Lab 3: Tokenization, Stemming, and Lemmatization

# 1. Sentence Tokenization

1.   
Natural Language Processing (NLP) is a crucial area of Artificial Intelligence that deals with the interaction between humans and machines using natural language.

2. The field of NLP includes several tasks such as language modeling, text classification, translation, information retrieval, and more.

3. Among the most essential preprocessing steps in NLP are tokenization, stemming, and lemmatization.

4. Tokenization involves breaking down the text into smaller units called tokens.

5. These can be sentences, words, characters, or even subwords.

6. Each type of tokenization has its own relevance.

7. For example, sentence tokenization helps understand sentence boundaries, word tokenization is crucial for word-level analysis,   
character-level tokenization is often used in neural language models, and subword tokenization is helpful in dealing with unknown or rare words.

8. Stemming is a technique where words are reduced to their base or root form.

9. However, the stemmed word may not be a valid English word.

10. For example, 'fishing', 'fished', and 'fisher' may all be reduced to 'fish'.

11. The Porter Stemmer is commonly used in many NLP applications due to its speed and simplicity.

12. Lemmatization, on the other hand, is a more sophisticated approach that uses a vocabulary and morphological analysis to return the dictionary form of a word.

13. For instance, the word 'better' is lemmatized to 'good' with proper context and part-of-speech tagging.

14. Lemmatization is more accurate than stemming but computationally heavier.

15. Stopwords are commonly used words in English such as 'is', 'in', 'the', which are often removed during preprocessing because they add little semantic value.

16. Removing stopwords helps in reducing noise and improving the performance of NLP models.

17. Let us consider some sample sentences for processing:  
  
1.

18. The children were playing joyfully in the playground.

19. 2.

20. She studies hard and has studied well for her exams.

21. 3.

22. The running water was soothing to hear.

23. 4.

24. He bought the best quality camera among all available models.

25. Each sentence showcases the use of various verbs, adverbs, and adjectives that demonstrate the need for contextual analysis in lemmatization.

26. This input file is structured to provide sufficient diversity for tokenization, stemming, and lemmatization demonstrations in NLP Lab Program 3.

# 2. Word Tokenization Techniques

Default Tokenizer:  
['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'is', 'a', 'crucial', 'area', 'of', 'Artificial', 'Intelligence', 'that', 'deals', 'with', 'the', 'interaction', 'between', 'humans', 'and', 'machines', 'using', 'natural', 'language', '.', 'The', 'field', 'of', 'NLP', 'includes', 'several', 'tasks', 'such', 'as', 'language', 'modeling', ',', 'text', 'classification', ',', 'translation', ',', 'information', 'retrieval', ',', 'and', 'more', '.', 'Among', 'the', 'most', 'essential', 'preprocessing', 'steps', 'in', 'NLP', 'are', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatization', '.', 'Tokenization', 'involves', 'breaking', 'down', 'the', 'text', 'into', 'smaller', 'units', 'called', 'tokens', '.', 'These', 'can', 'be', 'sentences', ',', 'words', ',', 'characters', ',', 'or', 'even', 'subwords', '.', 'Each', 'type', 'of', 'tokenization', 'has', 'its', 'own', 'relevance', '.', 'For', 'example', ',', 'sentence', 'tokenization', 'helps', 'understand', 'sentence', 'boundaries', ',', 'word', 'tokenization', 'is', 'crucial', 'for', 'word-level', 'analysis', ',', 'character-level', 'tokenization', 'is', 'often', 'used', 'in', 'neural', 'language', 'models', ',', 'and', 'subword', 'tokenization', 'is', 'helpful', 'in', 'dealing', 'with', 'unknown', 'or', 'rare', 'words', '.', 'Stemming', 'is', 'a', 'technique', 'where', 'words', 'are', 'reduced', 'to', 'their', 'base', 'or', 'root', 'form', '.', 'However', ',', 'the', 'stemmed', 'word', 'may', 'not', 'be', 'a', 'valid', 'English', 'word', '.', 'For', 'example', ',', "'fishing", "'", ',', "'fished", "'", ',', 'and', "'fisher", "'", 'may', 'all', 'be', 'reduced', 'to', "'fish", "'", '.', 'The', 'Porter', 'Stemmer', 'is', 'commonly', 'used', 'in', 'many', 'NLP', 'applications', 'due', 'to', 'its', 'speed', 'and', 'simplicity', '.', 'Lemmatization', ',', 'on', 'the', 'other', 'hand', ',', 'is', 'a', 'more', 'sophisticated', 'approach', 'that', 'uses', 'a', 'vocabulary', 'and', 'morphological', 'analysis', 'to', 'return', 'the', 'dictionary', 'form', 'of', 'a', 'word', '.', 'For', 'instance', ',', 'the', 'word', "'better", "'", 'is', 'lemmatized', 'to', "'good", "'", 'with', 'proper', 'context', 'and', 'part-of-speech', 'tagging', '.', 'Lemmatization', 'is', 'more', 'accurate', 'than', 'stemming', 'but', 'computationally', 'heavier', '.', 'Stopwords', 'are', 'commonly', 'used', 'words', 'in', 'English', 'such', 'as', "'is", "'", ',', "'in", "'", ',', "'the", "'", ',', 'which', 'are', 'often', 'removed', 'during', 'preprocessing', 'because', 'they', 'add', 'little', 'semantic', 'value', '.', 'Removing', 'stopwords', 'helps', 'in', 'reducing', 'noise', 'and', 'improving', 'the', 'performance', 'of', 'NLP', 'models', '.', 'Let', 'us', 'consider', 'some', 'sample', 'sentences', 'for', 'processing', ':', '1', '.', 'The', 'children', 'were', 'playing', 'joyfully', 'in', 'the', 'playground', '.', '2', '.', 'She', 'studies', 'hard', 'and', 'has', 'studied', 'well', 'for', 'her', 'exams', '.', '3', '.', 'The', 'running', 'water', 'was', 'soothing', 'to', 'hear', '.', '4', '.', 'He', 'bought', 'the', 'best', 'quality', 'camera', 'among', 'all', 'available', 'models', '.', 'Each', 'sentence', 'showcases', 'the', 'use', 'of', 'various', 'verbs', ',', 'adverbs', ',', 'and', 'adjectives', 'that', 'demonstrate', 'the', 'need', 'for', 'contextual', 'analysis', 'in', 'lemmatization', '.', 'This', 'input', 'file', 'is', 'structured', 'to', 'provide', 'sufficient', 'diversity', 'for', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatization', 'demonstrations', 'in', 'NLP', 'Lab', 'Program', '3', '.']

