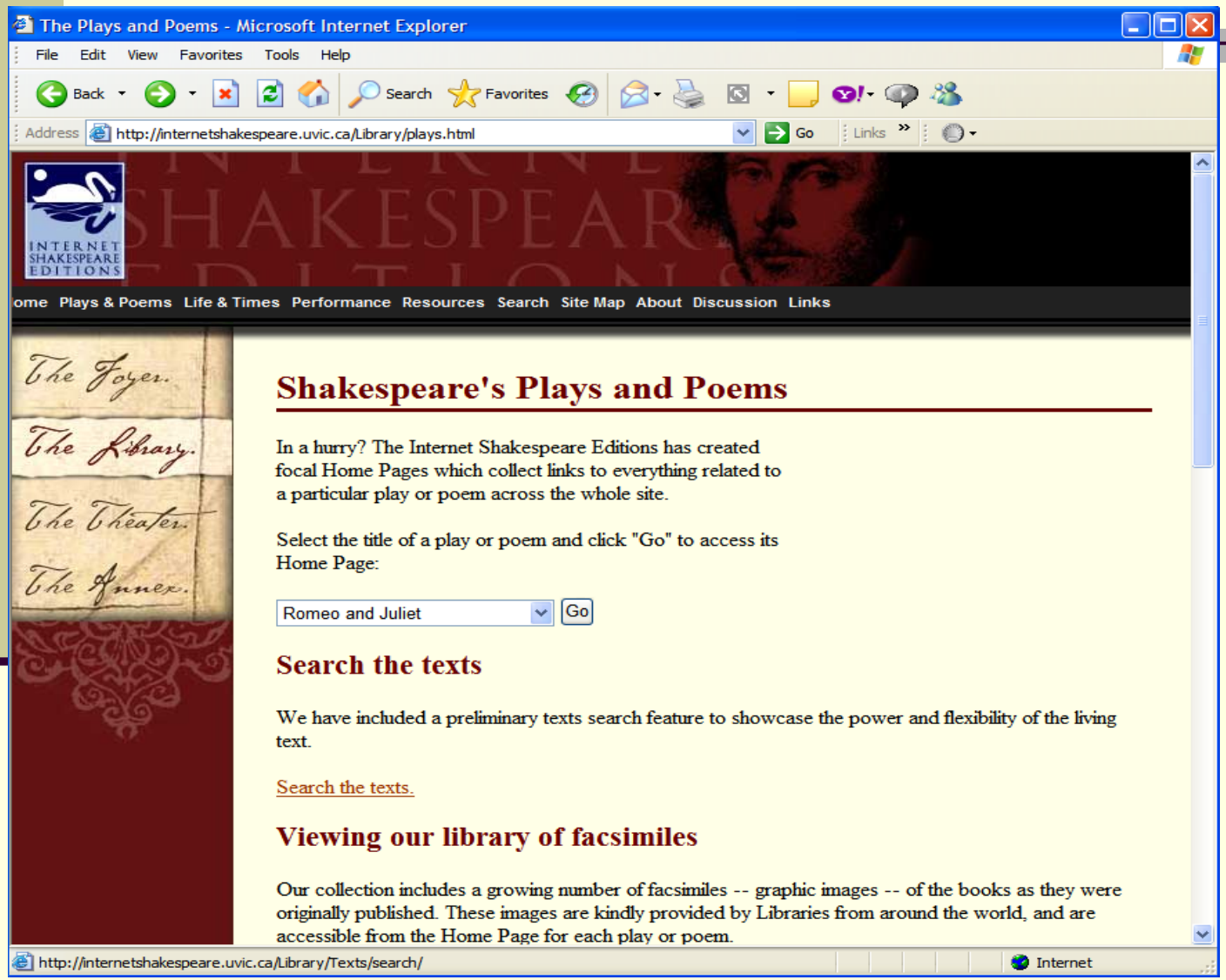
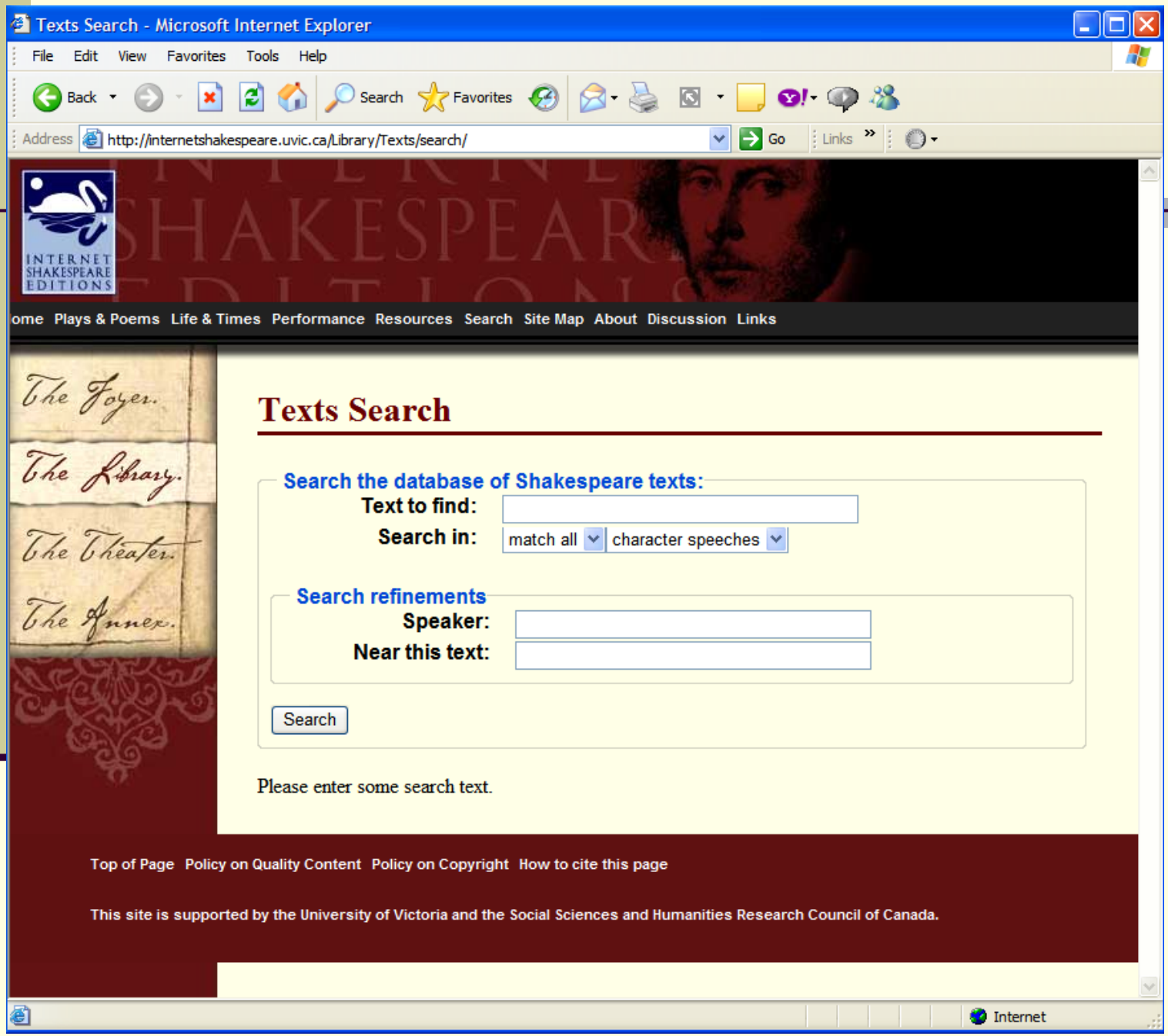
