Representation Re-Revisited

- Remember our sample document?
  - "If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck."

- Given our data pre-processing pipeline we transform the above into something like:
  - "look like duck swim like duck quack like duck probabl duck"

- Which we could represent as a document-term frequency matrix:

look	like	duck	swim	quack	probabl
1	3	4	1	1	1

N-grams

- Our representations so far have been single terms.
- These are known as *unigrams* or *1-grams*.
- Not surprisingly, there are also *bigrams*, *trigrams*, *4-grams*, *5-grams*, etc.
- N-grams allow us to extend the bags-of-words model to include word ordering!

Adding Bi-grams

- Let’s revisit our sample document and add bi-grams to the mix. Processed data:
  - “look like duck swim like duck quack like duck probabl duck”

- Now the bigrams based on the above:

look_like	like_duck	duck_swim	swim_like	duck_quack	quack_like	duck_probabl	probabl_duck
1	3	1	1	1	1	1	1

NOTE – We’ve more than doubled the total size of our matrix!!!