Treebank Tokenizer:  
['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'is', 'a', 'crucial', 'area', 'of', 'Artificial', 'Intelligence', 'that', 'deals', 'with', 'the', 'interaction', 'between', 'humans', 'and', 'machines', 'using', 'natural', 'language.', 'The', 'field', 'of', 'NLP', 'includes', 'several', 'tasks', 'such', 'as', 'language', 'modeling', ',', 'text', 'classification', ',', 'translation', ',', 'information', 'retrieval', ',', 'and', 'more.', 'Among', 'the', 'most', 'essential', 'preprocessing', 'steps', 'in', 'NLP', 'are', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatization.', 'Tokenization', 'involves', 'breaking', 'down', 'the', 'text', 'into', 'smaller', 'units', 'called', 'tokens.', 'These', 'can', 'be', 'sentences', ',', 'words', ',', 'characters', ',', 'or', 'even', 'subwords.', 'Each', 'type', 'of', 'tokenization', 'has', 'its', 'own', 'relevance.', 'For', 'example', ',', 'sentence', 'tokenization', 'helps', 'understand', 'sentence', 'boundaries', ',', 'word', 'tokenization', 'is', 'crucial', 'for', 'word-level', 'analysis', ',', 'character-level', 'tokenization', 'is', 'often', 'used', 'in', 'neural', 'language', 'models', ',', 'and', 'subword', 'tokenization', 'is', 'helpful', 'in', 'dealing', 'with', 'unknown', 'or', 'rare', 'words.', 'Stemming', 'is', 'a', 'technique', 'where', 'words', 'are', 'reduced', 'to', 'their', 'base', 'or', 'root', 'form.', 'However', ',', 'the', 'stemmed', 'word', 'may', 'not', 'be', 'a', 'valid', 'English', 'word.', 'For', 'example', ',', "'fishing", "'", ',', "'fished", "'", ',', 'and', "'fisher", "'", 'may', 'all', 'be', 'reduced', 'to', "'fish'.", 'The', 'Porter', 'Stemmer', 'is', 'commonly', 'used', 'in', 'many', 'NLP', 'applications', 'due', 'to', 'its', 'speed', 'and', 'simplicity.', 'Lemmatization', ',', 'on', 'the', 'other', 'hand', ',', 'is', 'a', 'more', 'sophisticated', 'approach', 'that', 'uses', 'a', 'vocabulary', 'and', 'morphological', 'analysis', 'to', 'return', 'the', 'dictionary', 'form', 'of', 'a', 'word.', 'For', 'instance', ',', 'the', 'word', "'better", "'", 'is', 'lemmatized', 'to', "'good", "'", 'with', 'proper', 'context', 'and', 'part-of-speech', 'tagging.', 'Lemmatization', 'is', 'more', 'accurate', 'than', 'stemming', 'but', 'computationally', 'heavier.', 'Stopwords', 'are', 'commonly', 'used', 'words', 'in', 'English', 'such', 'as', "'is", "'", ',', "'in", "'", ',', "'the", "'", ',', 'which', 'are', 'often', 'removed', 'during', 'preprocessing', 'because', 'they', 'add', 'little', 'semantic', 'value.', 'Removing', 'stopwords', 'helps', 'in', 'reducing', 'noise', 'and', 'improving', 'the', 'performance', 'of', 'NLP', 'models.', 'Let', 'us', 'consider', 'some', 'sample', 'sentences', 'for', 'processing', ':', '1.', 'The', 'children', 'were', 'playing', 'joyfully', 'in', 'the', 'playground.', '2.', 'She', 'studies', 'hard', 'and', 'has', 'studied', 'well', 'for', 'her', 'exams.', '3.', 'The', 'running', 'water', 'was', 'soothing', 'to', 'hear.', '4.', 'He', 'bought', 'the', 'best', 'quality', 'camera', 'among', 'all', 'available', 'models.', 'Each', 'sentence', 'showcases', 'the', 'use', 'of', 'various', 'verbs', ',', 'adverbs', ',', 'and', 'adjectives', 'that', 'demonstrate', 'the', 'need', 'for', 'contextual', 'analysis', 'in', 'lemmatization.', 'This', 'input', 'file', 'is', 'structured', 'to', 'provide', 'sufficient', 'diversity', 'for', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatization', 'demonstrations', 'in', 'NLP', 'Lab', 'Program', '3', '.']

Regex Tokenizer (words only):  
['Natural', 'Language', 'Processing', 'NLP', 'is', 'a', 'crucial', 'area', 'of', 'Artificial', 'Intelligence', 'that', 'deals', 'with', 'the', 'interaction', 'between', 'humans', 'and', 'machines', 'using', 'natural', 'language', 'The', 'field', 'of', 'NLP', 'includes', 'several', 'tasks', 'such', 'as', 'language', 'modeling', 'text', 'classification', 'translation', 'information', 'retrieval', 'and', 'more', 'Among', 'the', 'most', 'essential', 'preprocessing', 'steps', 'in', 'NLP', 'are', 'tokenization', 'stemming', 'and', 'lemmatization', 'Tokenization', 'involves', 'breaking', 'down', 'the', 'text', 'into', 'smaller', 'units', 'called', 'tokens', 'These', 'can', 'be', 'sentences', 'words', 'characters', 'or', 'even', 'subwords', 'Each', 'type', 'of', 'tokenization', 'has', 'its', 'own', 'relevance', 'For', 'example', 'sentence', 'tokenization', 'helps', 'understand', 'sentence', 'boundaries', 'word', 'tokenization', 'is', 'crucial', 'for', 'word', 'level', 'analysis', 'character', 'level', 'tokenization', 'is', 'often', 'used', 'in', 'neural', 'language', 'models', 'and', 'subword', 'tokenization', 'is', 'helpful', 'in', 'dealing', 'with', 'unknown', 'or', 'rare', 'words', 'Stemming', 'is', 'a', 'technique', 'where', 'words', 'are', 'reduced', 'to', 'their', 'base', 'or', 'root', 'form', 'However', 'the', 'stemmed', 'word', 'may', 'not', 'be', 'a', 'valid', 'English', 'word', 'For', 'example', 'fishing', 'fished', 'and', 'fisher', 'may', 'all', 'be', 'reduced', 'to', 'fish', 'The', 'Porter', 'Stemmer', 'is', 'commonly', 'used', 'in', 'many', 'NLP', 'applications', 'due', 'to', 'its', 'speed', 'and', 'simplicity', 'Lemmatization', 'on', 'the', 'other', 'hand', 'is', 'a', 'more', 'sophisticated', 'approach', 'that', 'uses', 'a', 'vocabulary', 'and', 'morphological', 'analysis', 'to', 'return', 'the', 'dictionary', 'form', 'of', 'a', 'word', 'For', 'instance', 'the', 'word', 'better', 'is', 'lemmatized', 'to', 'good', 'with', 'proper', 'context', 'and', 'part', 'of', 'speech', 'tagging', 'Lemmatization', 'is', 'more', 'accurate', 'than', 'stemming', 'but', 'computationally', 'heavier', 'Stopwords', 'are', 'commonly', 'used', 'words', 'in', 'English', 'such', 'as', 'is', 'in', 'the', 'which', 'are', 'often', 'removed', 'during', 'preprocessing', 'because', 'they', 'add', 'little', 'semantic', 'value', 'Removing', 'stopwords', 'helps', 'in', 'reducing', 'noise', 'and', 'improving', 'the', 'performance', 'of', 'NLP', 'models', 'Let', 'us', 'consider', 'some', 'sample', 'sentences', 'for', 'processing', '1', 'The', 'children', 'were', 'playing', 'joyfully', 'in', 'the', 'playground', '2', 'She', 'studies', 'hard', 'and', 'has', 'studied', 'well', 'for', 'her', 'exams', '3', 'The', 'running', 'water', 'was', 'soothing', 'to', 'hear', '4', 'He', 'bought', 'the', 'best', 'quality', 'camera', 'among', 'all', 'available', 'models', 'Each', 'sentence', 'showcases', 'the', 'use', 'of', 'various', 'verbs', 'adverbs', 'and', 'adjectives', 'that', 'demonstrate', 'the', 'need', 'for', 'contextual', 'analysis', 'in', 'lemmatization', 'This', 'input', 'file', 'is', 'structured', 'to', 'provide', 'sufficient', 'diversity', 'for', 'tokenization', 'stemming', 'and', 'lemmatization', 'demonstrations', 'in', 'NLP', 'Lab', 'Program', '3']