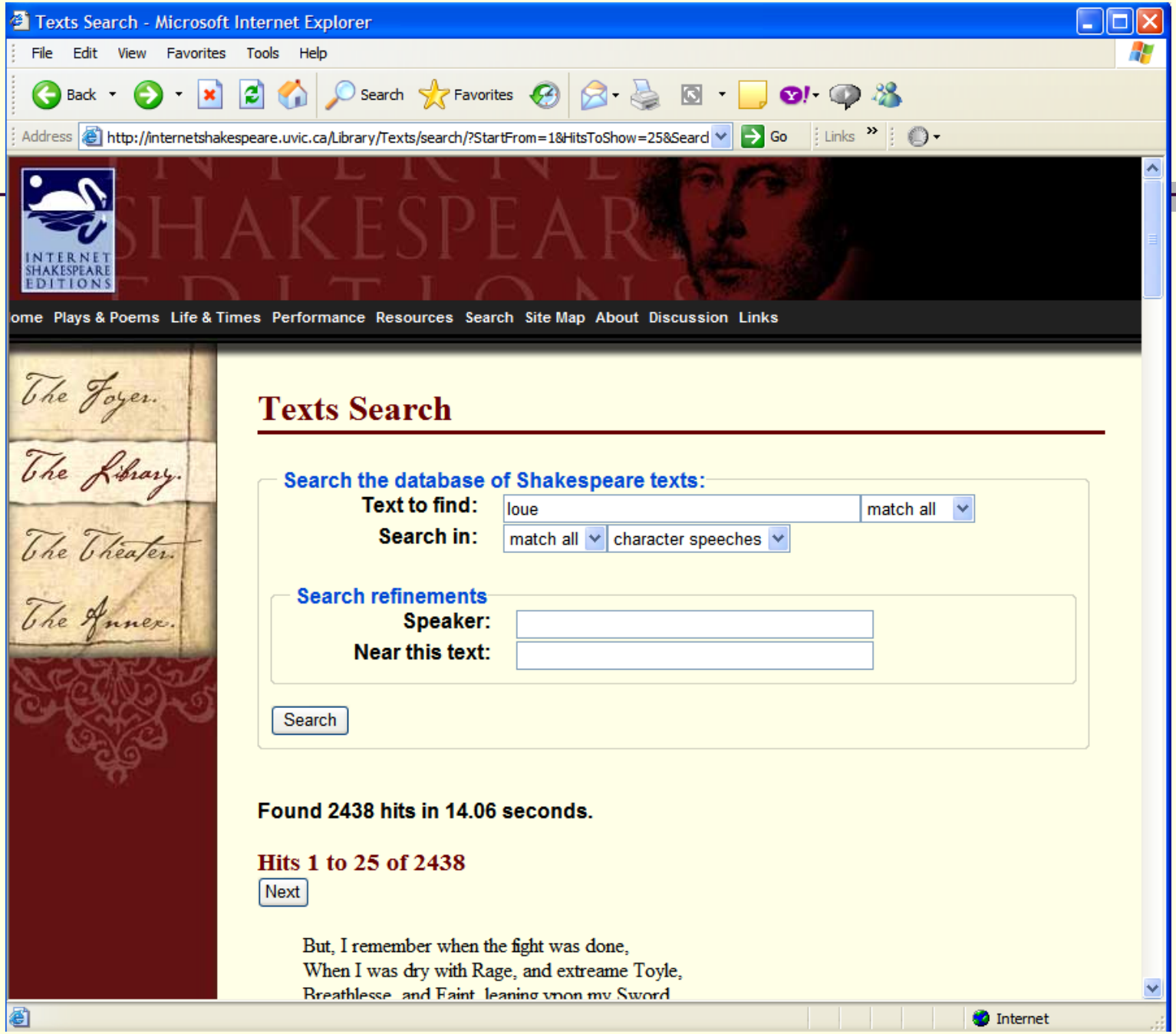
Preferential Infinitesimals for Information Retrieval

