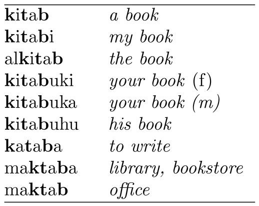
Stemming

- Many morphological variations of words
  - ***inflectional*** (plurals, tenses)
  - ***derivational*** (making verbs nouns etc.)
- In most cases, these have the same or very similar meanings
- Stemmers attempt to reduce morphological variations of words to a common stem
  - usually involves removing suffixes
- Can be done at indexing time or as part of query processing (like stopwords)

Stemming

- Generally a small but significant improvement in effectiveness
  - can be crucial for some languages
  - e.g., 5-10% improvement for English, up to 50% in Arabic



kitab	*a book*
kitabi	*my book*
alkitab	*the book*
kitabuki	*your book* (f)
kitabuka	*your book (m)*
kitabuhu	*his book*
kataba	*to write*
maktaba	*library, bookstore*
maktab	*office*

Words with the Arabic root **ktb**

Stemming

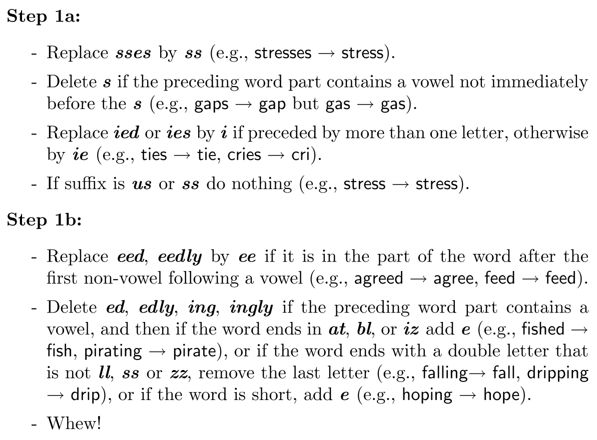
- Two basic types
  - **Dictionary-based:** uses lists of related words
  - **Algorithmic:** uses program to determine related words
- Algorithmic stemmers
  - ***suffix-s****:* remove ‘s’ endings assuming plural
    - » e.g., cats → cat, lakes → lake, wiis → wii
    - » Many *false positives*: supplies → supplie, ups → up
    - » Some *false negatives*: mice → mice (should be mouse)

Porter Stemmer

- Algorithmic stemmer used in IR experiments since the 70s
- Consists of a series of rules designed to strip off the longest possible suffix at each step
- Effective in TREC
- Produces ***stems*** not ***words***
- Makes a number of errors and difficult to modify

Porter Stemmer

- Example step (1 of 5)



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!

Let's try it

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.

Let's try it

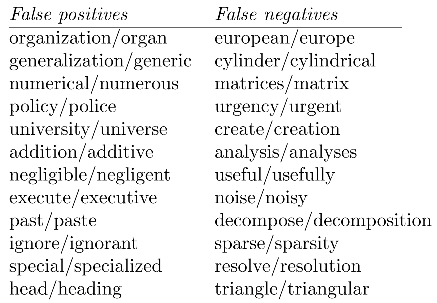
Original text:

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

Porter Stemmer



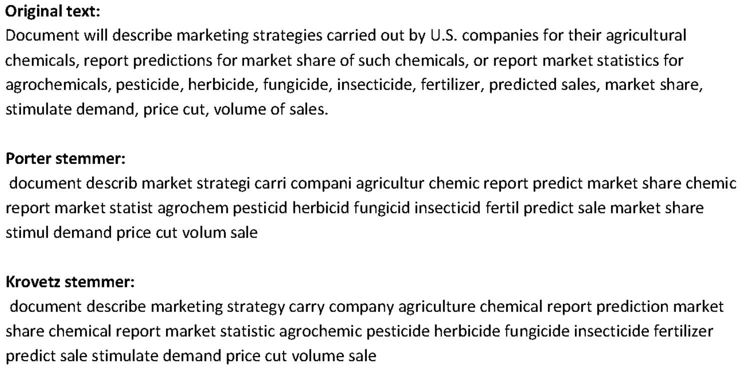
False positives	*False negatives*
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

