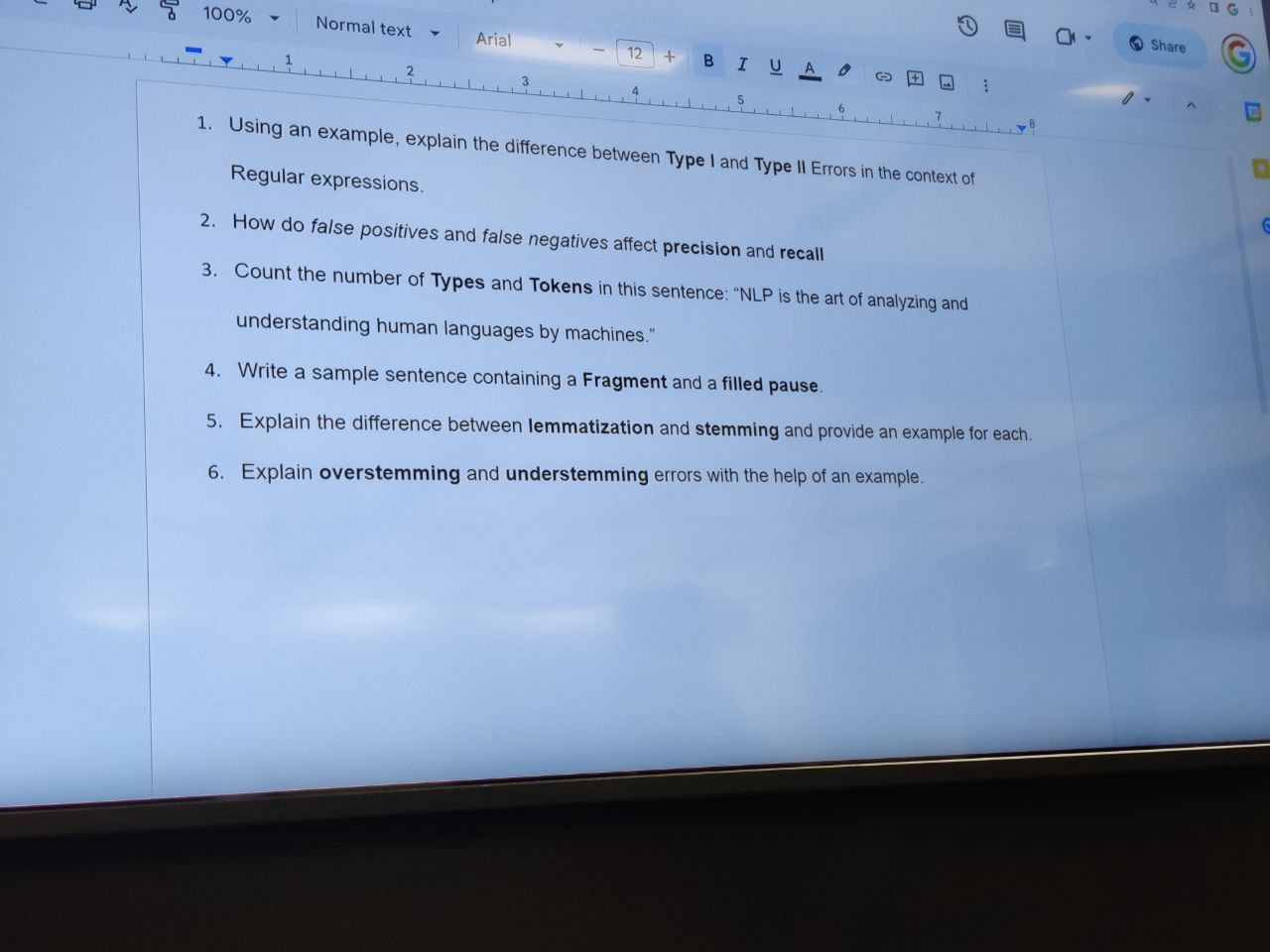
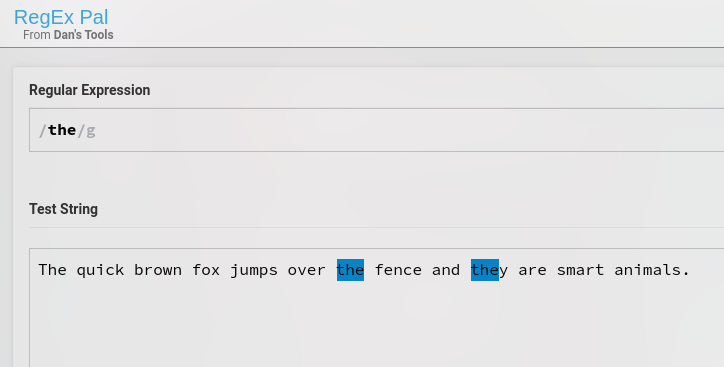
Regex, tokenisation, and lemmatisation



1. Type I errors refer to false positive conclusions while Type II errors refer to false negative conclusions. In the context of regex type I errors refer to tokens that are recognised by the regex pattern when it was not intended, and type II is when tokens are not recognised when it was intended to be so by the pattern. For example, searching the following text with the regex pattern, ‘the’ (quotes not included) will result in the following:

The quick brown fox jumps over the fence and they are smart animals.



So, here, the first word with the uppercase ‘T’ in ‘The’ is not recognised which is a type II error while a part of the word ‘they’ is recognised which is not intended and is a type I error.

1. Precision can be defined as the number of true positives divided by the sum of true positives and false positives while recall can be defined as the number of true positives divided by the sum of true positives and false negatives. The significance of these evaluation metrics is that the precision tells us the accuracy of positive predictions and recall measures the completeness of the positive predictions.
2. Tokens are the total number of individual words in a text corpus while types refer to the total number of unique words in the corpus. For the given sentence, the number of tokens and types is both 12 as all the words are unique.
3. Fragments are parts of words that are not fully formed and filled pauses are non-sensical words that are just used to fill any pauses during speech. Both these types of tokens usually appear in text when speech is converted to text. Examples are :
   1. For fragments: Ma..maybe, hel.. Hello
   2. For filled pauses: Errr, um
4. The main difference between the two is that lemmatization uses a vocabulary and morphological analysis of words to return the base or dictionary form of a word, which is known as the lemma, while stemming uses a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time.

For example, consider the words “running”, “ran”, and “runner”. A stemmer would reduce all three words to “run”, while a lemmatizer would reduce “running” and “runner” to “run” and “ran” to “run”.

1. Over stemming is when two unrelated words get reduced to the same common stem while under stemming is when two words which mean the same thing get stemmed to different stems.