WordPunct Tokenizer:  
['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'is', 'a', 'crucial', 'area', 'of', 'Artificial', 'Intelligence', 'that', 'deals', 'with', 'the', 'interaction', 'between', 'humans', 'and', 'machines', 'using', 'natural', 'language', '.', 'The', 'field', 'of', 'NLP', 'includes', 'several', 'tasks', 'such', 'as', 'language', 'modeling', ',', 'text', 'classification', ',', 'translation', ',', 'information', 'retrieval', ',', 'and', 'more', '.', 'Among', 'the', 'most', 'essential', 'preprocessing', 'steps', 'in', 'NLP', 'are', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatization', '.', 'Tokenization', 'involves', 'breaking', 'down', 'the', 'text', 'into', 'smaller', 'units', 'called', 'tokens', '.', 'These', 'can', 'be', 'sentences', ',', 'words', ',', 'characters', ',', 'or', 'even', 'subwords', '.', 'Each', 'type', 'of', 'tokenization', 'has', 'its', 'own', 'relevance', '.', 'For', 'example', ',', 'sentence', 'tokenization', 'helps', 'understand', 'sentence', 'boundaries', ',', 'word', 'tokenization', 'is', 'crucial', 'for', 'word', '-', 'level', 'analysis', ',', 'character', '-', 'level', 'tokenization', 'is', 'often', 'used', 'in', 'neural', 'language', 'models', ',', 'and', 'subword', 'tokenization', 'is', 'helpful', 'in', 'dealing', 'with', 'unknown', 'or', 'rare', 'words', '.', 'Stemming', 'is', 'a', 'technique', 'where', 'words', 'are', 'reduced', 'to', 'their', 'base', 'or', 'root', 'form', '.', 'However', ',', 'the', 'stemmed', 'word', 'may', 'not', 'be', 'a', 'valid', 'English', 'word', '.', 'For', 'example', ',', "'", 'fishing', "',", "'", 'fished', "',", 'and', "'", 'fisher', "'", 'may', 'all', 'be', 'reduced', 'to', "'", 'fish', "'.", 'The', 'Porter', 'Stemmer', 'is', 'commonly', 'used', 'in', 'many', 'NLP', 'applications', 'due', 'to', 'its', 'speed', 'and', 'simplicity', '.', 'Lemmatization', ',', 'on', 'the', 'other', 'hand', ',', 'is', 'a', 'more', 'sophisticated', 'approach', 'that', 'uses', 'a', 'vocabulary', 'and', 'morphological', 'analysis', 'to', 'return', 'the', 'dictionary', 'form', 'of', 'a', 'word', '.', 'For', 'instance', ',', 'the', 'word', "'", 'better', "'", 'is', 'lemmatized', 'to', "'", 'good', "'", 'with', 'proper', 'context', 'and', 'part', '-', 'of', '-', 'speech', 'tagging', '.', 'Lemmatization', 'is', 'more', 'accurate', 'than', 'stemming', 'but', 'computationally', 'heavier', '.', 'Stopwords', 'are', 'commonly', 'used', 'words', 'in', 'English', 'such', 'as', "'", 'is', "',", "'", 'in', "',", "'", 'the', "',", 'which', 'are', 'often', 'removed', 'during', 'preprocessing', 'because', 'they', 'add', 'little', 'semantic', 'value', '.', 'Removing', 'stopwords', 'helps', 'in', 'reducing', 'noise', 'and', 'improving', 'the', 'performance', 'of', 'NLP', 'models', '.', 'Let', 'us', 'consider', 'some', 'sample', 'sentences', 'for', 'processing', ':', '1', '.', 'The', 'children', 'were', 'playing', 'joyfully', 'in', 'the', 'playground', '.', '2', '.', 'She', 'studies', 'hard', 'and', 'has', 'studied', 'well', 'for', 'her', 'exams', '.', '3', '.', 'The', 'running', 'water', 'was', 'soothing', 'to', 'hear', '.', '4', '.', 'He', 'bought', 'the', 'best', 'quality', 'camera', 'among', 'all', 'available', 'models', '.', 'Each', 'sentence', 'showcases', 'the', 'use', 'of', 'various', 'verbs', ',', 'adverbs', ',', 'and', 'adjectives', 'that', 'demonstrate', 'the', 'need', 'for', 'contextual', 'analysis', 'in', 'lemmatization', '.', 'This', 'input', 'file', 'is', 'structured', 'to', 'provide', 'sufficient', 'diversity', 'for', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatization', 'demonstrations', 'in', 'NLP', 'Lab', 'Program', '3', '.']

