stemming

Stemming refers to the reduction of words to their root stems, which is helpful in information retrieval applications. For example:

connection, connected, connective, connecting	➔	connect
train, trains, train's, trains'	➔	train
the child's trains are connected	➔	the child train be connect

It is important to note that this analysis is entirely limited to English.

The Xiapian Project - Stemming

Porter's Stemming Algorithm (1980)

The algorithm comes from a central idea beginning with breaking each word into a group of vowels (a, e, i, o, u) and consonants (everything else). **V** is a group of vowels that are adjacent and **C** is a group of adjacent consonants:

V is one or more vowels (A, E, I, O, U)

C is one or more consonants

For example, if there are 2 vowels together, they form a **V** group; if there are 2 consonants together, then a **C** group is formed.

The idea behind Porter's algorithm is that all words are of the following form:

[C]VCVCVC…[V]

The above demonstrates an optional consonant, then alternating groups of vowels and consonants, followed by an optional vowel group at the end. The above is written alternatively as follows:

[C](VC){m}[V] ← with **m** copies of **VC**

The same optional consonant with m copies of alternating groups of vowels and consonants, ending with an optional vowel as before. For example:

m=0 KN, EE, KNEE, Y, BY ← m=0 because there are no V groups followed by C groups.

m=1 BUBBLE, OAKS, KNEES, IVY ←

m=2 BUBBLES, PRIVATE ←

The algorithm works with many rules that are manually scribed. Each rule checks whether a rule obeys a condition and depending on the outcome, the algorithm lengthens or shortens the word:

For each word, check whether it obeys a condition and shorten or lengthen it accordingly.

Porter's Algorithm, Step 1:

- SSES ➔ SS (e.g., caresses ➔ caress)
- IES ➔ I (e.g., ties ➔ ti, ponies ➔ poni)
- S ➔ <remove> (e.g, potatoes ➔ potato)
- IF m>0 Then EED ➔ EE (e. feed ➔ fee)
- If stem contains vowel and has ED ➔ <remove> (e.g, turned ➔ turn)
- If stem contains vowel and has ING ➔ <remove> (e.g, turning ➔ turn)
- If either of the green rules are successful,
  - AT ➔ ATE (conflat ➔ conflate)
  - BL ➔ BLE (troubl(ed) ➔ trouble)
  - IZ ➔ IZE (siz ➔ size)
- <vowel>Y ➔ I (e.g, happy ➔ happi , whereas sky ➔ sky not ski)

