WEB SEARCH Indexing:

Pre-processing

CT102
Information
Systems

How to represent/index WWW organic data?

For a web page (or any document to be searched) need to extract (programmatically) some abstract representation to support complex matching (between web page and query) and to speed up querying, i.e. full web page is not searched.

This abstract representation is typically created automatically and involves choosing a subset of words from the web page and giving these words certain weights that indicate their importance in describing the web page.

All HTML tags are ignored

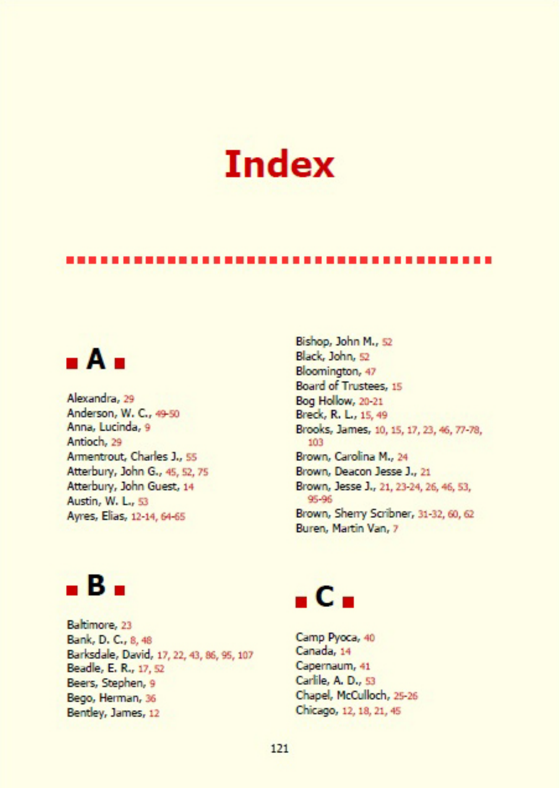
INDEXING OF “ORGANIC” WWW PAGES

An index associates a web page with one or more terms

A term may be associated with many web pages

Automatic indexing begins with no predefined set of index terms

These indexes are *dynamic* and stored on the web search engine *servers* in data stores



Index

■ A ■

Alexandra, 29
Anderson, W. C., 49-50
Anna, Lucinda, 9
Antioch, 29
Armentrout, Charles J., 55
Atterbury, John G., 45, 52, 75
Atterbury, John Guest, 14
Austin, W. L., 53
Ayres, Elias, 12-14, 64-65

■ B ■

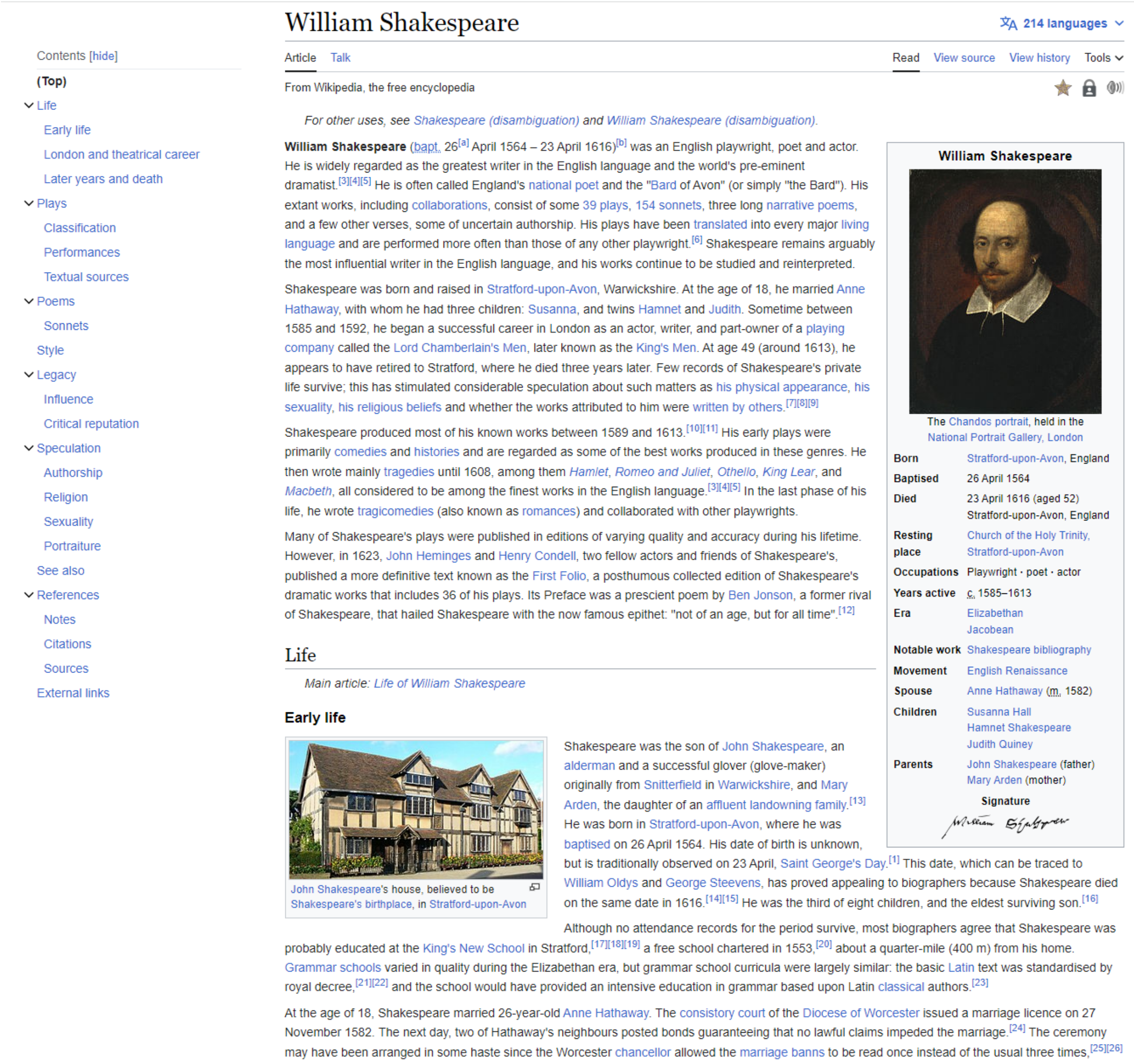
Baltimore, 23
Bank, D. C., 8, 48
Barksdale, David, 17, 22, 43, 86, 95, 107
Beadle, E. R., 17, 52
Beers, Stephen, 9
Bego, Herman, 36
Bentley, James, 12
Bishop, John M., 52
Black, John, 52
Bloomington, 47
Board of Trustees, 15
Bog Hollow, 20-21
Breck, R. L., 15, 49
Brooks, James, 10, 15, 17, 23, 46, 77-78, 103
Brown, Carolina M., 24
Brown, Deacon Jesse J., 21
Brown, Jesse J., 21, 23-24, 26, 46, 53, 95-96
Brown, Sherry Scribner, 31-32, 60, 62
Buren, Martin Van, 7