# 3. Character-Level Tokenization

['N', 'a', 't', 'u', 'r', 'a', 'l', 'L', 'a', 'n', 'g', 'u', 'a', 'g', 'e', 'P', 'r', 'o', 'c', 'e', 's', 's', 'i', 'n', 'g', '(', 'N', 'L', 'P', ')', 'i', 's', 'a', 'c', 'r', 'u', 'c', 'i', 'a', 'l', 'a', 'r', 'e', 'a', 'o', 'f', 'A', 'r', 't', 'i', 'f', 'i', 'c', 'i', 'a', 'l', 'I', 'n', 't', 'e', 'l', 'l', 'i', 'g', 'e', 'n', 'c', 'e', 't', 'h', 'a', 't', 'd', 'e', 'a', 'l', 's', 'w', 'i', 't', 'h', 't', 'h', 'e', 'i', 'n', 't', 'e', 'r', 'a', 'c', 't', 'i', 'o', 'n', 'b', 'e', 't', 'w', 'e', 'e', 'n', 'h', 'u', 'm', 'a', 'n', 's', 'a', 'n', 'd', 'm', 'a', 'c', 'h', 'i', 'n', 'e', 's', 'u', 's', 'i', 'n', 'g', 'n', 'a', 't', 'u', 'r', 'a', 'l', 'l', 'a', 'n', 'g', 'u', 'a', 'g', 'e', '.', 'T', 'h', 'e', 'f', 'i', 'e', 'l', 'd', 'o', 'f', 'N', 'L', 'P', 'i', 'n', 'c', 'l', 'u', 'd', 'e', 's', 's', 'e', 'v', 'e', 'r', 'a', 'l', 't', 'a', 's', 'k', 's', 's', 'u', 'c', 'h', 'a', 's', 'l', 'a', 'n', 'g', 'u', 'a', 'g', 'e', 'm', 'o', 'd', 'e', 'l', 'i', 'n', 'g', ',', 't', 'e', 'x', 't', 'c', 'l', 'a', 's', 's', 'i', 'f', 'i', 'c', 'a', 't', 'i', 'o', 'n', ',', 't', 'r', 'a', 'n', 's', 'l', 'a', 't', 'i', 'o', 'n', ',', 'i', 'n', 'f', 'o', 'r', 'm', 'a', 't', 'i', 'o', 'n', 'r', 'e', 't', 'r', 'i', 'e', 'v', 'a', 'l', ',', 'a', 'n', 'd', 'm', 'o', 'r', 'e', '.', 'A', 'm', 'o', 'n', 'g', 't', 'h', 'e', 'm', 'o', 's', 't', 'e', 's', 's', 'e', 'n', 't', 'i', 'a', 'l', 'p', 'r', 'e', 'p', 'r', 'o', 'c', 'e', 's', 's', 'i', 'n', 'g', 's', 't', 'e', 'p', 's', 'i', 'n', 'N', 'L', 'P', 'a', 'r', 'e', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', ',', 's', 't', 'e', 'm', 'm', 'i', 'n', 'g', ',', 'a', 'n', 'd', 'l', 'e', 'm', 'm', 'a', 't', 'i', 'z', 'a', 't', 'i', 'o', 'n', '.', 'T', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'i', 'n', 'v', 'o', 'l', 'v', 'e', 's', 'b', 'r', 'e', 'a', 'k', 'i', 'n', 'g', 'd', 'o', 'w', 'n', 't', 'h', 'e', 't', 'e', 'x', 't', 'i', 'n', 't', 'o', 's', 'm', 'a', 'l', 'l', 'e', 'r', 'u', 'n', 'i', 't', 's', 'c', 'a', 'l', 'l', 'e', 'd', 't', 'o', 'k', 'e', 'n', 's', '.', 'T', 'h', 'e', 's', 'e', 'c', 'a', 'n', 'b', 'e', 's', 'e', 'n', 't', 'e', 'n', 'c', 'e', 's', ',', 'w', 'o', 'r', 'd', 's', ',', 'c', 'h', 'a', 'r', 'a', 'c', 't', 'e', 'r', 's', ',', 'o', 'r', 'e', 'v', 'e', 'n', 's', 'u', 'b', 'w', 'o', 'r', 'd', 's', '.', 'E', 'a', 'c', 'h', 't', 'y', 'p', 'e', 'o', 'f', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'h', 'a', 's', 'i', 't', 's', 'o', 'w', 'n', 'r', 'e', 'l', 'e', 'v', 'a', 'n', 'c', 'e', '.', 'F', 'o', 'r', 'e', 'x', 'a', 'm', 'p', 'l', 'e', ',', 's', 'e', 'n', 't', 'e', 'n', 'c', 'e', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'h', 'e', 'l', 'p', 's', 'u', 'n', 'd', 'e', 'r', 's', 't', 'a', 'n', 'd', 's', 'e', 'n', 't', 'e', 'n', 'c', 'e', 'b', 'o', 'u', 'n', 'd', 'a', 'r', 'i', 'e', 's', ',', 'w', 'o', 'r', 'd', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'i', 's', 'c', 'r', 'u', 'c', 'i', 'a', 'l', 'f', 'o', 'r', 'w', 'o', 'r', 'd', '-', 'l', 'e', 'v', 'e', 'l', 'a', 'n', 'a', 'l', 'y', 's', 'i', 's', ',', 'c', 'h', 'a', 'r', 'a', 'c', 't', 'e', 'r', '-', 'l', 'e', 'v', 'e', 'l', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'i', 's', 'o', 'f', 't', 'e', 'n', 'u', 's', 'e', 'd', 'i', 'n', 'n', 'e', 'u', 'r', 'a', 'l', 'l', 'a', 'n', 'g', 'u', 'a', 'g', 'e', 'm', 'o', 'd', 'e', 'l', 's', ',', 'a', 'n', 'd', 's', 'u', 'b', 'w', 'o', 'r', 'd', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'i', 's', 'h', 'e', 'l', 'p', 'f', 'u', 'l', 'i', 'n', 'd', 'e', 'a', 'l', 'i', 'n', 'g', 'w', 'i', 't', 'h', 'u', 'n', 'k', 'n', 'o', 'w', 'n', 'o', 'r', 'r', 'a', 'r', 'e', 'w', 'o', 'r', 'd', 's', '.', 'S', 't', 'e', 'm', 'm', 'i', 'n', 'g', 'i', 's', 'a', 't', 'e', 'c', 'h', 'n', 'i', 'q', 'u', 'e', 'w', 'h', 'e', 'r', 'e', 'w', 'o', 'r', 'd', 's', 'a', 'r', 'e', 'r', 'e', 'd', 'u', 'c', 'e', 'd', 't', 'o', 't', 'h', 'e', 'i', 'r', 'b', 'a', 's', 'e', 'o', 'r', 'r', 'o', 'o', 't', 'f', 'o', 'r', 'm', '.', 'H', 'o', 'w', 'e', 'v', 'e', 'r', ',', 't', 'h', 'e', 's', 't', 'e', 'm', 'm', 'e', 'd', 'w', 'o', 'r', 'd', 'm', 'a', 'y', 'n', 'o', 't', 'b', 'e', 'a', 'v', 'a', 'l', 'i', 'd', 'E', 'n', 'g', 'l', 'i', 's', 'h', 'w', 'o', 'r', 'd', '.', 'F', 'o', 'r', 'e', 'x', 'a', 'm', 'p', 'l', 'e', ',', "'", 'f', 'i', 's', 'h', 'i', 'n', 'g', "'", ',', "'", 'f', 'i', 's', 'h', 'e', 'd', "'", ',', 'a', 'n', 'd', "'", 'f', 'i', 's', 'h', 'e', 'r', "'", 'm', 'a', 'y', 'a', 'l', 'l', 'b', 'e', 'r', 'e', 'd', 'u', 'c', 'e', 'd', 't', 'o', "'", 'f', 'i', 's', 'h', "'", '.', 'T', 'h', 'e', 'P', 'o', 'r', 't', 'e', 'r', 'S', 't', 'e', 'm', 'm', 'e', 'r', 'i', 's', 'c', 'o', 'm', 'm', 'o', 'n', 'l', 'y', 'u', 's', 'e', 'd', 'i', 'n', 'm', 'a', 'n', 'y', 'N', 'L', 'P', 'a', 'p', 'p', 'l', 'i', 'c', 'a', 't', 'i', 'o', 'n', 's', 'd', 'u', 'e', 't', 'o', 'i', 't', 's', 's', 'p', 'e', 'e', 'd', 'a', 'n', 'd', 's', 'i', 'm', 'p', 'l', 