Capstone Project 1: Data wrangling
Spring Board Data Science Career Track
~ By Zach Nguyen

Steps I performed for data cleaning:

- Importing nltk package for simple text transformation.
- Loading the raw dataframe of text.
- Applying the Tokenize function to every row.
- Removing punctuations.
- Removing stop words.
- Removing numbers.
- Lemmatizing the text to group words which have the same groups.
- Removing words that appear in the whole dataset less than a theshold t (currently set to t = 20). This is a solution to deal with the noise which might be cause by mispellings and typos which frequently appear in text data.
- Note: These are initial steps for data cleaning used to conduct EDA and answer the question: What type of contents are in ham rather than spam sms? However, we could also undo some of these transformation to take into account punctuations, numerics and a few interesting stop words as well.