Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Text: Bag of Words

Another very common data type we work with is **language**: bodies of text, online articles, corpora, etc.

Later in this course, we will discuss more powerful ways to analyze text, such as **sequential models**, and **transformers**.

Obligatory chatgpt reference.

There's a very simple encoding that we'll focus on here: the **bag of words** approach.

This approach is meant to be as simple as possible: for each word, we ask ourselves, "if this word in the text?", and answer yes (1) or no (0) for every single word.

Definition 1

The **bag of words** feature transformation takes a body of text, and creates a **feature** for every **word**: is that word in the text, or not?

$$\phi(x) = \begin{bmatrix} \text{Word 1 in } x \\ \text{Word 2 in } x \\ \vdots \\ \text{Word k in } x \end{bmatrix} \tag{1}$$

This approach is used for **bodies of text**.

Example: Consider the following sentence: "She read a book."

With the words: $\{\texttt{She}, \texttt{he}, \texttt{a}, \texttt{read}, \texttt{tired}, \texttt{water}, \texttt{book}\}$

We get:

$$\phi(\text{"She read a book."}) = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

A couple weaknesses to this approach:

- Ignores the order of words and syntax of the sentence.
- Doesn't encode meaning directly.
- Duplicate words are only included once.
- It doesn't create much structure for our model to use.

But, it's very easy to implement.