'i', 'c', 'i', 't', 'y', '.', 'L', 'e', 'm', 'm', 'a', 't', 'i', 'z', 'a', 't', 'i', 'o', 'n', ',', 'o', 'n', 't', 'h', 'e', 'o', 't', 'h', 'e', 'r', 'h', 'a', 'n', 'd', ',', 'i', 's', 'a', 'm', 'o', 'r', 'e', 's', 'o', 'p', 'h', 'i', 's', 't', 'i', 'c', 'a', 't', 'e', 'd', 'a', 'p', 'p', 'r', 'o', 'a', 'c', 'h', 't', 'h', 'a', 't', 'u', 's', 'e', 's', 'a', 'v', 'o', 'c', 'a', 'b', 'u', 'l', 'a', 'r', 'y', 'a', 'n', 'd', 'm', 'o', 'r', 'p', 'h', 'o', 'l', 'o', 'g', 'i', 'c', 'a', 'l', 'a', 'n', 'a', 'l', 'y', 's', 'i', 's', 't', 'o', 'r', 'e', 't', 'u', 'r', 'n', 't', 'h', 'e', 'd', 'i', 'c', 't', 'i', 'o', 'n', 'a', 'r', 'y', 'f', 'o', 'r', 'm', 'o', 'f', 'a', 'w', 'o', 'r', 'd', '.', 'F', 'o', 'r', 'i', 'n', 's', 't', 'a', 'n', 'c', 'e', ',', 't', 'h', 'e', 'w', 'o', 'r', 'd', "'", 'b', 'e', 't', 't', 'e', 'r', "'", 'i', 's', 'l', 'e', 'm', 'm', 'a', 't', 'i', 'z', 'e', 'd', 't', 'o', "'", 'g', 'o', 'o', 'd', "'", 'w', 'i', 't', 'h', 'p', 'r', 'o', 'p', 'e', 'r', 'c', 'o', 'n', 't', 'e', 'x', 't', 'a', 'n', 'd', 'p', 'a', 'r', 't', '-', 'o', 'f', '-', 's', 'p', 'e', 'e', 'c', 'h', 't', 'a', 'g', 'g', 'i', 'n', 'g', '.', 'L', 'e', 'm', 'm', 'a', 't', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'i', 's', 'm', 'o', 'r', 'e', 'a', 'c', 'c', 'u', 'r', 'a', 't', 'e', 't', 'h', 'a', 'n', 's', 't', 'e', 'm', 'm', 'i', 'n', 'g', 'b', 'u', 't', 'c', 'o', 'm', 'p', 'u', 't', 'a', 't', 'i', 'o', 'n', 'a', 'l', 'l', 'y', 'h', 'e', 'a', 'v', 'i', 'e', 'r', '.', 'S', 't', 'o', 'p', 'w', 'o', 'r', 'd', 's', 'a', 'r', 'e', 'c', 'o', 'm', 'm', 'o', 'n', 'l', 'y', 'u', 's', 'e', 'd', 'w', 'o', 'r', 'd', 's', 'i', 'n', 'E', 'n', 'g', 'l', 'i', 's', 'h', 's', 'u', 'c', 'h', 'a', 's', "'", 'i', 's', "'", ',', "'", 'i', 'n', "'", ',', "'", 't', 'h', 'e', "'", ',', 'w', 'h', 'i', 'c', 'h', 'a', 'r', 'e', 'o', 'f', 't', 'e', 'n', 'r', 'e', 'm', 'o', 'v', 'e', 'd', 'd', 'u', 'r', 'i', 'n', 'g', 'p', 'r', 'e', 'p', 'r', 'o', 'c', 'e', 's', 's', 'i', 'n', 'g', 'b', 'e', 'c', 'a', 'u', 's', 'e', 't', 'h', 'e', 'y', 'a', 'd', 'd', 'l', 'i', 't', 't', 'l', 'e', 's', 'e', 'm', 'a', 'n', 't', 'i', 'c', 'v', 'a', 'l', 'u', 'e', '.', 'R', 'e', 'm', 'o', 'v', 'i', 'n', 'g', 's', 't', 'o', 'p', 'w', 'o', 'r', 'd', 's', 'h', 'e', 'l', 'p', 's', 'i', 'n', 'r', 'e', 'd', 'u', 'c', 'i', 'n', 'g', 'n', 'o', 'i', 's', 'e', 'a', 'n', 'd', 'i', 'm', 'p', 'r', 'o', 'v', 'i', 'n', 'g', 't', 'h', 'e', 'p', 'e', 'r', 'f', 'o', 'r', 'm', 'a', 'n', 'c', 'e', 'o', 'f', 'N', 'L', 'P', 'm', 'o', 'd', 'e', 'l', 's', '.', 'L', 'e', 't', 'u', 's', 'c', 'o', 'n', 's', 'i', 'd', 'e', 'r', 's', 'o', 'm', 'e', 's', 'a', 'm', 'p', 'l', 'e', 's', 'e', 'n', 't', 'e', 'n', 'c', 'e', 's', 'f', 'o', 'r', 'p', 'r', 'o', 'c', 'e', 's', 's', 'i', 'n', 'g', ':', '1', '.', 'T', 'h', 'e', 'c', 'h', 'i', 'l', 'd', 'r', 'e', 'n', 'w', 'e', 'r', 'e', 'p', 'l', 'a', 'y', 'i', 'n', 'g', 'j', 'o', 'y', 'f', 'u', 'l', 'l', 'y', 'i', 'n', 't', 'h', 'e', 'p', 'l', 'a', 'y', 'g', 'r', 'o', 'u', 'n', 'd', '.', '2', '.', 'S', 'h', 'e', 's', 't', 'u', 'd', 'i', 'e', 's', 'h', 'a', 'r', 'd', 'a', 'n', 'd', 'h', 'a', 's', 's', 't', 'u', 'd', 'i', 'e', 'd', 'w', 'e', 'l', 'l', 'f', 'o', 'r', 'h', 'e', 'r', 'e', 'x', 'a', 'm', 's', '.', '3', '.', 'T', 'h', 'e', 'r', 'u', 'n', 'n', 'i', 'n', 'g', 'w', 'a', 't', 'e', 'r', 'w', 'a', 's', 's', 'o', 'o', 't', 'h', 'i', 'n', 'g', 't', 'o', 'h', 'e', 'a', 'r', '.', '4', '.', 'H', 'e', 'b', 'o', 'u', 'g', 'h', 't', 't', 'h', 'e', 'b', 'e', 's', 't', 'q', 'u', 'a', 'l', 'i', 't', 'y', 'c', 'a', 'm', 'e', 'r', 'a', 'a', 'm', 'o', 'n', 'g', 'a', 'l', 'l', 'a', 'v', 'a', 'i', 'l', 'a', 'b', 'l', 'e', 'm', 'o', 'd', 'e', 'l', 's', '.', 'E', 'a', 'c', 'h', 's', 'e', 'n', 't', 'e', 'n', 'c', 'e', 's', 'h', 'o', 'w', 'c', 'a', 's', 'e', 's', 't', 'h', 'e', 'u', 's', 'e', 'o', 'f', 'v', 'a', 'r', 'i', 'o', 'u', 's', 'v', 'e', 'r', 'b', 's', ',', 'a', 'd', 'v', 'e', 'r', 'b', 's', ',', 'a', 'n', 'd', 'a', 'd', 'j', 'e', 'c', 't', 'i', 'v', 'e', 's', 't', 'h', 'a', 't', 'd', 'e', 'm', 'o', 'n', 's', 't', 'r', 'a', 't', 'e', 't', 'h', 'e', 'n', 'e', 'e', 'd', 'f', 'o', 'r', 'c', 'o', 'n', 't', 'e', 'x', 't', 'u', 'a', 'l', 'a', 'n', 'a', 'l', 'y', 's', 'i', 's', 'i', 'n', 'l', 'e', 'm', 'm', 'a', 't', 'i', 'z', 'a', 't', 'i', 'o', 'n', '.', 'T', 'h', 'i', 's', 'i', 'n', 'p', 'u', 't', 'f', 'i', 'l', 'e', 'i', 's', 's', 't', 'r', 'u', 'c', 't', 'u', 'r', 'e', 'd', 't', 'o', 'p', 'r', 'o', 'v', 'i', 'd', 'e', 's', 'u', 'f', 'f', 'i', 'c', 'i', 'e', 'n', 't', 'd', 'i', 'v', 'e', 'r', 's', 'i', 't', 'y', 'f', 'o', 'r', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'a', 't', 'i', 'o', 'n', ',', 's', 't', 'e', 'm', 'm', 'i', 'n', 'g', ',', 'a', 'n', 'd', 'l', 'e', 'm', 'm', 'a', 't', 'i', 'z', 'a', 't', 'i', 'o', 'n', 'd', 'e', 'm', 'o', 'n', 's', 't', 'r', 'a', 't', 'i', 'o', 'n', 's', 'i', 'n', 'N', 'L', 'P', 'L', 'a', 'b', 'P', 'r', 'o', 'g', 'r', 'a', 'm', '3', '.']

