LaTeX submissions are mandatory. Submitting your assignment in another format will be graded no higher than R.

1 Lab Book S1

In this Lab Book you will work with text segmentation, the first step of converting written text to speech signal.

1.1 Assignment 1

[3pts] *Tokenizing a sentence*:

Task:

Using the definition of a "word" discussed in class (based on Taylor (2009)), tokenize and verbalize (without code) the following sentences into "words":

- Space is big
- You just won't believe how vastly, hugely, mind-bogglingly big it is.
- I'm gonna write to test@example.com if you pay me €42.42.
- In 1903 Polish physicist Marie Curie became the 1st woman to win a Nobel Prize, she was 36y.o. then.

Answer

1. Tokenizing
   - `<SPACE> <IS> <BIG>`
   - `<YOU> <JUST> <WON'T> <BELIEVE> <HOW> <VASTLY> <,> <HUGELY> <,> <MIND-BOGGLINGLY> <BIG> <IT> <IS> <.>`
   - `<I'M> <GONNA> <WRITE> <TO> <TEST@EXAMPLE.COM> <IF> <YOU> <PAY> <ME> <€42.42> <.>`
   - `<IN> <1903> <POLISH> <PHYSICIST> <MARIE> <CURIE> <BECAME> <THE> <1ST> <WOMAN> <TO> <WIN> <A> <NOBEL> <PRIZE> <,> <SHE> <WAS> <36Y.O.> <THEN> <.>`
2. Verbalizing
   - `<SPACE> <IS> <BIG>`
   - `<YOU> <JUST> <WON'T> <BELIEVE> <HOW> <VASTLY> <HUGELY> <MIND-BOGGLINGLY> <BIG> <IT> <IS>`
   - `<I'M> <GONNA> <WRITE> <TO> <TEST> <WORD_AT> <EXAMPLE> <WORD_DOT> <COM> <IF> <YOU> <PAY> <ME> <FORTY-TWO> <EUROS> <FORTY-TWO> <CENTS>`
   - `<IN> <NINTEEN> <OH> <THREE> <POLISH_COUNTRY> <PHYSICIST> <MARIE> <CURIE> <BECAME> <THE> <FIRST> <WOMAN> <TO> <WIN> <A> <NOBEL> <PRIZE> <SHE> <WAS> <THIRTYSIX> <YEARS> <OLD> <THEN>`

1.2 Assignment 1

[7pts] *Exploring text normalization*:

Preparation:

We will adopt an ad-hoc approach, and assume that the text that we are working with is mostly natural language. So we will use regular expressions trying to find a pattern. Please, write your own regular expression and use *only the standard/built in* regex libraries (e.g. `re` in Python). Don't use existing advanced text processing libraries such as Spacy or Keras.

Task:

Write a script that extracts the following features from the text:

- Phone numbers in international format, They begin with a `+` sign and contain up to 15 digits separated by spaces or dashes. For simplicity assume that only the `-` ("minus sign") will be used. Examples are: `+31503638004, +31 50 363 80 04, +31 50-363-80-04`. You can also choose to limit yourself to European countries: if you do, be transparent about it.
- Web links, such as `https://www.rug.nl/info/contact`, `https://en.wikipedia.org/wiki/Regular_expression#Syntax`, `https://www.google.com/search?q=python%20re`.

As always, write good comments in the code explaining how it works. Prepare a set of sentences to test your script and submit it together with your code to the GitHub repository. Make sure the code works "as is" and doesn't require editing paths, providing commandline argument and so on.

Answer

For your convenience, I've included the test content consisting of several sentences in the codes so you may not need to edit the path or commandline argument and so on.