Text Processing

Indexing

- An *index* is a list of things (keys) with pointers to other things (items).
  - Keywords -> catalog numbers (-> shelves).
  - Concepts -> page numbers.
  - Terms -> documents.
- Need for indexes:
  - Ease of use.
  - Speed.
  - Scalability.

Manual vs. Automatic Indexing

- Manual:
  - An "expert" assigns keys to each item.
  - Example: card catalog.
- Automatic:
  - Keys automatically identified and assigned.
  - Example: Google.
- Automatic as good as manual for most purposes.

Text Processing

- First step in automatic indexing.
- Converting documents into *index terms.*
- Terms are not just words.
  - Not all words are of equal value in a search.
  - Sometimes not clear where words begin and end.
    - Especially when not space-separated, e.g. Chinese, Korean.
  - Matching the exact words typed by the user doesn't work very well in terms of effectiveness.

Text Processing Steps

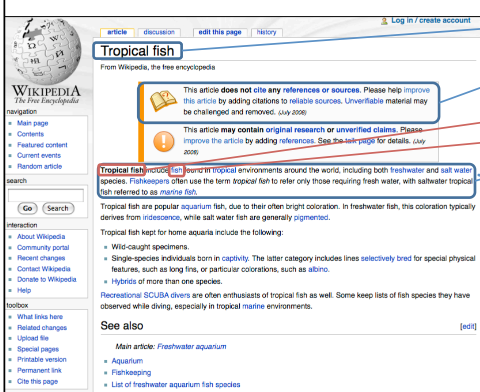
- For each document:
  - Parse it to locate the parts that are important.
  - Segment and tokenize the text in the important parts to get *words*.
  - Remove *stop words*.
  - *Stem* words to common roots.
- Advanced processing may included phrases, entity tagging, link-graph features, and more.

Parsing

- Some parts of a document are more important than others.
- Document parser recognizes structure using *markup* such as HTML tags.
  - Headers, anchor text, bolded text are likely to be important.
  - JavaScript, style information, navigation links less likely to be important.
  - Metadata can also be important.

Example Wikipedia Page

Wikipedia Markup



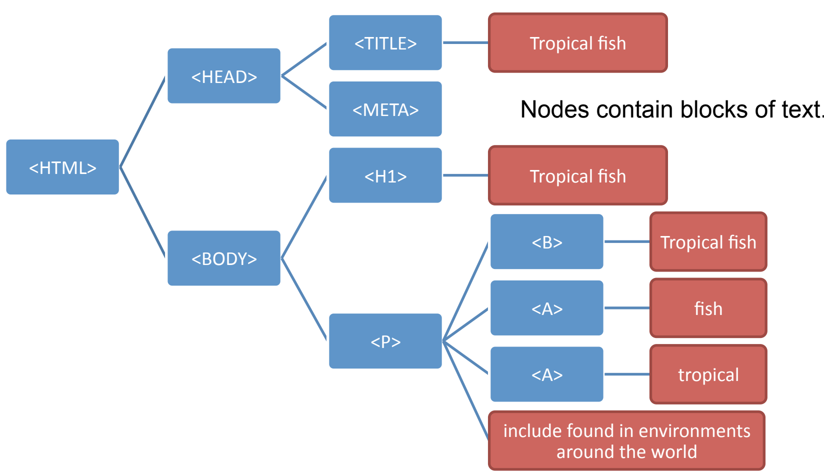
<title>Tropical fish</title>
<text>{{Unreferenced|date=July 2008}}
{{Original research|date=July 2008}}
'''Tropical fish''' include [[fish]] found in [[Tropics|topical]] environments around the world, including both [[fresh water|freshwater]] and [[sea water|salt water]] species. [[Fishkeeping|Fishkeepers]] often use the term ''tropical fish'' to refer only those requiring fresh water, with saltwater tropical fish referred to as ''[[list of marine aquarium fish species|marine fish]]''.
…

Wikipedia HTML


```
<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>
```

Document Parsing

- HTML pages organize into trees.



