CSE 494/598 INFORMATION RETRIEVAL, MINING AND INTEGRATION ON THE INTERNET

PROJECT PART – 3

REPORT

by

Manikandan Vijayakumar
1204624377

Task 1: K-means

Given a query, obtain the top-N documents using TF/IDF. Cluster the results using the simple K-means algorithm which randomly picks the initial k centroids.

1. Cluster the results of the queries given below, with
    a. Number of documents clustered, N = 50
    b. Number of clusters, k = 3
    c. Similarity algorithm = Vector similarity (TF-IDF) without PageRank

Submit a printout of the document numbers of the top-3 documents in each cluster.

Clusters (K=3) N=50 and top 3 docs in cluster	medic care	employee benefits	parking decal	admissions	languages
Cluster 1	[22816] [4432] [22443]	[223] [787] [4627]	[20828] [19597] [19375]	[938] [935] [992]	[16277] [1247] [16700]
Cluster 2	[23565] [23559] [23566]	[4599] [4591] [4592]	[2406] [2360] [2366]	[1043] [1048] [1049]	[1374] [14358] [14421]
Cluster 3	[20952] [359] [19875]	[4543] [775]	[649] [4595] [648]	[1075] [1077] [1073]	[20928] [14437] [20888]

2. **For each cluster you obtained above, determine short "summaries" of the clusters, using keywords that most distinguish those clusters from other clusters?**

    As the summaries of the clusters aren't mentioned to be in specific format, I considered using the hashtags to represent summaries for the clusters as it is more popular with the social media like twitter, facebook etc.,

    Following are the list of summaries for the clusters and the small snippet for the documents corresponding to the summaries to give a better understanding on why those documents were clustered together.

    Hashtags formed for the summaries are the best keywords that are found suitable for the clusters. Description and approach are defined further.

Table 1.0

medic care	
Cluster	**Cluster Summary**
Cluster 1	#workingatasu,#provost,#index,#care,#www.asu.edu,#nbsp,#index.html,#li, #child,#childcare
Documents	**Document Snippets**
[22816]	#40,#care,#child,#nodecor,#aaaaaa
[4432]	#strong,#expenses,#spending,#care,#fsa
[22443]	#provost,#animal,#index,#index.html,#www.asu.edu
Cluster	**Cluster Summary**
Cluster 2	#p7defmark,#child,#option,#care,#div,#studentaffairs,#ha,#begin,#leave, #dot_clear.gif
Documents	**Document Snippets**
[23565]	#child,#care,#dt,#dd,#p7defmark
[23559]	#child, #p7defmark,#care,#option,#studentaffairs
[23566]	#child,#p7defmark,#subsidy,#care,#ccampis
Cluster	**Cluster Summary**
Cluster 3	#care,#news,#health,#li,#nursing,#www.asu.edu,#instanceparam,#mhi,#leave,#value
Documents	**Document Snippets**
[20952]	#care,#mhi,#health,#news,#malloch
[359]	#leave,#care,#employee,#provider,#health
[19875]	#news,#li,#care,#decisions,#health

employee benefits	
Cluster	**Cluster Summary**
Cluster 1	#employee,#leave,#spphr.gif,#pay,#rule,#160,#spp,#horizontal,#146,#dt
Documents	**Document Snippets**
[223]	#employee,#interest,#relative,#rule,#acdhr.gif
[787]	#leave, #employee,#care,#spphr.gif,#child
[4627]	#hpr,#employee,#hrms,#packet,#payroll
Cluster	**Cluster Summary**
Cluster 2	#sunsans,#employee,#swiss,#geneva,#regular,#arial,#psciis.dll,#lms.asu.edu,#face,#stc
Documents	**Document Snippets**
[4599]	#benefits,#employee,#orientation,#your,#sunsans

