Good resource on preprocessing. Main steps for preprossessing:

1. Text cleaning. Convert to lowercase, remove links, remove non-word and non-whitespace characters (so punctuation, symbols, special characters), digits (somehow, digits and numbers don’t add significant meaning to NLP text analysis).
2. Tokenization. Break the text into words (or sub-words. Ex. “smart”, “er”).
3. Remove stopwords. Aka, remove the most commonly used words in the language. Ex. The, is, for.
4. ‘Rootification’. Commonly, Stemming or lemmatization.
   1. Stemming is cutting off the ends of words to convert to a common root form. Simple, fast, blind
   2. Lemmatization removes inflectional endings and aims to convert words to their base/dictionary form (aka, a ‘lemma’)

<https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9>

How much data do we need?

* Rough token counts:
  + 1k tokens: 4 page essay (double space)
  + 800k tokens: all of Shakespeare’s works
  + 1M tokens: the Brown corpus (a sample of about 1 million words from 1961. Classic language analysis corpus)
  + 500B tokens: GPT-3’s total corpus
* Task matters:
  + Classification < generation in terms of total data needed
  + Narrow scope needs less data
* Watch out for datasets that contain data that was collected illegally

<https://www.youtube.com/watch?v=gye07gsaMBc>

General tutorial. Runs from the ground all the way up to BERT

<https://www.youtube.com/watch?v=M7SWr5xObkA>

How to pre-train BERT

<https://discuss.huggingface.co/t/pre-train-bert-from-scratch/28318>