Maria Chowdhury, Alex Thomo and William Wadge
University of Victoria, Canada

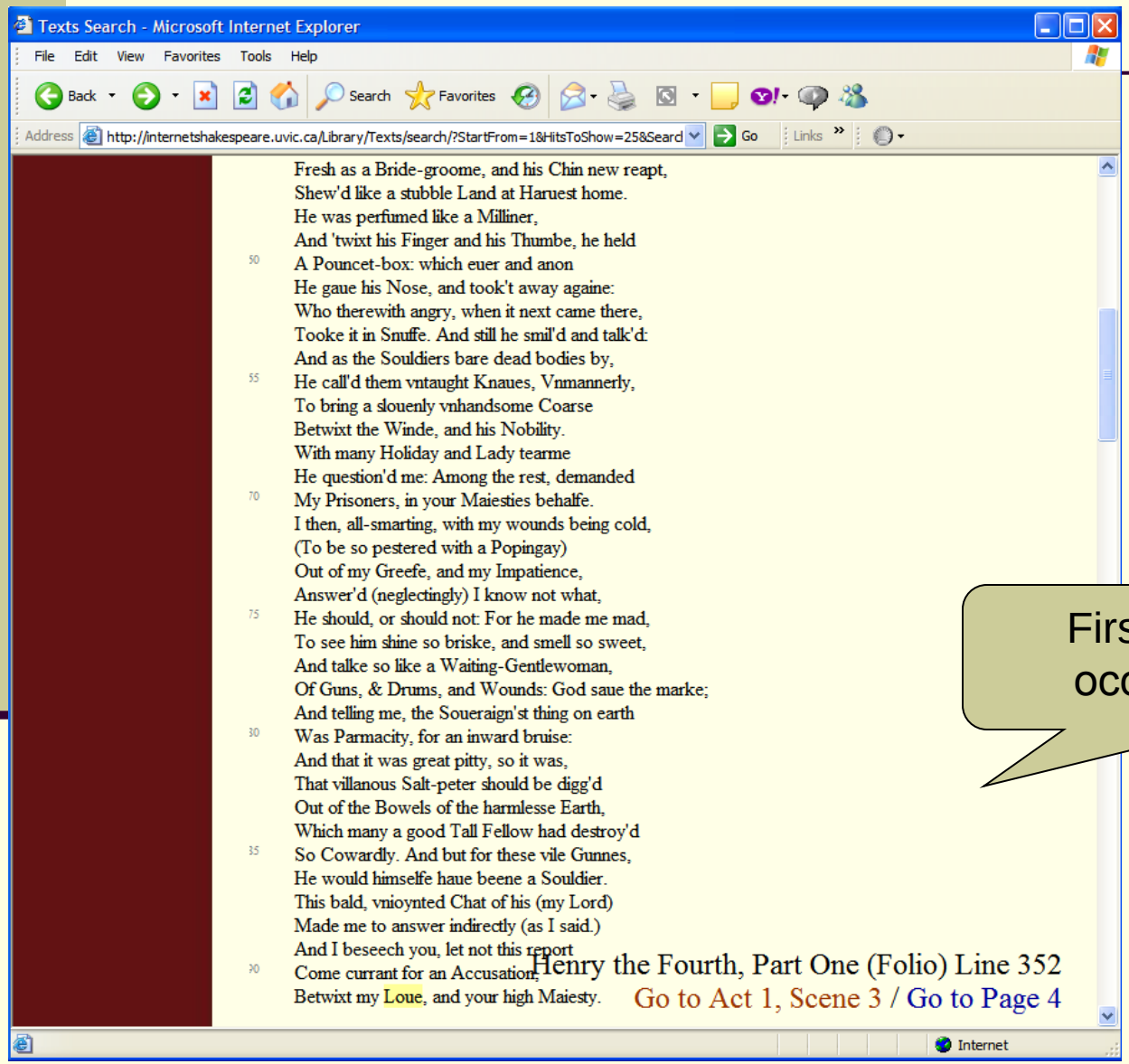
Searching


The Plays and Poems - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://internetshakespeare.uvic.ca/Library/plays.html
Go Links
INTERNET SHAKESPEARE EDITIONS
ome Plays & Poems Life & Times Performance Resources Search Site Map About Discussion Links
The Foyer.
The Library.
The Theater.
The Annex.
Shakespeare's Plays and Poems
In a hurry? The Internet Shakespeare Editions has created focal Home Pages which collect links to everything related to a particular play or poem across the whole site.
Select the title of a play or poem and click "Go" to access its Home Page:
Romeo and Juliet
Go
Search the texts
We have included a preliminary texts search feature to showcase the power and flexibility of the living text.
Search the texts.
Viewing our library of facsimiles
Our collection includes a growing number of facsimiles -- graphic images -- of the books as they were originally published. These images are kindly provided by Libraries from around the world, and are accessible from the Home Page for each play or poem.
http://internetshakespeare.uvic.ca/Library/Texts/search/
Internet


Texts Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://internetshakespeare.uvic.ca/Library/Texts/search/
Go Links
INTERNET SHAKESPEARE EDITIONS
ome Plays & Poems Life & Times Performance Resources Search Site Map About Discussion Links
The Foyer.
The Library.
The Theater.
The Annex.
Texts Search
Search the database of Shakespeare texts:
Text to find:
Search in:
match all
character speeches
Search refinements
Speaker:
Near this text:
Search
Please enter some search text.
Top of Page Policy on Quality Content Policy on Copyright How to cite this page
This site is supported by the University of Victoria and the Social Sciences and Humanities Research Council of Canada.
Internet


