Task 1.3

Whitespace

This tokenizer splits text on whitespace and isolates non-alphanumeric characters as separate tokens.

Example 1(works): `Hi, I am Tulasi.`

It is splitted into

```
["Hi", ",", "I", "am", "Tulasi", "."]
```

Example 2(works): `Би ном уншиж байна.` (Means `I am reading a book.`).

It is split into

```
["Би", "ном", "уншиж", "байна", "."]
```

Example 3(works): `This is a powerful tool.`

It is split into

```
["This", "is", "a", "powerful", "tool", "."]
```

Example 1(dosn't work): `It was uneventful.`

It is split into

```
["It", "was", "uneventful", "."]
```

Example 2(doesn't work): `The value is 3.14` It is split into

```
["The", "value", "is", "3", ".", "14"]
```

Example 3(doesn't work): `сургуулиудад` (Means `In schools`) It is split into

```
["сургуулиудад"]
```

Regex:

Expression:

```
\d+\.\d+ | \d+ | \w+ | [^\w\s]
```

This tokenizer explicitly separate numbers, words, and punctuation.

Example 1(works): `The value is 3.14.`

It is splitted into

```
["The", "value", "is", "3.14"]
```

Example 2(works): `Монгол хэл гайхамшигтай!` (Means `Mongolian language is wonderful!`)

It is split into

```
["Монгол", "хэл", "гайхамшигтай", "!"]
```

Example 3(works): `Hello, world!`

It is split into

```
["Hello", ",", "world", "!"]
```

Example 1(dosn't work): `don’t do that`

It is split into

```
["don", "’", "t", "do", "that"]
```

Example 2(doesn't work): `state-of-the-art` It is split into

```
["state", "-", "of", "-", "the", "-", "art"]
```

Example 3(doesn't work): `сургуулиудад` (Means `In schools`)

It is split into

```
["сургуулиудад"]
```

BPE

It is trained on continuous character sequences. It learns which subwords merge frequently. **Example 1(works):**

`internationalization.`

It is splitted into

```
["inter", "nation", "al", "ization", "</w>"]
```

Example 2(works): `сургуулиудад` (Means `In schools`)

It is split into

```
["сур", "гуу", "ли", "удад", "</w>"]
```

Example 3(works): `counterintuitively`

It is split into

```
["counter", "intuitive", "ly", "</w>"]
```

Example 1(dosn't work): `govern`

It is split into

```
["go", "ver", "n", "</w>"]
```

Example 2(doesn't work): `It is The truth is important.` It is split into

```
["It", "</w>", "is", "</w>", "th", "e", "</w>", "tr", "uth", "</w>"]
```

Example 3(doesn't work): `гайхамшигтай` (Means `Wonderful`) It is split into

```
["г", "ай", "хам", "шиг", "тай", "</w>"]
```

Task 2.3 Perplexity scores:

	Whitespace	**Regex**	**BPE**
base	inf	inf	inf
Kneser-ney	412.8	413.7	37.5
Witten-bell	11283.9	11299.9	70.9

Task 2.4

1. Regex

Base

Prompt: `in the United`

Output: `in the United States .`

Prompt: `the economy is`

Output: `the economy is doing well .`

Prompt: `climate change is`

Output: `climate change is a foot away from the city center )`

Explanation:

Works well when the exact 4 gram continuation is present in training data, producing fluent and correct completions.

Kneser ney

Prompt: `the president of the`

Output: `the president of the United States .`

Prompt: `deep learning models`

Output: `deep learning models . . .`

Prompt: `the results of the`

Output: `the results of the first to know about the . . .`

Explanation:

Kneser ney smoothing avoids zero probabilities, allowing the model to generalize unseen 4 grams. However, the regex tokenizer has frequent punctuations causing the model to reproduce them.

Witten bell

Prompt: `quantum computing will`

Output: `quantum computing will Hygienic Hygienic Hygienic ...`

Prompt: `artificial intelligence ethics`

Output: `artificial intelligence ethics Hygienic Hygienic Hygienic ...`

Prompt: `the theory of relativity`

Output: `the theory of relativity .`

Explanation:

Witten bell smoothing assigns probability mass to unseen continuations, but rare tokens are frequent in the backoff distribution. This results in repetition. But still sometimes produces correct small completions.

---

2. Whitespace

Base

Prompt: `one of the most`

Output:

`one of the most important thing is to make sure that you are not a good idea to have a good time to start thinking about what`

Prompt: `the number of people`

Output: `the number of people who have been in the past .`

Prompt: `machine learning is`

Output: `machine learning is`

Explanation:

Whitespace tokenization increases sparsity else similar to regex. Without smoothing, it can only continue very frequent phrases and fails for unseen 4-gram,.

Kneser ney

Prompt: `the number of people`

Output: `the number of people who have been in the past .`

Prompt: `climate mitigation strategies`

Output: `climate mitigation strategies . . .`

Prompt: `entropy regularization improves`

Output: `entropy regularization improves . . .`

Explanation:

Smoothing allows continuation of unseen contexts, but whitespace tokenization provides no morpheme information. So, technical or rare phrases often back off to punctuation-heavy continuations.

Witten bell

Prompt: `the results of the`

Output: `the results of the first things you will have to be a part of the world .`

Prompt: `protein folding process`

Output: `protein folding process Hygienic Hygienic Hygienic ...`

Prompt: `the united nations`

Output: `the united nations Hygienic Hygienic Hygienic ...`

Explanation:

Similar to regex, it produces repetitive outputs for rare scientific terms and other named entities.

---

3. BPE

Base

Prompt: `govern`

Output:

`go ver n the tr u th is th at the s am e ti me as se ar ch ing for a s im il ar`

Prompt: `internation`

Output:

`inter n a tion of the s am e ti me as se ar ch ing for a s im il ar to the s am`

Prompt: `electro`

Output:

`el ec tr o m ag az in e arti cl e and you would like to see a m an who ha d been a t`

Explanation:

BPE avoids out of vocabulary issues but, without smoothing, the model depends on subword transitions. This results in fragmented text.

Kneser ney

Prompt: `international trade agre`

Output:

`inter n a tion al tr ad e ag r e s ti ll have to ch ec k out the l as t year`

Prompt: `hyperparameter optimization`

Output:

`hyper par am et er op ti mi z a tion of the s am e ti me as se ar ch ing for a s im`

Prompt: `counterintuitively the`

Output:

`c ou n ter intu i tiv el y the s am e ti me as se ar ch ing for a s im il ar`

Explanation:

BPE handles rare and unseen words and kneser ney smoothing provides good probability estimates. Outputs are coherent at the subword level but due to the small context window, no sentence level meaning.

Witten bell

Prompt: `photosynthesis occurs`

Output:

`photo syn th es is oc cur s wh en you ar e a fe w m on ths ag o`

Prompt: `electroencephalogram`

Output:

`el ec t ro enc ep h al o gr am is a g re at de al wi th the s am e ti`

Prompt: `unbeliev`

Output:

`un b el i ev you b ec om e a m em b er of the s am e ti me as se ar ch`

Explanation:

Witten bell backoff over subwords leads to repetitive continuations. The model often relevance and incoherent outputs.