[4591]	#orientation,#benefits,#arial,#employee,#face
[4592]	#nbsp,#benefits,#employee,#sunsans,#margin
Cluster	**Cluster Summary**
Cluster 3	#employee,#retirement,#certificates,#recognition,#classification,#system,#images_advisor, #mm_nbgroup,#extraordinary,#nonexempt
Documents	**Document Snippets**
[4543]	#employee,#retirement,#certificates,#recognition,#classification
[775]	#extraordinary,#classification,#nonexempt,#404,#employee

parking decal	
Cluster	**Cluster Summary**
Cluster 1	#parking,#news,#li,#lot,#spaces,#www.asu.edu,#transit,#structure,#helton,#lots
Documents	**Document Snippets**
[20828]	#parking,#news,#li,#rates,#force
[19597]	#parking,#news,#helton,#li,#spaces
[19375]	#parking,#lot,#debate,#news,#october
Cluster	**Cluster Summary**
Cluster 2	#parking,#pts,#transit,#projects_files,#dps,#decal,#tbody,#improvement,#maintenance, #style4
Documents	**Document Snippets**
[2406]	#style4,#parking,#decal,#transit,#projects_files
[2360]	#disabled,#pts,#dps,#decal,#placard
[2366]	#pts,#dps,#decal,#parking,#disabled
Cluster	**Cluster Summary**
Cluster 3	#decal,#parking,#ptshr.gif,#margin,#pts,#subheader,#gate,#decals,#vehicle,#downtown
Documents	**Document Snippets**
[649]	#decal,#gate,#card,#stolen,#replacement
[4595]	#decal,#parking,#sunsans,#swiss,#geneva
[648]	#decal,#gate,#ptshr.gif,#stolen,#card

admissions	
Cluster	**Cluster Summary**
Cluster 1	#admissions,#menulink,#undoclass,#changeclass,#titles,#div,#whychooseasu, #onmouseout,#onmouseover,#_blank
Documents	**Document Snippets**
[938]	#admissions,#undoclass,#menulink,#changeclass,#titles

[935]	#admissions,#menulink,#undoclass,#changeclass,#titles
[992]	#admissions,#menulink,#undoclass,#changeclass,#_blank
Cluster	**Cluster Summary**
Cluster 2	#admissions,#steps,#undoclass,#menulink,#changeclass,#titles,#onmouseout,#onmouseover,#div,#delm.style.display
Documents	**Document Snippets**
[1043]	#admissions,#steps,#menulink,#undoclass,#changeclass
[1048]	#admissions,#steps,#menulink,#undoclass,#changeclass
[1049]	#admissions,#steps,#undoclass,#menulink,#changeclass
Cluster	**Cluster Summary**
Cluster 3	#admissions,#undoclass,#menulink,#changeclass,#tempecampus,#visitcampus,#titles,#div,#_blank,#onmouseout
Documents	**Document Snippets**
[1075]	#admissions,#tempecampus,#menulink,#undoclass,#changeclass
[1077]	#admissions,#tempecampus,#visitcampus,#menulink,#undoclass
[1073]	#admissions,#tempecampus,#visitcampus,#undoclass,#menulink

languages	
Cluster	**Cluster Summary**
Cluster 1	#blockquote,#labriola,#lib,#hayden,#dc,#languages,#language,#palgen,#libraries,#strong
Documents	**Document Snippets**
[16277]	#blockquote,#labriola,#hayden,#language,#indian
[1247]	#990000,#strong,#asian,#tel,#casimages
[16700]	#palgen,#lib,#dc,#southeast,#se_asia
Cluster	**Cluster Summary**
Cluster 2	#dll,#style13,#clas,#ffcc33,#menu,#style4,#udm,#class,#style12,#span
Documents	**Document Snippets**
[1374]	#dll,#style13,#clas,#japanese,#languages
[14358]	#style13,#ideograph,#align:justify,#dll,#justify:inter
[14421]	#dll,#style13,#clas,#ffcc33,#menu
Cluster	**Cluster Summary**
Cluster 3	#cli,#languages,#news,#slavic,#russian,#li,#reesc,#div,#blocktext,#literatures
Documents	**Document Snippets**
[20928]	#cli,#news,#li,#armenian,#languages
[14437]	#reesc, #slavic,#russian,#languages,#blocktext
[20888]	#news,#li,#study,#nsep,#languages

Explain how you obtained these summaries. You are free to come up with your own strategy for finding these summaries, there is no set algorithm that you have to use.

