Text Preprocessing Report

1. Original Texts

English Text:

"Artificial intelligence is the future. Artificial intelligence improves human life. The future of Artificial intelligence is bright."

Arabic Text:

"الذكاء الاصطناعي هو المستقبل. الذكاء الاصطناعي يعزز حياة الإنسان. المستقبل الذكاء."

2. Text After Preprocessing Steps

2.1. Tokenization

- **English Tokens**:

```
['Artificial', 'intelligence', 'is', 'the', 'future', '.',
'Artificial', 'intelligence', 'improves', 'human', 'life', '.', 'The',
'future', 'of', 'Artificial', 'intelligence', 'is', 'bright', '.']
```

- **Arabic Tokens**:

```
['الذكاء', 'الاصطناعي', 'هو', 'المستقبل', '،', 'الذكاء',
'الاصطناعي', 'يعزز', 'حياة', 'الإنسان', '،', 'المستقبل', 'الذكاء']
```

2.2. Stopword Removal

- **Filtered English Tokens**:

```
['Artificial', 'intelligence', 'future', 'Artificial', 'intelligence',
'improves', 'human', 'life', 'future', 'Artificial', 'intelligence',
'bright']
```

- **Filtered Arabic Tokens:**

```
['الذكاء', 'الاصطناعي', 'المستقبل', 'الذكاء', 'الاصطناعي', 'يعزز',
'حياة', 'الإنسان', 'المستقبل', 'الذكاء']
```

2.3. Noise Removal

- **Cleaned English Tokens:**

```
['Artificial', 'intelligence', 'future', 'Artificial', 'intelligence',
'improves', 'human', 'life', 'future', 'Artificial', 'intelligence',
'bright']
```

- **Cleaned Arabic Tokens:**

```
['الذكاء', 'الاصطناعي', 'المستقبل', 'الذكاء', 'الاصطناعي', 'يعزز',
'حياة', 'الإنسان', 'المستقبل', 'الذكاء']
```

2.4. Normalization

- **Normalized English Tokens:**

```
['artificial', 'intelligence', 'future', 'artificial', 'intelligence',
'improves', 'human', 'life', 'future', 'artificial', 'intelligence',
'bright']
```

- **Normalized Arabic Tokens:**

```
['الذكاء', 'الاصطناعي', 'المستقبل', 'الذكاء', 'الاصطناعي', 'يعزز',
'حياة', 'الإنسان', 'المستقبل', 'الذكاء']
```

2.5. POS Tagging

- **English POS Tags**:

```
[('artificial', 'ADJ'), ('intelligence', 'NOUN'), ('future', 'NOUN'),
('artificial', 'ADJ'), ('intelligence', 'NOUN'), ('improves', 'VERB'),
('human', 'ADJ'), ('life', 'NOUN'), ('future', 'NOUN'), ('artificial',
'ADJ'), ('intelligence', 'NOUN'), ('bright', 'ADJ')]
```

- **Arabic POS Tags**:

```
[('الذكاء', 'noun'), ('الاصطناعي', 'adj'), ('المستقبل', 'noun'),
('الذكاء', 'noun'), ('الاصطناعي', 'adj'), ('يعزز', 'verb'), ('حياة',
'noun'), ('الإنسان', 'noun'), ('المستقبل', 'noun'), ('الذكاء',
'noun')]
```

3. Python Code

For the Python code part, I couldn’t include the actual code cells in the word document or else it would be such a mess so a solution I came up with is to provide an external PDF format for the notebook itself, thanks for the understanding 😁.

4. Results and Observations

Observations

- **Arabic Challenges**:

    - Preprocessing Arabic is more complex due to diacritics, special characters, and word inflections.
    - Tools like **Camel Tools** are essential for Arabic-specific tasks like POS tagging.
- **English Challenges**:
    - Handling contractions (e.g., "isn't") during tokenization and normalization required careful adjustments.

Key Benefits of Preprocessing

1. **Tokenization**: Makes text manageable for further analysis.
2. **Stopword Removal**: Simplifies text by removing irrelevant words.
3. **Noise Removal**: Ensures only meaningful content remains.
4. **Normalization**: Standardizes text for consistency.
5. **POS Tagging**: Adds linguistic structure for advanced NLP tasks.