Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

1. Exercise for "Sprachverarbeitung und Text Mining"

05.11.2021

1 Knowledge Questions

1. Name six typical pre-processing steps of NLP! Which of them relate to semantics, which to syntax?

a) Syntax

- Tokenisation and sentence recognition
- Word type recognition (POS tagging)
- Phrase recognition and parsing

a) Semantics

- Entity Recognition
- Relation Recognition
- Entity Resolution

Are other steps needed to apply any of these? Give reasons for your answer!

- Some preprocessing steps require data that has already been processed. For example, the methods for stemming presented in the lecture require a prior division into tokens.

2. Which preprocessing step is typically performed together with end-of-sentence detection?

   - End-of-sentence detection is strongly associated with tokenisation.

3. What meanings can the following sentences have:

   a) This is a big drop.

      - A big drop of fluid.
      - A big decline.

   b) I greeted the woman with the hat.

      - I took off my hat to greet the woman.
      - I greeted the woman who had a hat.

   For which applications do the ambiguities of the respective sentences pose difficulties?

   - This is a big drop. (Lexical Disambiguation: Machine Translation)
   - I greeted the woman with the hat. (Syntax: Parser)

   Which three levels of text normalisation are typical for NLP tasks? Give an example of each.

4. - segmentation / tokenisation of words in the text e.g. “The cat plays with cats.” ⇒ `The|cat|plays|with|cats`
   - Normalisation of word forms
     e.g. “The cat plays with cats.” ⇒ `the|cat|play|with|cat|.`
   - Segmentation of sentences
     e.g. “The cat plays with cats. The dog plays with dogs.”
     ⇒ `<s>The cat plays with cats.</s>`
     `<s>The dog plays with dogs.</s>`

5. Name three typical (language-dependent) difficulties in tokenisation.

- Apostrophised compounds (Justice's intervention)
- Hyphenated or compound words (justice-based)
- Abbreviations (e.g. e.g.)
- Terms consisting of several words (United States of America)
- Languages without explicit word division

6. Explain lemmatisation and stemming and the difference between the two.

- Lemmatisation: reduction of word variations to a basic form of the word
- Stemming: reduction of word variations by pruning the word to a root word
- In lemmatization, words are reduced to valid base form words regardless of whether they have the same root. In stemming words are reduced to a stem that is not necessarily a valid word. Word variations that change the stem are not reduced to a common base form.

Which three stemming techniques have been introduced in the lecture?

- look-up table
- Suffix-stripping algorithms
- Stochastic algorithms

2 Regular Expressions

We4bee would like to make it possible to predict the need for beekeeping action (e.g. feeding time, swarming time, breeding condition etc.) and environmental events (e.g. earthquake) with the help of high-tech sensors developed for use in beehives and a connected BigData analysis as well as machine learning. In addition, the data analysis should contribute to a better understanding of bee behaviour, ensure the preservation of the honey bee and thus make an important contribution to environmental protection.

Bees pollinate 80% of all useful and wild plants and thus have a worldwide economic benefit of 265 billion euros. Monocultures and the associated use of pesticides threaten the honey bee and have already led to the disappearance of this animal in parts of China. Since then, migrant workers there have had to laboriously pollinate fruit trees by hand. In order to counteract this alarming development and ensure the preservation of the honey bee, we must better understand its behaviour and needs.

Therefore, we4bee would like to lend TopBar hives equipped with high-tech sensors to schools, universities and young beekeepers in order to collect and analyse data transmitted on temperature, humidity, air pressure, weight, sound/vibration, light intensity and fine dust pollution. The aim is to place half of the hives in urban areas and the other half in rural areas in Germany in order to determine any differences between urban and rural bees.

In order to improve the understanding for the importance of the honey bee for humans and our environment, the educational facilities are supplied besides with interdisciplinary teaching material approximately around the bee. A web and a mobile app also allow users to clearly display the data collected for each beehive. In addition, the students/users/participants are given an insight into the fascinating world of embedded computer and high-tech environmental sensors, which are also used in automotive engineering at Audi.

