- The most fundamental process in **NLP** is **counting n-grams** in documents.

- All collection of documents that contain various forms of n-grams is defined as a **corpus**. (Plural of corpus is a **corpora**.)

- A **document** isn't an actual document *per se*, as "document" here can be an individual tweet amongst millions of tweets (corpus), a Yelp review (document) amongst thousands of reviews (corpus), etc.

- We often want to count terms in the **documents** which they are contained.
- However, just counting how many times a word appears in a corpus often **does not always** capture whether this term is important.
- So, how do we resolve this?

However, just counting how many times a word appears in a corpus often **does not always** capture whether this term is important.

- So, how do we resolve this?