):
    reduced = defaultdict(int)
    for word, count in mapped_data:
        reduced[word] += count
    return reduced

if __name__ == "__main__":
    # Define stopwords
    stopwords = {"the", "and", "of", "to", "a", "in",
                 "is", "it", "that", "for", "on",
                 "with", "as", "by", "at"}

```

```
    # Process all text files
    all_files = glob.glob("*.txt")
    mapped_data = []
    for file_path in all_files:
        mapped_data.extend(mapper(file_path, stopwords))

    reduced_data = reducer(mapped_data)

    # Top 5 least frequent
    least_frequent = sorted(reduced_data.items(), key=lambda x: x[1])
        [:5]
    print("\nTop 5 Least Frequent Words:")
    for word, freq in least_frequent:
        print(f"{word}: {freq}")
```

3) Solution Output

Top 5 Least Frequent Words

servers: 1
storage: 1
designed: 1
handle: 1
failures: 1