Texts Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://internetshakespeare.uvic.ca/Library/Texts/search/?StartFrom=1&HitsToShow=25&Search
Go Links
INTERNET SHAKESPEARE EDITIONS
ome Plays & Poems Life & Times Performance Resources Search Site Map About Discussion Links
The Foyer.
The Library.
The Theater.
The Annex.
Texts Search
Search the database of Shakespeare texts:
Text to find: loue
match all
Search in: match all
character speeches
Search refinements
Speaker:
Near this text:
Search
Found 2438 hits in 14.06 seconds.
Hits 1 to 25 of 2438
Next
But, I remember when the fight was done,
When I was dry with Rage, and extreame Toyle,
Breathlesse, and Faint, leaning vpon my Sword
Internet

(No) Ranking …



First hit has only one occurrence of “loue”

(No) Ranking …

Well, do not then. For since you loue me not,
I will not loue my selfe. Do you not loue me?
Nay, tell me if thou speak'st in iest, or no.

Henry the Fourth, Part One (Folio) Line 942
Go to Act 2, Scene 3 / Go to Page 8

This hit with three occurrences of "loue" appears 10th in the list.

IR Ranking

- Term frequency (TF)
- Inverse document frequency (IDF)
- TF-IDF weighting Scheme

Term frequency (TF)

- Normalized term frequency of t_i in d_j is

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, \ldots, f_{mj}\}}$$

maximum is computed over the terms that appear in document d_j

Inverse document frequency (IDF)

- Inverse document frequency of term t_i is

$$idf_i = \log \frac{n}{n_i}$$

TF×IDF weighting scheme

- TF×IDF weighting scheme assigns to term t_i a weight in document d_j given by

$$tf\text{-}idf_{ij} = tf_{ij} \times idf_i$$

- Then each document is represented as a vector of tf-idf values.
- A query, e.g. loue, woman, is also represented as vector.
- Document-query similarities are computed using cosine similarity.

Search the database of Shakespeare texts:

Text to find: loue+woman | match all

Search in: match all | character speeches

Search refinements

Speaker:

Near this text:

Search

Found 42 hits in 2.6109999999999998 seconds.