- Porter2 stemmer addresses some of these issues
- Approach has been used with other languages

Krovetz Stemmer

- Hybrid algorithmic-dictionary-based method
  - Word checked in dictionary
    - If present, either left alone or stemmed based on its manual “exception” entry
    - If not present, word is checked for suffixes that could be removed
    - After removal, dictionary is checked again
- Produces words not stems
- Comparable effectiveness
- Lower false positive rate, 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

Next

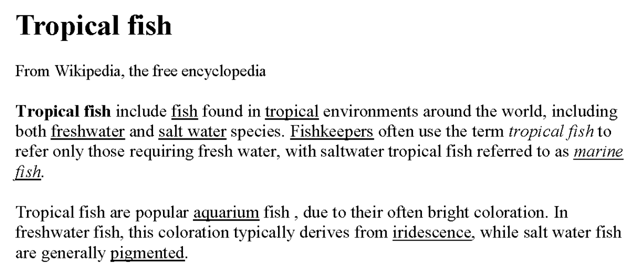
- Phrases
- Document structure
- Link analysis

- We’ll skip “phrases” until the next class.

Document Structure and Markup

- Some parts of documents are more important than others
- Document parser recognizes structure using markup, such as HTML tags
  - Headers, anchor text, bolded text all likely to be important
  - Metadata can also be important
  - Links used for ***link analysis***

Example Web Page



Tropical fish

From Wikipedia, the free encyclopedia

Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species. Fishkeepers often use the term *tropical fish* to refer only those requiring fresh water, with saltwater tropical fish referred to as *marine fish*.

Tropical fish are popular aquarium fish , due to their often bright coloration. In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Example Web Page

```
<html>
<head>
<meta name="keywords" content="Tropical fish, Airstone, Albinism, Algae eater,
Aquarium, Aquarium fish feeder, Aquarium furniture, Aquascaping, Bath treatment
(fishkeeping),Berlin Method, Biotope" />
…
<title>Tropical fish - Wikipedia, the free encyclopedia</title>
</head>
<body>
 …
<h1 class="firstHeading">Tropical fish</h1>
…
<p><b>Tropical fish</b> include <a href="/wiki/Fish" title="Fish">fish</a> found in <a
href="/wiki/Tropics" title="Tropics">tropical</a> environments around the world,
including both <a href="/wiki/Fresh_water" title="Fresh water">freshwater</a> and <a
href="/wiki/Sea_water" title="Sea water">salt water</a> species. <a
href="/wiki/Fishkeeping" title="Fishkeeping">Fishkeepers</a> often use the term
<i>tropical fish</i> to refer only those requiring fresh water, with saltwater tropical fish
referred to as <i><a href="/wiki/List_of_marine_aquarium_fish_species" title="List of
marine aquarium fish species">marine fish</a></i>.</p>
<p>Tropical fish are popular <a href="/wiki/Aquarium" title="Aquarium">aquarium</a>
fish , due to their often bright coloration. In freshwater fish, this coloration typically
derives from <a href="/wiki/Iridescence" title="Iridescence">iridescence</a>, while salt
water fish are generally <a href="/wiki/Pigment" title="Pigment">pigmented</a>.</p>
…
</body></html>
```

Link Analysis

- Links are a key component of the Web
- Important for navigation, but also for search
  - e.g., <a href="http://example.com" >Example website</a>
  - "Example website" is the **anchor text**
  - "http://example.com" is the **destination link**
  - both are used by search engines

Anchor Text

- Used as a description of the content of the ***destination*** *page*
  - i.e., collection of anchor text in all links pointing to a page used as an additional text field
- Anchor text tends to be short, descriptive, and similar to query text
- Retrieval experiments have shown that anchor text has significant impact on effectiveness for ***some types of queries***
  - i.e., more than PageRank

