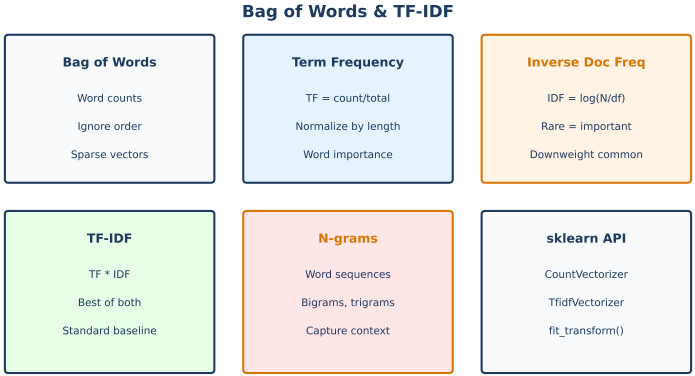
Lesson 38 Summary: BOW & TF-IDF

Data Science with Python – Key Concepts

Data Science Program

BOW & TF-IDF Overview



Convert text to numerical vectors for ML

Bag of Words

Count-based representation:

- **Vocabulary:** All unique words
- **Vector:** Word counts per document
- **Sparse:** Most entries are zero

Ignores word order, captures frequency

TF-IDF Formula

Weight by importance:

- **TF:** Term frequency in document
- **IDF:** $\log(\frac{N}{df})$ penalizes common words
- **TF-IDF:** TF $\times$ IDF

Rare words get higher weights

N-grams

Capture word sequences:

- **Unigrams:** Single words
- **Bigrams:** Word pairs ("not good")
- **Trigrams:** Word triples

---

ngram_range=(1,2) includes uni and bigrams

sklearn Vectorizers

Two main classes:

- **CountVectorizer:** Raw counts (BOW)
- **TfidfVectorizer:** TF-IDF weights
- **API:** fit_transform(docs)

Same interface, different weighting

Quick Reference

Essential Commands:

Task	Code
BOW	`CountVectorizer().fit_transform(docs)`
TF-IDF	`TfidfVectorizer().fit_transform(docs)`
Bigrams	`TfidfVectorizer(ngram_range=(1,2))`
Max features	`TfidfVectorizer(max_features=1000)`
Get names	`vec.get_feature_names_out()`

TF-IDF is the standard baseline for text classification