Porter's Algorithm: Step 2

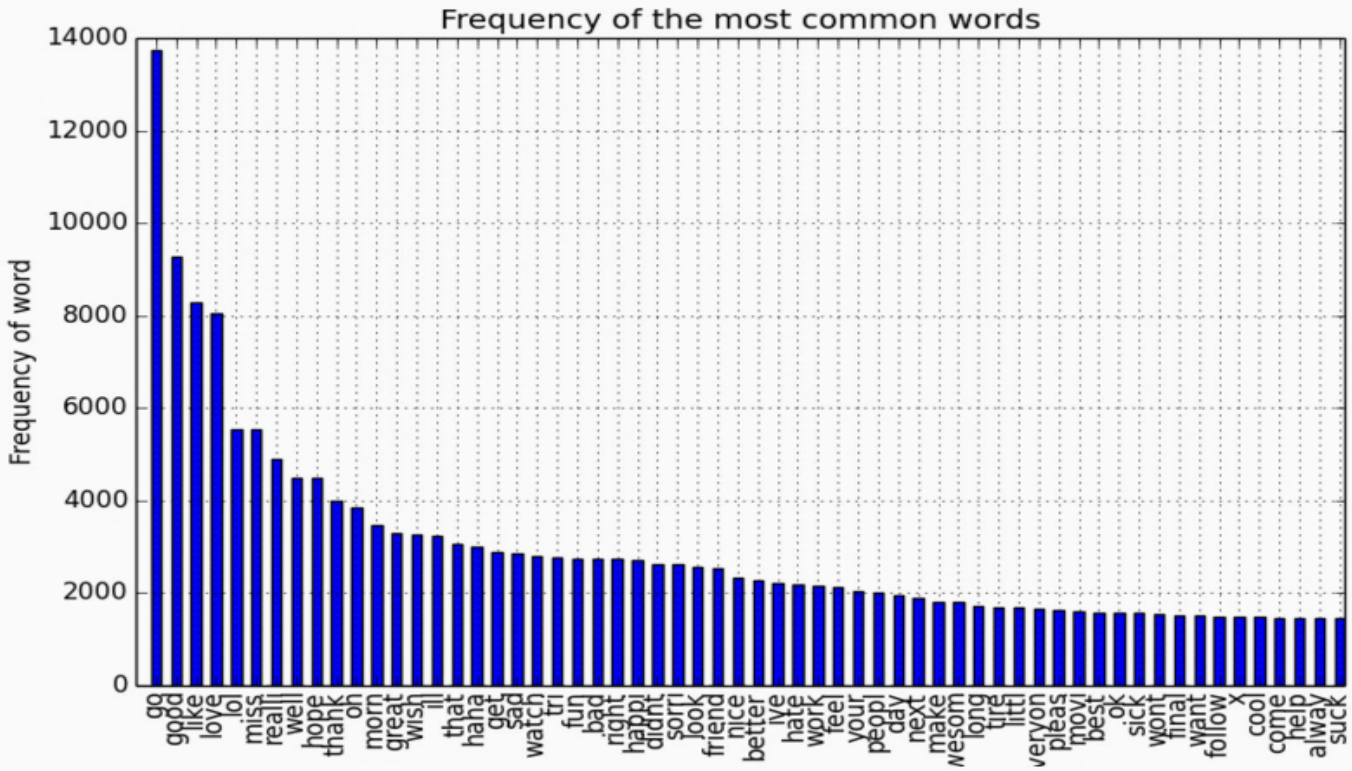
(m>0) ATIONAL	➔	ATE	relational	➔	relate
(m>0) TIONAL	➔	TION	conditional	➔	condition
(m>0) ENCI	➔	ENCE	valenci	➔	valence
(m>0) ANCI	➔	ANCE	hesitanci	➔	hesitance
(m>0) IZER	➔	IZE	digitizer	➔	digitize
(m>0) ABLI	➔	ABLE	conformabli	➔	conformable
(m>0) ALLI	➔	AL	radicalli	➔	radical
(m>0) ENTLI	➔	ENT	differentli	➔	different
(m>0) ELI	➔	E	vileli	➔	vile
(m>0) OUSLI	➔	OUS	analogousli	➔	analogous
(m>0) IZATION	➔	IZE	vietnamization	➔	vietnamize
(m>0) ATION	➔	ATE	predication	➔	predicate
(m>0) ATOR	➔	ATE	operator	➔	operate
(m>0) ALISM	➔	AL	feudalism	➔	feudal
(m>0) IVENESS	➔	IVE	decisiveness	➔	decisive
(m>0) FULNESS	➔	FUL	hopefulness	➔	hopeful
(m>0) OUSNESS	➔	OUS	callousness	➔	callous
(m>0) ALITI	➔	AL	formaliti	➔	formal
(m>0) IVITI	➔	IVE	sensitiviti	➔	sensitive
(m>0) BILITI	➔	BLE	sensibiliti	➔	sensible

Steps 3, 4, and 5 are similar sets of rules as illustrated above.
Porter's Algorithm is not perfect but often improves information retrieval performance.

The following are examples are errors the Porter's Algorithm is known for:

- severing vs. several ➔ sever
- university vs. universe ➔ univers
- iron vs. ironic ➔ iron

Most Popular Words



After the raw text data has been preprocessed to remove stopwords, non-alpha characters, converted text to lowercase, removed white spaces, and finally stemmed (etc.), a Pareto chart will plot the result.