CSSE 315 – Natural Language Processing
Rose-Hulman Institute of Technology

Worksheet 02

Name (Print):_____ Date:_____

1. Warm-up. How would you split this sentence into words (tokens)? Could be more than just one way...

   I'm moving to Winston-Salem.

2. Definitions

Term	Definition
	The minimal unit that the machine can process
	Splitting the raw text into smaller chunks
Stop words	Examples:

3. Provide at least 2 reasons for tokenization:

   1.
   2.

4. Complete the following:

```
# Convert to lowercase
sample_text.____________
# Split text by comma
sample_text.____________
# Join words into a string using a whitespace
" ".____________(sample_text)
# Replace # by empty string
sample_text.________________
# Combine lowercase and strip in the same line
sample_text.______________________________
```

5. Stopwords

```
from sklearn.feature_extraction.text import ____________________

```

6. Tokenizers

Term	Definition
	Splits a sentence into individual words
	Breaks the text into individual characters
	Splits a word into subwords

7. How would you split “unwanted”, “football” using subword tokenizer?

   unwanted football

8. Test BPE tokenizer with your own words, and provide the output. Share any interesting/-surprising observations

Word Examples	BPE Tokenizer Output

9. Which tokenizer is most commonly used?

   - Word-level
   - Character-level
   - Subword-level

10. Provide the names of at least one model for each tokenizer

Tokenizer	Model
Word-Level	
Character-Level	
BPE/Subword	