CS-584

Class 2--In Class Activity

9/18/24

Wyatt Blair

What is a word?  
**1. Come up with a clear definition of a word which, when violated,  
strictly excludes strings of symbols that are NOT words. Think  
about examples where it’s not so clear-cut, such as compound terms like  
”New York” or abbreviations like ”Dr.”.**

A word is a distinct and meaningful string of characters. In some cases, certain strings encompass the same meaning as other strings, but they are in fact the same word. Cases include “Dr.” and “doctor”. This is separate from entity disambiguation, where something like “he” might refer to “Wyatt” and is instead more akin to the “h8r” and “hater” example from class. In the case of “New York” I would argue this is a combination of words which, when combined, create a new concept which was not present without the union.

**2. What properties should something have in order to be classified  
as a word? Break down the definition you just made. Brainstorm as  
many properties as possible. For example:  
• Must consist of alphabetic characters  
• Must not contain spaces (or should it?)  
• Can it include punctuation (e.g., ”I’m”)?  
• Should it have a minimum or maximum length?  
• Can numbers (e.g., ”2024”) or symbols (e.g., ””) be considered words?**

* Must be unique in its meaning
* Must not have any spaces in it (as I said before, “New York” is a phrase which creates a distinct concept out of two individual words which have different meanings when taken separately).
* Must contain characters which are intelligible to the language (doesn’t necessarily have to be from the language (like resumé) but it does need to be widely understood within the language, unlike: 猫.)

**3. How would you implement the properties you’ve discussed in a  
program? In other words, how would you write code to identify and  
extract words from a large text based on these properties? Think about  
which properties would need to be encoded as rules.**

* Unique Meaning: Use an LLM to perform entity disambiguation on two proposed “words” to determine if it is truly a unique/distinct word or if they are the same
* No Spaces: Can just do something like: `words = sentence.split(‘ ’)`
* Intelligible Characters: Use a whitelist and either throw an error or remove words in the case of them being “unrecognizable”

**At the end of the activity, you should submit to my email jkarell@stevens.edu,  
the following in either a word doc, or CLEAR photo of your CLEARLY hand-  
written work.:**