■ C ■

Camp Pyoca, 40
Canada, 14
Capernaum, 41
Carlile, A. D., 53
Chapel, McCulloch, 25-26
Chicago, 12, 18, 21, 45

121

INTRODUCING A SAMPLE TEXT *from* wikipedia



William Shakespeare

From Wikipedia, the free encyclopedia

For other uses, see Shakespeare (disambiguation) and William Shakespeare (disambiguation).

William Shakespeare (bapt. 26[a] April 1564 – 23 April 1616)[b] was an English playwright, poet and actor. He is widely regarded as the greatest writer in the English language and the world's pre-eminent dramatist.[3][4][5] He is often called England's national poet and the "Bard of Avon" (or simply "the Bard"). His extant works, including collaborations, consist of some 39 plays, 154 sonnets, three long narrative poems, and a few other verses, some of uncertain authorship. His plays have been translated into every major living language and are performed more often than those of any other playwright.[6] Shakespeare remains arguably the most influential writer in the English language, and his works continue to be studied and reinterpreted.

Shakespeare was born and raised in Stratford-upon-Avon, Warwickshire. At the age of 18, he married Anne Hathaway, with whom he had three children: Susanna, and twins Hamnet and Judith. Sometime between 1585 and 1592, he began a successful career in London as an actor, writer, and part-owner of a playing company called the Lord Chamberlain's Men, later known as the King's Men. At age 49 (around 1613), he appears to have retired to Stratford, where he died three years later. Few records of Shakespeare's private life survive; this has stimulated considerable speculation about such matters as his physical appearance, his sexuality, his religious beliefs and whether the works attributed to him were written by others.[7][8][9]

Shakespeare produced most of his known works between 1589 and 1613.[10][11] His early plays were primarily comedies and histories and are regarded as some of the best works produced in these genres. He then wrote mainly tragedies until 1608, among them *Hamlet*, *Romeo and Juliet*, *Othello*, *King Lear*, and *Macbeth*, all considered to be among the finest works in the English language.[3][4][5] In the last phase of his life, he wrote tragicomedies (also known as romances) and collaborated with other playwrights.

Many of Shakespeare's plays were published in editions of varying quality and accuracy during his lifetime. However, in 1623, John Heminges and Henry Condell, two fellow actors and friends of Shakespeare's, published a more definitive text known as the First Folio, a posthumous collected edition of Shakespeare's dramatic works that includes 36 of his plays. Its Preface was a prescient poem by Ben Jonson, a former rival of Shakespeare, that hailed Shakespeare with the now famous epithet: "not of an age, but for all time".[12]

William Shakespeare

The Chandos portrait, held in the National Portrait Gallery, London

Born	Stratford-upon-Avon, England
Baptised	26 April 1564
Died	23 April 1616 (aged 52) Stratford-upon-Avon, England
Resting place	Church of the Holy Trinity, Stratford-upon-Avon
Occupations	Playwright · poet · actor
Years active	c. 1585–1613
Era	Elizabethan Jacobean
Notable work	Shakespeare bibliography
Movement	English Renaissance
Spouse	Anne Hathaway (m. 1582)
Children	Susanna Hall Hamnet Shakespeare Judith Quiney
Parents	John Shakespeare (father) Mary Arden (mother)

