
NLP

Text Similarity

Morphological Similarity: Stemming

Morphological Similarity

- Words with the same root:
  - scan (base form)
  - scans, scanned, scanning (inflected forms)
  - scanner (derived forms, suffixes)
  - rescan (derived forms, prefixes)
  - rescanned (combinations)

Stemming

- To stem a word is to reduce it to a base form, called the *stem*, after removing various suffixes and endings and, sometimes, performing some additional transformations
- Examples
  - *scanned* → *scan*
  - *indication* → *indicate*
- In practice, prefixes are sometimes preserved, so *rescan* will not be stemmed to *scan*

Porter's Stemming Method

- Porter's stemming method is a rule-based algorithm introduced by Martin Porter in 1980
- The paper ("An algorithm for suffix stripping") has been cited more than 7,000 times according to Google Scholar
- The input is an individual word. The word is then transformed in a series of steps to its stem
- The method is not always accurate

Porter's Algorithm

- Example 1:
  - Input = *computational*
  - Output = *comput*
- Example 2:
  - Input = *computer*
  - Output = *comput*
- The two input words end up stemmed the same way

Porter's Algorithm

- The *measure* of a word is an indication of the number of syllables in it
  - Each sequence of consonants is denoted by C
  - Each sequence of vowels is denoted as V
  - The initial C and the final V are optional
  - So, each word is represented as [C]VCVC ... [V], or [C](VC){k}[V], where *k* is its *measure*

Examples of Measures

- k=0: I, AAA, CNN, TO, GLEE
- k=1: OR, EAST, BRICK, STREET, DOGMA
- k=2: OPAL, EASTERN, DOGMAS
- k=3: EASTERNMOST, DOGMATIC

Porter's Algorithm

- The initial word is then checked against a sequence of transformation patterns, in order.
- An example pattern is:
  - `(m>0) ATION -> ATE` `medication -> medicate`
- Note that this pattern matches *medication* and *dedication*, but not *nation*.
- Whenever a pattern matches, the word is transformed and the algorithm restarts from the beginning of the list of patterns with the transformed word.
- If no pattern matches, the algorithm stops and outputs the most recently transformed version of the word.

Example Rules

- Step 1a

SSES	->	SS	presses	->	press
IES	->	I	lies	->	li
SS	->	SS	press	->	press
S	->	ø	lots	->	lot

- Step 1b

(m>0) EED -> EE refereed -> referee

(doesn't apply to bleed since m('BL')=0)

Example Rules

- Step 2

(m>0)	ATIONAL	-> ATE	inflational	->	inflate
(m>0)	TIONAL	-> TION	notional	->	notion
(m>0)	IZER	-> IZE	nebulizer	->	nebulize
(m>0)	ENTLI	-> ENT	intelligentli	->	intelligent
(m>0)	OUSLI	-> OUS	analogousli	->	analogous
(m>0)	IZATION	-> IZE	realization	->	realize
(m>0)	ATION	-> ATE	predication	->	predicate
(m>0)	ATOR	-> ATE	indicator	->	indicate
(m>0)	IVENESS	-> IVE	attentiveness	->	attentive
(m>0)	ALITI	-> AL	realiti	->	real
(m>0)	BILITI	-> BLE	abiliti	->	able

Example Rules

- Step 3

(m>0) ICATE -> IC	replicate -> replic
(m>0) ATIVE -> ø	informative -> inform
(m>0) ALIZE -> AL	realize -> real
(m>0) ICAL -> IC	electrical -> electric
(m>0) FUL -> ø	blissful -> bliss
(m>0) NESS ->	tightness -> tight

- Step 4

(m>1) AL -> ø	appraisal -> apprais
(m>1) ANCE -> ø	conductance -> conduct
(m>1) ER -> ø	container -> contain
(m>1) IC -> ø	electric -> electr
(m>1) ABLE -> ø	countable -> count
(m>1) IBLE -> ø	irresistible -> irresist
(m>1) EMENT -> ø	displacement -> displac
(m>1) MENT -> ø	investment -> invest
(m>1) ENT -> ø	respondent -> respond

Examples

- Example 1:
  - Input = *computational*
  - Step 2: replace *ational* with *ate*: *computate*
  - Step 4: replace *ate* with *ø: comput*
  - Output = *comput*
- Example 2:
  - Input = *computer*
  - Step 4: replace *er* with *ø*: *comput*
  - Output = *comput*
- The two input words end up stemmed the same way

External Pointers

- Online demo
  - http://text-processing.com/demo/stem/
- Martin Porter's official site
  - http://tartarus.org/martin/PorterStemmer/

Quiz

- How will the Porter stemmer stem these words?

```
construction    ?
increasing      ?
unexplained     ?
differentiable  ?
```

- Check the Porter paper (or the code for the stemmer) in order to answer these questions.
- Is the output what you expected? If not, explain why.

Answers to the Quiz

construction	?
increasing	?
unexplained	?
differentiable	?

construction	**construct**
increasing	**increas**
unexplained	**unexplain**
differentiable	**differenti**

NACLO Problem

- Thorny Stems, NACLO 2008 problem by Eric Breck
  - http://www.naclo.cs.cmu.edu/assets/problems/NACLO08h.pdf

Solution to the NACLO Problem

- Thorny Stems
  - http://www.naclo.cs.cmu.edu/problems2008/N2008-HS.pdf


NLP