Since I am using hashtags to represent the summaries, I focused on getting the best hashtags by finding the best or highly important keywords for the clusters and their documents.

Strategy:

I use hashmaps and class objects as data structures to store the information about the clusters which holds the term and corresponding weights in key value pairs. Similarly, the documents have the same along with the document id and distance to centroid. Once the clusters are formed, the cluster holds the list of all possible terms and their term weights. Having this, I iterate through all terms and filter the **top 10 terms in the cluster based on their term weights to get a meaningful and most relevant keyword for a cluster.**

I planned to use the top 10 keywords for the cluster summary as it well defines the reason for to be a cluster also we have the documents in the clusters sharing the same keywords. In case of document snippets I use the top 5 keywords as summary for the documents in the cluster as atleast top 2 keywords in the documents must be present in the summary of the cluster which helps to easily make the user understand the reason for having those documents in the same cluster.

3. **Pick any two queries from the set given below. Change the value of 'k' between 3 and 10. What do you observe? Why?**

Note: For this experiment, the initial centroids are picked randomly through a random number generator function.

I have picked the two following queries for a reason that they have maximum variations in the document sets so that they can be clustered for larger k values. Also one query has two terms and the other has one term in it which helps to find even enough variations in them:

- Medic care
- Languages

How does execution time change?

Execution time varies w.r.t to picking the initial centroids, and size of K. Based on picking the right centroid and the total number of clusters to be formed, the total number of iterations need to compute clustering of documents increases. This eventually leads to increase in time.

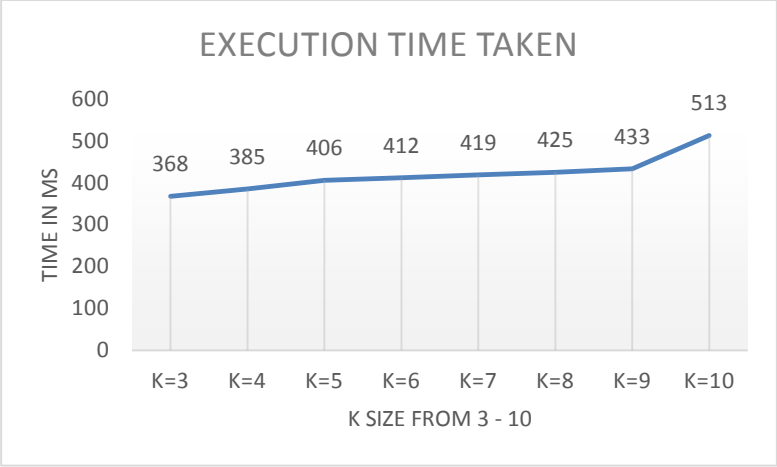
Query/k		K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Medi care	Iterations	2	3	5	6	6	6	5	4
	Time ms	368	385	406	412	419	425	433	513
languages	Iterations	3	3	3	5	5	6	5	5
	Time ms	32	56	66	65	67	75	87	93

Now I ran the same experiment for the query languages, keeping the query K=3 default and noticed the total number of iterations and time taken by randomly picking the centroid every time.

Query/k		K=3	K=3	K=3	K=3	K=3	K=3	K=3	K=3
Experiment id run count		1	2	3	4	5	6	7	8
languages	Iterations	3	2	4	3	2	3	7	6
	Time ms	43	31	47	26	21	28	55	47

From the above experiment, it's clear that the time taken depends on choosing the right centroid which eventually leads to more computation time.

