20-00-0546-iv Foundations of Language Technology

Homework 6
Processing Raw Text

10. December 2020

In case your submission consists of several files, compress these to a `zip`-file. Indicate clearly which submission corresponds to which question. Include comments in your program code to make it easier readable. It is very important that you submit your solution as a Jupyter Notebook file (`.ipnyb`).The deadline for the homework is **Thursday, 17.12.20 23:59 CET after the practice class**.

6.1 Homework

Homework 6.1 *(10 points) Implement an SMS decoder. Similar to the T9 system on mobile phones, your decoder should translate from digit sequences to words:*

(a) Choose at least one appropriate corpus and discuss, why you chose this corpus. You will use the corpus to estimate which word is more frequent and should be a preferred output.

(b) Implement a function `get_T9_word(digits)` which for a given sequence of digits, e.g. "252473", returns the most likely word, e.g. "Claire".

(c) Apply the decoder to each digit sequence in this "sentence":
`['43556','69','374363','73837','4','26','3463']`
The original sentence was: "hello my friend peter i am fine"
Is the output readable? What errors have been made?

Homework 6.2 *This is a quite complicated optional task that you might explore if you are interested. It will not count as part of the official homework assignment.*
Improve the SMS decoder of homework 6.1 as follows:

(a) Take the context into account, guess the word using the bigram probability of the previous entered word with the function $get\_T9\_word(prevWord, number)$. *Test the improvement with the* $(context\_word, digit)$ *tuples in the following list:*

```
print(get_T9_word('i',     '26'))
print(get_T9_word('its',   '26'))
print(get_T9_word('a',     '3463'))
print(get_T9_word('will',  '3463'))
print(get_T9_word('the',   '1111'))
```

(b) Apply the decoder to each digit sequence in this "sentence":
`['43556','69','374363','73837','4','26','3463']`
The original sentence was: "hello my friend peter i am fine"
Is the output readable? What errors have been made?