Text Representation/ Vectorization

School of Information Studies
Syracuse University

What Is Text Mining?

Knowledge discovery from text data.

Computers do not understand human language; they just count.

Three-step process:

1. Convert text documents to numeric vectors
2. Find patterns in the numeric vectors
3. Explain the patterns' semantic meaning

What to Count?
Text Representation

Which six words/phrases did you choose to describe yourself?

If you were a text document, then the six words are a representation of you as a document.

Are your description words similar or different from other students' words?

Text Representation

A document can have many properties

Which property to represent?

- Topic
- Sentiment and opinion
- Genre
- Author
- Writing style
- Confidence
- …

Bag-of-Words (BoW) representation

How to Count? Vectorization

Step 1: Create a dictionary of all unique words.

1. "glasses"
2. "smart"
3. "tired"
4. …

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

	"glasses"	"smart"	"tired"	…
Jack	1	1	0	
Jill	0	1	0	
Ben	1	1	1	