# Name : Rajashekar Mudigonda, A20556922.
# CS585 WA-1

## Objectives:
1. (5 points) Demonstrate your understanding of tokenization and lemmatization.
2. (15 points) Demonstrate your understanding of the Byte-Pair Encoding algorithm.

## Problem 1
Your task is to manually tokenize and lemmatize selected English sentences (if you are using a specific resource, for example a specific dictionary, mention it and / or provide a link).

A) **[1 pt]** Tokenize and lemmatize the following English sentence:

*I really like going out to theaters.*

| **Tokenized and lemmatized sentence (use commas to separate individual tokens):** |
|---|
| Tokenized : I, really, like, going, out, to, theatres, .<br>Lemmatized: I really like go out to theater<br>Sources:<br>http://text-processing.com/demo/tokenize/<br>https://cst.dk/online/lemmatiser/uk/<br>https://seohorsesense.com/free/lemmatization.php |

B) **[1 pt]** Tokenize and lemmatize the following English sentence:

*His birthday was supposed to be a surprise!*

| **Tokenized and lemmatized sentence (use commas to separate individual tokens):** |
|---|
| Tokenized : His, birthday, was, supposed, to, be, a surprise, .<br>Lemmatized : His birthday be suppose to be a surprise |

C) **[1 pt]** Tokenize and lemmatize the following English sentence:

*This was a fantastic day to be alive.*

| **Tokenized and lemmatized sentence (use commas to separate individual tokens):** |
|---|
| Tokenized : This, was, a, fantastic, day, to, be, alive, .<br><br>Lemmatized: This be a fantastic day to be alive. |

D) **[1 pt]** Tokenize and lemmatize the following English sentence:

*What was your favorite movie among all movie?*

| **Tokenized and lemmatized sentence (use commas to separate individual tokens):** |
| --- |
| Tokenized: What, was, your, favorite, movie, among, all, movie<br>Lemmatized: What be your favorite movie among all movie |

E) **[1 pt]** Tokenize and lemmatize the following English sentence:

*I am seeing a lot of mice.*

| **Tokenized and lemmatized sentence (use commas to separate individual tokens):** |
| --- |
| Tokenized : I, am, seeing, a, lot, of, mice, .<br>Lemmatized : I be see a lot of mouse |

## Problem 2

Consider the following training corpus $C$:

*this there that hat what when bat mere her here are hare*

and initial vocabulary $V$:

$$V = \{a, b, e, h, i, m, n, r, s, t, w\}$$

Use Byte-Pair Encoding (**just the token learning phase**) with hyperparameter $k$ set to 4 to expand the vocabulary $V$. Show all steps below. **In case of ties, merge the byte pair that appears first in the corpus** $C$.

| Splits | Pairs | Merge( r + e -> re) | Vocabulary |
|---|---|---|---|
| t h i s _ | t + h (3) | t h i s _ | *a, b, e, h, i, m, n, r, s, t, w* |
| t h e r e _ | h + e (4) | t h e re _ | re |
| t h a t _ | e + r (4) | t h a t _ | |
| h a t _ | **r + e (5)** | h a t _ | |
| w h a t _ | e + _ (5) | w h a t _ | |
| w h e n _ | h + a(4) | w h e n _ | |
| b a t _ | a + t(4) | b a t _ | |
| m e r e _ | w + h(2) | m e re _ | |
| h e r _ | b + a(2) | h e r _ | |
| h e r e _ | m + e(1) | h e re _ | |
| a r e _ | a + r(2) | a re _ | |
| h a r e _ | e + n(1) | h a re _ | |

| Splits | Pairs | Merge(re + _ -> re_) | Vocabulary |
|---|---|---|---|
| t h i s _ | t + h (3) | t h i s _ | *a, b, e, h, i, m, n, r, s, t, w* |
| t h e re _ | h + e (4) | t h e re_ | re , re_ |
| t h a t _ | | t h a t _ | |
| h a t _ | **re + _ (5)** | h a t _ | |
| w h a t _ | | w h a t _ | |
| w h e n _ | h + a(4) | w h e n _ | |
| b a t _ | a + t(4) | b a t _ | |
| m e re _ | w + h(2) | m e re_ | |
| h e r _ | b + a(2) | h e r _ | |
| h e re _ | m + e(1) | h e re_ | |
| a r e _ | a + re(2) | a re_ | |
| h a r e _ | e + n(1) | h a re_ | |

| Splits | Pairs | Merge ( h +e -> he) | Vocabulary |
|---|---|---|---|
| t h i s _ | t + h (3) | t h i s _ | *a, b, e, h, i, m, n, r, s, t, w* |
| t h e re_ | **h + e (4)** | **t he re_** | **re, re_** |
| t h a t _ | e + re_ (2) | t h a t _ | he |
| h a t _ | h + a (4) | h a t _ | |
| w h a t _ | a + t (4) | w h a t _ | |
| w h e n _ | w + h (2) | w he n _ | |
| b a t _ | b + a (1) | b a t _ | |
| m e re_ | m + e (1) | m e re_ | |
| h e r _ | e + n (1) | he re_ | |
| h e re_ | a + re_(2) | he re_ | |
| a re_ | | a re_ | |
| h a re_ | | h a re_ | |

| **Splits** | **Pairs** | Merge( h + a -> ha) | **Vocabulary** |
|---|---|---|---|
| t h i s _ | t + h (3) | t h i s _ | *a, b, e, h, i, m, n, r, s, t, w* |
| t he re_ | | t he re_ | re, re_ |
| t h a t _ | e + re_ (2) | t ha t _ | he |
| h a t _ | **h + a (4)** | ha t _ | ha |
| w h a t _ | a + t (4) | w ha t _ | |
| w he n _ | w + h (2) | w he n _ | |
| b a t _ | b + a (1) | b a t _ | |
| m e re_ | m + e (1) | m e re_ | |
| he r _ | e + n (1) | he re_ | |
| he re_ | a + re_(2) | he re_ | |
| a re_ | | a re_ | |
| h a re_ | | ha re_ | |

Final Vocabulary : *a, b, e, h, i, m, n, r, s, t, w, re, re_, he, ha*