Signature

William Shakspere

Life

Main article: Life of William Shakespeare

Early life

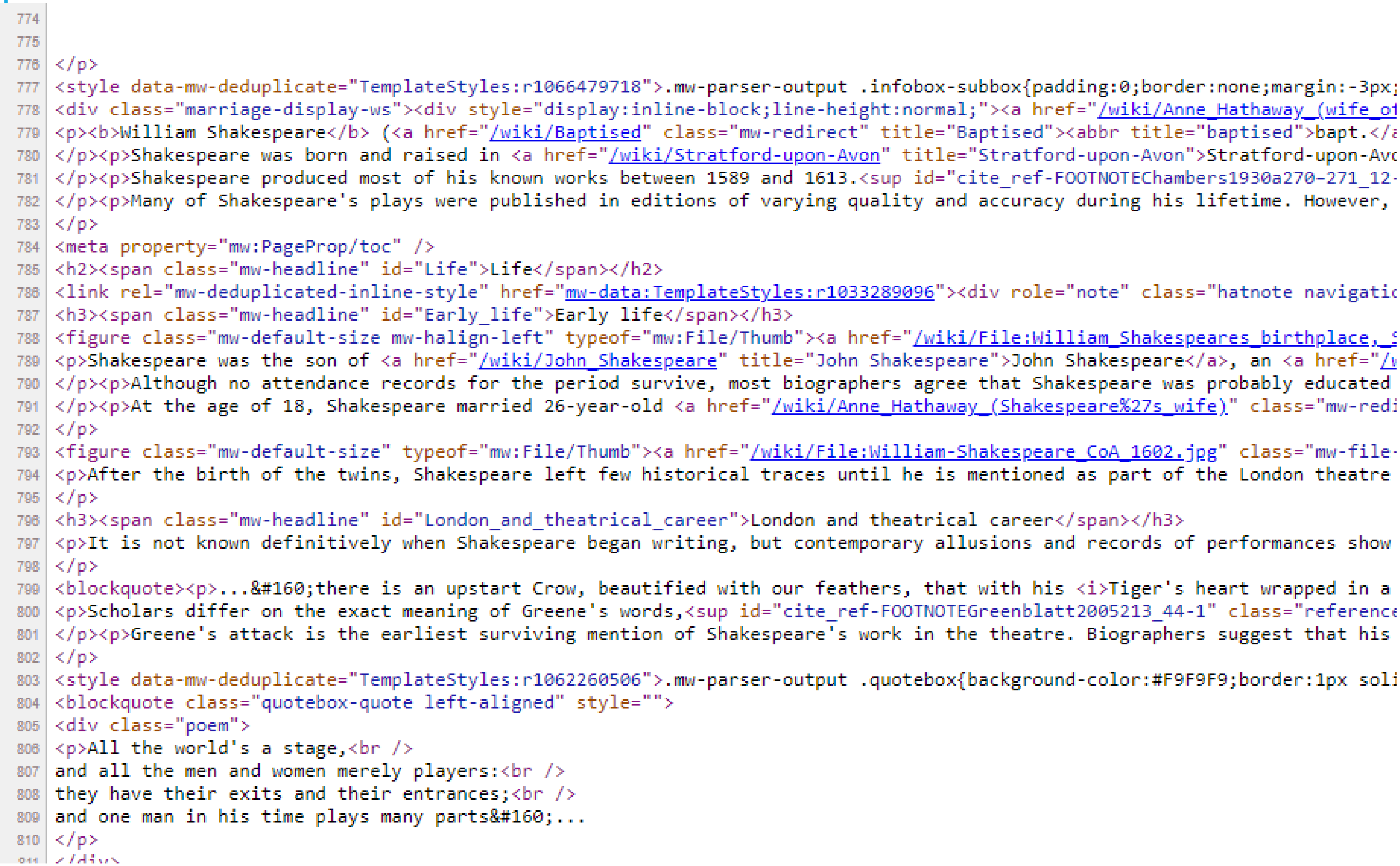
John Shakespeare's house, believed to be Shakespeare's birthplace, in Stratford-upon-Avon

Shakespeare was the son of John Shakespeare, an alderman and a successful glover (glove-maker) originally from Snitterfield in Warwickshire, and Mary Arden, the daughter of an affluent landowning family.[13] He was born in Stratford-upon-Avon, where he was baptised on 26 April 1564. His date of birth is unknown, but is traditionally observed on 23 April, Saint George's Day.[1] This date, which can be traced to William Oldys and George Steevens, has proved appealing to biographers because Shakespeare died on the same date in 1616.[14][15] He was the third of eight children, and the eldest surviving son.[16]

Although no attendance records for the period survive, most biographers agree that Shakespeare was probably educated at the King's New School in Stratford,[17][18][19] a free school chartered in 1553,[20] about a quarter-mile (400 m) from his home. Grammar schools varied in quality during the Elizabethan era, but grammar school curricula were largely similar: the basic Latin text was standardised by royal decree,[21][22] and the school would have provided an intensive education in grammar based upon Latin classical authors.[23]

At the age of 18, Shakespeare married 26-year-old Anne Hathaway. The consistory court of the Diocese of Worcester issued a marriage licence on 27 November 1582. The next day, two of Hathaway's neighbours posted bonds guaranteeing that no lawful claims impeded the marriage.[24] The ceremony may have been arranged in some haste since the Worcester chancellor allowed the marriage banns to be read once instead of the usual three times,[25][26]

WHAT A CRAWLER DOWNLOADS



```html


</p>
<style data-mw-deduplicate="TemplateStyles:r1066479718">.mw-parser-output .infobox-subbox{padding:0;border:none;margin:-3px;
<div class="marriage-display-ws"><div style="display:inline-block;line-height:normal;"><a href="/wiki/Anne_Hathaway_(wife_o
<p><b>William Shakespeare</b> (<a href="/wiki/Baptised" class="mw-redirect" title="Baptised"><abbr title="baptised">bapt.</
</p><p>Shakespeare was born and raised in <a href="/wiki/Stratford-upon-Avon" title="Stratford-upon-Avon">Stratford-upon-Av
</p><p>Shakespeare produced most of his known works between 1589 and 1613.<sup id="cite_ref-FOOTNOTEChambers1930a270-271_12
</p><p>Many of Shakespeare's plays were published in editions of varying quality and accuracy during his lifetime. However,
</p>
<meta property="mw:PageProp/toc" />
<h2><span class="mw-headline" id="Life">Life</span></h2>
<link rel="mw-deduplicated-inline-style" href="mw-data:TemplateStyles:r1033289096"><div role="note" class="hatnote navigati
<h3><span class="mw-headline" id="Early_life">Early life</span></h3>
<figure class="mw-default-size mw-halign-left" typeof="mw:File/Thumb"><a href="/wiki/File:William_Shakespeares_birthplace,_
<p>Shakespeare was the son of <a href="/wiki/John_Shakespeare" title="John Shakespeare">John Shakespeare</a>, an <a href="/
</p><p>Although no attendance records for the period survive, most biographers agree that Shakespeare was probably educated
</p><p>At the age of 18, Shakespeare married 26-year-old <a href="/wiki/Anne_Hathaway_(Shakespeare%27s_wife)" class="mw-redi
</p>
<figure class="mw-default-size" typeof="mw:File/Thumb"><a href="/wiki/File:William-Shakespeare_CoA_1602.jpg" class="mw-file-
<p>After the birth of the twins, Shakespeare left few historical traces until he is mentioned as part of the London theatre
</p>
<h3><span class="mw-headline" id="London_and_theatrical_career">London and theatrical career</span></h3>
<p>It is not known definitively when Shakespeare began writing, but contemporary allusions and records of performances show
</p>
<blockquote><p>...&#160;there is an upstart Crow, beautified with our feathers, that with his <i>Tiger's heart wrapped in a
<p>Scholars differ on the exact meaning of Greene's words,<sup id="cite_ref-FOOTNOTEGreenblatt2005213_44-1" class="reference
</p><p>Greene's attack is the earliest surviving mention of Shakespeare's work in the theatre. Biographers suggest that his
</p>
<style data-mw-deduplicate="TemplateStyles:r1062260506">.mw-parser-output .quotebox{background-color:#F9F9F9;border:1px soli
<blockquote class="quotebox-quote left-aligned" style="">
<div class="poem">
<p>All the world's a stage,<br />
and all the men and women merely players:<br />
they have their exits and their entrances;<br />
and one man in his time plays many parts&#160;...
</p>
</div>
```

