Handling Text with Python

→ strings → tokens → characters

→ documents / files.

→ len (text 1) → text 1. split (" ")

→ [w for w in text1 if len(w) > 3]

→ [w for w in text1 if w.istitle()]

→ [w for w in text1 if w.endswith('s')]

→ len (set (text))

→ len (set ([w. lower() for w in text]))

→ w. startswith ('t')

→ w. endswith ('t')

→ t in w

→ s. isupper(), s. islower(),

s. istitle()

→ s. isalpha(), s. isdigit(),

→ s.lower(), s.upper(), s.title()

→ s.split('_')

→ s.splitlines()

→ s.join('t')

→ s.strip(), s.rstrip()

→ s.find(t), s.rfind(t)

→ s.replace(u, v)

Handling larger Texts

```
f = open('file.txt', 'r')
f.readline()  # reads first line
```

```
{ f.seek(0)
  f.read()        full file
```

```
text.splotlines()   # separated
                      by \n
```

```
for line in f:
    doSomething(line).
```

```
f.write(_).
f.close()
f.closed        # checks.
```

Regular Expressions

```
→ [w for w in text1 if w.startswith('#')].
                                     @
→ [w for w in text1 if re.search(r'@
                         [a-z A-Z 0-9_]+',
                                      w))]
```

```
.   ^   $   [ ]   [a-z]

[^abc]   a|b   ( )   \   \b

\d   \D   \s   \S   \w   \W

*   +   ?   {n}   {,n}

{n,}   {m,n}
```

```
re.search(r'@\w+', w)
```

→ `re.findall(r'[aeiou]', w)`

→ For regular expression of dates, Refer jupyter notebook.

→ `extractall()`