# Tokenization Questions

Shreyasi Deshmukh

December 6, 2022

## 1 Segmentation

1. Sentences with a semi-colon can be both, complete sentences as well as partial sentences. For example: "I went to the library: my sister went to make lunch." This can be segmented into two different but complete sentences. However, if we consider the following sentence: My recipe requires: a cabbage, and pasta. This is an incomplete sentence. Hence it depends upon the context.

2. Ellipses are three dots in the middle of a sentence, usually leading to the next part of the sentences, they should be treated as a single sentence.

3. Exclamation after first word - not a seperate sentence
Comma - not a seperate sentence.

4. Hard Tasks for Segmenter- In languages without the use of many punctuation, identifying a sentence boundary is difficult. (Source: https://tm-town-nlp-resources.s3.amazonaws.com/ch2.pdf)

## 2 Tokenization

1. Punctuation is a seperate entity in a sentence, therefore we cannot tokenize it with words.

2. Both should be treated as a single token, since seperating them will change the meaning entirely.

3. Case suffixes should be treated the same as punctuation marks in a sentence.

4. Contrations and Clitics should be a single token.