Vectorization (How to Count)

School of Information Studies
Syracuse University

How to Count Tokens?

Convert documents into word vectors

Bag of Words (BoW)

- Boolean
- Term frequency
- Normalized term frequency
- Tf*idf

Vectorization

Step 1: Create a dictionary of unique words.

1. "vector"
2. "number"
3. "text"
4. …

	"vector"	"number"	"text"	…
Doc1	1	0	0	
Doc2	1	1	1	
doc3	1	0	1	

Step 2: Represent every document as a word vector: each word is an attribute/feature.

Boolean Vectors

Word presence or absence

	"vector"	"number"	"text"	…
Doc1	1	0	0	
Doc2	1	1	1	
Doc3	1	0	1	

Frequency Vectors

Word frequency: the number of word occurrences

	"vector"	"number"	"text"	…
Doc1	5	0	0	
Doc2	1	3	6	
Doc3	2	0	8	

Normalized Frequency Vectors

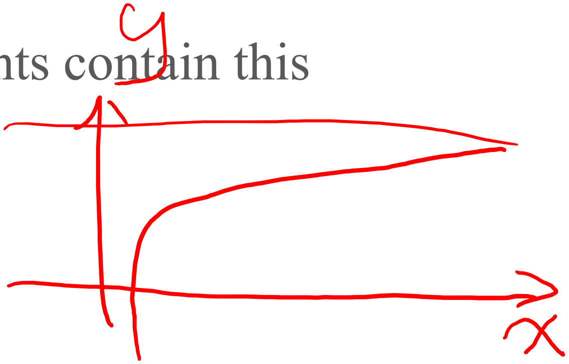
Normalized word frequency: word frequency normalized by the document length

	"vector"	"number"	"text"	…
Doc1	0.51	0.02	0.01	
Doc2	0.12	0.15	0.35	
Doc3	0.02	0.13	0.43	

TF*IDF Vectors

Tf*idf weighting

- Tf: term (word) frequency
- Df: document frequency; i.e, how many documents contain this term; e.g., 2 out of 3 documents -> 2/3
- Idf: inversed-document frequency, 3/2 = 1.5
- Tfidf=tf*log(idf)



	"vector"	"number"	"text"
Doc1	1	0	0
Doc2	0.1	0.3	0.6
Doc3	0.2	0	0.8

3/3 → 1   1/3   2/3

	"vector"	"number"	"text"
Doc1	0	0	0
Doc2	0	0.3*log3	0.6*log1.5
Doc3	0	0	0.8*log1.5

History of TF*IDF

A concept borrowed from information retrieval

A “blind” weighting strategy for text classification