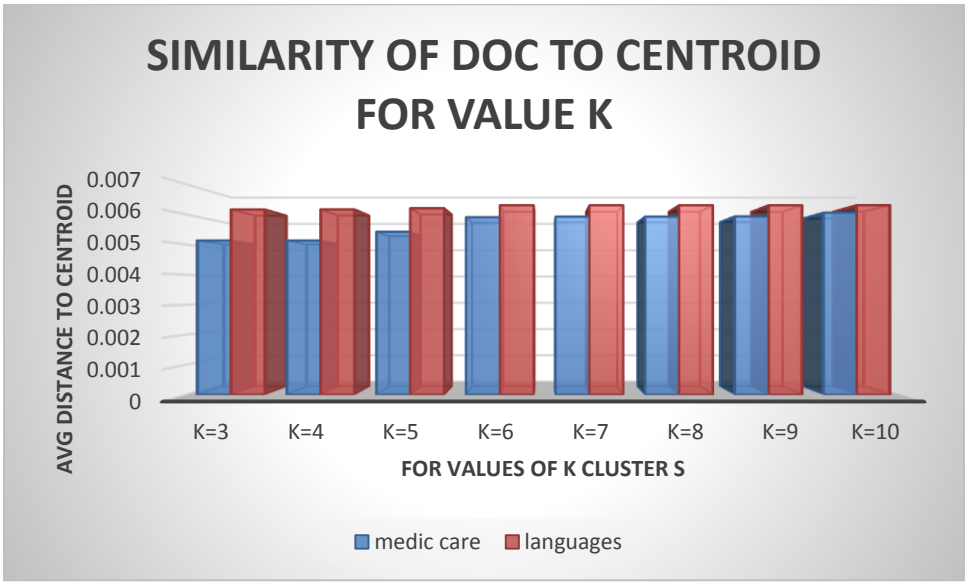
Also, as the number of clusters K increases, the computation time also gets increases. **For example, if K=3, then the distance is computed for three centroids with all the documents (3*N), say for K=10, in this case we need to compute distance for 10 clusters with all same all documents (10*N). This clearly says the execution time increase with increase the number of clusters.**



How does the similarity of the document to the centroid of the cluster change?

As the value of K changes, the average distance (similarity) of the documents in the cluster to the centroid varies as below for the query "medic care" and "languages"

K clusters		K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Average Distance	**Medic care**	0.005134	0.005134	0.005426	0.005909	0.005926	0.005926	0.005934	0.006058
	languages	0.006165	0.006168	0.006207	0.006304	0.006304	0.006304	0.006304	0.006307



From the above experiment, we can make a claim that the similarity of cluster increases for higher value of k.

How did the value of k affect the clustering? Justify with a couple of examples.

For the query, medic care:

Cluster:
#care,#news,#li,#health,#nursing,#www.asu.edu,#instanceparam,#mhi,#leave,#value

K	3	4	5	6	7	8	9	10
Documents In the cluster	[20952] [359] [19875]	[20952] [19875] [20590]	[20952] [19875] [20590]	[20952] [19875] [21023]	[20952] [20590] [21023]	[20952] [20590] [21023]	[20952] [20590] [21023]	[20952] [20590] [21023]
Summary of doc 1	#care,#mhi, #health #news ,#malloch,	#care,#mhi, #health, #news, #malloch,	#care, #mhi, #health, #news, #malloch,	#care,#mhi, #health, #news, #malloch,	#care,#mhi, #health, #news, #malloch	#care,#mhi, #health, #news, #malloch	#care,#mhi, #health, #news, #malloch	#care,#mhi, #health, #news, #malloch
Summary of doc 2	#leave,#care, #employee, #provider, #health,	#news,#li, #care, #decisions, #health,	#news, #li,#care, #decision #health,	#news,#li, #care, #decisions, #health,	#care,#ebp, #evidence, #patient, #news,	#care,#ebp, #evidence, #patient, #news,	#care,#ebp, #evidence, #patient, #news,	#care,#ebp, #evidence, #patient, #news,
Summary of doc 3	#news,#li, #care, #decisions, #health,	#care,#ebp, #evidence, #patient, #news,	#care, #ebp, #evidence, #patient, #news,	#chir,#news, #health, #care, #li,	#chir,#news, #health, #care, #li,	#chir,#news, #health, #care, #li,	#chir,#news, #health, #care, #li,	#chir,#news, #health, #care, #li,