NEXT QUESTION: How is this represented after indexing?

Need techniques which automatically (i.e. (programmatically) find the words or combination of words which best represent the meaning of the text.

William Shakespeare

From Wikipedia, the free encyclopedia

This article is about the poet and playwright. For other persons of the same name, see William Shakespeare (disambiguation)*. For other uses of "Shakespeare", see* Shakespeare (disambiguation)*.*

William Shakespeare (bapt. 26 April 1564 – 23 April 1616)[a] was an English poet, playwright, and actor, widely regarded as the greatest writer in the English language and the world's greatest dramatist.[2][3][4] He is often called England's national poet and the "Bard of Avon".[5][b] His extant works, including collaborations, consist of some 39 plays,[c] 154 sonnets, two long narrative poems, and a few other verses, some of uncertain authorship. His plays have been translated into every major living language and are performed more often than those of any other playwright.[7]

Shakespeare was born and raised in Stratford-upon-Avon, Warwickshire. At the age of 18, he married Anne Hathaway, with whom he had three children: Susanna and twins Hamnet and Judith. Sometime between 1585 and 1592, he began a successful career in London as an actor, writer, and part-owner of a playing company called the Lord Chamberlain's Men, later known as the King's Men. At age 49 (around 1613), he appears to have retired to Stratford, where he died three years later. Few records of Shakespeare's private life survive; this has stimulated considerable speculation about such matters as his physical appearance, his sexuality, his religious beliefs, and whether the works attributed to him were written by others.[8][9][10] Such theories are often criticised for failing to adequately note that few records survive of most commoners of the period.

Shakespeare produced most of his known works between 1589 and 1613.[11][12][d] His early plays were primarily comedies and histories and are regarded as some of the best work produced in these genres. Until about 1608, he wrote mainly tragedies, among

WHAT IS THE TEXT ABOUT?

William Shakespeare (bapt. 26 April 1564 – 23 April 1616)[
widely regarded as the greatest writer in the English language
often called England's national poet and the "Bard of Avon".[5][b] His extant works, including collaborations,
consist of some 39 plays,[c] 154 sonnets, two long narrative poems, and a few other verses, some of uncertain authorship. His plays have been translated into every major living language and are performed more often than those of any other playwright.[7]

Shakespeare was born a
Hathaway, with whom he
1592, he began a succes
the Lord Chamberlain's M
retired to Stratford, whe
stimulated considerable s
beliefs, and whether the
criticised for failing to ac

What words are most important in understanding what the text is about?

At the age of 18, he married Anne
nd Judith. Sometime between 1585 and
-owner of a playing company called
ound 1613), he appears to have
espeare's private life survive; this has
earance, his sexuality, his religious
[9][10] Such theories are often
mmoners of the period.

Shakespeare produced most of his known works between 1589 and 1613.[11][12][d] His early plays were primarily comedies and histories and are regarded as some of the best work produced in these genres. Until about 1608, he wrote mainly tragedies, among them Hamlet, Othello, King Lear, and Macbeth, all considered to be among the finest works in the English language.[2][3][4] In the last phase of his life, he wrote tragicomedies
(also known as romances) and collaborated with other playwr

Do you think word frequencies can tell us about the meaning?

Many of Shakespeare's plays were published in editions of va
er,
in 1623, two fellow actors and friends of Shakespeare's, John
definitive text known as the First Folio, a posthumous collected
included all but two of his plays.[13] The volume was preface
presciently hails Shakespeare in a now-famous quote as "not o

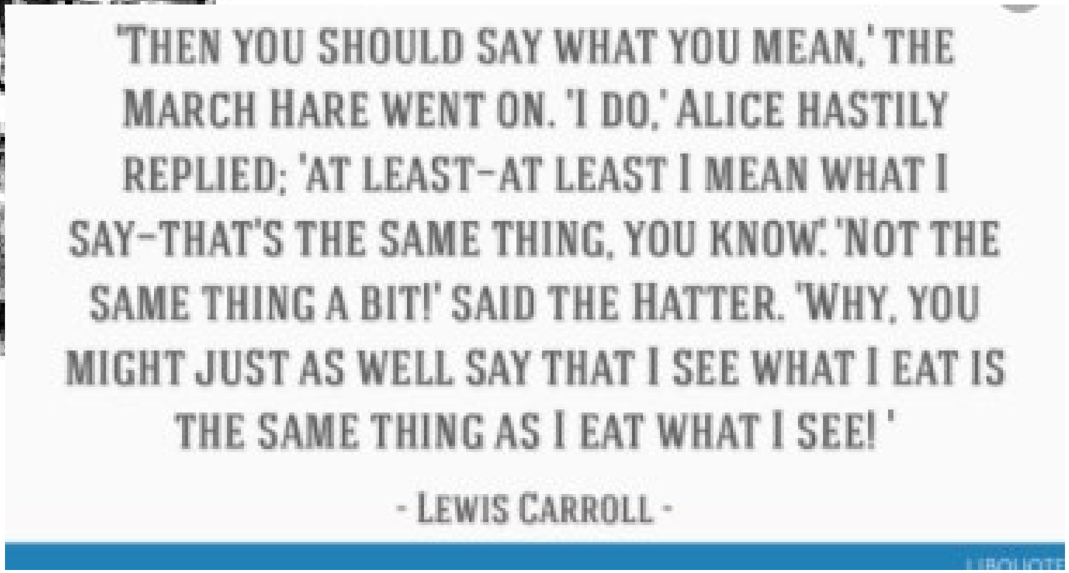
Throughout the 20th and 21st centuries, Shakespeare's works
by new movements in scholarship and performance. His plays remain popular and are studied, performed, and reinterpreted through various cultural and political contexts around the world.

