Boolesches Retrieval

Anhang: Regeln des Porter-Stemmers

Porter-Stemmer

Porter-Stemmer (als Beispiel eines "Affix Removal"-Verfahrens)

Der Algorithmus des Porter-Stemmers besteht aus (Bedingung:Aktion)-Regeln.

Bedingungen:

Vokale sind hier *a*, *e*, *i*, *o*, *u* und, falls ein Konsonant vorausgeht, auch *y*.

Sei *C* eine Folge von Konsonanten, *V* ein Folge von Vokalen.

Das Maß *m* eines Stamms *S* ist definiert durch die Anzahl der *VC*-Sequenzen in *S*:

$$[C]\,(VC)^m\,[V]$$

Beispiel:

m = 0:	TR, EE, TREE, Y, BY
m = 1:	TROUBLE, OATS, TREES, IVY
m = 2:	TROUBLES, PRIVATE, OATEN

Porter-Stemmer

- * <*X*> : der Stamm endet mit einem bestimmten Buchstaben *X*
- * v * : der Stamm enthält einen Vokal
- *d : der Stamm endet mit einem Doppel-Konsonant
- *o : der Stamm endet mit einer Konsonant-Vokal-Konsonant-Folge und der letzte Konsonant ist kein *w*, *x* oder *y*

Suffix-Bedingungen: (current_suffix = = pattern)
Regel-Bedingungen: (rule which was used)

Aktionen: (old_suffix → new_suffix)

Porter-Stemmer

Algorithmus:

step1a (word);
step1b (stem);
if (second or third rule of step 1b was used) step1b1 (stem);
step1c (stem);
step2 (stem);
step3 (stem);
step4 (stem);
step5a (stem);
step5b (stem);

Porter-Stemmer

Step 1a

Conditions	Suffix	Replacement	Examples
NULL	sses	ss	caresses → caress
NULL	ies	i	ponies → poni
			ties → ti
NULL	ss	ss	carress → carress
NULL	s	NULL	cats → cat

Die Regeln innerhalb eines "Steps" werden nacheinander auf Anwendbarkeit überprüft, nur eine wird angewendet. Es wird das größt mögliche Suffix ersetzt.

Porter-Stemmer

Step 1b

Conditions	Suffix	Replacement	Examples
(m > 1)	eed	ee	feed → feed agreed → agree
(*v*)	ed	NULL	plastered → plaster bled → bled
(*v*)	ing	NULL	motoring → motor sing → sing

Porter-Stemmer

Step 1b1

Conditions	Suffix	Replacement	Examples
NULL	at	ate	conflat(ed) → conflate
NULL	bl	ble	troubl(ing) → trouble
NULL	iz	ize	siz(ed) → size
(*d ∧ ¬(*<L>∨*<S>∨*<Z>)	NULL	single letter	hopp(ing) → hop tann(ed) → tan fall(ing) → fall
(m=1 ∧ *o)	NULL	e	fail(ing) → fail fil(ing) → file

Porter-Stemmer

Step 1c

Conditions	Suffix	Replacement	Examples
(*v*)	y	i	happy → happi sky → sky

Porter-Stemmer

Step 2

Conditions	Suffix	Replacement	Examples
(m>0)	ational	ate	relational → relate
(m>0)	tional	tion	conditional → condition rational → rational
(m>0)	enci	ence	valenci → valence
(m>0)	anci	ance	hesitanci → hesitence
(m>0)	izer	ize	digitizer → digitize

(m>0): bli → ble, alli → al, entli → ent, eli → e, ousli → ous, logi → log,
ization → ize, ation → ate, ator → ate, alism → al, iveness → ive,
fulness → ful, ousness → ous, aliti → al, iviti → ive, biliti → ble

Porter-Stemmer

Step 3

Conditions	Suffix	Replacement	Examples
(m>0)	icate	ic	triplicate → triplic
(m>0)	ative	NULL	formative → form
(m>0)	alize	al	formalize → formal
(m>0)	iciti	ic	electricity → electric
(m>0)	ical	ic	electrical → electric
(m>0)	ful	NULL	hopeful → hope
(m>0)	ness	NULL	goodness → good

Porter-Stemmer

Step 4

Conditions	Suffix	Replacement	Examples
(m>1)	al	NULL	revival → reviv
(m>1)	ance	NULL	allowance → allow
(m>1)	ence	NULL	inference → infer
(m>1)	er	NULL	airliner → airlin
(m>1)	ic	NULL	gyroscopic → gyroscop
(m>1)	able	NULL	adjustable → adjust
(m>1)	ible	NULL	defensible → defens
(m>1)	ant	NULL	irritant → irrit
(m>1)	ement	NULL	replacement → replac
(m>1)	ment	NULL	adjustment → adjust
...			

