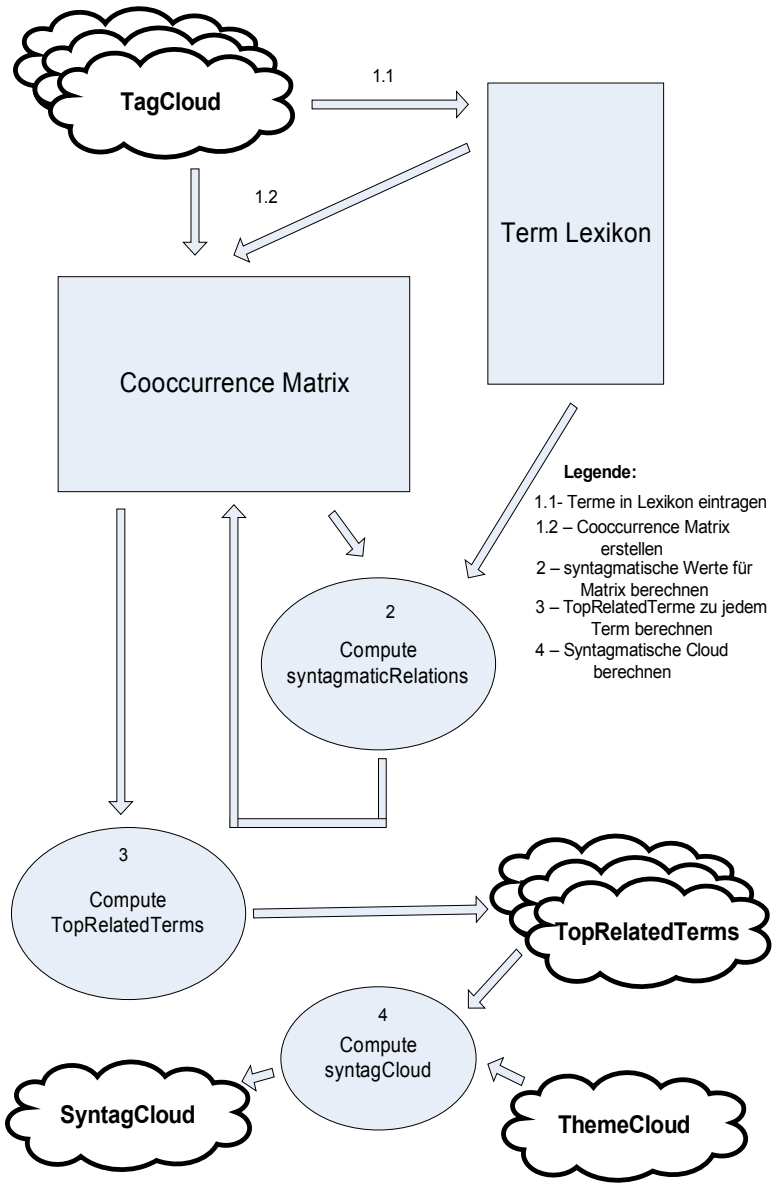
Erzeugung von ThemeClouds unter Nutzung von syntagmatischen Clouds:

Eine syntagmatische Cloud ist im TagHandler eine Menge von Termen, die anhand einer weiteren Menge von Termen und Parametern berechnet wird.
Die berechneten Terme der syntagmatischen Cloud sollen zu den angebenen Termen passen und so die Erzeugung von sogenannten ThemeClouds erleichtern.
Eine ThemeCloud ist eine vom Nutzer erstellte Menge von Termen, die in einem Zusammenhang zueinander stehen. Einer ThemeCloud muss ein Bezeichner vergeben werden. Terme der ThemeCloud können bewertet und als nicht geeignet (Blacklist) markiert werden.

