

# COUNTING N-GRAMS

- 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.