End Result of Parsing

- Blocks of text from important parts of page.
  – 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”.
- Next step: segmenting and tokenizing.

Tokenizing

- Forming words from sequence of characters in blocks of text.
- Surprisingly complex in English, can be harder in other languages.
- Early IR systems:
  – Any sequence of alphanumeric characters of length 3 or more.
  – Terminated by a space or other special character.
  – Upper-case changed to lower-case.

Tokenizing

- Example:
  – "Bigcorp's 2007 bi-annual report showed profits rose 10%." becomes
  – "bigcorp 2007 annual report showed profits rose"
- Too simple for search applications or even large-scale experiments
- Why? Too much information lost
  – Small decisions in tokenizing can have major impact on effectiveness of some queries

Tokenizing Problems

- Small words can be important in some queries, usually in combinations
    - xp, ma, pm, ben e king, el paso, master p, gm, j lo, world war II
- Both hyphenated and non-hyphenated forms of many words are common
  – Sometimes hyphen is not needed
    - e-bay, wal-mart, active-x, cd-rom, t-shirts
  – At other times, hyphens should be considered either as part of the word or a word separator
    - winston-salem, mazda rx-7, e-cards, pre-diabetes, t-mobile, spanish-speaking

Tokenizing Problems

- Special characters are an important part of tags, URLs, code in documents
- Capitalized words can have different meaning from lower case words
  - Bush, Apple
- Apostrophes can be a part of a word, a part of a possessive, or just a mistake
  - rosie o'donnell, can't, don't, 80's, 1890's, men's straw hats, master's degree, england's ten largest cities, shriner's

Tokenizing Problems

- Numbers can be important, including decimals
  - nokia 3250, top 10 courses, united 93, quicktime 6.5 pro, 92.3 the beat, 288358
- Periods can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., cis.udel.edu
- Note: tokenizing steps for queries must be identical to steps for documents

Tokenizing Process

- Assume we have used the parser to find blocks of important text.
- A word may be any sequence of alphanumeric characters terminated by a space or special character.
  - everything converted to lower case.
  - everything indexed.
- Defer complex decisions to other components
  - example: 92.3 → 92 3 but search finds documents with 92 and 3 adjacent
  - incorporate some rules to reduce dependence on query transformation components

End Result of Tokenization

- List of words in blocks of text.
  - 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
- Next step: stopping.
- But first: text statistics.

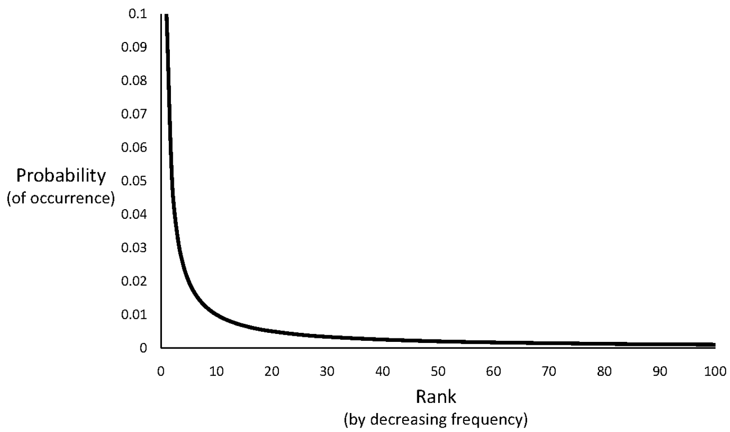
Text Statistics

- Huge variety of words used in text but
- Many statistical characteristics of word occurrences are predictable
  - e.g., distribution of word counts
- Retrieval models and ranking algorithms depend heavily on statistical properties of words
  - e.g., important words occur often in documents but are not high frequency in collection