So, increase of K causes clustering of more similar documents clustered together. From the above experiment, we could see the documents with more similar cluster remains the same with the increase of value K.

Also at k=3 the cluster with [20590] Is moved to different cluster and the cluster with document [20952], [19875], [21023] remain in the same cluster because of its similarity with respect to the keywords in the summary.

So, based on this the better value of K is 7. As the cluster remains the same for the previous iteration. It is one of the best heuristic way to compute the better centroid.

For the query, languages:

Cluster: #blockquote,#labriola,#hayden,#language,#i,#indian,#navajo,#languages,#b,#native,

For K=3;

For the summary with #russian #language, the below documents where in different clusters

Cluster 1: [14437],[20928],[20888] and another cluster containing: [20928],[14441],[20888]

[14437] - #reesc, #slavic, #russian, #languages, #blocktext

[14441] - #blocktext,#navtext,#russian,#div,#reesc

For K=6:

Both the documents in different clusters are clustered together:
[14437], [14439], [14441]

This experiment supports the previous claim that as the K cluster varies, it helps to cluster the most similar/relevant documents together.

Do the clusters seem to roughly correspond to the natural category of the pages?

Yes. The cluster seem to likely correspond to the natural category as the document summary and the cluster summary share more co-relation between them w.r.t to the Task 1 part 2 summary table. Based on the above Table 1.0, we can claim that the cluster seem to roughly correspond to the natural clusters

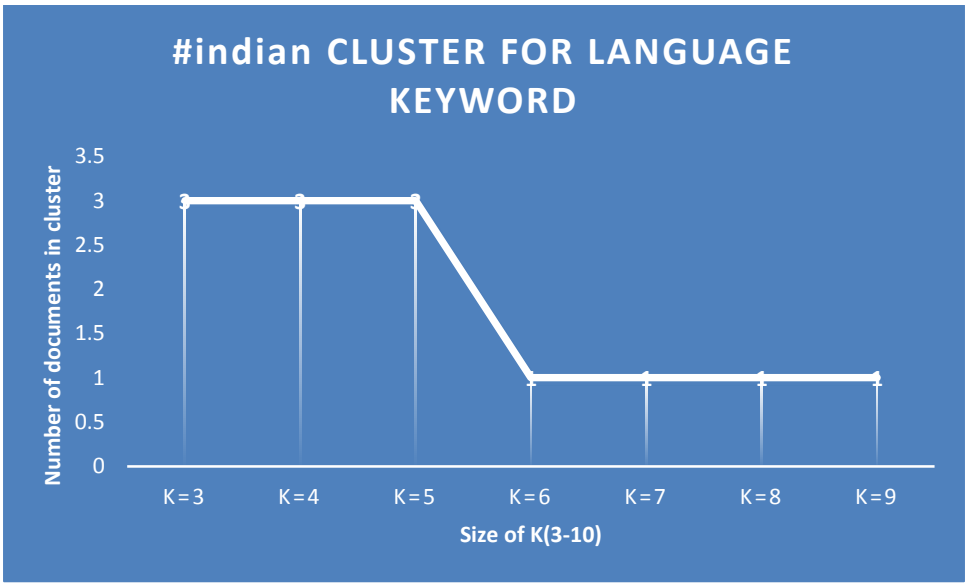
Did the value of k affect this? Mention any other observations you have?

Yes. It does affects by making the **natural category to specific category** of pages to the cluster. This can be apparently proved by following observation for the given query "languages and the cluster with the terms containing the summary as #indian and #languages as the maximum term weights in the cluster centroid.

