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

Worksheet 03

Name (Print):________________________ Date:____________

1. Warm-up. Play a game: https://roft.io/

   Are you able to identify real vs fake?

2. What is the typical Text preprocessing pipeline

   1. 2. 3. 4.

3. Lemmatization vs Stemming

Term	Definition
	A rule-based technique to reduce a word to a stem
	Reduces the word to its canonical dictionary form

4. Read about NLP/NLG/NLU (see link on the course schedule)

Term	Definition	Application
NLP		
NLG		
NLU		

5. Provide definitions:

Term	Definition
Feature	
Feature Engineering	
Vectorization	

6. Count Frequency (frequency of occurrence)

   it was the best of times it was the worst of times

   - Write the dictionary (list of unique words).
   - Provide a Count Frequency for each unique token

7. What is the main difference between Count Frequency and Term Frequency?

8. N-Grams. Write 1 example for unigran, bigram, trigram

   Tim Cook is the CEO of Apple

   -
   -
   -

9. Write the formula for TF-IDF

10. Calculate TF-IDF for for Example (see https://obscrivn.github.io/mynewbook/features.html#tf-idf