Good resource. 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>