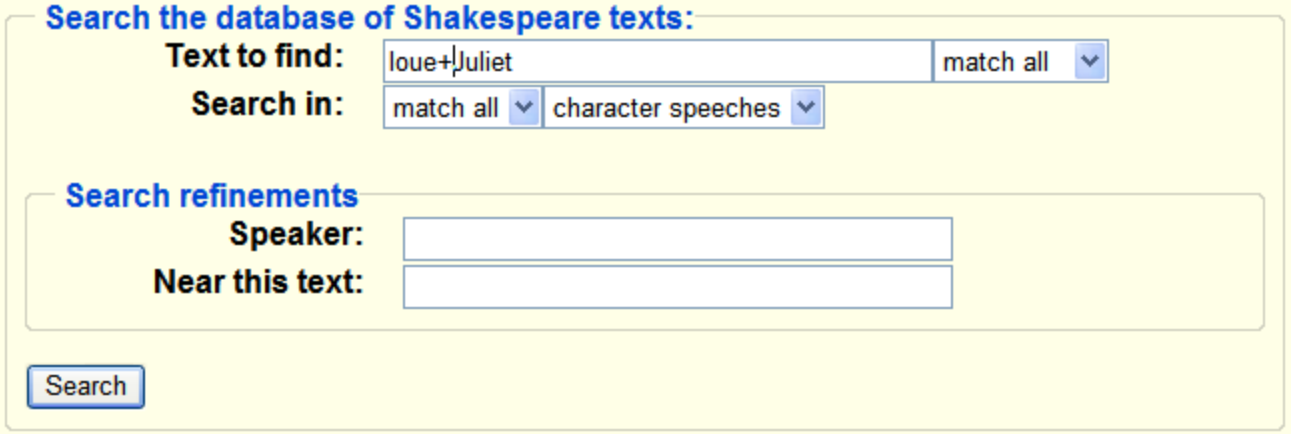
Hits 1 to 25 of 42

Next

And when I am a horsebacke, I will sweare
I loue thee infinitely. But hearke you *Kate*,
I must not haue you henceforth, question me,
Whether I go: nor reason whereabout.
Whether I must, I must: and to conclude,
This Euening must I leaue thee, gentle *Kate*.
I know you wise, but yet no further wise
Then *Harry Percies* wife. Constant you are,
But yet a woman: and for secrecie,
No Lady closer. For I will beleeue
Thou wilt not vtter what thou do'st not know,
And so farre wilt I trust thee, gentle Kate.

Henry the Fourth, Part One (Folio) Line 946

Go to Act 2, Scene 3 / Go to Page 8



Found 46 hits in 2.002 seconds.

Hits 1 to 25 of 46

Next

Romeo and Juliet (Folio) Line 721
Go to Act 1, Scene 5 / Go to Page 6

Iul. My onely Loue sprung from my onely hate,

Romeo and Juliet (Folio) Line 722
Go to Act 1, Scene 5 / Go to Page 6

Too early seene, vnknowne, and knowne too late,
Prodigious birth of Loue it is to me,
That I must loue a loathed Enemie.

Romeo and Juliet (Folio) Line 828
Go to Act 2, Scene 2 / Go to Page 7

Denie thy Father and refuse thy name:
Or if thou wilt not, be but sworne my Loue,
And Ile no longer be a *Capulet*.

Preferences

loue &woman

Semantics:

- Rank documents w.r.t. “loue” first.
- Among documents ranked equally w.r.t. “loue” those with more occurrences of “woman” should be ranked higher.
- Don’t ignore documents with “loue” occurrences, but without “woman” occurrences.

Google

- Suppose a user wants to retrieve documents about “image-information-retrieval” and among those, he would be interested in documents mentioning “google-search” and “google-ranking”.

- What would happen if the user gives the following query:

