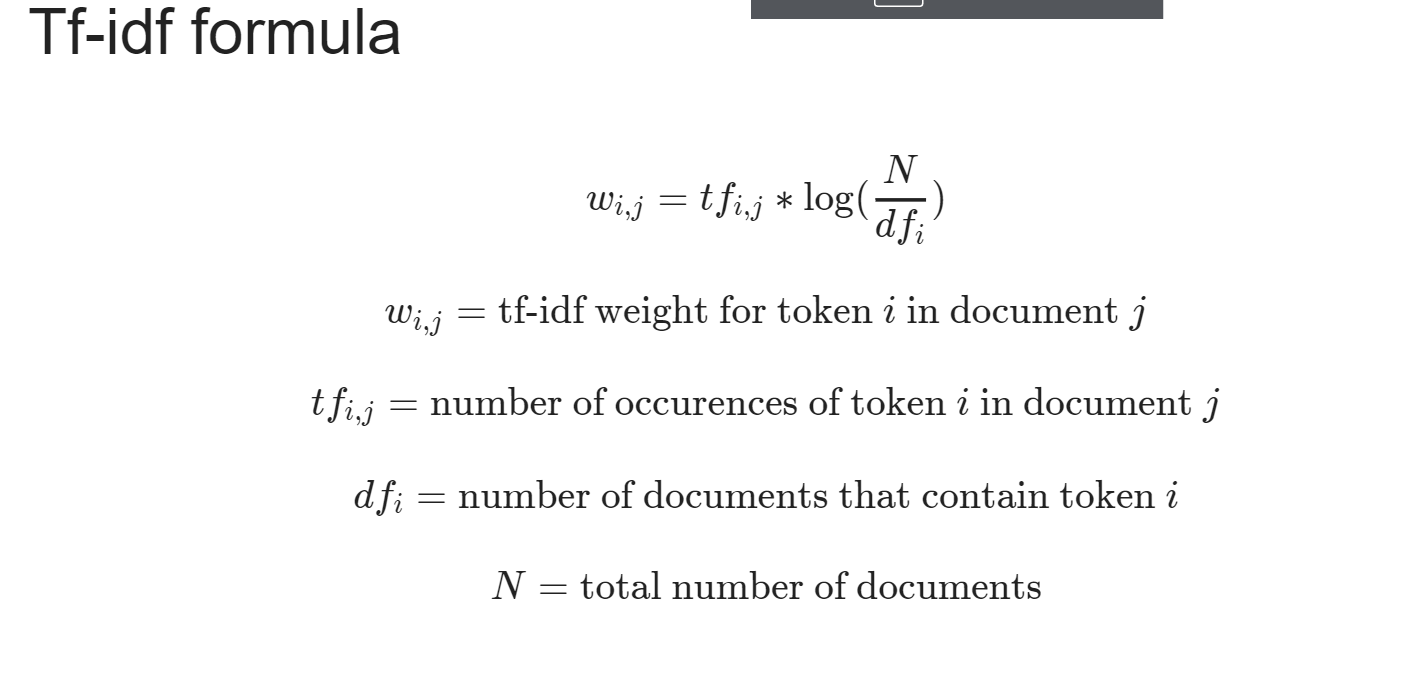
**What is tf-idf?**

* Term frequency - inverse document frequency
* Allows you to determine the most important words in each document
* Each corpus may have shared words beyond just stopwords
* These words should be down-weighted in importance
* Example from astronomy: "Sky"
* Ensures most common words don't show up as key words
* Keeps document specific frequent words weighted high



**Tf-idf with genism**

In [10]: from gensim.models.tfidfmodel import TfidfModel

In [11]: tfidf = TfidfModel(corpus)