One of the observations as example below to support claim:

Following table illustrates how the cluster size varies w.r.t to K value of cluster in our case I use the cluster with which has #indian in the summary to define it.

Query : Language								
Size of K	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Clusters	[16277] [70] [19929]	[16277] [5370] [16700]	[16277] [5370] [19929]	[16277]	[16277]	[16277]	[16277]	[16277]
#indian Cluster Size (# of docs)	3	3	3	1	1	1	1	1



Task 2: Similarity Pages

"Similar pages" feature, as seen on Google.

Code -> Attached in the project file

Implementation:

For the similar pages, a feature is provided to select the document id from the top k results to get the similar results w.r.t to the document id rather with the query given by user. When the user picks the document id, the document having the top k terms which are found to be the most important keywords of the document is passed as the query and the top k results are computed.

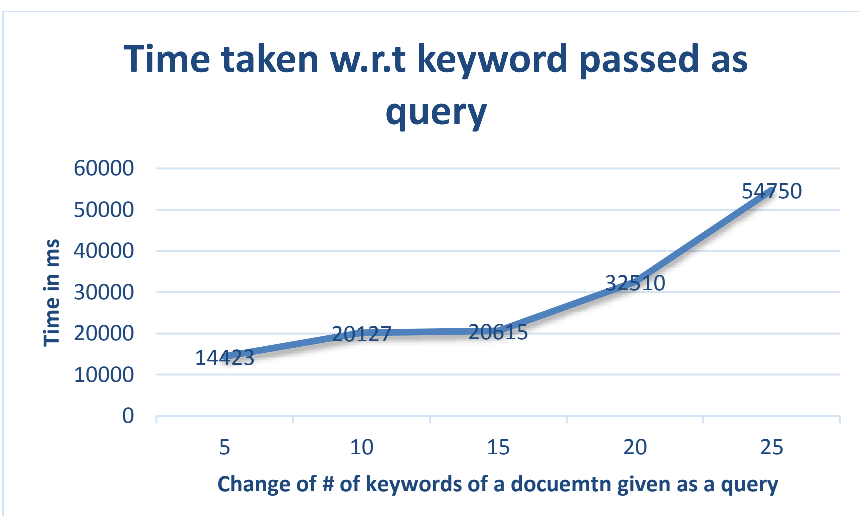
Challenges:

1. Time- time to compute the similar top k pages, as complete document cannot be passed as a query as it consume a lot time to get the results.
2. Choosing the number of keywords to be passed as a query to the system to get top k results. Picking the number of keywords from a document and the most relevant one is another major challenge.

Better the keywords chosen, to pass as a query, the precision will be high. And minimum the number of keywords, lesser the time taken to find the top k similar documents.

Following experiment on finding the precision and time taken for varied K value(Number of keywords to be passed as a input) for a query “language” and selecting the same document to get the similar documents, we can able to find a threshold on total number of keywords to be sent as a query.

Query : language					
Similar page Document selected: www.asu.edu%%languages%%slav%%index.html **[14437] -#reesc, #slavic,#russian,#languages,#blocktext**					
k->#of keywords passed as a query	5	10	15	20	25
Time	14423	20127	20615	32510	54750
Precision	8/10 =0.8	8/10 =0.8	8/10 =0.8	8/10 =0.8	8/10 =0.8
Top k IDF results	[14437] [14438] [14441] [14440] [14439] [14435] [14436] [14351] [14453] [14384]	[14437] [14438] [14441] [14440] [14439] [14435] [14436] [14351] [14453] [14384]	[14437] [14438] [14441] [14440] [14439] [14435] [14436] [14351] [14453] [14384]	[14437] [14438] [14441] [14440] [14439] [14435] [14436] [14351] [14453] [14384]	[14437] [14438] [14441] [14440] [14439] [14435] [14436] [14351] [14453] [14384]
Top K cluster results	[14437] [14438] [14441] [14440] [14439]	[14437] [14438] [14441] [14440] [14439]	[14437] [14438] [14441] [14440] [14439]	[14437] [14438] [14441] [14440] [14439]	[14437] [14438] [14441] [14440] [14439]