Zipf's Law

- Distribution of word frequencies is very *skewed*
  - a few words occur very often, many words hardly ever occur
  - e.g., two most common words ("the", "of") make up about 10% of all word occurrences in text documents
- Zipf's "law":
  - observation that rank (*r*) of a word times its frequency (*f*) is approximately a constant (*k*)
    - assuming words are ranked in order of decreasing frequency
  - i.e., $r.f \approx k$ or $r.P_r \approx c$, where P_r is probability of word occurrence and $c \approx 0.1$ for English

Zipf's Law



Wikipedia Statistics (wiki000 subset)

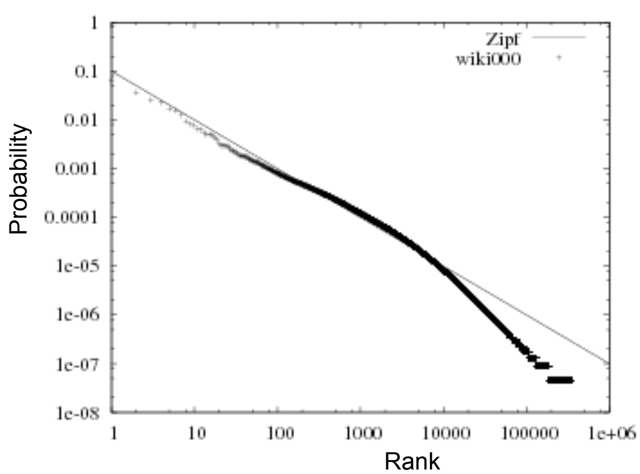
Total documents	5,001
Total word occurrences	22,545,922
Vocabulary size	348,436
Words occurring > 1000 times	2,751
Words occurring once	163,404

Word	Freq	r	Pr (%)	r.Pr
politician	5096	510	0.023	0.116
contractor	100	14,852	$4.4 \cdot 10^{-4}$	0.066
kickboxer	10	56,125	$4.4 \cdot 10^{-5}$	0.025
comdedian	1	185,035	$4.4 \cdot 10^{-6}$	0.008

Top 50 Words from wiki000 Subset

Word	*Freq.*	r	$P_r(\%)$	$r.P_r$	*Word*	*Freq.*	r	$P_r(\%)$	$r.P_r$
the	1424390	1	0.063	0.063	this	63076	26	0.002	0.072
of	832458	2	0.036	0.073	american	55582	27	0.002	0.066
and	579392	3	0.025	0.077	were	53033	28	0.002	0.065
in	505530	4	0.022	0.089	also	52137	29	0.002	0.067
to	376854	5	0.016	0.083	not	50731	30	0.002	0.067
a	357149	6	0.015	0.095	have	48903	31	0.002	0.067
ref	282192	7	0.012	0.087	has	48627	32	0.002	0.069
is	211077	8	0.009	0.074	new	45595	33	0.002	0.066
s	183617	9	0.008	0.073	his	43413	34	0.001	0.065
as	170823	10	0.007	0.075	united	41976	35	0.001	0.065
for	144690	11	0.006	0.070	its	41625	36	0.001	0.066
by	142221	12	0.006	0.075	other	41310	37	0.001	0.067
was	119216	13	0.005	0.068	first	40469	38	0.001	0.068
on	113523	14	0.005	0.070	their	40364	39	0.001	0.069
with	112296	15	0.004	0.074	d	40129	40	0.001	0.071
that	111534	16	0.004	0.079	one	40080	41	0.001	0.072
are	104115	17	0.004	0.078	states	38991	42	0.001	0.072
from	95831	18	0.004	0.076	b	38882	43	0.001	0.074
or	87157	19	0.003	0.073	1	38535	44	0.001	0.075
it	73110	20	0.003	0.064	but	36375	45	0.001	0.072
an	67643	21	0.003	0.063	such	35077	46	0.001	0.071
at	66623	22	0.002	0.065	world	34491	47	0.001	0.071
which	66570	23	0.002	0.067	most	33929	48	0.001	0.072
name	65350	24	0.002	0.069	city	33369	49	0.001	0.072
be	64169	25	0.002	0.071	all	32466	50	0.001	0.071