Words occurring most frequently?

an	44
and	26
his	13
in	32
of	21

play*	13
shakespeare	9
the	26
wrote/write	4
work	7

DO YOU THINK WORD FREQUENCIES CAN TELL US SOMETHING ABOUT THE *MEANING* OF THE TEXT?



'THEN YOU SHOULD SAY WHAT YOU MEAN,' THE MARCH HARE WENT ON. 'I DO,' ALICE HASTILY REPLIED; 'AT LEAST–AT LEAST I MEAN WHAT I SAY–THAT'S THE SAME THING, YOU KNOW.' 'NOT THE SAME THING A BIT!' SAID THE HATTER. 'WHY, YOU MIGHT JUST AS WELL SAY THAT I SEE WHAT I EAT IS THE SAME THING AS I EAT WHAT I SEE!'

- LEWIS CARROLL -

Looking at a smaller portion of the paragraph …

Shakespeare produced most of his known work between 1590 and 1613. Shakespeare's early plays were mainly comedies and histories. After which, Shakespeare wrote mainly tragedies until about 1608, including Hamlet, King Lear, and Macbeth. In his last phase, Shakespeare wrote tragicomedies and collaborated with other playwrights.

Many of Shakespeare's plays were published in editions of varying quality and accuracy during his lifetime, and in 1623, two of Shakespeare's former theatrical colleagues were involved in publishing the First Folio, a collected edition of Shakespeare's dramatic works that included all but two of the plays now recognised as Shakespeare's.

WORDS OCCURRING MOST FREQUENTLY?

Shakespeare produced most of his known work between 1590 and 1613. Shakespeare's early plays were mainly comedies and histories. After which, Shakespeare wrote mainly tragedies until about 1608, including Hamlet, King Lear, and Macbeth. In his last phase, Shakespeare wrote tragicomedies and collaborated with other playwrights.

Many of Shakespeare's plays were published in editions of varying quality and accuracy during his lifetime, and in 1623, two of Shakespeare's former theatrical colleagues were involved in publishing the First Folio, a collected edition of Shakespeare's dramatic works that included all but two of the plays now recognised as Shakespeare's.

HOW TO DEFINE IMPORTANT WORDS? … in terms of *meaning*

<u>Questions to consider:</u>

Is each type of word equally important?

Are upper and lower case words different (in terms of meaning)?

Are plural and singular words and different tenses of words *very* different in terms of meaning?

Is Punctuation significant (in terms of meaning)?

How to deal with different words for same meaning?

How to deal with a word that has multiple meanings

WHAT IS AN “IMPORTANT” WORD?

An important word is one that give us the most information about what the web page is about and helps us distinguish between different web pages in terms of meaning

That is, the word that tell us about the *meaning of the content of the web page*

HOW TO DEFINE “IMPORTANT” WORDS?

Is Each unique word important (in terms of meaning)?

No – just Nouns and Verbs (mostly)

Are Upper and Lower case words different (in terms of meaning)?

No - upper and lower case versions of the same word should be treated as the same (except for proper nouns)

Are Plural and Singular/Tenses different (in terms of meaning)

No - should be treated as the same word

HOW TO DEFINE "IMPORTANT" WORDS?

Is Punctuation significant (in terms of meaning?)

Mostly, no - should not be considered to give different meaning

How to deal with different words for same meaning?

Need a thesaurus

How to deal with a word that has multiple meanings

Need to use the other words surrounding the word to disambiguate the word

…. Words become *terms*

In automatic indexing, due to many versions of a word being considered the same, the terminology of term is used to encompass all versions of a word.

e.g.,

term = rain

Sample words = rains, raining, rained

Indexing: finding the best terms automatically aka *pre-processing*

For each web page (or fragment) a number of *pre-processing* steps are carried out:

- Case folding: words are changed to lowercase (may be special cases for proper nouns)
- Punctuation is removed (punctuation removal)
- "stop words" are removal (stop word removal)
- "Stemming" or "Lemmatization" is performed

CASE FOLDING:
WORDS ARE CHANGED TO LOWERCASE
words are changed to lowercase

In computing, unless strings are **exactly** the same they will not be considered equal

e.g.,

- 'Example' and 'example' are not the same
- 'eXample' and 'example' are not the same

However there is no difference in meaning between the uppercase and lowercase versions.

Therefore, *in general* all strings should be changed to one case – lowercase is the convention ("case folding")

Exceptions are added for proper nouns

Punctuation is removed

Simple punctuation, such as , . ; - gives little meaning

Other punctuation is a short-hand version of two words, e.g. “she’s”, “they’ll”

Other punctuation is more complex and relates to the word following the punctuation e.g., “shakespeare’s plays”

In general, it is too costly in terms of computation effort to distinguish between different types of punctuation and so it is usually removed and replaced with a space.

N.B. As part of punctuation removal, any “trailing” letters left behind are removed as part of stop word removal (rather than being augmented)

e.g.

she’s → she s want she

they’ll → they ll want they

DEALING WITH PROPER NOUNS

In English, we know that the first word at start of every sentence begins with a capital letter.