Based on the above experiment to threshold for K(number of keywords to be passed as a query to similar pages to the document, in our case I queried to get pages similar to Russian languages, I found the top k result documents obtained, remains the same even the k value is increased. So, sending a **top 5 keywords** of the document which is determined based on their idf value, is a better threshold to pass the keywords as a query to get the most similar pages w.r.t to that document. Not alone that, the time taken increases as the number of keyword passed as a query increases as shown in the graph, so it is better to pass **top 5 keywords of the document as a query to get the similar pages of the document**.

Conclusion: To get most similar pages for a document, pass the top 5 keywords as a query.

Extra Credit: "Phrase Search"

As mentioned in @159 post in piazza, providing support for searching with Shingles.

"EITHER: when someone enters a frequent shingle as a query, find pages that have that phrase and rank them higher than the regular TF/IDF results;
OR: when someone enters a phrase as a query with double quotes "hayden library" then only show documents that have that phrase."

Code: Submitted along with the project

Implementation:

To support frequent singles as a query, I used the second approach that was mentioned, by showing the top k documents with the phrase hayden and library in it. To do this, while fetching the documents with those term, I limit the documents that containing only those two phrases rather the documents that contains any one of them. Say the dcouments only with phrase "hayden libray" rather "hayden" or "library" By this I could able to fetch the most relevant documents with containing both the terms and rank the one with its phrase occurrence. I use the previous tf-idf similarity to rank those documents.

Experiment:

Query	Normal Search	Phrase Search
	Computer Science	"Computer Science"
# of documents fetched	6504	704
Time taken	267	22
Precision	2/10=0.2	8/10 = 0.8

For Normal IDF search: Computer Science

[20473]	www.asu.edu%%news%%research%%politicization_discussion_093004.htm
Snippet	ASU to Host Discussion on the Politicization of Science
[17727]	www.asu.edu%%lib%%resources%%db%%lnotecsi.htm
Snippet	Lecture Notes in Computer Science
[20327]	www.asu.edu%%news%%research%%brickyard_042503.htm
Snippet	ASU creates new computer, information sciences institute
[17398]	www.asu.edu%%lib%%noble%%scirefrm%%earlylit.htm
Snippet	Early Literature Sources
[17832]	www.asu.edu%%lib%%resources%%db%%scionlin.htm
Snippet	Science Online
[23746]	www.asu.edu%%studentaffairs%%reslife%%resnet.htm
Snippet	Residential Life provides and maintains 13 residence hall computer labs for student usage.
[20920]	www.asu.edu%%news%%stories%%200606%%20060621_morrison.htm
Snippet	Study shows Arizonans recognize benefits of science, technology
[4041]	www.asu.edu%%graduate%%gapd%%gradcouncil%%minutes%%tempe%%062706.htm
Snippet	TEMPE GRADUATE COUNCIL
[19916]	www.asu.edu%%news%%faculty_students%%capitol_scholars_041703.htm
Snippet	Capitol Scholars Program welcomes 21 undergraduates
[18382]	www.asu.edu%%lib%%systems%%newsletter%%jannews.htm
Snippet	How to Protect Your Computer

For Phrase search: "Computer Science"

