Methodology of Kren.R

Jonathan Fung

June 25, 2022

Lets take the first verse of fromis_9's DM as an example.

[Verse 1: Nagyung, Seoyeon]
Hey, you, 지금 뭐 해?
잠깐 밖으로 나올래?
네가 보고 싶다고
거울 속의 난 so perfect
새로 산 신발도, check it, okay

First, the section title is parsed out and stored with `clean_section`.

```
clean_section("[Verse 1: Nagyung, Seoyeon]")
```

```
[1] "Verse 1"
```

Then, we need to determine if the lyrics bytes come from the English or Korean Unicode blocks.

```
detect_lang("Hey, you 지금 뭐 해?")
```

```
[[1]]
[1] "Hey, you 지금 뭐 해?"

[[2]]
 [1] NA NA NA NA NA NA NA NA NA  1  1 NA  1 NA  1 NA
```

We can then use this information to determine if a word is Korean or not. Each word is determined by space-separation.

```
construct_hangulp("Hey, you 지금 뭐 해?")
```

```
  words hangulp
1  Hey,   FALSE
2   you   FALSE
3  지금    TRUE
4    뭐    TRUE
5   해?    TRUE
```

Then, features such as Korean proportion over the line, and syllable count per word can be extracted. `lyrics_tree_data` applies this function over a whole song.

```
extract_features(construct_hangulp("Hey, you 지금 뭐 해?"))
```

```
[[1]]
  kr_dist syllables
1       0         1
2       0         1
3       1         2
4       1         1
5       1         1

[[2]]
[[2]]$kr_word_prop
[1] 0.6

[[2]]$kr_cnt
[1] 3

[[2]]$word_cnt
[1] 5
```

English syllables are calculated using the `syllables` library, which scrapes poetrysoup. Since Korean is a syllabic language, the syllable count is simply the number of blocks.

```
apply(construct_hangulp("거울 속의 난 so perfect"), 1, count_syllable)
```

```
[1] 2 2 1 1 2
```