Zipf’s Law for wiki000 Subset



Zipf's Law

- What is the proportion of words with a given frequency?
  - Word that occurs *n* times has rank $r_n = k/n$
  - Number of words with frequency *n* is
    - $r_n - r_{n+1} = k/n - k/(n+1) = k/n(n+1)$
  - Proportion found by dividing by total number of words = highest rank = *k*
  - So, proportion with frequency *n* is $1/n(n+1)$

Zipf's Law

- Example word frequency ranking

Rank	Word	Freq
4999	objective	494
5000	albany	494
5001	defend	494
5002	appeals	493
5003	125	493
5004	lasting	493
5005	png	493

- To compute number of words with frequency 493
  - rank of "png" minus the rank of "defend"
  - 5005 − 5001 = 4

Example

Num. occurrences (n)	Predicted proportion (1/n(n+1))	Actual proportion	Actual number of words
1	.500	.469	163,404
2	.167	.151	52,672
3	.083	.070	24,272
4	.050	.045	15,685
5	.033	.030	10,437
6	.024	.022	7,832
7	.018	.017	5,962
8	.014	.014	4,890
9	.011	.011	3,886
10	.009	.009	3,291

- Proportions of words occurring *n* times in 5,001 Wikipedia documents
- Vocabulary size is 348,436.

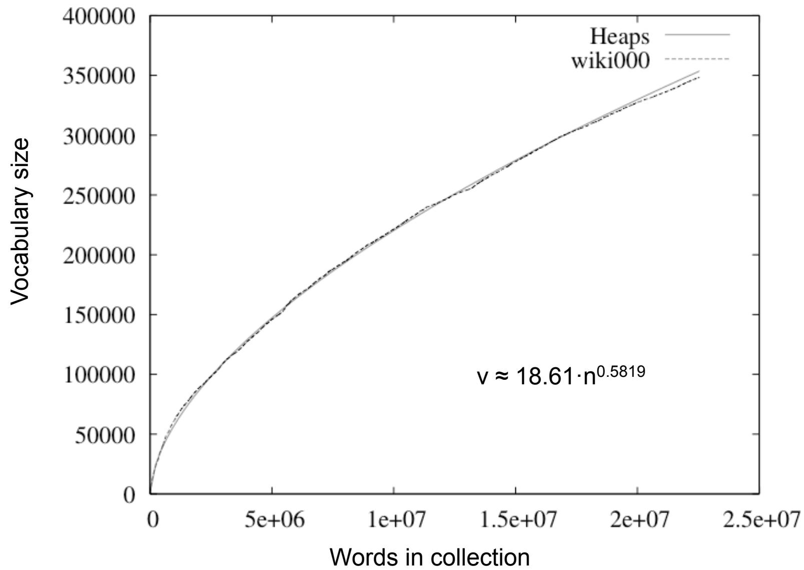
Vocabulary Growth

- As corpus grows, so does vocabulary size
  - Fewer new words when corpus is already large
- Observed relationship (*Heaps' Law*):

$$v = k.n^{\beta}$$

where *v* is vocabulary size (number of unique words), *n* is the number of words in corpus, *k*, *β* are parameters that vary for each corpus (typical values given are $10 \leq k \leq 100$ and $\beta \approx 0.5$)

wiki000 Subset Example



Heaps' Law Predictions

- Predictions for TREC collections are accurate for large numbers of words
  - e.g., first 22,545,922 words of wiki000 scanned
  - prediction is 353,587 unique words
  - actual number is 348,436
- Predictions for small numbers of words (i.e. < 1000) are much worse

Heaps' Law Predictions

- Heaps' Law works with very large corpora
  - new words occurring even after seeing 30 million!
- New words come from a variety of sources
  - spelling errors, invented words (e.g. product, company names), code, other languages, email addresses, etc.
- Search engines must deal with these large and growing vocabularies