# 4. Subword-Level Tokenization

Naively split long words into two halves:

['Nat', 'ural', 'Lang', 'uage', 'Proce', 'ssing', 'NLP', 'is', 'a', 'cru', 'cial', 'area', 'of', 'Artif', 'icial', 'Intell', 'igence', 'that', 'deals', 'with', 'the', 'inter', 'action', 'bet', 'ween', 'humans', 'and', 'mach', 'ines', 'using', 'nat', 'ural', 'lang', 'uage', 'The', 'field', 'of', 'NLP', 'incl', 'udes', 'sev', 'eral', 'tasks', 'such', 'as', 'lang', 'uage', 'mode', 'ling', 'text', 'classif', 'ication', 'trans', 'lation', 'infor', 'mation', 'retr', 'ieval', 'and', 'more', 'Among', 'the', 'most', 'esse', 'ntial', 'prepro', 'cessing', 'steps', 'in', 'NLP', 'are', 'tokeni', 'zation', 'stem', 'ming', 'and', 'lemmat', 'ization', 'Tokeni', 'zation', 'invo', 'lves', 'brea', 'king', 'down', 'the', 'text', 'into', 'sma', 'ller', 'units', 'called', 'tokens', 'These', 'can', 'be', 'sent', 'ences', 'words', 'chara', 'cters', 'or', 'even', 'subw', 'ords', 'Each', 'type', 'of', 'tokeni', 'zation', 'has', 'its', 'own', 'rele', 'vance', 'For', 'exa', 'mple', 'sent', 'ence', 'tokeni', 'zation', 'helps', 'under', 'stand', 'sent', 'ence', 'bound', 'aries', 'word', 'tokeni', 'zation', 'is', 'cru', 'cial', 'for', 'word', 'level', 'anal', 'ysis', 'char', 'acter', 'level', 'tokeni', 'zation', 'is', 'often', 'used', 'in', 'neural', 'lang', 'uage', 'models', 'and', 'sub', 'word', 'tokeni', 'zation', 'is', 'hel', 'pful', 'in', 'dea', 'ling', 'with', 'unk', 'nown', 'or', 'rare', 'words', 'Stem', 'ming', 'is', 'a', 'tech', 'nique', 'where', 'words', 'are', 'red', 'uced', 'to', 'their', 'base', 'or', 'root', 'form', 'How', 'ever', 'the', 'ste', 'mmed', 'word', 'may', 'not', 'be', 'a', 'valid', 'Eng', 'lish', 'word', 'For', 'exa', 'mple', 'fis', 'hing', 'fished', 'and', 'fisher', 'may', 'all', 'be', 'red', 'uced', 'to', 'fish', 'The', 'Porter', 'Ste', 'mmer', 'is', 'comm', 'only', 'used', 'in', 'many', 'NLP', 'applic', 'ations', 'due', 'to', 'its', 'speed', 'and', 'simpl', 'icity', 'Lemmat', 'ization', 'on', 'the', 'other', 'hand', 'is', 'a', 'more', 'sophis', 'ticated', 'appr', 'oach', 'that', 'uses', 'a', 'vocab', 'ulary', 'and', 'morpho', 'logical', 'anal', 'ysis', 'to', 'return', 'the', 'dicti', 'onary', 'form', 'of', 'a', 'word', 'For', 'inst', 'ance', 'the', 'word', 'better', 'is', 'lemma', 'tized', 'to', 'good', 'with', 'proper', 'con', 'text', 'and', 'part', 'of', 'speech', 'tag', 'ging', 'Lemmat', 'ization', 'is', 'more', 'accu', 'rate', 'than', 'stem', 'ming', 'but', 'computa', 'tionally', 'hea', 'vier', 'Stop', 'words', 'are', 'comm', 'only', 'used', 'words', 'in', 'Eng', 'lish', 'such', 'as', 'is', 'in', 'the', 'which', 'are', 'often', 'rem', 'oved', 'during', 'prepro', 'cessing', 'bec', 'ause', 'they', 'add', 'little', 'sema', 'ntic', 'value', 'Remo', 'ving', 'stop', 'words', 'helps', 'in', 'redu', 'cing', 'noise', 'and', 'impr', 'oving', 'the', 'perfo', 'rmance', 'of', 'NLP', 'models', 'Let', 'us', 'cons', 'ider', 'some', 'sample', 'sent', 'ences', 'for', 'proce', 'ssing', '1', 'The', 'chil', 'dren', 'were', 'pla', 'ying', 'joyf', 'ully', 'in', 'the', 'playg', 'round', '2', 'She', 'stu', 'dies', 'hard', 'and', 'has', 'stu', 'died', 'well', 'for', 'her', 'exams', '3', 'The', 'run', 'ning', 'water', 'was', 'soot', 'hing', 'to', 'hear', '4', 'He', 'bought', 'the', 'best', 'qua', 'lity', 'camera', 'among', 'all', 'avai', 'lable', 'models', 'Each', 'sent', 'ence', 'show', 'cases', 'the', 'use', 'of', 'var', 'ious', 'verbs', 'adv', 'erbs', 'and', 'adjec', 'tives', 'that', 'demon', 'strate', 'the', 'need', 'for', 'conte', 'xtual', 'anal', 'ysis', 'in', 'lemmat', 'ization', 'This', 'input', 'file', 'is', 'struc', 'tured', 'to', 'pro', 'vide', 'suffi', 'cient', 'dive', 'rsity', 'for', 'tokeni', 'zation', 'stem', 'ming', 'and', 'lemmat', 'ization', 'demonst', 'rations', 'in', 'NLP', 'Lab', 'Pro', 'gram', '3']