Porter-Stemmer

Step 4

Conditions	Suffix	Replacement	Examples
...			
(m>1)	ent	NULL	dependent → depend
(m>1) ∧ (*<T>)	ion	NULL	adoption → adopt
(m>1) ∧ (*<S>)	ion	NULL	
(m>1)	ou	NULL	homologou → homolog
(m>1)	ism	NULL	communism → commun
(m>1)	ate	NULL	activate → activ
(m>1)	iti	NULL	angulariti → angular
(m>1)	ous	NULL	homologous → homolog
(m>1)	ive	NULL	effective → effect
(m>1)	ize	NULL	bowdlerize → bowdler

Porter-Stemmer

Step 5a

Conditions	Suffix	Replacement	Examples
(m>1)	e	NULL	probate → probat rate → rate
(m=1 ∧ ¬ *o)	e	NULL	cease → ceas

Step 5b

Conditions	Suffix	Replacement	Examples
(m>1 ∧ *d ∧ *<L>)	NULL	single letter	controll → control roll → roll

Porter-Stemmer

Infos und Quellcodes:

- zu Snowball ("a small string processing language for creating stemmers") Stemmer-Varianten für verschiedene Sprachen:

  http://snowball.tartarus.org/

- speziell zu Porter-Stemmer:

  http://www.tartarus.org/~martin/PorterStemmer/

Boolesches Retrieval

Anhang: Herausforderungen für einen deutschen Stemmer

Preprocessing: Deutsch-"Stemmer"

Flektierte Verben

(ich)	**zahle**, **zahlte**, **gezahlt**
(du)	**zahlst**, **zahltest**, gezahlt
(sie, er)	**zahlt**, **zahlte**, gezahlt
(wir)	**zahlen**, **zahlten**, gezahlt
(ihr)	zahlt, **zahltet**, gezahlt
(sie)	zahlen, zahlten, gezahlt

(ich)	zahle an, zahlte an, **anzahlen**, **angezahlt**, **anzuzahlen**
...	...

→ Präfixe, Infixe, Suffixe

Preprocessing: Deutsch-"Stemmer"

Unregelmäßige Verben

nehmen, nahm, genommen, nähmest, etc.

schwimmen, schwamm, geschwommen, schwömme, schwämme, etc

gehen, ging, gegangen, etc.

→ Stammbildung nur über Vollformenlexikon möglich

Preprocessing: Deutsch-"Stemmer"

Präfixe

nehmen, ab/nehmen, zu/nehmen, weg/nehmen, ver/nehmen

→ Präfixe können bedeutungstragend sein und dürfen nicht grundsätzlich entfernt werden

Durch Entfernung können gegensätzliche Begriffe gleichgemacht werden:

zu/nehmen → nehmen
ab/nehmen → nehmen

Preprocessing: Deutsch-"Stemmer"

Präfixe

geben, ge/geben
fahren, ge/fahren

→ "ge" als Präfixe entsteht auf Grund der Flexion und kann prinzipiell entfernt werden.

aber:
Gefahren sind wir 2 Stunden ...
Gefahren gab es viele ...

Preprocessing: Deutsch-"Stemmer"

Infixe

abnehmen, abzunehmen
abgeben, abzugeben

abgeben, abgegeben
abtragen, abgetragen
abnehmen, abgenommen
anziehen, angezogen

→ Die Infixe "zu" und "ge" entstehen auf Grund der Flexion und können prinzipiell entfernt werden.
("zu": einfach, "ge": schwieriger)

Preprocessing: Deutsch-"Stemmer"

Algorithmus

Vokale:

a e i o u y ä ö ü

Vorbereitung:

Ersetze "ß" durch "ss".

