Modeling Text

Abdus Salam Azad

Vectorize text

- Most of the models work with numbers

- So if we us a numeric vector for texts will be more convenient

A Simplified Example

- Training Documents:
  - D1:
    - **Life**
    - We have a purpose
  - D2:
    - **Death**
    - We will die
- Test Documents
  - D3:
    - **???**
    - A real purpose
- Here each **word** is a feature
- We represent each **document** as a vector

Topic

Body

Bag of word models

$$\begin{bmatrix} we \\ have \\ purpose \\ will \\ die \\ real \end{bmatrix} \text{D1:}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{D2:}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \text{D3:}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

- Each element corresponds to one word of the dictionary
  - Dictionary: all the words in all the documents

Bag of word models

$$\begin{bmatrix} we \\ have \\ purpose \\ will \\ die \\ real \end{bmatrix} \text{D1:} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{D2:} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \text{D3:} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

- For each word for a document we can use
  - Is the word present or not ? [Word Occurrence Vector]
  - Count of the word [Count Vector]
  - Tf-idf weight

TF-IDF weights

- Let **w** be a word, **d** be a document, **N(d,w)** be the number of occurrences of **w** in **d**
- **TF(d,w)** = **N(d,w)** / **W(d)**
  - where **W(d)** is the total number of words in **d**
- **IDF(d,w)** = log(**D** / **C(w)**)
  - where **D** is the total number of documents
  - **C(w)** is the total number of documents that contains the word **w**
- The TF-IDF weight for **w** in **d** is **TF(d,w)*IDF(d,w)**

STOP Words

- The words which appear in nearly every document
  - Am, is, are
  - Was, were
  - A, an, the

- Does not have effect of classification