In addition, proper nouns which can occur anywhere in a sentence, have the first letter in capitals, e.g. placenames, people's names, etc.
It is often important to treat proper nouns as a special case and not to change them to lowercase.

- Punctuation, and the position/location of a word in a sentence can be used to distinguish these special cases.

Note that abbreviations (e.g., EU, USA, HEA, etc.) will generally all be in uppercase and may also remain in uppercase.

- These can be distinguished by the fact that they are all uppercase or that they contain "non-standard" punctuation occurrences, e.g., U.S.A.

Task: Carry out the 1st two steps for 1st paragraph of Shakespeare example with no special case for proper nouns ….

Shakespeare produced most of his known work between 1590 and 1613. Shakespeare’s early plays were mainly comedies and histories. After which, Shakespeare wrote mainly tragedies until about 1608, including Hamlet, King Lear, and Macbeth. In his last phase, Shakespeare wrote tragicomedies and collaborated with other playwrights.

shakespeare produced most of his known work between 1590 and 1613 shakespeare early plays were mainly comedies and histories after which shakespeare wrote mainly tragedies until about 1608 including hamlet king lear and macbeth in his last phase shakespeare wrote tragicomedies and collaborated with other playwrights

Stop word removal

Stop words are words that do not provide any extra information about the meaning of a document

Stop words are very common (frequently occur) in a document and often have a small number of letters

Examples are language specific. In English: the, a, and

Stop words are removed to save storage space and to speed up searches

The tendency now is to have a quite small list of stop words

No common set is used – depends on domain – different stop words would be used for Twitter data than for web page data

SAMPLE ENGLISH STOP WORD LIST (stopwords1.txt)

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

from: http://www.textfixer.com/resources/common-english-words.txt

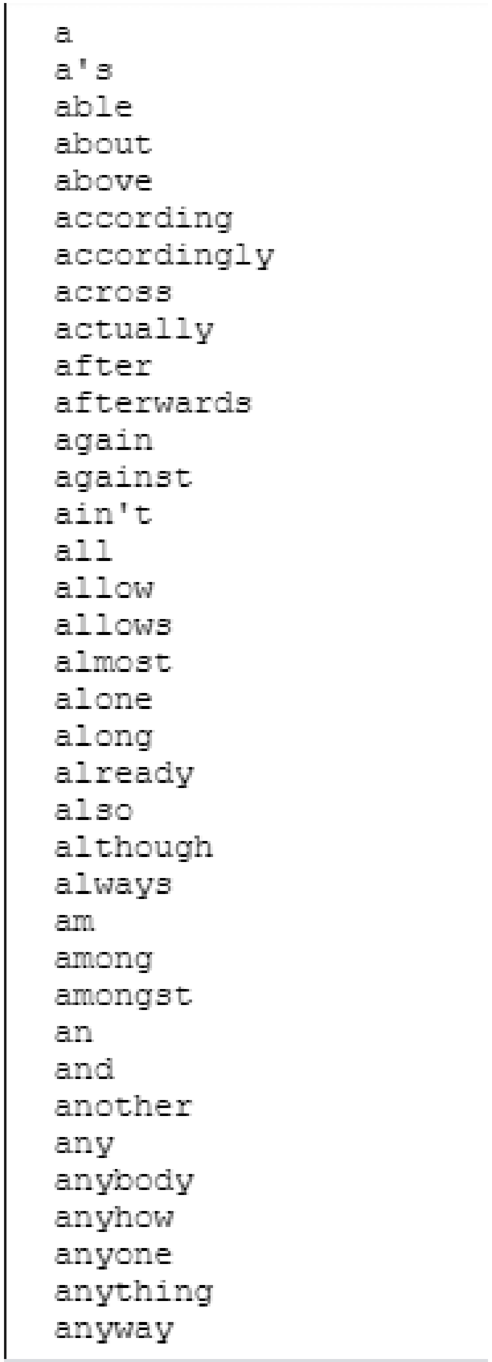
LIST POSSIBLY USED BY GOOGLE (stopwords2.txt)

- a
- about
- above
- an
- and
- are
- as
- at
- be
- by
- for
- from
- how
- i
- if
- in
- is
- it
- not
- of
- often
- on
- or
- than
- that
- the
- these
- they
- this
- to
- very
- via
- was
- what
- when
- where
- whether
- who
- will
- with

KEVIN BOUGE STOP WORD LIST

A much longer list of stop words and available in many languages - Arabic, Armenian, Brazilian, Bulgarian, Chinese, Czech, Danish, Dutch, English, Farsi, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Latvian, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish.

https://sites.google.com/site/kevinbouge/stopwords-lists



```
a
a's
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
ain't
all
allow
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
```

APPROACH FOR STOP WORD REMOVAL:

- When a document is initially processed, each word is checked against a stop word list. If the word is not on list it is output to new file; if word is found then it is not output
- Each query should also be processed against a stop list
- High level algorithmic steps:

```
while not EOF do:
    read in line
    for each word in line:
        if word not in stop list:
            write word to new file
```

IMPROVED APPROACH:

Before the stop word list is checked, find the length of each word (`len(word)`)

Remove all words of length 1 and 2

This is easy to implement and means that a much shorter stop word list can be used if words of length 1, 2 (and maybe 3) do not have to be checked against the stop word list.

