Tokenization

- Was umfasst ein Wort?
  - “they'll want to leave”
  - Format/Abkürzungen: U.S.A, $200
- Regulärer Ausdruck: String mit Muster

```
pattern = r'\w+|[^\w\s]+'
```

  - split(): trennt String bei Leerzeichen
  - nltk.tokenize.regexp_tokenize(plain_text, pattern)

Satz Extraktion

- Was umfasst einen Satz?
- “Have you called me last week?”
- The method is quite novel: You put two potatoes in one pan.
- nltk.tokenize.regexp_tokenize(plain_text, pattern)

Satz Extraktion

- Punkt Tokenizer
  - trennt Text in einzelne Sätze
  - unüberwachtes Lernverfahren, trainiert auf einem großen Textkorpus
  - 'tokenizers/punkt/english.pickle'
  - http://nltk.org/doc/guides/tokenize.html

Satz Extraktion

```
>>> import nltk.data
>>> text = """
... Punkt knows that the periods in Mr. Smith and Johann S. Bach
... do not mark sentence boundaries.  And sometimes sentences
... can start with non-capitalized words.  i is a good variable
... name.
... """
>>> tokenizer = nltk.data.load('tokenizers/punkt/english.pickle')
>>> print '\n-----\n'.join(tokenizer.tokenize(text))
Punkt knows that the periods in Mr. Smith and Johann S. Bach
do not mark sentence boundaries.
-----
And sometimes sentences
can start with non-capitalized words.
-----
i is a good variable
name.
```

Evaluation: Metriken

- Eigenes Ergebnis womit vergleichen?
  - nltk.corpus.treebank.words()
  - nltk.corpus.treebank.sents()
- set(reference_set).intersection(test_set) / float(len(set(reference_set)))
- import math (for floating point operations)