Handout Lemmatisierung

Das Wort Lemma stammt aus dem altgriechischen und bedeutet so viel wie „das Angenommene“ oder „die Prämisse“. Es bezeichnet das Wort im lexikalischen Sinne, also die Wortform welche im Lexikon abgebildet wird und ist immer abhängig von dem jeweiligen Anwendungsgebiet.

Lemma = gehen Stamm = geh

Lemmatisierung ist die Reduktion der Wortform auf ihre Grundform und wird auch lexikonbasiertes Stemming genannt. Dies wird dadurch erreicht, dass die erreichte Grundform in einem elektronischen Wörterbuch nachgeschlagen wird. In einem Vollformenlexikon steht jede Wortform, welches die Suche effizienter macht aber mehr Speicherplatz erfordert. Handelt es sich bei dem Nachschlagewerk um ein Grundformenlexikon wird die Wortform durch morphologische Regeln auf eine potenzielle Grundform reduziert und dann nachgeschlagen. Der Vorgang verringert die Gesamtzahl unterschiedlicher Wortformen in Sprachdaten und erhöht die Stabilität und Signifikanz statischer Analysen. Es wird sowohl zum korpusbasierten Lernen als auch als Basis semantischer Analyse benutzt. Lemmatisierer arbeiten mit bereits tokenisierten und POS-getaggten Daten.

Nachteile der Lemmatisierung sind die umfangreichen elektronischen Wörterbücher, welche aufwendig in der Erstellung und Wartung sind. Ebenfalls problematisch sind Wortformen, die nicht im Lexikon vorkommen. In der Regel findet keine Derivationsanalyse statt.
Lemmatisierung in der deutschen Sprache weist das Problem der Kompositazerlegung auf, welche notwendig wäre, aber erschwert wird durch verschiedene korrekte Zerlegungsmöglichkeiten.

Beim Stemming werden Wörter auf einen gemeinsamen Wortstamm zurückgeführt. Die Stammformreduktion benutzt kein Lexikon, sondern als erstes wird die Anzahl der Vokale-Konsonanten-Sequenzen bestimmt. Danach werden Verkürzungsregeln angewendet, also Paare von Bedinungen und Ableitungen für verschiedene Suffixe. Durch die verschiedenen Durchgänge, in denen jeweils unterschiedliche Suffixe entfernt werden, wird die Abfolge von Flexions- und Derivationssuffixen dargestellt.

Allerdings kann die Entfernung von Suffixen durch Bedingungen an den verbleibenden Stamm eingeschränkt werden und künstliche Ambiguitäten können auftreten. Ebenfalls kann es beim Stemming vorkommen, dass verwandete Wortformen nicht identifiziert werden können und es zur Erzeugung nicht existenter Stämme kommen kann. Auch die aufwendige Testphase ist ein Nachteil des Stemmings.

Stemming Algorithmen sind oftmals deutlich schneller aber auch weniger umfangreich als Lemmatisierer. Beide haben ihre Nachteile vor allem in Anwendung mit der deutschen Sprache. Stemming wird hier durch wortinterne Prozesse, Komposita und ihre Zerlegung behindert. Bei der Lemmatisierung ist es die Frage, was als Lemma bei der systematischen Konversion angenommen wird. Ein weiteres Problem stell hier die Komposition dar, da es nicht möglich ist alle Komposita im Lexikon aufzunehmen. Das macht ihre Zerlegung so wichtig. Neben diesen beiden Verfahren sind Suchmaschinen mit Trunkierungen als weitere Option vertreten. Diese sind allerdings am simpelsten und haben das Problem alles aufzuzeigen ohne Relevanz zu überprüfen.