image-information-retrieval, google-search, google-ranking

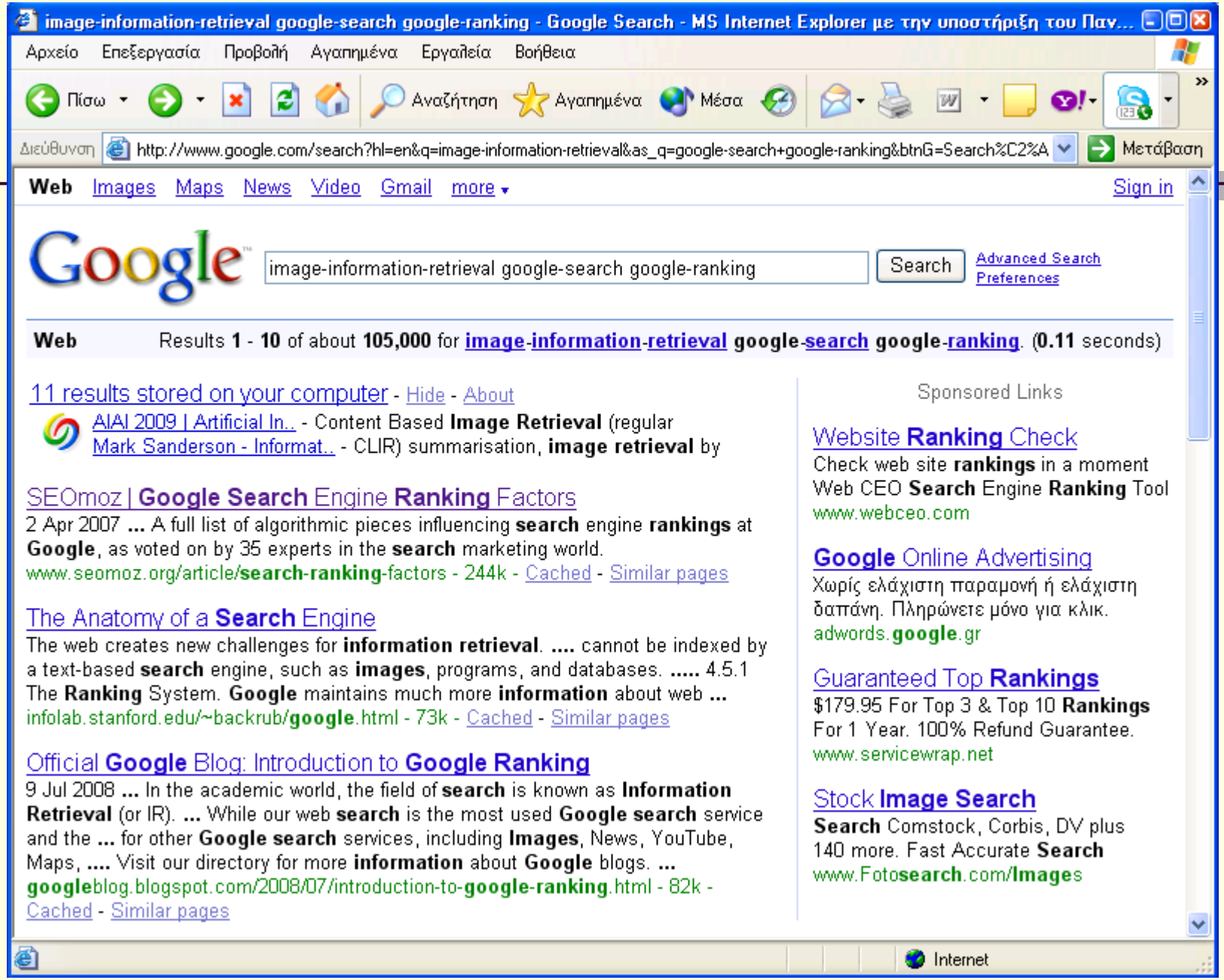


image-information-retrieval google-search google-ranking - Google Search - MS Internet Explorer με την υποστήριξη του Παν...
Αρχείο Επεξεργασία Προβολή Αγαπημένα Εργαλεία Βοήθεια
Πίσω
Αναζήτηση
Αγαπημένα
Μέσα
Διεύθυνση
http://www.google.com/search?hl=en&q=image-information-retrieval&as_q=google-search+google-ranking&btnG=Search%C2%A
Μετάβαση
Web Images Maps News Video Gmail more
Sign in
Google
image-information-retrieval google-search google-ranking
Search
Advanced Search
Preferences
Web Results 1 - 10 of about 105,000 for image-information-retrieval google-search google-ranking. (0.11 seconds)
11 results stored on your computer - Hide - About
AIAI 2009 | Artificial In.. - Content Based Image Retrieval (regular
Mark Sanderson - Informat.. - CLIR) summarisation, image retrieval by
SEOmoz | Google Search Engine Ranking Factors
2 Apr 2007 ... A full list of algorithmic pieces influencing search engine rankings at Google, as voted on by 35 experts in the search marketing world.
www.seomoz.org/article/search-ranking-factors - 244k - Cached - Similar pages
The Anatomy of a Search Engine
The web creates new challenges for information retrieval. .... cannot be indexed by a text-based search engine, such as images, programs, and databases. ..... 4.5.1 The Ranking System. Google maintains much more information about web ...
infolab.stanford.edu/~backrub/google.html - 73k - Cached - Similar pages
Official Google Blog: Introduction to Google Ranking
9 Jul 2008 ... In the academic world, the field of search is known as Information Retrieval (or IR). ... While our web search is the most used Google search service and the ... for other Google search services, including Images, News, YouTube, Maps, ... Visit our directory for more information about Google blogs. ...
googleblog.blogspot.com/2008/07/introduction-to-google-ranking.html - 82k - Cached - Similar pages
Sponsored Links
Website Ranking Check
Check web site rankings in a moment Web CEO Search Engine Ranking Tool
www.webceo.com
Google Online Advertising
Χωρίς ελάχιστη παραμονή ή ελάχιστη δαπάνη. Πληρώνετε μόνο για κλικ.
adwords.google.gr
Guaranteed Top Rankings
$179.95 For Top 3 & Top 10 Rankings For 1 Year. 100% Refund Guarantee.
www.servicewrap.net
Stock Image Search
Search Comstock, Corbis, DV plus 140 more. Fast Accurate Search
www.Fotosearch.com/Images
Internet