Stopping

- Function words (determiners, prepositions) have little meaning on their own
- High occurrence frequencies
  - Top 6 words: *the, of, and, in, to, a*
- Treated as *stopwords* (i.e. removed)
  - reduce index space, improve response time, improve effectiveness
- Can be important in combinations
  - e.g., "to be or not to be"

Stopping

- Keep track of all very common words in a *stopwords list*.
- During text processing, ignore any word on the list.
- Stopword list can be created from high-frequency words or based on a standard list
- Lists are customized for applications, domains, and even parts of documents
  – e.g., "click" is a good stopword for anchor text

Stopping

- When storage space is not a concern, it can be better to not stop.
  – Queries are less restricted.
  – Remove stop words at query time unless user says to include them.
- Google does not stop.
  – "to be or not to be" returns results.
  – +the returns results (over 14 billion).

End Result of Stopping

- List of words minus those on the stop list.
  - tropical fish include fish found tropical environments around world including both freshwater salt water species fishkeepers often use term tropical fish refer only those requiring fresh water saltwater tropical fish referred marine fish
- Next step: stemming.

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

kita**b**	*a book*
kita**b**i	*my book*
al**kit**a**b**	*the book*
kita**b**uki	*your book* (f)
kita**b**uka	*your book (m)*
kita**b**uhu	*his book*
ka**t**a**b**a	*to write*
ma**kt**a**b**a	*library, bookstore*
ma**kt**a**b**	*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
    - Many *false negatives*: supplies → supplie
    - Some *false positives*: ups → up

Porter Stemmer

- Algorithmic stemmer used in IR experiments since the 70s
- Consists of a series of rules designed to the longest possible suffix at each step
- Provably effective
- 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!

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
  - Word checked in dictionary
    - If present, either left alone or replaced with "exception"
    - 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

End Result of Stemming

- List of stemmed terms:
  - tropic fish include fish found tropic environ around world include both freshwat salt water speci fishkeep often use term tropic fish refer onli those requir fresh water saltwat tropic fish refer marin fish
  - (from Porter2 stemmer)
- Next step: advanced processing, or indexing.

Advanced Text Processing

- Part-of-speech tagging.
- Sense disambiguation.
- Synonym classification.
- Named entity tagging.
- Phrase identification.
- Referent resolution.
- Sentence segmentation.
- Translation.
- Speech recognition.

```
Martin Hall, 49, head of public
    policy and external affairs
    at the London Stock
    Exchange, is to leave at the
    end of June.
…
The departure of Hall, who had
    been in the running to be
    head of corporate affairs at
    the BBC, appears to have
    been prompted by the
    decision of the new chief
    executive, Michael Lawrence,
    to split Hall's job in two
    and take the public policy
    element under his own wing.
```

```
<person id=pe1>Martin Hall</person>,
    49, <sense num=2>head</sense> of
    <owl1>public policy</owl1> and
    external affairs at the <corp
    id=co1>London Stock Exchange</
    corp>, is to <syn grp=1>leave</
    syn> at the end of June.
…
The <syn grp=1>departure</syn> of
    <person id=pe1>Hall</person>,
    <ref to=pe1>who</ref> had been
    in the running to be head of
    corporate affairs at the <corp
    id=co2>BBC</corp>, appears to
    have been prompted by the
    decision of the new chief
    executive, <person
    id=pe2>Michael Lawrence</
    person>, to split <person
    id=pe1>Hall's</person> job in
    two and take the public policy
    element under <ref to=pe1>his</
    ref> own wing.
```

Text Processing Errors

- All text processing is errorful.
  - Design decisions produce segmentation errors, stopping errors, stemming errors.
  - False positives and false negatives.
  - More advanced methods → more difficult processing → more errors.
- Does the benefit outweigh the cost?
  - Segmentation & stemming: definitely.
  - POS tagging, NE tagging: depends on domain.
  - Synonym classes: maybe not.