Text: Unstructured Data

- Structured: Data is organized into pre-defined structure like a table of database - with rows and columns.
- UnStructured Data: Data does not have a pre-defined structure. Think of a collection of emails, a bunch of satellite images or the entire text of speeches from the british parliament since 1803.

Modeling/representing text

- Bag of words - Documents simply represented by the words in the document and their frequencies. Disregards grammar and word order
  - Bayesian SPAM filter
- Semantic - mapping natural language rules to get a formal representation of the meaning of the text
  - Name entity identification

Bag of words

- Corpus:
  - A: John likes to play soccer
  - B: John is reading a book

	John	likes	soccer	play	book	reading	a	is	to
A	1	1	1	1					1
B	1				1	1	1	1	

n-gram model

- The Bag-of-words model is an orderless document representation. Only the counts of words matter.
- We could do this also by choosing consecutive pairs (2-gram) and representing each pair
- A: John likes to play soccer
- B: John is reading a book
- 2-gram (bigram):

	John likes	likes to	play soccer	to play	John is	is reading	reading a	a book
A	1	1	1	1				
B					1	1	1	1

Cleaning text

- Stop words: Common words that are not useful in providing value or context. Eg: ‘the’, ‘an’, ‘in’ etc.
- Stemming: Returning words to their original stem. Eg: ‘Chopping’, ‘Chopped’ are all replaced with ‘Chop’
- Lower case conversion
- Remove punctuations
- Strip extra white spaces
- Remove numbers

Example

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors.

word <fctr>	freq <dbl>
the	4
and	3
dursley	3
they	3
very	3
was	3
were	3
mrs	2
much	2
neck	2

Example

Mr. Mrs. Dursley, number four, Privet Drive, proud say perfect normal, thank much. They last peopl expect involv anyth strang mysterious, just hold nonsense. Mr. Dursley director firm call Grunnings, made drills. He big, beefi man hard neck, although larg mustache. Mrs. Dursley thin blond near twice usual amount neck, came use spent much time crane garden fences, spi neighbors.

word <fctr>	freq <dbl>
dursley	3
mrs	2
much	2
neck	2
although	1
amount	1
anyth	1
beefi	1
big	1
blond	1