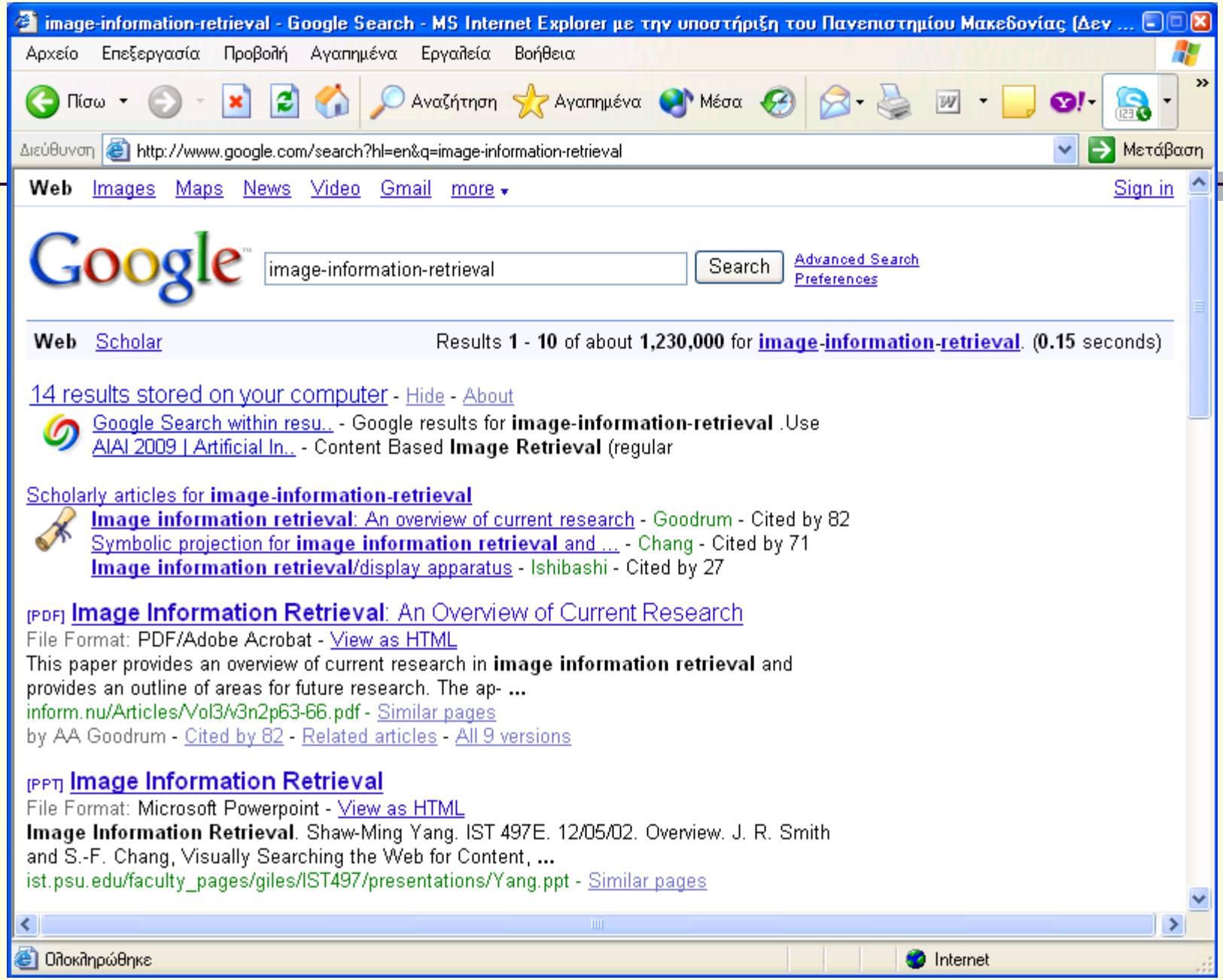


image-information-retrieval - Google Search - MS Internet Explorer με την υποστήριξη του Πανεπιστημίου Μακεδονίας (Δεν ...
Αρχείο Επεξεργασία Προβολή Αγαπημένα Εργαλεία Βοήθεια
Πίσω
Αναζήτηση
Αγαπημένα
Μέσα
Διεύθυνση
http://www.google.com/search?hl=en&q=image-information-retrieval
Μετάβαση
Web Images Maps News Video Gmail more
Sign in
Google
image-information-retrieval
Search
Advanced Search
Preferences
Web Scholar
Results 1 - 10 of about 1,230,000 for image-information-retrieval. (0.15 seconds)
14 results stored on your computer - Hide - About
Google Search within resu.. - Google results for image-information-retrieval .Use
AIAI 2009 | Artificial In.. - Content Based Image Retrieval (regular
Scholarly articles for image-information-retrieval
Image information retrieval: An overview of current research - Goodrum - Cited by 82
Symbolic projection for image information retrieval and ... - Chang - Cited by 71
Image information retrieval/display apparatus - Ishibashi - Cited by 27
[PDF] Image Information Retrieval: An Overview of Current Research
File Format: PDF/Adobe Acrobat - View as HTML
This paper provides an overview of current research in image information retrieval and provides an outline of areas for future research. The ap- ...
inform.nu/Articles/Vol3/v3n2p63-66.pdf - Similar pages
by AA Goodrum - Cited by 82 - Related articles - All 9 versions
[PPT] Image Information Retrieval
File Format: Microsoft Powerpoint - View as HTML
Image Information Retrieval. Shaw-Ming Yang. IST 497E. 12/05/02. Overview. J. R. Smith and S.-F. Chang, Visually Searching the Web for Content, ...
ist.psu.edu/faculty_pages/giles/IST497/presentations/Yang.ppt - Similar pages
Ολοκληρώθηκε
Internet

Google

music-information-retrieval:100, google-search, google-ranking

- "100 times more important" seems quite convincing in colloquial talking!
- However, what if, *according to Google*, documents about google-search were 1000 times more important than documents about music-information-retrieval?

Infinitesimals

- ε is said to be infinitely small or infinitesimal iff $-a < \varepsilon < a$ for every $a \in R^+$.
  - *Ref*. Jerome Keisler. *Infinitesimal Calculus*.
- For $a, b, r, s \in R^+$, we have
  - $a\varepsilon^r < b\varepsilon^s$ iff $r > s$
  - $a\varepsilon^r < b\varepsilon^r$ iff $a < b$.
- Examples
  - $10\,\varepsilon^2 < \varepsilon$
  - $1{,}000{,}000\,\varepsilon^2 < \varepsilon$
  - $5\varepsilon + 7\varepsilon^2 + 3\varepsilon^3 < 6\varepsilon + 100\varepsilon^2$

Google

music-information-retrieval, google-search:ε, google-ranking:ε^2.