It goes without saying that all the data collected will be made accessible to the general public.

Project lead: we4bee
Dr. Max Mustermann
Universitätsbibliothek Würzburg
Am Hubland
97074 Würzburg
Tel. 0931 / 31 12345
mustermax@bibliothek.uni-wuerzburg.de

Specify regular expressions (if possible) to extract exactly the following data from the above document:

1. all instances of the name of the project “we4bee”

   `[W|w]e4bee`

2. all words

   `[\wäü]+`

3. all sentences

   `[^.]+.`
   here the definition of the term sentence is important. This naive regex also matches abbreviation points, which presumably should not be interpreted as sentence boundaries. This is where Regex reaches its limits and one has to resort to more complex procedures.

4. the first word of each line

   `^[\wöäüß]+` *Requires multiline processing in regex parser*

5. all words followed by a whitespace (excluding the whitespace)

   `[\wäü]+(?= )`

6. all words containing a hyphen

   `([\wäü]+[-])+[\wäü]+(?= )`

7. postcode and city

   `\d{5} .+`

8. telephone number

```
\d[ \d]*/[\d ]+
```

9. email address

```
.+@.+
```

10. all verbs in infinitive

```
(like|predict|use|ensure|have|counteract|understand|lend|
collect|analyse|place|determine|improve|allow|display)(?= )
```

With regex, without further preprocessing, only the enumeration of the searched words remains.

Note: You can use your regular expressions online for example on the website `http://www.regexr.com/`.

3 Tokenisation and Evaluation

Given is a tokenizer described by the following procedure:

- Replace each space with a word boundary.
- Insert a word boundary before each character "."

Apply the tokenizer described above to the input text below.

> With the emergence of Corona, Prof. Dr. Med. Wurst measured an increase in patients by ca. 50%.

Determine the number of true-positives, false-positives and false-negatives by comparing the output of the system with the desired solution. Desired is the tokenisation result marked as gold standard.

Note: A true-positive is a token that appears in both the gold standard and the solution of the system. A false positive is a token that appears only in the system solution. A false negative is a token that appears only in the gold standard.

Gold standard: ("_" corresponds to word boundary)
With_the_emergence_of_Corona_,_Prof._Dr._Med._Wurst_measured_an_increase_in_patients_by_ca._50%_.

> System:
> With_the_emergence_of_Corona,_Prof_._Dr_._Med_._Wurst_measured_an_increase_in_patients_by_ca_._50%_.
>
> TP: Search all tokens that occur in both lists (green).
>
> GLD: With_the_emergence_of_Corona_,_Prof.__Dr.__Med.__
> SYS: With_the_emergence_of_Corona,__Prof_._Dr_._Med_._
>
> GLD: Wurst_measured_an_increase_in_patients_by_ca.__50%_._
> SYS: Wurst_measured_an_increase_in_patients_by_ca_._50%_._
>
> FN: Search all tokens that occur in gold but not in system.
>
> GLD: With_the_emergence_of_Corona_,_Prof.__Dr.__Med.__
> SYS: With_the_emergence_of_Corona,__Prof_._Dr_._Med_._
>
> GLD: Wurst_measured_an_increase_in_patients_by_ca.__50%_._
> SYS: Wurst_measured_an_increase_in_patients_by_ca_._50%_._
>
> FP: Search all tokens that occur in system but not in gold.
>
> GLD: With_the_emergence_of_Corona_,_Prof.__Dr.__Med.__
> SYS: With_the_emergence_of_Corona,__Prof_._Dr_._Med_._
>
> GLD: Wurst_measured_an_increase_in_patients_by_ca.__50%_._
> SYS: Wurst_measured_an_increase_in_patients_by_ca_._50%_._

> Number tp: 13
> Number of fp: 9
> Number fn: 6

How would you solve the problems you found?

> Use a more sophisticated tokenizer.