Ersetze "u" und "y" zwischen Vokalen durch "U" und "Y"
(damit Handhabung als Konsonant, z.B.: baUer)

Bestimme die Regionen R1 und R2

Preprocessing: Deutsch-"Stemmer"

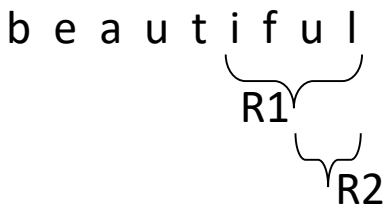
Regionen R1 und R2

- R1 ist die Region nach dem ersten Konsonanten, der einem Vokal folgt, oder die Null-Region am Ende des Wortes, falls kein derartiger Konsonant existiert.
  Die Region ***vor*** R1 muss mindestens 3 Zeichen lang sein.

- R2 ist die Region nach dem ersten Konsonanten, der in R1 einem Vokal folgt, oder die Null-Region am Ende des Wortes, falls kein derartiger Konsonant existiert.

Preprocessing: Deutsch-"Stemmer"

Beispiele für R1 und R2:



„t“ ist der erste Konsonant nach einem Vokal → R1: „iful“
„f“ ist der erste Konsonant nach einem Vokal in R1 → R2: „ul“

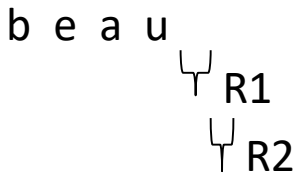
Preprocessing: Deutsch-"Stemmer"

Beispiele für R1 und R2:

„t“ ist der erste Konsonant nach einem Vokal → R1: „y“
R1 enthält keinen Konsonanten → R2: Null-Region

Preprocessing: Deutsch-"Stemmer"

Beispiele für R1 und R2:



R1 und R2 sind Null-Regionen

Preprocessing: Deutsch-"Stemmer"

Führe jeden der folgenden 3 Schritte aus:

1. Schritt:

Suche nach dem längsten der folgenden Suffixe:

(a) "e" "em" "en" "ern" "er" "es"
(b) "s" (mit einer vorangehenden gültigen "s"-Endung)

und lösche falls in R1.

Gültige "s"-Endungen: b, d, f, g, h, k, l, m, n, r, t
Gültige "st"-Endungen: b, d, f, g, h, k, l, m, n, t

z.B.: äckern -> äck, ackers -> acker, armes -> arm

Preprocessing: Deutsch-"Stemmer"

2. Schritt:

Suche nach dem längsten der folgenden Suffixe:

(a) "en" "er" "est"
(b) "st" (mit einer vorangehenden gültigen "st"-Endung ,
ihrerseits mit mindestens 3 vorangehenden Zeichen)

und lösche falls in R1.

z.B.: derbsten -> derbst (Schritt 1), und
derbst -> derb
(Schritt 2, "b" ist eine gültige "st"-Endung, der 3 Zeichen vorausgehen)

Preprocessing: Deutsch-"Stemmer"

3. Schritt:

Suche nach dem längsten der folgenden Suffixe und führe die entsprechende Aktion aus:

"end" "ung" (belebend, Beurteilung, Beendigung)

Lösche wenn in R2.
Wenn zudem "ig" aber nicht "eig" vorausgeht,
lösche wenn in R2.

"ig" "ik" "isch" (wackelig, Grammatik, wählerisch,
aber: Blätterteig, Rindfleisch)

Lösche wenn in R2 und kein "e" vorausgeht

Preprocessing: Deutsch-"Stemmer"

noch 3. Schritt:

"lich" "heit" (absonderlich, wunderlich, Gesundheit, Geborgenheit)
Lösche wenn in R2.
Falls "er" oder "en" vorausgeht, lösche wenn in R1.

"keit" (Einsamkeit, Freundlichkeit, Friedfertigkeit)
Lösche wenn in R2.
Falls "lich" oder "ig" vorausgehen, lösche wenn in R1.

Preprocessing: Deutsch-"Stemmer"

letzter Schritt:

Ersetze "U" und "Y" durch "y" und "u"

Ersetze "ä" "ö" "ü" durch "a" "o" "u"

Preprocessing: Deutsch-"Stemmer"

nehme	nehm
nehmen	nehm
nehmend	nehmend
nehmenden	nehmend
nehmet	nehmet
nehmt	nehmt

aufeinanderfolge	aufeinanderfolg
aufeinanderfolgen	aufeinanderfolg
aufeinanderfolgend	aufeinanderfolg
aufeinanderfolgende	aufeinanderfolg
aufeinanderfolgenden	aufeinanderfolg
aufeinanderfolgender	aufeinanderfolg
aufeinanderfolgt	aufeinanderfolgt
aufeinanderfolgten	aufeinanderfolgt