Search Engine

Prashanth Reddy Madi

Basic Work

- 3000 webpages in UNT domain using crawler
- Preprocessing steps

  Html parser

  Tokenization

  Stop word removal

  Stemming (porter stemmer)

Method

- Weighting scheme : tfidf

  $w_{ij} = tf_{ij}\, idf_{i} = tf_{ij} \log_2 (N/ df_i)$

- Vector space model

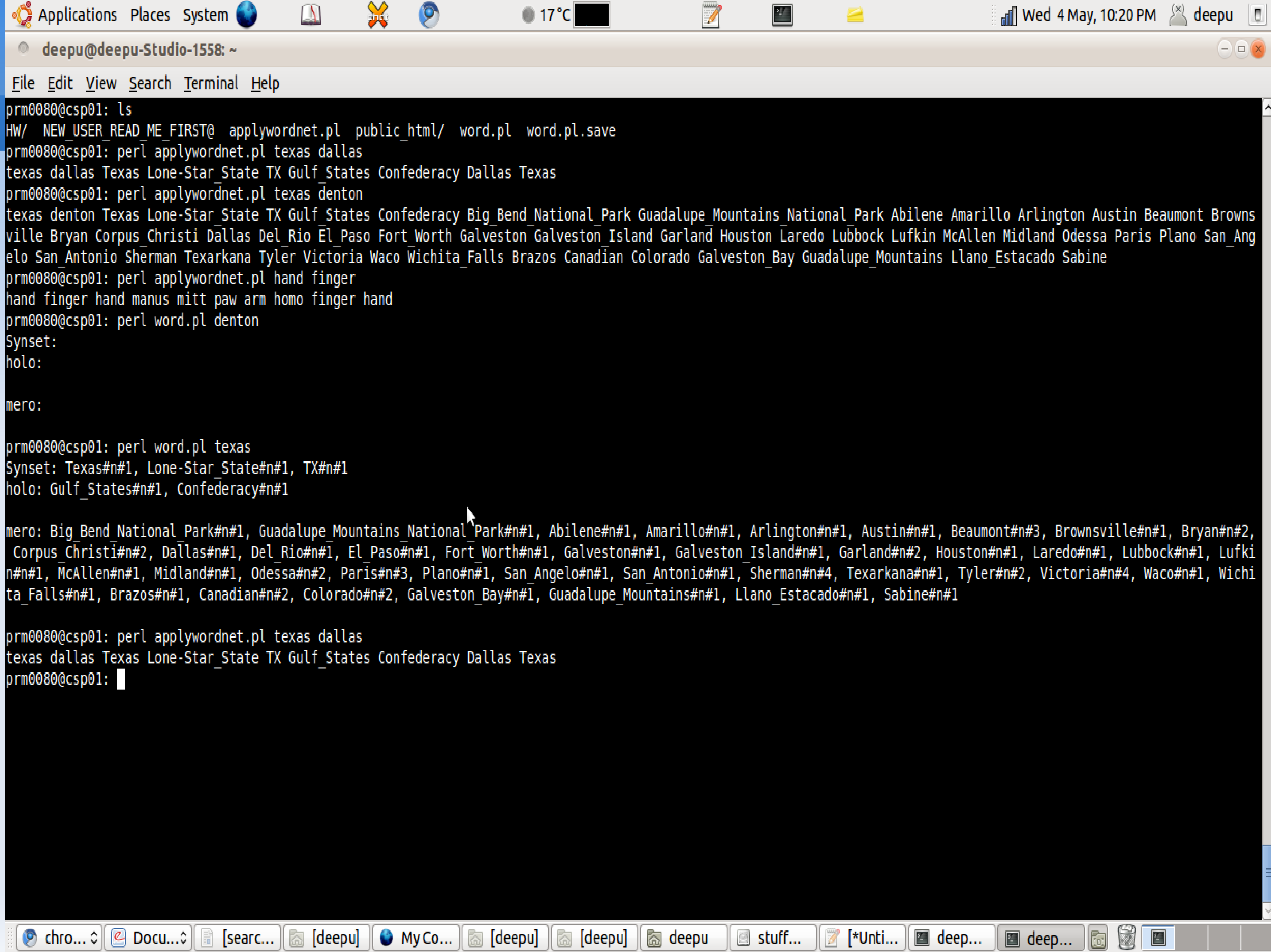
$$sim(d_j, d_k) = \frac{\vec{d_j} \cdot \vec{d_k}}{|d_j||d_k|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$

