block6b-nltk-textprocessing

September 24, 2018

1 Block 6b: NLTK

```
In [ ]: import collections

        freq = collections.Counter(text)
```

```
In [ ]: print(freq.most_common(25))
```

```
In [ ]: import string

        def remove_punctuation(text):
            punctuation = string.punctuation
            for marker in punctuation:
                text = text.replace(marker, "")
            return text
```

```
In [ ]: import nltk

        with open("grundgesetz.txt", encoding="utf-8") as infile:
            text_raw = infile.read()

        text = text_raw.lower()
        text = remove_punctuation(text)
        text = text.split()
        text = nltk.Text(text)
```

```
In [ ]: text.concordance("freiheit")
```

```
In [ ]: text.similar("freiheit")
```

```
In [ ]: text.dispersion_plot(["artikel", "gesetz", "freiheit"])
```

1.1 NLTK-Beispiel

(aus: http://www.nltk.org/book/ch02.html)

```
In [ ]: import nltk
        from nltk.corpus import inaugural

        cfd = nltk.ConditionalFreqDist(
                  (target, fileid[:4])
                  for fileid in inaugural.fileids()
                    for w in inaugural.words(fileid)
                       for target in ['america', 'citizen']
                           if w.lower().startswith(target))
        cfd.plot()
```