# Cleaning and Tokenization

Before implementing any natural language processing algorithm, you might want to clean the data and tokenize it. Here are a few things to keep track of when handling your data.

- Letter case “The” == “the” == “THE” → *lowercase / upper case*
- Punctuation , ! . ? → . “ ‘ « » ’ ” → ∅ … !! ??? → .
- Numbers 1 2 3 5 8 → ∅ 3.14159 90210 → *as is/<NUMBER>*
- Special characters ∇ $ € § ¶ ** → ∅
- Special words 😊 #nlp → :happy: #nlp

You can clean data using python as follows:

```
corpus = 'Who ❤️ "word embeddings" in 2020? I do!!!'

data = re.sub(r'[,!?;-]+', '.', corpus)
data = nltk.word_tokenize(data)  # tokenize string to words
data = [ ch.lower() for ch in data
         if ch.isalpha()
         or ch == '.'
         or emoji.get_emoji_regexp().search(ch)
       ]
```

```
→ ['who', '❤️', 'word', 'embeddings', 'in', '.', 'i', 'do', '.']
```

You can add as many conditions as you want in the lines corresponding to the green rectangle above.