**Ausgangsdaten:**

* Liste der Wörter (fast 2 Mio mit Formen)
* Duden (library in python):
  + Bedeutungsübersicht
  + Synonyme
  + Antonyme
  + Typische Verbindungen
  + Bedeutungen, Beispiele, Wendungen
* Child Corpora (mit CRAN)

**Anzuschauen:**

* Häufigkeiten jedes Präfixes, wie viele Wörter mit gegebenem Präfix
* (opt. eher mit Substantive oder Verb)
* Eher mit welcher Präposition
* Die nächsten Präfixe (in den Synonymen)
* Die unterschiedlichsten Präfixe (in den Antonymen)
* Sentiment Analysis (positive/ negative/ neutrale Bedeutung)
* Term frequency – inverse document frequency (TF–IDF looks at how frequently a word appears in a document and its importance relative to the whole set of documents)
* Wie viele Varianten gewöhnlich ein Wurzel hat (mit wie vielen Präfixen und welchen) (und wenn eine gibt es – welche W^keit, dass eine andere kommt. Bayes, conditional probability)
* Wie unterschiedlich die Bedeutung desselben Wurzels mit verschiedenen Präfixen
* Topic modelling (probabilistic latent semantic indexing)