

# Statistical NLP

Einführung in die Computerlinguistik – Übung 20.11.2015

# Write

1. Open console (Use Putty on Windows)

2. Only if you are on your private computer

```
ssh username@remote.cip.ifi.lmu.de
```

3. Login to my machine

```
ssh tbd
```

4. Enable chat

```
msg y
```

5. Send answers or questions

```
echo "bla bla" | write rothes
```

# Independence Test

1. Download the script "munge.py"

2. Go to the folder and run:

```
python munge.py surprisemultiple 10 10 1 100
```

3. What does each of these commands/arguments mean?

# Independence test

4. Use Google counts and find words that are statistical dependent in Wikipedia.

`site:de.wikipedia.org München`

Report the surprise value

5. Find words that are statistical independent

6. Find words that are statistical dependent and negative correlated

# Independence test

7. Modify the script so that it takes the probabilities of  $w_1$ ,  $w_2$  and joint as input (instead of the counts). You do not have to add or remove lines. Please don't do so, so we can refer to line numbers.

- Line 83 reads 4 Integer arguments. We need 3 Floats.
- Line 58-60 show you how to read Ints or Floats
- The arguments are written into local variables in line 6-9
- The script calculates probabilities (e.g. in line 15). We don't have to do this anymore.

# NLTK

8. Start interactive python session

```
python
```

9. Import nltk

```
import nltk
```

10. Tokenize a sentence of your choice

```
nltk.tokenize.word_tokenize("bla bla")
```

11. POS Tag the result

```
nltk.pos_tag(text)
```

# NLTK

12. Write a Python script that returns the POS tagged sentence given this input

```
python posTagger.py "bla bla"
```

13. POS Tag the following sentences:

- book
- the book is great
- book a flight
- the representative put chairs on the table

# Google Translate

Find Google Translate errors for each of the following categories:

14. Sayings

15. Compounds

16. Other