Extra Work

- Wordnet
- clustering

Wordnet

- Wordnet::querydata (perl package)
- Most of the words are ambiguous. Each word has many meanings.
- Inefficient (compared to statistical approach)
- Geographical data is less ambiguous
- Wordnet (Synset, Meronymy and Holonymy)
- Extended Query

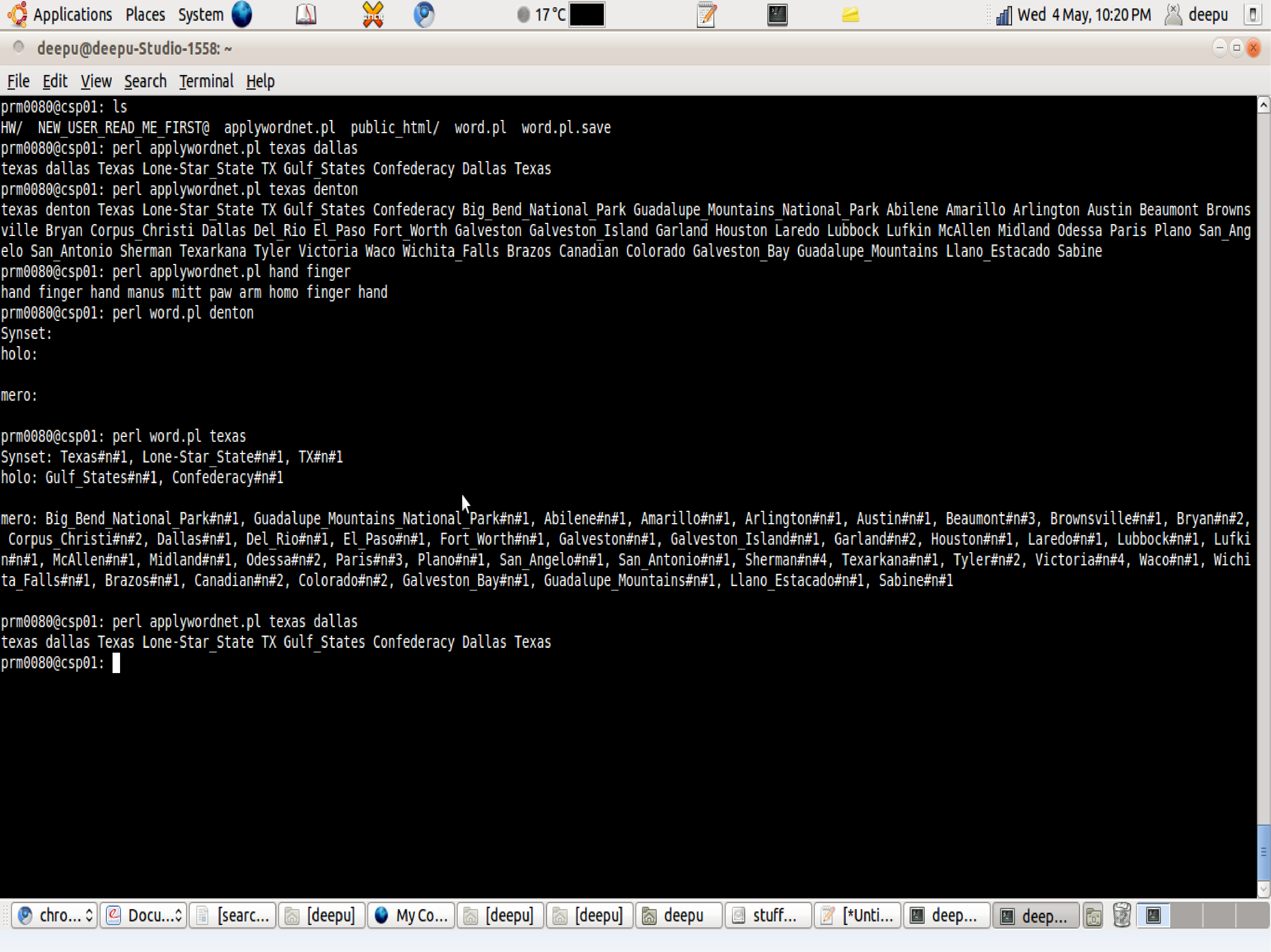


```
prm0080@csp01: ls
HW/  NEW_USER_READ_ME_FIRST@  applywordnet.pl  public_html/  word.pl  word.pl.save
prm0080@csp01: perl applywordnet.pl texas dallas
texas dallas Texas Lone-Star_State TX Gulf_States Confederacy Dallas Texas
prm0080@csp01: perl applywordnet.pl texas denton
texas denton Texas Lone-Star_State TX Gulf_States Confederacy Big_Bend_National_Park Guadalupe_Mountains_National_Park Abilene Amarillo Arlington Austin Beaumont Browns
ville Bryan Corpus_Christi Dallas Del_Rio El_Paso Fort_Worth Galveston Galveston_Island Garland Houston Laredo Lubbock Lufkin McAllen Midland Odessa Paris Plano San_Ang
elo San_Antonio Sherman Texarkana Tyler Victoria Waco Wichita_Falls Brazos Canadian Colorado Galveston_Bay Guadalupe_Mountains Llano_Estacado Sabine
prm0080@csp01: perl applywordnet.pl hand finger
hand finger hand manus mitt paw arm homo finger hand
prm0080@csp01: perl word.pl denton
Synset:
holo:

mero:

prm0080@csp01: perl word.pl texas
Synset: Texas#n#1, Lone-Star_State#n#1, TX#n#1
holo: Gulf_States#n#1, Confederacy#n#1

mero: Big_Bend_National_Park#n#1, Guadalupe_Mountains_National_Park#n#1, Abilene#n#1, Amarillo#n#1, Arlington#n#1, Austin#n#1, Beaumont#n#3, Brownsville#n#1, Bryan#n#2,
 Corpus_Christi#n#2, Dallas#n#1, Del_Rio#n#1, El_Paso#n#1, Fort_Worth#n#1, Galveston#n#1, Galveston_Island#n#1, Garland#n#2, Houston#n#1, Laredo#n#1, Lubbock#n#1, Lufki
n#n#1, McAllen#n#1, Midland#n#1, Odessa#n#2, Paris#n#3, Plano#n#1, San_Angelo#n#1, San_Antonio#n#1, Sherman#n#4, Texarkana#n#1, Tyler#n#2, Victoria#n#4, Waco#n#1, Wichi
ta_Falls#n#1, Brazos#n#1, Canadian#n#2, Colorado#n#2, Galveston_Bay#n#1, Guadalupe_Mountains#n#1, Llano_Estacado#n#1, Sabine#n#1

prm0080@csp01: perl applywordnet.pl texas dallas
texas dallas Texas Lone-Star_State TX Gulf_States Confederacy Dallas Texas
prm0080@csp01: 
```

