Text Analytics

Data Science Dojo

Overview

- What is text analytics?
- Fundamentals
  - Tokenization
  - Stemming and Lemmatization
  - Document Vectors
- Term Frequency (TF) and Inverse Document Frequency (IDF)
  - Creating and inverted index and retrieving documents from a query

Text Analytics

- How do we turn unstructured data into structured data?
  - Create columns based on document content
  - Each **term** in document creates a column
    - Column types: binary, word count, TF-IDF
  - Do we want to count every word?
    - Stop words

TF-IDF Matrix

	Beijing	Dish	Duck	Rabbit	Recipe
D1	0	0	0.097	0	0
D2	0.199	0.199	0.097	0	0
D3	0	0	0.097	0.199	0.111
D4	0	0	0	0.398	0.222
D5	0.398	0.398	0.097	0	0.222

End of Sample Slides

4 of 25 slides in presentation