Or say…

music-information-retrieval, google-search:2ε, google-ranking:ε.

Document structure

paper → preamble body

preamble → title author+ abstract keywords

body → introduction section* related-work? References

Document structure with weights

paper → (preamble:3) (body:1)

preamble → (title:2) (author:1)+ (abstract:1) (keywords:10)

body → (introduction:2) (section:1)* (related-work:ε)? (references:ε^2)

Normalizing weights

- Since an annotated element can be nested inside other elements, which can be annotated as well, the question is: How to compute the actual weight of an element in a DTD?
- Multiply weights along ancestor path?
- What we want is “an element to never be more important than its parent.”

Normalizing weights

paper → (preamble : 1) (body : 1/3)

preamble → (title : 1/5) (author : 1/10)+ (abstract : 1/10)
(keywords : 1)

body → (introduction : 1) (section : 1/2)*
(related-work : $\varepsilon/2$)? (references : $\varepsilon^2/2$)

TF revisited

- Suppose that term t_i occurs f_{ijk} times in element e_k of document d_j

$$tf_{ij} = \frac{\sum_k w_k f_{ijk}}{max\{\sum_k w_k f_{1jk}, \ldots, \sum_k w_k f_{mjk}\}}$$

TF revisited – Example

- Suppose that t_i occurs
  - once in the keywords element,
  - twice in the abstract element,
  - three times in the section elements,
  - four times in the related-work element, and
  - twice in the references element

  of document d_j.

- Then, the numerator of the tf_{ij} fraction will be

  $1 \cdot 1 \cdot 1 + 1 \cdot (1/10) \cdot 2 + (1/3) \cdot (1/2) \cdot 3 + (1/3) \cdot (\varepsilon/2) \cdot 4 + (1/3) \cdot (\varepsilon^2/2) \cdot 2 =$

  $$1.7 + (2/3) \cdot \varepsilon + (1/3) \cdot \varepsilon^2.$$

IDF revisited

- For an element-weight pair (e_h,w_h), let
  - n^h be the total number of such elements in the XML documents in collection.
- Suppose that a term t_i occurs in n_{hi} of them.
- Then, we define the IDF of t_i wrt these elements as

$$idf_{hi} = \log \frac{n_h}{n_{hi}}$$

IDF revisited

- Next, we define the IDF score of a term t_i with respect to the whole document collection as

$$idf_i = \frac{\sum_h w_h \cdot idf_{hi}}{\sum_h w_h}$$

TF×IDF weighting scheme

- TF×IDF weighting scheme assigns to term t_i a weight in document d_j given by

$$tf\text{-}idf_{ij} = tf_{ij} \times idf_i$$

- Then each document is represented as a vector of tf-idf values.
- A query, is also represented as vector.
  - The values are exactly those hyperreal numbers specified by the user multiplied by the IDF scores of the terms.

Documents and queries

- We rank the documents by computing their similarly score with respect to a query q.

$$cosine(\mathbf{w}_j, \mathbf{w}_q) = \frac{\langle \mathbf{w}_j, \mathbf{w}_q \rangle}{||\mathbf{w}_j|| \times ||\mathbf{w}_q||} = \frac{\sum_{i=1}^{m} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{m} w_{ij}^2} \times \sqrt{\sum_{i=1}^{m} w_{iq}^2}}$$

Experiments

- Corpus I: On-line Internet Shakespeare Edition of the English Department, University of Victoria
  - 33,000 speeches.
- Corpus II: An INEX (INitiative for the Evaluation of XML retrieval) corpus.
  - Numerous XML documents of moderate size.
  - Topics of documents vary from climate change to space exploration.
  - We preferentially annotated the DTD of this collection.
- Representative queries given in the full version:
  - http://www.cs.uvic.ca/~thomo/publications/aiai09.pdf

Example

Q: Norway climate:ε information: ε^2,

Our System:

<title>Climate in Norway< /title>

<description>Find information about the climate in Norway in summer.< /description>

<narrative>I would like to travel to Norway in July, but I have no idea about the weather. I don't know which clothes to put in my bag. To be relevant, a paragraph or a document should let me know the mean average temperature in this season and the precipitation level, or just give me an information like continental climate or polar climate...

< /narrative>

Example

Q: Norway climate:ε information: ε^2,

Classical System:

<title>Ontology< /title>

<description>Find **information** about ontology.</description>

<narrative>An ontology is typically a hierarchical data structure containing all the relevant entities and their relationships and rules within that domain (e.g., a domain ontology). …For it plays a very important role in **information** extraction, entity recognition etc., I would like to learn more **information** about the introduction of it and how it works. Besides, I expect to find relevant **information** as elements in larger documents …

< /narrative>

Thank you!

References

Maria Chowdhury, Alex Thomo, William W. Wadge. Preferential Infinitesimals for Information Retrieval. AIAI 2009: 113-125

Maryam Khezrzadeh, Alex Thomo, William W. Wadge. Harnessing the power of "favorites" lists for recommendation systems. RecSys 2009: 289-292