VECTORIZATION

SYRACUSE UNIVERSITY
School of Information Studies

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”
…

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

	“vector”	“number”	“text”	…
Doc1	1	0	0	
Doc2	1	1	1	
Doc3	1	0	1	

VALUES OF WORD FEATURES

Boolean value: Word presence or absence

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

VALUES OF WORD FEATURES

Word frequency: The number of word occurrences

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

VALUES OF WORD FEATURES

Normalized word frequency: Word frequency normalized by the document length

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

VALUES OF WORD FEATURES

Tf–idf weighting

Tf: Term (word) frequency

Df: Document frequency, i.e, how many documents contain this term (e.g., 8 out of 100 documents -> 8/100)

Idf: Inverse document frequency, 100/8

Tf–idf = tf*log(idf)

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

	"vector"	"number"	"text"
Doc1	0	0	0
Doc2	0	0.3*log3	0.6*log 1.5
Doc3	0	0	0.8*log 1.5

TF–IDF

Concept borrowed from information retrieval

A “blind” weighting strategy for text classification