**University of Central Missouri**

**Department of Computer Science & Cybersecurity**

**CS5760 Natural Language Processing**

**Fall 2025**

**Homework 1.**

**Student name:**

**Submission Requirements:**

* Once finished your assignment push your source code to your repo (GitHub) and explain the work through the ReadMe file properly. Make sure you add your student info in the ReadMe file.
* Submit your GitHub link on the Bright Space.
* Comment your code appropriately ***IMPORTANT.***
* Any submission after provided deadline is considered as a late submission.

**Q1. Regex**

**Task:** Write a regex to find

1. **U.S. ZIP codes (disjunction + token boundaries)**  
   Match 12345 **or** 12345-6789 **or** 12345 6789 (hyphen **or** space allowed for the +4 part). Make sure you only match whole tokens (not inside longer strings).
2. **Negation in disjunction (word start rules)**  
   Find all **words** that **do not** start with a capital letter. Words may include internal apostrophes/hyphens like don’t, state-of-the-art.
3. **Convenient aliases (numbers, a bit richer)**  
   Extract all numbers that may have:
   1. optional sign (+/-),
   2. optional thousands separators (commas),
   3. optional decimal part,
   4. optional scientific notation (e.g., 1.23e-4).
4. **More disjunction (spelling variants)**  
   Match any spelling of “email”: email, e-mail, or e mail. Accept either a **space** or a **hyphen** (including en‐dash –) between e and mail, and be case-insensitive.
5. **Wildcards, optionality, repetition (with punctuation)**  
   Match the interjection go, goo, gooo, … (one or more o), **as a word**, and allow an optional trailing punctuation mark ! . , ? (e.g., gooo!).
6. **Anchors (line/sentence end with quotes)**  
   Match lines that **end with a question mark** possibly followed only by closing quotes/brackets like ")”’] and spaces.

**Q2. Programming Question:**

**1. Tokenize a paragraph**  
Take a short paragraph (3–4 sentences) in **your language** (e.g., from news, a story, or social media).

* Do **naïve space-based tokenization**.
* Manually correct the tokens by handling punctuation, suffixes, and clitics.  
  Submit both versions and highlight differences.

**2. Compare with a Tool**  
Run the paragraph through an NLP tool that supports your language (e.g., **NLTK, spaCy**, or any open-source tokenizer if available).

* Compare tool output vs. your manual tokens.
* Which tokens differ? Why?

**3. Multiword Expressions (MWEs)**  
Identify at least **3 multiword expressions (MWEs)** in your language. Example:

* Place names, idioms, or common fixed phrases.
* Explain why they should be treated as single tokens.

**4. Reflection (5–6 sentences)**

* What was the hardest part of tokenization in your language?
* How does it compare with tokenization in English?
* Do punctuation, morphology, and MWEs make tokenization more difficult?

**Q3. Manual BPE on a toy corpus**

3.1 Using the same corpus from class:

low low low low low lowest lowest newer newer newer newer newer newer wider wider wider new new

1. Add the end-of-word marker \_ and write the *initial vocabulary* (characters + \_).
2. Compute bigram counts and perform the **first three merges** by hand:
   * Step 1: most frequent pair → merge → updated corpus snippet (show at least 2 lines).
   * Step 2: repeat.
   * Step 3: repeat.
3. After each merge, list the new token and the updated vocabulary.

3.2 — Code a mini-BPE learner

1. Use the classroom code above (or your own) to learn BPE merges for the toy corpus.
   * Print the top pair at each step and the evolving vocabulary size.
2. Segment the words: new, newer, lowest, widest, and one word you invent (e.g., newestest).
   * Include the subword sequence produced (tokens with \_ where applicable).
3. In 5–6 sentences, explain:
   * How subword tokens solved the OOV (out-of-vocabulary) problem.
   * One example where subwords align with a meaningful morpheme (e.g., er\_ as English agent/comparative suffix).

3.3 — *Your language* (or English if you prefer)

Pick one short paragraph (4–6 sentences) in *your own language* (or English if that’s simpler).

1. Train BPE on that paragraph (or a small file of your choice).
   * Use end-of-word \_.
   * Learn at least 30 merges (adjust if the text is very small).
2. Show the five most frequent merges and the resulting five longest subword tokens.
3. Segment 5 different words from the paragraph:
   * Include one rare word and one derived/inflected form.
4. Brief reflection (5–8 sentences):
   * What kinds of subwords were learned (prefixes, suffixes, stems, whole words)?
   * Two concrete pros/cons of subword tokenization for your language.

**Q4. Word Pair:**

Sunday → Saturday

Tasks:

1. Find the minimum edit distance between *Sunday* and *Saturday* under both models:
   * Model A (Sub = 1, Ins = 1, Del = 1)
   * Model B (Sub = 2, Ins = 1, Del = 1)
2. Write out at least one valid edit sequence (step by step).
3. Reflect (4–5 sentences):
   * Did both models give the same distance?
   * Which operations (insert/delete/substitute) were most useful here?
   * How would the choice of model affect applications like spell check vs. DNA alignment?