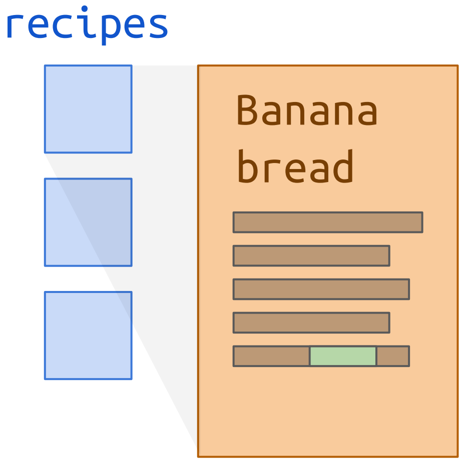
text analysis

COGS 108 Discussion Lab 7 (Week 8)

TF-IDF

“Term frequency - Inverse Document Frequency” is a metric of how relevant a *word* is to a *document* in a *corpus* given that some words appear more frequently than others.



word	idf
banana	0.8
bread	0.4
...	
bake	0.1
chicken	0.0

How we'll compute it…

1. **Grab a corpus of documents**
2. **Clean up the corpus to just be the words**
   (remove punctuation, etc.)
3. **Make a `TfidfVectorizer (from sklearn)`**
4. **Fit the `TfidfVectorizer` to your corpus**

…

Go over to the notebook to show these steps
for a corpus of recipes…