Practical 2 – Report

1. Segmentation

Q. How should you segment sentences with semi-colon? As a single sentence or as two sentences? Should it depend on context?

A. Semi-colon more often than not implies that the sentence isn't over however but it also separates different clauses within the sentence. So, in case we need to work on the clauses level of a sentence and not the complete sentence, we ought to portion sentences with semi-colons into two sentences.

Q. Should sentences with ellipsis… be treated as a single sentence or as several sentences?

A. Sentences with ellipsis conclusion with a path of dabs. In a few cases, the ellipsis can happen within the middle of the sentence as well. Hence, they ought to be treated as a single token and/or sentence and not numerous sentences.

Q. If there is an exclamation after the first word in the sentence should it be a separate sentence? How about if there is a comma?

A. On the off chance that there's an exclamation mark after the primary word, it ought to be treated as a single token and not a partitioned sentence. The comma does not demonstrate a partitioned sentence

Q. Can you think of some hard tasks for the segmenter?

A. Understanding the composing framework for the dialect being handled can be troublesome due to the assortment of dialects.

2. Tokenization

Q. Why should we split punctuation from the token it goes with ?

A. Accentuation could be a partitioned substance in a sentence; hence we cannot tokenize it with words.

Q. Should abbreviations with space in them be written as a single token or two tokens ? How about numerals like 134 000 ?

A. Ought to be treated as a single token since separating them will change the meaning totally.

Q. If you have a case suffix following punctuation, how should it be tokenised ?

A. For certain accentuations such as hyphens, case addition ought to be treated as the same token as the punctuation mark.

Q. Should contractions and clitics be a single token or two (or more) tokens

A. They should be one token.