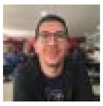
BUZZ CHRONICLES > ALL
Saved by @ankitsrihbti
See On Twitter

Twitter Thread by Andres Segura-Tinoco

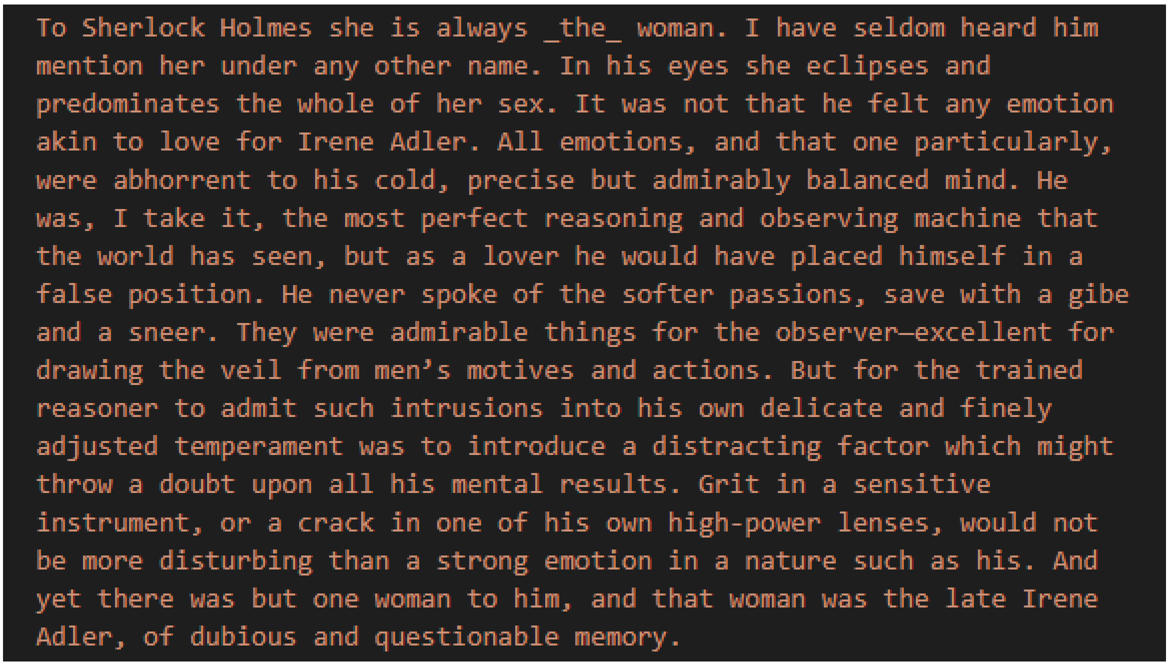


Andres Segura-Tinoco
@SeguraAndres7

Text Normalization (TN) are techniques in the field of #NLP that are used to prepare text, sentences, and words for further processing or analysis.

Two of the most common TN techniques are Stemming and Lemmatization. In the next thread■ I will briefly tell you about them.

1/5



```
To Sherlock Holmes she is always _the_ woman. I have seldom heard him
mention her under any other name. In his eyes she eclipses and
predominates the whole of her sex. It was not that he felt any emotion
akin to love for Irene Adler. All emotions, and that one particularly,
were abhorrent to his cold, precise but admirably balanced mind. He
was, I take it, the most perfect reasoning and observing machine that
the world has seen, but as a lover he would have placed himself in a
false position. He never spoke of the softer passions, save with a gibe
and a sneer. They were admirable things for the observer—excellent for
drawing the veil from men's motives and actions. But for the trained
reasoner to admit such intrusions into his own delicate and finely
adjusted temperament was to introduce a distracting factor which might
throw a doubt upon all his mental results. Grit in a sensitive
instrument, or a crack in one of his own high-power lenses, would not
be more disturbing than a strong emotion in a nature such as his. And
yet there was but one woman to him, and that woman was the late Irene
Adler, of dubious and questionable memory.
```

The aim of both methods (Stemming and Lemmatization) is the same: to reduce the inflectional forms of each word/term into a common base or root.

So what is the difference between them?

