In [1]:

```python
import os
import re
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
```

In [2]:

```python
# Initialize stemmer and stopwords
ps = PorterStemmer()
stop_words = set(stopwords.words("english"))
```

In [3]:

```python
folder_path = "./input"
```

In [4]:

```python
def preprocess_and_conflate(file_path):
    with open(file_path, "r") as f:
        text = f.read().lower()

    # Remove non-alphabetic characters
    words = re.findall(r"\b[a-z]+\b", text)

    # Remove stopwords and apply stemming
    stemmed_words = [ps.stem(w) for w in words if w not in stop_words]

    # Build document representative (unique stems with frequency)
    freq = {}
    for w in stemmed_words:
        freq[w] = freq.get(w, 0) + 1

    return freq
```

In [5]:

```python
# Process all documents
documents = {}
for fname in ["doc1.txt", "doc2.txt", "doc3.txt", "doc4.txt"]:
    file_path = os.path.join(folder_path, fname)
    documents[fname] = preprocess_and_conflate(file_path)
```

In [6]:

```python
# Print results
for doc, rep in documents.items():
    print(f"\nDocument Representative for {doc}:")
    print(rep)
```

```
Document Representative for doc1.txt:
{'machin': 1, 'learn': 2, 'subset': 1, 'artifici': 1, 'intellig': 1, 'algorithm':
1, 'build': 1, 'mathemat': 1, 'model': 2, 'base': 1, 'sampl': 1, 'data': 2, 'know
n': 1, 'train': 1, 'make': 1, 'predict': 1, 'decis': 1, 'without': 1, 'explicitl
i': 1, 'program': 1}

Document Representative for doc2.txt:
{'artifici': 1, 'intellig': 3, 'simul': 1, 'human': 1, 'process': 2, 'machin': 1,
'ai': 1, 'applic': 1, 'includ': 1, 'natur': 1, 'languag': 1, 'speech': 1, 'recogn
it': 1, 'comput': 1, 'vision': 1, 'abil': 1, 'learn': 1, 'reason': 1, 'self': 1,
'correct': 1}

Document Representative for doc3.txt:
{'data': 1, 'mine': 2, 'process': 1, 'discov': 1, 'pattern': 1, 'larg': 1, 'datas
et': 1, 'combin': 1, 'statist': 1, 'artifici': 1, 'intellig': 1, 'databas': 1, 's
ystem': 1, 'use': 1, 'inform': 1, 'help': 1, 'decis': 1, 'make': 1}

Document Representative for doc4.txt:
{'comput': 1, 'network': 3, 'connect': 1, 'devic': 1, 'enabl': 1, 'commun': 2, 'i
nvolv': 1, 'protocol': 1, 'topolog': 1, 'transmiss': 1, 'media': 1, 'internet':
1, 'largest': 1, 'support': 1, 'global': 1}
```

In [7]:

```
import json

# Save all document representatives into one JSON file
with open("documents.json", "w") as f:
    json.dump(documents, f, indent=4)

print("Document representatives saved to documents.json")
```

```
Document representatives saved to documents.json
```

In [ ]: