Stemming

VSM, session 10

Northeastern University
College of Computer and Information Science

Slides by: Jesse Anderton

Stemming

A *stemming* algorithm converts words to their root forms (“stems”) in order to focus on their underlying meaning.

- It works well in English or in Arabic, with many nouns and verbs deriving from a common root
- It works worse in German or Turkish, which compose very long words with complex meanings.

It’s common to add the root to the query’s term vector (while leaving the unstemmed form present).

ki**t**a**b**	*a book*
ki**t**a**b**i	*my book*
al**k**i**t**a**b**	*the book*
ki**t**a**b**uki	*your book (f)*
ki**t**a**b**uka	*your book (m)*
ki**t**a**b**uhu	*his book*
ka**t**a**b**a	*to write*
ma**kt**a**b**a	*library, bookstore*
ma**kt**a**b**	*office*

Arabic words that stem to **ktb**

çekoslovakyalilaştiramadiklarimizdanmişsiniz

“(it is speculated that) you had been one of those whom we could not convert to a Czechoslovakian.”

–Common example of a Turkish word demonstrating agglutinative languages.

Stemming Algorithms

Two major families of stemming algorithms exist:

- *Dictionary-based stemmers* use lists of related words.
- *Algorithmic stemmers* use some algorithm to derive related words.

A simple algorithmic stemmer for English may remove the suffix -s:

- cats → cat, lakes → lake, plays → play
- But many false negatives: supplies → supplie
- And some false positives: ups → up

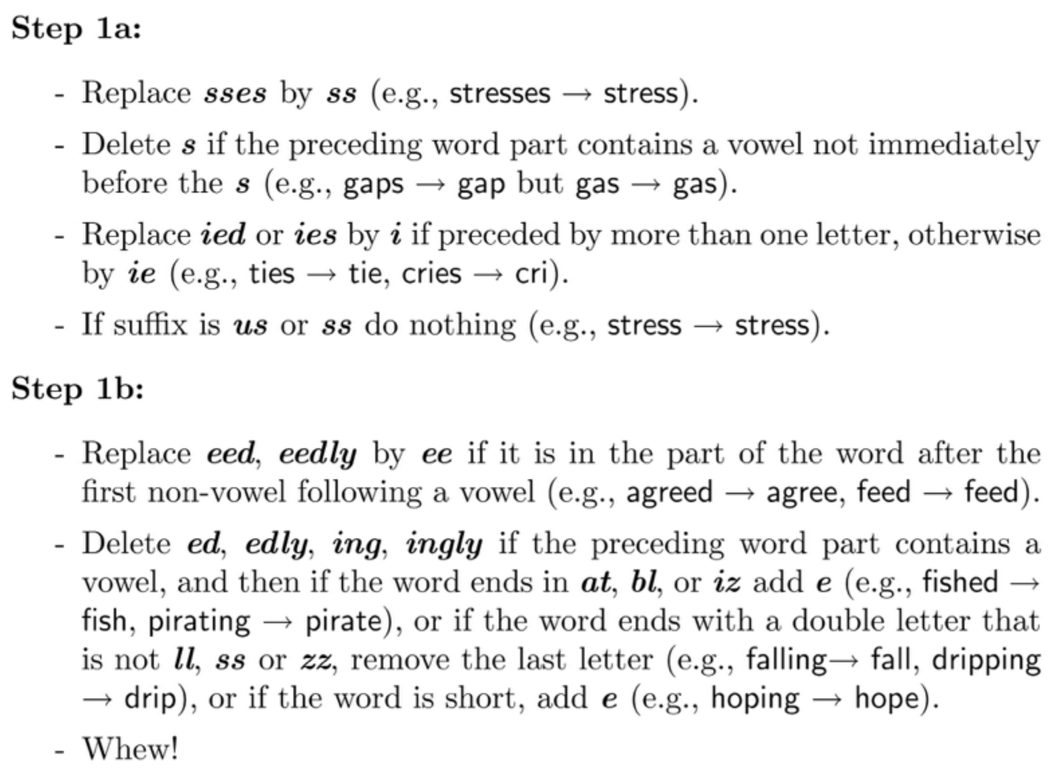
Producing high quality rules is very challenging.

Porter Stemmer

The *Porter Stemmer* was developed in the 70's, and consists of a large series of rules to repeatedly apply until only the stem is left.

It is fairly effective, though makes many categorical errors. Its complexity makes it hard to modify, though the porter2 stemmer fixes some of its problems.

It outputs stems, not recognizable words.



Step 1a:

- Replace ***sses*** by ***ss*** (e.g., stresses → stress).
- Delete ***s*** if the preceding word part contains a vowel not immediately before the ***s*** (e.g., gaps → gap but gas → gas).
- Replace ***ied*** or ***ies*** by ***i*** if preceded by more than one letter, otherwise by ***ie*** (e.g., ties → tie, cries → cri).
- If suffix is ***us*** or ***ss*** do nothing (e.g., stress → stress).

Step 1b:

- Replace ***eed***, ***eedly*** by ***ee*** if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed → agree, feed → feed).
- Delete ***ed***, ***edly***, ***ing***, ***ingly*** if the preceding word part contains a vowel, and then if the word ends in ***at***, ***bl***, or ***iz*** add ***e*** (e.g., fished → fish, pirating → pirate), or if the word ends with a double letter that is not ***ll***, ***ss*** or ***zz***, remove the last letter (e.g., falling→ fall, dripping → drip), or if the word is short, add ***e*** (e.g., hoping → hope).
- Whew!

Porter Stemmer, step 1 of 5

Krovetz Stemmer

The *Krovetz Stemmer* is a hybrid of dictionary and algorithmic methods.

It first checks the dictionary. If not found, it tries to remove suffixes and then checks the dictionary again.

It produces recognizable words, unlike the Porter stemmer.

Its effectiveness is comparable to the Porter stemmer. It has a lower false positive rate, but somewhat higher false negative.

Stemmer Comparison

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Porter stemmer:

document describ market strategi carri compani agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale market share stimul demand price cut volum sale

Krovetz stemmer:

document describe marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer predict sale stimulate demand price cut volume sale

Stem Classes

A given stemming algorithm creates *stem classes* of words which are stemmed to the same root.

These classes are generally too large and varied in meaning to use for query expansion, but they can be narrowed down using term co-occurrence statistics.

The assumption is that those terms which tend to appear in the same document are more likely to be related (or interchangeable).

/bank banked banking bankings banks
/ocean oceaneering oceanic oceanics oceanization oceans
/polic polical polically police policeable policed
-policement policer policers polices policial
-policically policier policiers policies policing
-policization policize policly policy policying policys

/policies policy
/police policed policing
/bank banking banks

Stem classes, before and after term co-occurrence thinning is applied

Wrapping Up

Adding words from query terms' stem class is an effective way to improve document matching.

Many stemming algorithms exist; the Porter and Krovetz are commonly used, but there are many other popular stemmers (e.g. the Snowball stemmer, with variants for many languages).

Next, we'll discuss term co-occurrence statistics, which can be used to fix stem classes and identify other related words to add to the query vector.