PageRank

- Billions of web pages, some more informative than others
- Links can be viewed as information about the ***popularity*** (***authority***?) of a web page
  - can be used by ranking algorithm
- ***Inlink*** count could be used as simple measure
- Link analysis algorithms like PageRank provide more reliable ratings
  - less susceptible to link spam

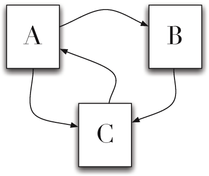
Random Surfer Model

- Browse the Web using the following algorithm:
  - Choose a random number r between 0 and 1
  - If $r < \lambda$:
    - » Go to a random page
  - If $r \geq \lambda$:
    - » Click a link at random on the current page
  - Start again
- **PageRank** of a page is the probability that the "random surfer" will be looking at that page
  - links from popular pages will increase PageRank of pages they point to

Dangling Links

- Random jump guarantees that all pages on the Internet will eventually be reached
  - prevents getting stuck on pages that
    - » do not have links
    - » contain only links that no longer point to other pages
    - » have links forming a loop
- Links that point to the first two types of pages are called *dangling links*
- Each web page has a PageRank

PageRank



- Ignoring the “surprise me” button, PageRank (*PR*) of page C = *PR*(A)/2 + *PR*(B)/1
- More generally,

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  – where B_u is the set of pages that point to *u*, and L_v is the number of outgoing links from page *v* (not counting duplicate links)

PageRank

- Don’t know PageRank values at start
- Assume equal values (1/3 in this case), then iterate:
  – first iteration: *PR*(C) = 0.33/2 + 0.33 = 0.5, *PR*(A) = 0.33, and *PR*(B) = 0.17
  – second: *PR*(C) = 0.33/2 + 0.17 = 0.33, *PR*(A) = 0.5, *PR*(B) = 0.17
  – third: *PR*(C) = 0.42, *PR*(A) = 0.33, *PR*(B) = 0.25
- Converges to *PR*(C) = 0.4, *PR*(A) = 0.4, and *PR*(B) = 0.2

PageRank

- Taking random page jump into account, 1/3 chance of going to any page when $r < \lambda$
- *PR*(C) = λ/3 + (1 − λ) · (*PR*(A)/2 + *PR*(B)/1)
- More generally,

$$PR(u) = \frac{\lambda}{N} + (1 - \lambda). \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

  – where *N* is the number of pages, λ typically 0.15

```
procedure PageRank($G$)
        ▷ $G$ is the web graph, consisting of vertices (pages) and edges (links).
    $(P, L) \leftarrow G$                              ▷ Split graph into pages and links
    $I \leftarrow$ a vector of length $|P|$              ▷ The current PageRank estimate
    $R \leftarrow$ a vector of length $|P|$   ▷ The resulting better PageRank estimate
    for all entries $I_i \in I$ do
        $I_i \leftarrow 1/|P|$                    ▷ Start with each page being equally likely
    end for
    while $R$ has not converged do
        for all entries $R_i \in R$ do
            $R_i \leftarrow \lambda/|P|$ ▷ Each page has a $\lambda/|P|$ chance of random selection
        end for
        for all pages $p \in P$ do
            $Q \leftarrow$ the set of pages   such that $(p, q) \in L$ and $q \in P$
            if $|Q| > 0$ then
                for all pages $q \in Q$ do
                    $R_q \leftarrow R_q + (1 - \lambda)I_p/|Q|$      ▷ Probability $I_p$ of being at
page $p$
                end for
            else
                for all pages $q \in P$ do
                    $R_p \leftarrow R_q + (1 - \lambda)I_p/|P|$
                end for
            end if
            $I \leftarrow R$                         ▷ Update our current PageRank estimate
        end for
    end while
    return $R$
end procedure
```