**Natural Language Processing**

Bag of Word

* Count each word and use the count as an independent variable
* May need to remove punctuation
* Stop words: the, is which, at – unlikely to improve understanding, little meaning on their own – normally you exclude them
* But ‘The Who’ could be valid and significant – so be careful
* Stemming can reduce argue, arguing, argument, …
  + You can build a database
  + You can write a rule to take out ing, ed, ly, s
  + It can miss many like child and children
  + Porter Stemmer is often used
* You want to make sure you do not have a very sparse matrix of words