Corpus
Tokenization
- Simple Transformation
- Lowercase
- Remove Numbers
- Remove Punctuations
- Remove English Stop Words
- Remove Own Stop Words
- Strip Whitespace
- Specific Transformations

Stemming
Document Term Matrix
Sparse Term