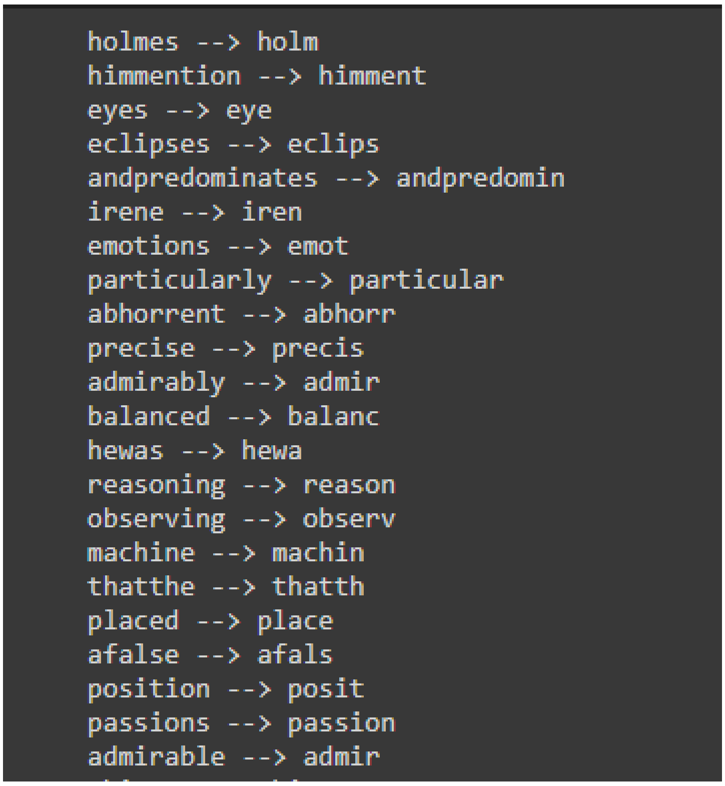
2/5

Stemming: process in which terms are transformed to their root in order to reduce the size of the vocabulary. It is carried by applying word reduction rules.

Two of the most common stemming algorithms are:

■■Porter

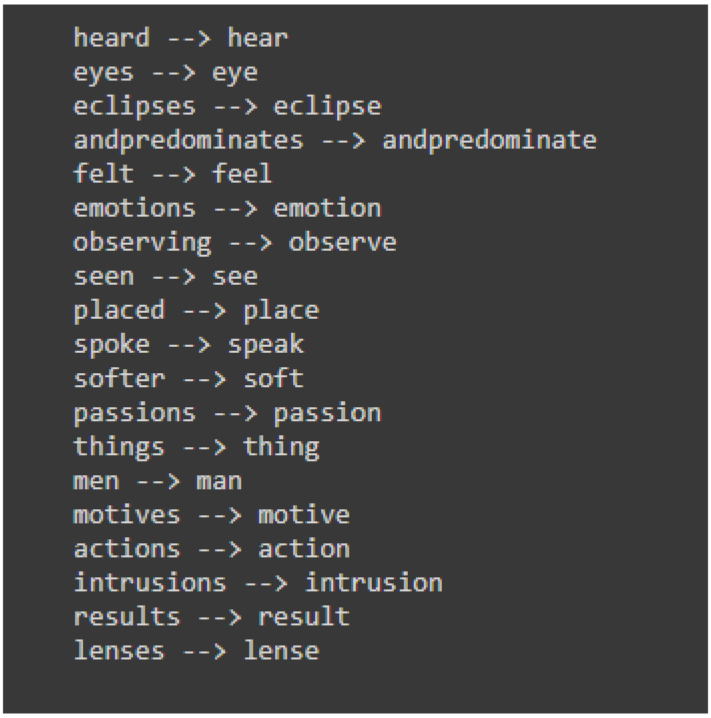
■■Snowball



```
holmes --> holm
himmention --> himment
eyes --> eye
eclipses --> eclips
andpredominates --> andpredomin
irene --> iren
emotions --> emot
particularly --> particular
abhorrent --> abhorr
precise --> precis
admirably --> admir
balanced --> balanc
hewas --> hewa
reasoning --> reason
observing --> observ
machine --> machin
thatthe --> thatth
placed --> place
afalse --> afals
position --> posit
passions --> passion
admirable --> admir
```

Lemmatization: it performs a morphological analysis using reference dictionaries to create equivalence classes between words.

For example, for the token “eclipses”, a stemming rule would return the term “eclips“, while through lemmatization we would get the term “eclipse“.



```
heard --> hear
eyes --> eye
eclipses --> eclipse
andpredominates --> andpredominate
felt --> feel
emotions --> emotion
observing --> observe
seen --> see
placed --> place
spoke --> speak
softer --> soft
passions --> passion
things --> thing
men --> man
motives --> motive
actions --> action
intrusions --> intrusion
results --> result
lenses --> lense
```

Finally, let me share a quick example on the use of these two NLP techniques (with spaCy and Python):

https://t.co/Qm0Fa4cGaV

5/5