CAB203 Problem solving

Shridhar Thorat: n10817239

June 16, 2022

1 Regular languages and finite state automata

In order to censor the articles: '*a*', '*an*' and '*the*', and append a student's number in their emails, a Python function **censor(s)** was developed. This function applies the required censoring to a string **s**, replacing each letter in an article with the pound sign: #, and appends a student number (including a single whitespace) if the string was censored.

Syntax and patterns:

This function took advantage of the Python package **re**, which provides regular expression syntax and functions. The syntax used were; \b and ( ). The syntax \b defines the beginning and ending of a word boundary, while ( ) defines a capture group.

The regex patterns for the three articles in Python were chosen as below. Where r simply defines a raw string.

- r"\b(a)\b" only matches the word "*a*"
- r"\b(an)\b" only matches the word "*an*"
- r"\b(the)\b" only matches the word "*the*"

These regex patterns are checked in the string s from left to right. The patterns can be interpreted as below.

- The first \b sets the starting word boundary.
- The article in () defines the character capture group. Where only characters in this group are matches.
- The last \b sets the end of the word boundary.

Thus the patterns only match words in the capture group. However, the articles must be matched regardless of capitalisation - which isn't done by the pattern. Instead of modifying the pattern, the flag re.IGNORECASE is used.

Flags used:

The flag re.IGNORECASE (re.I) is used as a function parameter. As the name suggests, it performs case-*in*sensitive matching for a pattern. For instance, the pattern r"\b(the)\b" would match "*The*", "*ThE*", etc, when the flag is used in any function.

Now we require functions which can determine if a match (article) exists in a string and all the matches for each of the patterns.

Functions used:

In order to find the first location where a pattern produces a match, the regular function **re.search(*pattern, string, flags=re.I*)** can be used. As the description, this function is used to determine whether any of the patterns exist, regardless of capitalisation (as defined by the flag). If the patterns exists, it can then be censored. However, this requires determining their location within the string.

This can be done by using the function **re.finditer(*pattern, string, flags=re.I*)**which scans the string left-to-right and returns an iterator that yields *match* objects which match the regex pattern.

I.e, it returns every match to the pattern, as a regex *match* object. These objects have 2 attributes useful in this problem; the **match.start()** and **match.end()**. These give the starting and ending index of a match respectively. Given the start and end indices, the string can be censored with a pound # between these indices - thus censoring the article. Applying this for each match (i.e. article), the whole string can be censored.

2 Linear algebra