CSE556 - Natural Language Processing
Assignment 2 Report

Aditya Chetan (2016217)

August 23, 2018

1 Assumptions

- I have considered that a paragraph must end with at least a 2-length sequence of \n,\t, \r.
- I have assumed that the sentence may end with a period, exclamation mark, question mark or closing brackets. This mark if followed by a white space, and then the new sentence starts.
- I have considered abbreviations, hyphenated words, URLs and numbers to be single words.
- For a word to start a sentence, it must either be in all caps or capitalised.
- I am only considering the occurence of the word to be counted if it is in lowecase, uppercase or capitalised. Other cases like camel case, etc. are ignored.
- Word count only works for normal dictionary type words (no hyphenations, contractions, etc.)

2 Methodology

2.1 Regex for paragraphs

```
[\"?!\.\)\]\}](\n|\r|\t| {4,}){2,}\t*
```

- A paragraph would typically optionally end on a punctuation mark used to end a sentence (such a period, quote, parentheses, etc.) as indicated by the first disjunction.
- Then I have assumed that a new paragraph must be preceded by some combination of newline or carriage return or tab (eg. `\n\r, \n\n, \n\t`
- After this combination, it may be followed by as many tabs as needed.

2.2 Regex for sentences

```
([a-zA-Z\.]{3,}?|[0-9]+)[\.\'\"\?!\}\)\]]+\s+[\"\(\[\{A-Z0-9\n]
```

- The first group `([a-zA-Z\.]{3,}?|[0-9]+)` handles the abbreviations and sentences ending in numbers.
- After this, we have the punctuation mark that might end a sentence in the next disjunction.
- After the punctuation mark we must have at least one whitespace character (this is an assumption).
- This is followed by a Capital letter, or a number or an opening bracket or inverted comma. It might also be a newline.

2.3 Regex for words

```
[^\[\{ \}\.\"\)\(\]\!\?\b\s,]*[A-Z0-9a-z]+,?[^\[\{ \}\'\"\)\(\]\!\?\b\s,]*
```

- First come the punctuation marks or spaces that may precede a word. These must not be matched.
- This is followed by the alphabets or numbers that may make up the word. This is matched by greedy kleene plus matching. We have an optional comma.
- Then we match one or more punctuation marks/spaces.