Stop word removal for portion of Shakespeare example using stopwords1.txt

shakespeare produced most of his known work between 1590 and 1613 shakespeare early plays were mainly comedies and histories after which shakespeare wrote mainly tragedies until about 1608 including hamlet king lear and macbeth in his last phase shakespeare wrote tragicomedies and collaborated with other playwrights

shakespeare produced known work between 1590 1613 shakespeare early plays mainly comedies histories shakespeare wrote mainly tragedies until 1608 including hamlet king lear macbeth last phase shakespeare wrote tragicomedies collaborated playwrights

NOTE: Reduction in number of terms

Original paragraph has 46 words

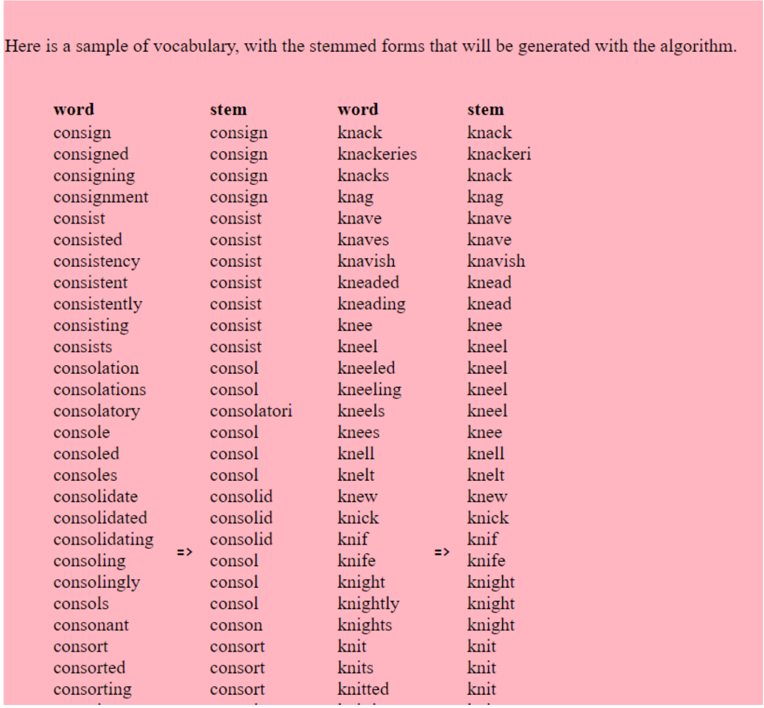
After stop word removal, there are 31 words left

shakespeare produced most of his known work between 1590 and 1613 shakespeare early plays were mainly comedies and histories after which shakespeare wrote mainly tragedies until about 1608 including hamlet king lear and macbeth in his last phase shakespeare wrote tragicomedies and collaborated with other playwrights

shakespeare produced known work between 1590 1613 shakespeare early plays mainly comedies histories shakespeare wrote mainly tragedies until 1608 including hamlet king lear macbeth last phase shakespeare wrote tragicomedies collaborated playwrights

STEMMING

- Stemming tries to find the “stem” of each word.
- A stem represents variant forms of a word which share a common meaning.
- The approach used is language specific.
- Assuming words are written left to right (as in English), then the stem is on the left and letters are often removed on the right.
- As part of stemming, zero or more suffixes may also be added on the right.



Here is a sample of vocabulary, with the stemmed forms that will be generated with the algorithm.

word		**stem**	**word**		**stem**
consign		consign	knack		knack
consigned		consign	knackeries		knackeri
consigning		consign	knacks		knack
consignment		consign	knag		knag
consist		consist	knave		knave
consisted		consist	knaves		knave
consistency		consist	knavish		knavish
consistent		consist	kneaded		knead
consistently		consist	kneading		knead
consisting		consist	knee		knee
consists		consist	kneel		kneel
consolation		consol	kneeled		kneel
consolations		consol	kneeling		kneel
consolatory		consolatori	kneels		kneel
console		consol	knees		knee
consoled		consol	knell		knell
consoles		consol	knelt		knelt
consolidate		consolid	knew		knew
consolidated		consolid	knick		knick
consolidating		consolid	knif		knif
consoling	=>	consol	knife	=>	knife
consolingly		consol	knight		knight
consols		consol	knightly		knight
consonant		conson	knights		knight
consort		consort	knit		knit
consorted		consort	knits		knit
consorting		consort	knitted		knit

FOR EXAMPLE: Stem of these terms?

connected

connection

connecting

connections

connect

computing

computers

computed

computations

compute

comput

worried

worries

worrying

worri

HOW DOES STEMMING WORK?

- Consists of many set of rules that are checked in a certain order
- Terms are usually stemmed as part of pre-processing (after stop word removal) to avoid stemming stop words
- The commonly-used stemming algorithms (for English) are called Porter's Stemming Algorithm, Snowball Stemmer (Porter 2) and Lancaster Stemming algorithm
- Stemming does not work for all languages (e.g. Chinese)
- Is it used? Yes … widely

SAMPLE RULES (1 OF 2)

```
if (word ends in ‘ies’) :
        remove ‘ies’
        add ‘y’
```

e.g., pastries → pastry
ponies → pony
berries → berry

SAMPLE RULES (2 OF 2)

If (word ends ‘es’ but not in ‘oes’):

remove ‘s’

e.g.,

files → file

ceases → cease

potatoes →

banjoes →

TRY IT ONLINE …

Interactive version:

Snowball (and others):

http://text-processing.com/demo/nstem/

People mostly use existing implementations and do not re-code it (due to complexity of rules):

See:

http://tartarus.org/~martin/PorterStemmer/

http://snowball.tartarus.org/algorithms/english/stemmer.html

Try: Stemming for portion of Shakespeare example with Snowball English stemmer from `http://text-processing.com/demo/stem/`

shakespeare produced known work between 1590 1613
shakespeare early plays mainly comedies histories
shakespeare wrote mainly tragedies until1608 including
hamlet king lear macbeth last phase shakespeare wrote
tragicomedies collaborated playwrights

shakespear produc known work between 1590 1613
shakespear earli play main comedi histori shakespear
wrote main tragedi until1608 includ hamlet king lear
macbeth last phase shakespear wrote tragicomedi
collabor playwright

LEMMATISATION

A lemma is a base form (core) of a word and it is what we look up in a dictionary

Lemmatisation is the conversion of a word to its lemma

e.g.,

walking → walk

walked → walk

goose → goose (stem: goos)