Beispiel:

1. ThemeCloud : leer, Nutzer setzt Bezeichner: „Objektorientierte Programmiersprachen“
2. Nutzer fügt den Term „java“ zu ThemeCloud hinzu – ThemeCloud: „java“
3. syntagmatische Cloud wird berechnet : syntagCloud: „c++, c, Delphi“
4. Nutzer fügt zu ThemeCloud hinzu: „c++“ und „Delphi“; er fügt „c“ der Blacklist hinzu
5. ThemeCloud: „java, c++, Delphi“, berechnete syntagmatische Cloud: „Perl, Ruby, Groovy, Object Pascal“
6. u.s.w.



Berechnung einer syntagmatischen Cloud:

1. Voraussetzungen:

- Term Lexikon
- Cooccurrence-Matrix
- topRelatedTerms

Es wurde ein Term-Lexikon angelegt, in dem jeder Term mit einer eindeutigen ID versehen ist. Das Lexikon basiert auf einer Menge von TagClouds, von denen angenommen wird, das die meisten Tags innerhalb dieser TagClouds etwas miteinander zu tun haben.
Es wird gezählt, wie oft jeder Tag insgesamt aufgetreten ist („termFrequency“) und wie oft Tag-Paare gemeinsam in einer Tag-Cloud gefunden wurden (Cooccurrence, kurz „cooc“).
Zu jedem Term wird mit Hilfe der Cooccurrence-Matrix berechnet, welche anderen Terme gut zu ihm passen („topRelatedTerms“). Zu jedem topRelatedTerm wird ein syntagmatischer Wert unter Nutzung eines statistischen Verfahrens, z.b. der LogLikelihood-Formel, berechnet.

2. Berechnung:

2.1 Berechnung der syntagmatischen Cloud

Folgende Parameter werden genutzt:

a) Modifikatoren für die Formel $\sum(1..n)\ 1 / (a * topRelatedTermRank\_n) + b)$:
- a (eine Gleitkommazahl)
- b (eine Gleitkommazahl)
- topRelatedTermRank (wie gut ist der topRelatedTerm, Rang == 1 ist der beste Rang) Der Rang ergibt sich aus dem berechneten syntagmatischen Wert zu dem topRelatedTerm.

b) minSyntag (der syntagmatische Wert der topRelatedTerme darf diesen Wert nicht unterschreiten)
c) syntagmaticEntityTermFactor (max. Verhältnis Anzahl gefundener syntagCloud-Terme zu Anzahl Terme in themeCloud)
d) maxTopRelatedTerms (so viele topRelatedTerme werden maximal für die Berechnung verwendet)

Für die Berechnung der syntagmatischen Cloud werden nur Terme genutzt, die topRelatedTerme zu den Termen der ThemeCloud sind und deren syntagmatischer Wert, der bei der topRelatedTerm-Berechnung ermittelt wurde, nicht unter den minSyntag fällt. Aus den syntagmatischen Werten ergeben sich die Ränge der topRelatedTerme (topRelatedTermRank), wobei der höchste Rang dem höchsten syntagmatischen Wert entspricht. Es werden maximal so viele topRelatedTerme verwendet, wie im Parameter „maxTopRelatedTerms“ angegeben wurden.

Mit der Formel ∑(1..n) 1 / (a * topRelatedTermRank_n) + b) werden nun Scores zu jedem topRelatedTerm berechnet. Die topRelatedTerme werden nach ihrem Score sortiert.
Ist die Anzahl der gefunden Terme größer als die Anzahl der ThemeCloud Terme multipliziert mit dem syntagmaticEntityTermFactor, so wird die Liste der gefunden Terme gekürzt, so das sie maximal um den gegebenen Factor größer ist als die Liste der ThemeCloud- Terme.
Diese Liste der gefunden Terme ergibt die syntagmatische Cloud.

2.2 Berechnung der topRelatedTerms:

Die Terme, die gut zu einem einzelnen Term passen, werden mittels eines statistischen Verfahrens ermittelt. Momentan wird dazu der LogLikelihood Algorithmus (Likelihood-Quotienten-Test) genutzt. Das Ergebnis ist ein syntagmatischer Wert für jeden einzelnen Term bezogen auf den Term, zu dem er passen könnte. Je höher dieser Wert ist, umso besser passt der Term.