

Example 2 - Stopwords

This example continues on Example 1. Make sure you start off with the code of Example 1.

We need to get rid of stop words. Here are some examples of stop words: “the”, “of”, “to”, “in”.

NLTK comes with stop words lists for most languages. To get English stop words, you can use this code.

At the top of the script, along with the imports, do:

```
from nltk.corpus import stopwords
```

Just for the sake of it, let's check which words in English are considered by nltk to be stop words.

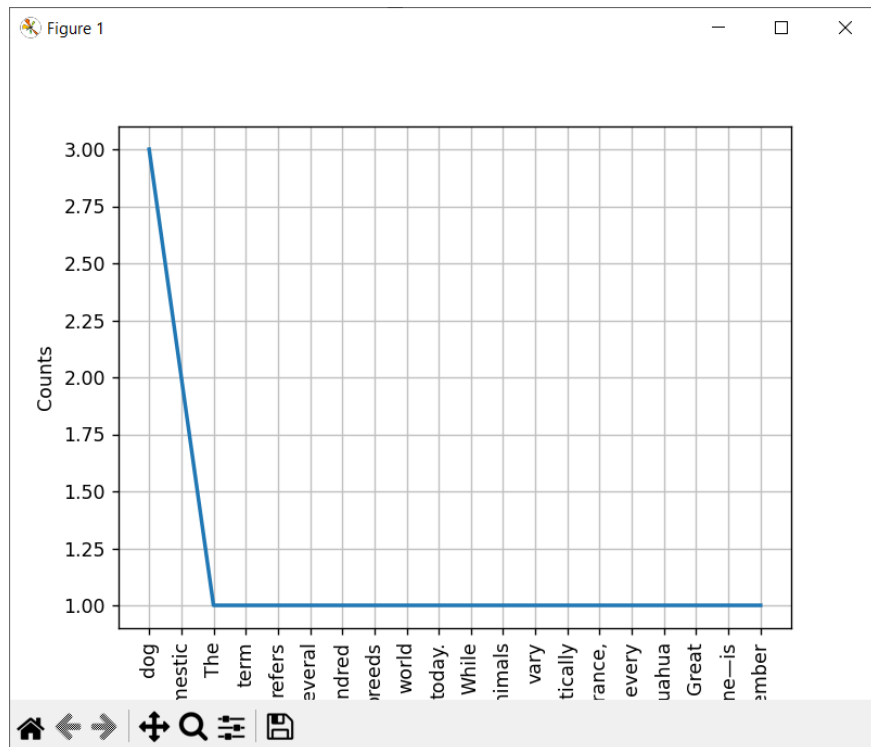
```
print(stopwords.words('english'))
```

Now let's put it to use. Right after we tokenise, and before counting the frequencies:

```
clean_tokens = tokens[:]
for token in tokens:
    if token in stopwords.words('english'):
        clean_tokens.remove(token)
```

We are making a copy of the token list, and then going through all tokens and removing them from the new token list, if they are a stopword. The frequency distribution function should now be modified to work on the **clean_tokens**.

Check out the new output.



Why is “The” still showing in the chart despite being a stop word? Try to fix the code to remove this problem.

Challenge

The program besides displaying the graph should also output what the text is about. For example: “The text is about: dog”