# README

## Tokenizer

The `Token` class contains all the information about each token that we need to write in the spreadsheet.
The `tokenizer()` function first does some preprocessing on the text:\

- removes `-\n` so that the words are not split into two tokens
- replaces different versions of quotes with standard forms (' and ")
- puts a space after and before punctuations, except for hyphens
- joins apostrophe to the second token if it is a contraction (for eg- man's -> man + 's')
- fixes whitespaces so that multiple spaces or newlines are replaced by a single space

Then it finds tokens using the regex pattern: `/\w+-?\w+|\S+/`

- `\w+` matches any word character (alphanumeric & underscore)
- `-?` matches zero or one hyphen
- `\w+` matches any word character (alphanumeric & underscore)
- `|` is the OR operator
- `\S+` matches any non-whitespace character

So the pattern matches any word with or without a hyphen, or any non-whitespace character.\

If the `sentence` parameter is `True`, then the function returns 3 things: tokens, list of tokenized sentences, list of sentences.
To tokenize sentences, the function uses the regex pattern: `/(?<=[.?])\s+/`

- `(?<=[.?])` is a positive lookbehind assertion that matches a space (since the following pattern is `\s+`) that is preceded by a period or a question mark
- `\s+` matches one or more whitespace characters

PROF

## pos tagger

I have used the `nltk` library to tag the tokens with the Universal POS tagset.\

## Simplified POS

## Corrections

## Rendering tsv as xlsx

I first replaced all double quotes (") with (") in the tsv, and after opening it in excel, I replaced it back to double quotes.