[17727]	www.asu.edu%%lib%%resources%%db%%lnotecsi.htm
Snippet	Lecture Notes in Computer Science
[20327]	www.asu.edu%%news%%research%%brickyard_042503.htm
Snippet	ASU creates new computer, information sciences institute
[4041]	www.asu.edu%%graduate%%gapd%%gradcouncil%%minutes%%tempe%%062706.htm
Snippet	TEMPE GRADUATE COUNCIL
[17857]	www.asu.edu%%lib%%resources%%db%%synth.htm
Snippet	Synthesis Digital Library
[17382]	www.asu.edu%%lib%%noble%%library%%bestind.htm
Snippet	Best Indexes and Databases for Science, Health and Engineering Topics
[20594]	www.asu.edu%%news%%stories%%200510%%20051024_DNAcomputer.htm
Snippet	Frasch: DNA could power computer calculations
[22010]	www.asu.edu%%provost%%articulation%%chksheets%%03-04%%03ckl-phs.htm
Snippet	Certificate in Symbolic Systems

[17336]	www.asu.edu%%lib%%noble%%eng%%bioengin.htm
Snippet	Bioengineering
[20145]	www.asu.edu%%news%%faculty_students%%stephenyau_042103.htm
Snippet	Professor receives honor for distributed computing work
[17344]	www.asu.edu%%lib%%noble%%eng%%el_engineer.htm
Snippet	Electrical Engineering

By limiting the documents only with his has the phrase rather just the word, we could achieve fetching few documents than before and also the precision is highly increases by being more specific.

Extra Credit: "Snippet generation"

For the snippet generation, I am using the library "**jericho-html-3.3.jar**" to parse the html file to get the title and important keywords by accessing the metadata of the html page.

I parse the HTML title element and HTML metadata tag to fetch the keywords and titles of a document which I am suing it as a snippet.

The time taken is very less as I am parsing only the title and metadata tag ignoring the other HTML elements.

Example Snippet for a query "asu" looks below:

query> asu

[21144] 0.12939808431628005 Url www.asu.edu%%news%%stories%%200611%%20061113_googlestart.htm
Document title:
ASU News > ASU and Google offer personalized start page for students
Snippets:
{comma separated list of keywords here}

[487] 0.12817426920509736 Url www.asu.edu%%aad%%manuals%%fin%%fin301-02.html
Document title:
FIN 301–02: Financially Related Organization Deposits
Snippets:
financial, services, comptroller's, comptrollers, office, comptroller's office, gifts, asu foundation, asu alumni association, sun angel foundation, endowment, research park, collegiate golf foundation, checks, solicitations, sales tax, expenditures

[223] 0.12522002108213767 Url www.asu.edu%%aad%%manuals%%acd%%acd204-08.html
Document title:

ACD 204–08: Conflict of Interest
Snippets:
academic affairs, contracts, purchases, sales, substantial interest, remote interest, disclosure, competitive bids, relative, prosecution, competitive bids, remedy

[**17928] 0.12159791372921432** Url www.asu.edu%%lib%%sfx%%sfx-faq.htm
Document title:
ASU Libraries: Get It! FAQ
Snippets:
sfx, Get It!, ASU

[19786] 0.12115815981808453 Url www.asu.edu%%news%%community%%clubasu_091103.htm
Document title:
ASU News & Information from the Office of Media Relations and Public Information
Snippets:
(none)

[20621] 0.11566621193519586 Url
www.asu.edu%%news%%stories%%200511%%20051115_campusenrollment.htm
Document title:
ASU News > ASU?s Tempe campus nation?s largest
Snippets:
enrollment, students

[3542] 0.11479448365758017 Url www.asu.edu%%fa%%advantage%%release2.html
Document title:
The ASU Advantage - Student Financial Assistance Office
Snippets:
(none)

[827] 0.11262785808704655 Url www.asu.edu%%aad%%manuals%%usi%%usi104-02table.html
Document title:
USI 104–02 Table
Snippets:
(none)

[20792] 0.11255769759099218 Url
www.asu.edu%%news%%stories%%200603%%20060310_google.htm
Document title:
ASU News > Google to locate a temporary facility at ASU
Snippets:
google, partnership

[20547] 0.11254689112131414 Url
www.asu.edu%%news%%stories%%200509%%20050908_assiststudents.htm
Document title:
ASU News > ASU News > ASU assists students
Snippets:(none)