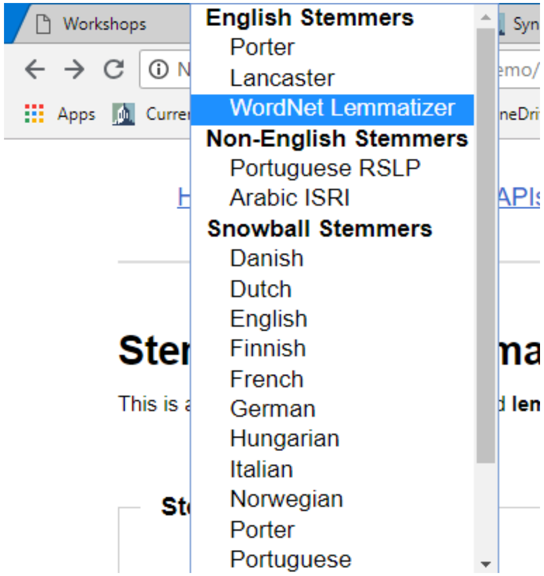
geese → goose (stem: gees)

Finding the lemma of a word is much harder (automatically) than finding a stem

TRY IT ONLINE ...

This interactive version has English Lemmatisation also:

http://text-processing.com/demo/nstem/



SHAKESPEARE EXAMPLE AGAIN:

shakespeare produced known work between 1590 1613 shakespeare early plays mainly comedies histories shakespeare wrote mainly tragedies until1608 including hamlet king lear macbeth last phase shakespeare wrote tragicomedies collaborated playwrights

WordNet Lemmatizer:

shakespeare produced known work between 1590 1613 shakespeare early play mainly comedy history shakespeare wrote mainly tragedy until1608 including hamlet king lear macbeth last phase shakespeare wrote tragicomedy collaborated playwright

COMPARING RESULTS:

WordNet Lemmatizer:

shakespeare produced known work between 1590 1613 shakespeare early play mainly comedy history shakespeare wrote mainly tragedy until1608 including hamlet king lear macbeth last phase shakespeare wrote tragicomedy collaborated playwright

Snowball English stemmer:

shakespear produc known work between 1590 1613 shakespear earli play main comedi histori shakespear wrote main tragedi until1608 includ hamlet king lear macbeth last phase shakespear wrote tragicomedi collabor playwright

THESAURUS

Synonyms are different words with identical or very similar meanings

Often important to identify terms which have synonyms

Examples:

- cry/weep/lament
- ill/sick
- thesis/dissertation
- holiday/vacation
- mail/post
- student/pupil

IMPLEMENTATION

Two approaches to include synonyms where a thesaurus can be used:

- To replace each term in a document with its variants (based on the thesaurus)
- To broaden a query by including variants of terms in the query (much more efficient approach)

Online at:

http://thesaurus.com/

Looking at all these pre-processing steps for following two Shakespeare paragraphs:

(*Note*: 97 words)

Shakespeare produced most of his known work between 1590 and 1613. Shakespeare’s early plays were mainly comedies and histories. After which, Shakespeare wrote mainly tragedies until about 1608, including Hamlet, King Lear, and Macbeth. In his last phase, Shakespeare wrote tragicomedies and collaborated with other playwrights.

Many of Shakespeare’s plays were published in editions of varying quality and accuracy during his lifetime, and in 1623, two of Shakespeare’s former theatrical colleagues were involved in publishing the First Folio, a collected edition of Shakespeare’s dramatic works that included all but two of the plays now recognised as Shakespeare's.

Looking at all these pre-processing steps for the two Shakespeare paragraphs:

(stopwords2.txt & Porter Stemmer)

(*Note*: 71 terms)

shakespear produc most known work between 1590 1613 shakespear earli play were mainli comedi histori after which shakespear wrote mainli tragedi until1608 includ hamlet king lear macbeth last phase shakespear wrote tragicomedi collabor other playwright mani shakespear play were publish edit vari qualiti accuraci dure lifetime1623 shakespear former theatric colleagu were involv publish first folio collect edit shakespear dramat work includ play recognis shakespear

TERMS THAT OCCUR MORE THAN ONCE:

shakespear	8
play	3
were	3
edit	2
hi	2
include	2
mainli	2
publish	2
two	2
work	2
write	2

TERMS THAT OCCUR ONCE
… also important!

1590
1608
1613
1623
accuraci
between
collabor
colleagu
collect
comedi
dramat
dure
earli
first
folio
former
hamlet
histori
involv
king
known
last
lear
lifetim
macbeth
mani
now
phase
playwright
produc
qualiti
recognis
theatric
tragedi
tragicomedi
until
vari

CLASS WORK … QUESTION

For each sentence given show how a pre-processing stage, involving case change, punctuation removal, stop word removal and stemming, produces a new representation of each sentence.

Indicate clearly the approaches you are using, listing the stop words you are using and the approach and the general type of stemming rules used.

* You may use an online stemmer (use Snowball) and stopwords2.txt and do not have any special rules for Proper Nouns.

SENTENCES…

3setences.txt on blackboard

Consider the following three short sentences, s1, s2 and s3, and their contents:

s1: Python is a very powerful programming language.

s2: Python is often compared to the programming languages Perl, Ruby, Scheme and Java.

s3: Python, Perl, Ruby, Scheme, Java- what's the difference and is Python the best?

stopwords2.txt

- a
- about
- above
- an
- and
- are
- as
- at
- be
- by
- for
- from
- how
- i
- if
- in
- is
- it
- not
- of
- often
- on
- or
- than
- that
- the
- these
- they
- this
- to
- very
- via
- was
- what
- when
- where
- whether
- who
- will
- with

PRE-PROCESSING SUMMARY

Indexing automatically scans the web page downloaded by the crawlers for the most important words and converts these to terms following a sequence of steps involving:

- case folding/change
- punctuation removal
- stop word removal
- stemming or lemmatisation

- These words are then weighted (next topic) and stored as the *representation of the web page*