# 5. Stemming using Porter Stemmer

Natural ➝ natur

Language ➝ languag

Processing ➝ process

NLP ➝ nlp

is ➝ is

a ➝ a

crucial ➝ crucial

area ➝ area

of ➝ of

Artificial ➝ artifici

Intelligence ➝ intellig

that ➝ that

deals ➝ deal

with ➝ with

the ➝ the

interaction ➝ interact

between ➝ between

humans ➝ human

and ➝ and

machines ➝ machin

using ➝ use

natural ➝ natur

language ➝ languag

The ➝ the

field ➝ field

of ➝ of

NLP ➝ nlp

includes ➝ includ

several ➝ sever

tasks ➝ task

such ➝ such

as ➝ as

language ➝ languag

modeling ➝ model

text ➝ text

classification ➝ classif

translation ➝ translat

information ➝ inform

retrieval ➝ retriev

and ➝ and

more ➝ more

Among ➝ among

the ➝ the

most ➝ most

essential ➝ essenti

preprocessing ➝ preprocess

steps ➝ step

in ➝ in

NLP ➝ nlp

are ➝ are

tokenization ➝ token

stemming ➝ stem

and ➝ and

lemmatization ➝ lemmat

Tokenization ➝ token

involves ➝ involv

breaking ➝ break

down ➝ down

the ➝ the

text ➝ text

into ➝ into

smaller ➝ smaller

units ➝ unit

called ➝ call

tokens ➝ token

These ➝ these

can ➝ can

be ➝ be

sentences ➝ sentenc

words ➝ word

characters ➝ charact

or ➝ or

even ➝ even

subwords ➝ subword

Each ➝ each

type ➝ type

of ➝ of

tokenization ➝ token

has ➝ ha

its ➝ it

own ➝ own

relevance ➝ relev

For ➝ for

example ➝ exampl

sentence ➝ sentenc

tokenization ➝ token

helps ➝ help

understand ➝ understand

sentence ➝ sentenc

boundaries ➝ boundari

word ➝ word

tokenization ➝ token

is ➝ is

crucial ➝ crucial

for ➝ for

word ➝ word

level ➝ level

analysis ➝ analysi

character ➝ charact

level ➝ level

tokenization ➝ token

is ➝ is

often ➝ often

used ➝ use

in ➝ in

neural ➝ neural

language ➝ languag

models ➝ model

and ➝ and

subword ➝ subword

tokenization ➝ token

is ➝ is

helpful ➝ help

in ➝ in

dealing ➝ deal

with ➝ with

unknown ➝ unknown

or ➝ or

rare ➝ rare

words ➝ word

Stemming ➝ stem

is ➝ is

a ➝ a

technique ➝ techniqu

where ➝ where

words ➝ word

are ➝ are

reduced ➝ reduc

to ➝ to

their ➝ their

base ➝ base

or ➝ or

root ➝ root

form ➝ form

However ➝ howev

the ➝ the

stemmed ➝ stem

word ➝ word

may ➝ may

not ➝ not

be ➝ be

a ➝ a

valid ➝ valid

English ➝ english

word ➝ word

For ➝ for

example ➝ exampl

fishing ➝ fish

fished ➝ fish

and ➝ and

fisher ➝ fisher

may ➝ may

all ➝ all

be ➝ be

reduced ➝ reduc

to ➝ to

fish ➝ fish

The ➝ the

Porter ➝ porter

Stemmer ➝ stemmer

is ➝ is

commonly ➝ commonli

used ➝ use

in ➝ in

many ➝ mani

NLP ➝ nlp

applications ➝ applic

due ➝ due

to ➝ to

its ➝ it

speed ➝ speed

and ➝ and

simplicity ➝ simplic

Lemmatization ➝ lemmat

on ➝ on

the ➝ the

other ➝ other

hand ➝ hand

is ➝ is

a ➝ a

more ➝ more

sophisticated ➝ sophist

approach ➝ approach

that ➝ that

uses ➝ use

a ➝ a

vocabulary ➝ vocabulari

and ➝ and

morphological ➝ morpholog

analysis ➝ analysi

to ➝ to

return ➝ return

the ➝ the

dictionary ➝ dictionari

form ➝ form

of ➝ of

a ➝ a

word ➝ word

For ➝ for

instance ➝ instanc

the ➝ the

word ➝ word

better ➝ better

is ➝ is

lemmatized ➝ lemmat

to ➝ to

good ➝ good

with ➝ with

proper ➝ proper

context ➝ context

and ➝ and

part ➝ part

of ➝ of

speech ➝ speech

tagging ➝ tag

Lemmatization ➝ lemmat

is ➝ is

more ➝ more

accurate ➝ accur

than ➝ than

stemming ➝ stem

but ➝ but

computationally ➝ comput

heavier ➝ heavier

Stopwords ➝ stopword

are ➝ are

commonly ➝ commonli

used ➝ use

words ➝ word

in ➝ in

English ➝ english

such ➝ such

as ➝ as

is ➝ is

in ➝ in

the ➝ the

which ➝ which

are ➝ are

often ➝ often

removed ➝ remov

during ➝ dure

preprocessing ➝ preprocess

because ➝ becaus

they ➝ they

add ➝ add

little ➝ littl

semantic ➝ semant

value ➝ valu

Removing ➝ remov

stopwords ➝ stopword

helps ➝ help

in ➝ in

reducing ➝ reduc

noise ➝ nois

and ➝ and

improving ➝ improv

the ➝ the

performance ➝ perform

of ➝ of

NLP ➝ nlp

models ➝ model

Let ➝ let

us ➝ us

consider ➝ consid

some ➝ some

sample ➝ sampl

sentences ➝ sentenc

for ➝ for

processing ➝ process

The ➝ the

children ➝ children

were ➝ were

playing ➝ play

joyfully ➝ joy

in ➝ in

the ➝ the

playground ➝ playground

She ➝ she

studies ➝ studi

hard ➝ hard

and ➝ and

has ➝ ha

studied ➝ studi

well ➝ well

for ➝ for

her ➝ her

exams ➝ exam

The ➝ the

running ➝ run

water ➝ water

was ➝ wa

soothing ➝ sooth

to ➝ to

hear ➝ hear

He ➝ he

bought ➝ bought

the ➝ the

best ➝ best

quality ➝ qualiti

camera ➝ camera

among ➝ among

all ➝ all

available ➝ avail

models ➝ model

Each ➝ each

sentence ➝ sentenc

showcases ➝ showcas

the ➝ the

use ➝ use

of ➝ of

various ➝ variou

verbs ➝ verb

adverbs ➝ adverb

and ➝ and

adjectives ➝ adject

that ➝ that

demonstrate ➝ demonstr

the ➝ the

need ➝ need

for ➝ for

contextual ➝ contextu

analysis ➝ analysi

in ➝ in

lemmatization ➝ lemmat

This ➝ thi

input ➝ input

file ➝ file

is ➝ is

structured ➝ structur

to ➝ to

provide ➝ provid

sufficient ➝ suffici

diversity ➝ divers

for ➝ for

tokenization ➝ token

stemming ➝ stem

and ➝ and

lemmatization ➝ lemmat

demonstrations ➝ demonstr

in ➝ in

NLP ➝ nlp

Lab ➝ lab

Program ➝ program

# 6. Lemmatization (Basic - noun-only)

Natural ➝ Natural

Language ➝ Language

Processing ➝ Processing

NLP ➝ NLP

is ➝ is

a ➝ a

crucial ➝ crucial

area ➝ area

of ➝ of

Artificial ➝ Artificial

Intelligence ➝ Intelligence

that ➝ that

deals ➝ deal

with ➝ with

the ➝ the

interaction ➝ interaction

between ➝ between

humans ➝ human

and ➝ and

machines ➝ machine

using ➝ using

natural ➝ natural

language ➝ language

The ➝ The

field ➝ field

of ➝ of

NLP ➝ NLP

includes ➝ includes

several ➝ several

tasks ➝ task

such ➝ such

as ➝ a

language ➝ language

modeling ➝ modeling

text ➝ text

classification ➝ classification

translation ➝ translation

information ➝ information

retrieval ➝ retrieval

and ➝ and

more ➝ more

Among ➝ Among

the ➝ the

most ➝ most

essential ➝ essential

preprocessing ➝ preprocessing

steps ➝ step

in ➝ in

NLP ➝ NLP

are ➝ are

tokenization ➝ tokenization

stemming ➝ stemming

and ➝ and

lemmatization ➝ lemmatization

Tokenization ➝ Tokenization

involves ➝ involves

breaking ➝ breaking

down ➝ down

the ➝ the

text ➝ text

into ➝ into

smaller ➝ smaller

units ➝ unit

called ➝ called

tokens ➝ token

These ➝ These

can ➝ can

be ➝ be

sentences ➝ sentence

words ➝ word

characters ➝ character

or ➝ or

even ➝ even

subwords ➝ subwords

Each ➝ Each

type ➝ type

of ➝ of

tokenization ➝ tokenization

has ➝ ha

its ➝ it

own ➝ own

relevance ➝ relevance

For ➝ For

example ➝ example

sentence ➝ sentence

tokenization ➝ tokenization

helps ➝ help

understand ➝ understand

sentence ➝ sentence

boundaries ➝ boundary

word ➝ word

tokenization ➝ tokenization

is ➝ is

crucial ➝ crucial

for ➝ for

word ➝ word

level ➝ level

analysis ➝ analysis

character ➝ character

level ➝ level

tokenization ➝ tokenization

is ➝ is

often ➝ often

used ➝ used

in ➝ in

neural ➝ neural

language ➝ language

models ➝ model

and ➝ and

subword ➝ subword

tokenization ➝ tokenization

is ➝ is

helpful ➝ helpful

in ➝ in

dealing ➝ dealing

with ➝ with

unknown ➝ unknown

or ➝ or

rare ➝ rare

words ➝ word

Stemming ➝ Stemming

is ➝ is

a ➝ a

technique ➝ technique

where ➝ where

words ➝ word

are ➝ are

reduced ➝ reduced

to ➝ to

their ➝ their

base ➝ base

or ➝ or

root ➝ root

form ➝ form

However ➝ However

the ➝ the

stemmed ➝ stemmed

word ➝ word

may ➝ may

not ➝ not

be ➝ be

a ➝ a

valid ➝ valid

English ➝ English

word ➝ word

For ➝ For

example ➝ example

fishing ➝ fishing

fished ➝ fished

and ➝ and

fisher ➝ fisher

may ➝ may

all ➝ all

be ➝ be

reduced ➝ reduced

to ➝ to

fish ➝ fish

The ➝ The

Porter ➝ Porter

Stemmer ➝ Stemmer

is ➝ is

commonly ➝ commonly

used ➝ used

in ➝ in

many ➝ many

NLP ➝ NLP

applications ➝ application

due ➝ due

to ➝ to

its ➝ it

speed ➝ speed

and ➝ and

simplicity ➝ simplicity

Lemmatization ➝ Lemmatization

on ➝ on

the ➝ the

other ➝ other

hand ➝ hand

is ➝ is

a ➝ a

more ➝ more

sophisticated ➝ sophisticated

approach ➝ approach

that ➝ that

uses ➝ us

a ➝ a

vocabulary ➝ vocabulary

and ➝ and

morphological ➝ morphological

analysis ➝ analysis

to ➝ to

return ➝ return

the ➝ the

dictionary ➝ dictionary

form ➝ form

of ➝ of

a ➝ a

word ➝ word

For ➝ For

instance ➝ instance

the ➝ the

word ➝ word

better ➝ better

is ➝ is

lemmatized ➝ lemmatized

to ➝ to

good ➝ good

with ➝ with

proper ➝ proper

context ➝ context

and ➝ and

part ➝ part

of ➝ of

speech ➝ speech

tagging ➝ tagging

Lemmatization ➝ Lemmatization

is ➝ is

more ➝ more

accurate ➝ accurate

than ➝ than

stemming ➝ stemming

but ➝ but

computationally ➝ computationally

heavier ➝ heavier

Stopwords ➝ Stopwords

are ➝ are

commonly ➝ commonly

used ➝ used

words ➝ word

in ➝ in

English ➝ English

such ➝ such

as ➝ a

is ➝ is

in ➝ in

the ➝ the

which ➝ which

are ➝ are

often ➝ often

removed ➝ removed

during ➝ during

preprocessing ➝ preprocessing

because ➝ because

they ➝ they

add ➝ add

little ➝ little

semantic ➝ semantic

value ➝ value

Removing ➝ Removing

stopwords ➝ stopwords

helps ➝ help

in ➝ in

reducing ➝ reducing

noise ➝ noise

and ➝ and

improving ➝ improving

the ➝ the

performance ➝ performance

of ➝ of

NLP ➝ NLP

models ➝ model

Let ➝ Let

us ➝ u

consider ➝ consider

some ➝ some

