Web Intelligence Course

Lecture 2: NLP-based Knowledge Extraction
Exercises

1. Perform all the preprocessing steps for data tokenisation, normalisation, and stopword removal in the following text. Further determine the type/token ratio, and rank the words based on their frequencies

   "Hi there! Did you hear about the new café that just opened downtown?" asked Sarah.
   "No, I haven't. What's so special about it?" replied John, looking curious.
   "It's supposed to have the best espresso in town—people are raving about it!" Sarah responded enthusiastically.
   "Really? I'm always on the lookout for good coffee. When should we check it out?" John asked.
   "How about this Saturday? I heard they have a live jazz band in the evening," Sarah suggested.
   "That sounds fantastic! I'm in. Let's meet there at 7 PM, then," John agreed.
   "Perfect! I'll text you the address later. By the way, did you see the new movie trailer for 'The Last Horizon'? It looks amazing, don't you think?" Sarah inquired.
   "Oh, I did! It looks intense. I can't wait to see it," John said with excitement.

2. Stem the following words using Porter Stemmer
   1. *Excitement*
   2. *Misunderstanding*
   3. *Transcontinental*
   4. *Substitutional*

3. Determine the Levenshtein distance for the following:

   "Anthropomorphization" and "Anthropomorphism"
   "Overcompensation" and "Overcompensate"
   "Counterproductive" and "Counterproductivity"

4. Determine the Hamming distances for the following:

   M4ch1n3L3arn1ng and M4ch1n3L3arN1ng
   GATTACA and GATCTAC