sample ➝ sample

sentences ➝ sentence

for ➝ for

processing ➝ processing

The ➝ The

children ➝ child

were ➝ were

playing ➝ playing

joyfully ➝ joyfully

in ➝ in

the ➝ the

playground ➝ playground

She ➝ She

studies ➝ study

hard ➝ hard

and ➝ and

has ➝ ha

studied ➝ studied

well ➝ well

for ➝ for

her ➝ her

exams ➝ exam

The ➝ The

running ➝ running

water ➝ water

was ➝ wa

soothing ➝ soothing

to ➝ to

hear ➝ hear

He ➝ He

bought ➝ bought

the ➝ the

best ➝ best

quality ➝ quality

camera ➝ camera

among ➝ among

all ➝ all

available ➝ available

models ➝ model

Each ➝ Each

sentence ➝ sentence

showcases ➝ showcase

the ➝ the

use ➝ use

of ➝ of

various ➝ various

verbs ➝ verb

adverbs ➝ adverb

and ➝ and

adjectives ➝ adjective

that ➝ that

demonstrate ➝ demonstrate

the ➝ the

need ➝ need

for ➝ for

contextual ➝ contextual

analysis ➝ analysis

in ➝ in

lemmatization ➝ lemmatization

This ➝ This

input ➝ input

file ➝ file

is ➝ is

structured ➝ structured

to ➝ to

provide ➝ provide

sufficient ➝ sufficient

diversity ➝ diversity

for ➝ for

tokenization ➝ tokenization

stemming ➝ stemming

and ➝ and

lemmatization ➝ lemmatization

demonstrations ➝ demonstration

in ➝ in

NLP ➝ NLP

Lab ➝ Lab

Program ➝ Program

# 7. Lemmatization (Advanced - POS-aware)

Natural ➝ Natural

Language ➝ Language

Processing ➝ Processing

NLP ➝ NLP

is ➝ be

a ➝ a

crucial ➝ crucial

area ➝ area

of ➝ of

Artificial ➝ Artificial

Intelligence ➝ Intelligence

that ➝ that

deals ➝ deal

with ➝ with

the ➝ the

interaction ➝ interaction

between ➝ between

humans ➝ human

and ➝ and

machines ➝ machine

using ➝ use

natural ➝ natural

language ➝ language

The ➝ The

field ➝ field

of ➝ of

NLP ➝ NLP

includes ➝ include

several ➝ several

tasks ➝ task

such ➝ such

as ➝ a

language ➝ language

modeling ➝ model

text ➝ text

classification ➝ classification

translation ➝ translation

information ➝ information

retrieval ➝ retrieval

and ➝ and

more ➝ more

Among ➝ Among

the ➝ the

most ➝ most

essential ➝ essential

preprocessing ➝ preprocessing

steps ➝ step

in ➝ in

NLP ➝ NLP

are ➝ be

tokenization ➝ tokenization

stemming ➝ stem

and ➝ and

lemmatization ➝ lemmatization

Tokenization ➝ Tokenization

involves ➝ involve

breaking ➝ break

down ➝ down

the ➝ the

text ➝ text

into ➝ into

smaller ➝ small

units ➝ unit

called ➝ call

tokens ➝ token

These ➝ These

can ➝ can

be ➝ be

sentences ➝ sentence

words ➝ word

characters ➝ character

or ➝ or

even ➝ even

subwords ➝ subwords

Each ➝ Each

type ➝ type

of ➝ of

tokenization ➝ tokenization

has ➝ have

its ➝ it

own ➝ own

relevance ➝ relevance

For ➝ For

example ➝ example

sentence ➝ sentence

tokenization ➝ tokenization

helps ➝ help

understand ➝ understand

sentence ➝ sentence

boundaries ➝ boundary

word ➝ word

tokenization ➝ tokenization

is ➝ be

crucial ➝ crucial

for ➝ for

word ➝ word

level ➝ level

analysis ➝ analysis

character ➝ character

level ➝ level

tokenization ➝ tokenization

is ➝ be

often ➝ often

used ➝ use

in ➝ in

neural ➝ neural

language ➝ language

models ➝ model

and ➝ and

subword ➝ subword

tokenization ➝ tokenization

is ➝ be

helpful ➝ helpful

in ➝ in

dealing ➝ deal

with ➝ with

unknown ➝ unknown

or ➝ or

rare ➝ rare

words ➝ word

Stemming ➝ Stemming

is ➝ be

a ➝ a

technique ➝ technique

where ➝ where

words ➝ word

are ➝ be

reduced ➝ reduce

to ➝ to

their ➝ their

base ➝ base

or ➝ or

root ➝ root

form ➝ form

However ➝ However

the ➝ the

stemmed ➝ stemmed

word ➝ word

may ➝ may

not ➝ not

be ➝ be

a ➝ a

valid ➝ valid

English ➝ English

word ➝ word

For ➝ For

example ➝ example

fishing ➝ fishing

fished ➝ fish

and ➝ and

fisher ➝ fisher

may ➝ may

all ➝ all

be ➝ be

reduced ➝ reduce

to ➝ to

fish ➝ fish

The ➝ The

Porter ➝ Porter

Stemmer ➝ Stemmer

is ➝ be

commonly ➝ commonly

used ➝ use

in ➝ in

many ➝ many

NLP ➝ NLP

applications ➝ application

due ➝ due

to ➝ to

its ➝ it

speed ➝ speed

and ➝ and

simplicity ➝ simplicity

Lemmatization ➝ Lemmatization

on ➝ on

the ➝ the

other ➝ other

hand ➝ hand

is ➝ be

a ➝ a

more ➝ more

sophisticated ➝ sophisticated

approach ➝ approach

that ➝ that

uses ➝ use

a ➝ a

vocabulary ➝ vocabulary

and ➝ and

morphological ➝ morphological

analysis ➝ analysis

to ➝ to

return ➝ return

the ➝ the

dictionary ➝ dictionary

form ➝ form

of ➝ of

a ➝ a

word ➝ word

For ➝ For

instance ➝ instance

the ➝ the

word ➝ word

better ➝ better

is ➝ be

lemmatized ➝ lemmatized

to ➝ to

good ➝ good

with ➝ with

proper ➝ proper

context ➝ context

and ➝ and

part ➝ part

of ➝ of

speech ➝ speech

tagging ➝ tag

Lemmatization ➝ Lemmatization

is ➝ be

more ➝ more

accurate ➝ accurate

than ➝ than

stemming ➝ stem

but ➝ but

computationally ➝ computationally

heavier ➝ heavy

Stopwords ➝ Stopwords

are ➝ be

commonly ➝ commonly

used ➝ use

words ➝ word

in ➝ in

English ➝ English

such ➝ such

as ➝ a

is ➝ be

in ➝ in

the ➝ the

which ➝ which

are ➝ be

often ➝ often

removed ➝ remove

during ➝ during

preprocessing ➝ preprocessing

because ➝ because

they ➝ they

add ➝ add

little ➝ little

semantic ➝ semantic

value ➝ value

Removing ➝ Removing

stopwords ➝ stopwords

helps ➝ help

in ➝ in

reducing ➝ reduce

noise ➝ noise

and ➝ and

improving ➝ improve

the ➝ the

performance ➝ performance

of ➝ of

NLP ➝ NLP

models ➝ model

Let ➝ Let

us ➝ u

consider ➝ consider

some ➝ some

sample ➝ sample

sentences ➝ sentence

for ➝ for

processing ➝ process

The ➝ The

children ➝ child

were ➝ be

playing ➝ play

joyfully ➝ joyfully

in ➝ in

the ➝ the

playground ➝ playground

She ➝ She

studies ➝ study

hard ➝ hard

and ➝ and

has ➝ have

studied ➝ study

well ➝ well

for ➝ for

her ➝ her

exams ➝ exam

The ➝ The

running ➝ running

water ➝ water

was ➝ be

soothing ➝ soothe

to ➝ to

hear ➝ hear

He ➝ He

bought ➝ buy

the ➝ the

best ➝ best

quality ➝ quality

camera ➝ camera

among ➝ among

all ➝ all

available ➝ available

models ➝ model

Each ➝ Each

sentence ➝ sentence

showcases ➝ showcases

the ➝ the

use ➝ use

of ➝ of

various ➝ various

verbs ➝ verb

adverbs ➝ adverb

and ➝ and

adjectives ➝ adjective

that ➝ that

demonstrate ➝ demonstrate

the ➝ the

need ➝ need

for ➝ for

contextual ➝ contextual

analysis ➝ analysis

in ➝ in

lemmatization ➝ lemmatization

This ➝ This

input ➝ input

file ➝ file

is ➝ be

structured ➝ structure

to ➝ to

provide ➝ provide

sufficient ➝ sufficient

diversity ➝ diversity

for ➝ for

tokenization ➝ tokenization

stemming ➝ stemming

and ➝ and

lemmatization ➝ lemmatization

demonstrations ➝ demonstration

in ➝ in

NLP ➝ NLP

Lab ➝ Lab

Program ➝ Program

# 8. After Removing Stopwords

['Natural', 'Language', 'Processing', 'NLP', 'crucial', 'area', 'Artificial', 'Intelligence', 'deals', 'interaction', 'humans', 'machines', 'using', 'natural', 'language', 'field', 'NLP', 'includes', 'several', 'tasks', 'language', 'modeling', 'text', 'classification', 'translation', 'information', 'retrieval', 'Among', 'essential', 'preprocessing', 'steps', 'NLP', 'tokenization', 'stemming', 'lemmatization', 'Tokenization', 'involves', 'breaking', 'text', 'smaller', 'units', 'called', 'tokens', 'sentences', 'words', 'characters', 'even', 'subwords', 'type', 'tokenization', 'relevance', 'example', 'sentence', 'tokenization', 'helps', 'understand', 'sentence', 'boundaries', 'word', 'tokenization', 'crucial', 'word', 'level', 'analysis', 'character', 'level', 'tokenization', 'often', 'used', 'neural', 'language', 'models', 'subword', 'tokenization', 'helpful', 'dealing', 'unknown', 'rare', 'words', 'Stemming', 'technique', 'words', 'reduced', 'base', 'root', 'form', 'However', 'stemmed', 'word', 'may', 'valid', 'English', 'word', 'example', 'fishing', 'fished', 'fisher', 'may', 'reduced', 'fish', 'Porter', 'Stemmer', 'commonly', 'used', 'many', 'NLP', 'applications', 'due', 'speed', 'simplicity', 'Lemmatization', 'hand', 'sophisticated', 'approach', 'uses', 'vocabulary', 'morphological', 'analysis', 'return', 'dictionary', 'form', 'word', 'instance', 'word', 'better', 'lemmatized', 'good', 'proper', 'context', 'part', 'speech', 'tagging', 'Lemmatization', 'accurate', 'stemming', 'computationally', 'heavier', 'Stopwords', 'commonly', 'used', 'words', 'English', 'often', 'removed', 'preprocessing', 'add', 'little', 'semantic', 'value', 'Removing', 'stopwords', 'helps', 'reducing', 'noise', 'improving', 'performance', 'NLP', 'models', 'Let', 'us', 'consider', 'sample', 'sentences', 'processing', 'children', 'playing', 'joyfully', 'playground', 'studies', 'hard', 'studied', 'well', 'exams', 'running', 'water', 'soothing', 'hear', 'bought', 'best', 'quality', 'camera', 'among', 'available', 'models', 'sentence', 'showcases', 'use', 'various', 'verbs', 'adverbs', 'adjectives', 'demonstrate', 'need', 'contextual', 'analysis', 'lemmatization', 'input', 'file', 'structured', 'provide', 'sufficient', 'diversity', 'tokenization', 'stemming', 'lemmatization', 'demonstrations', 'NLP', 'Lab', 'Program']

# 9. Observations

Stemming cuts suffixes and often results in non-English root forms.  
Simple lemmatization always assumes noun POS which is not always accurate.  
Advanced lemmatization considers POS tags and returns more meaningful lemmas.  
This output shows how important contextual understanding is in text preprocessing.