Address Resolution

- Texas

  denton,dallas,irving.....
- 1811, Maple st, denton, Texas.
- Finding places in Texas → Texas
- Finding particular place in Texas → denton,texas


Applications Places System
17 °C
Wed 4 May, 10:20 PM deepu
deepu@deepu-Studio-1558: ~
File Edit View Search Terminal Help
chro... Docu... [searc... [deepu] My Co... [deepu] [deepu] deepu stuff... [*Unti... deep... deep...

```
prm0080@csp01: ls
HW/  NEW_USER_READ_ME_FIRST@  applywordnet.pl  public_html/  word.pl  word.pl.save
prm0080@csp01: perl applywordnet.pl texas dallas
texas dallas Texas Lone-Star_State TX Gulf_States Confederacy Dallas Texas
prm0080@csp01: perl applywordnet.pl texas denton
texas denton Texas Lone-Star_State TX Gulf_States Confederacy Big_Bend_National_Park Guadalupe_Mountains_National_Park Abilene Amarillo Arlington Austin Beaumont Browns
ville Bryan Corpus_Christi Dallas Del_Rio El_Paso Fort_Worth Galveston Galveston_Island Garland Houston Laredo Lubbock Lufkin McAllen Midland Odessa Paris Plano San_Ang
elo San_Antonio Sherman Texarkana Tyler Victoria Waco Wichita_Falls Brazos Canadian Colorado Galveston_Bay Guadalupe_Mountains Llano_Estacado Sabine
prm0080@csp01: perl applywordnet.pl hand finger
hand finger hand manus mitt paw arm homo finger hand
prm0080@csp01: perl word.pl denton
Synset:
holo:

mero:

prm0080@csp01: perl word.pl texas
Synset: Texas#n#1, Lone-Star_State#n#1, TX#n#1
holo: Gulf_States#n#1, Confederacy#n#1

mero: Big_Bend_National_Park#n#1, Guadalupe_Mountains_National_Park#n#1, Abilene#n#1, Amarillo#n#1, Arlington#n#1, Austin#n#1, Beaumont#n#3, Brownsville#n#1, Bryan#n#2,
 Corpus_Christi#n#2, Dallas#n#1, Del_Rio#n#1, El_Paso#n#1, Fort_Worth#n#1, Galveston#n#1, Galveston_Island#n#1, Garland#n#2, Houston#n#1, Laredo#n#1, Lubbock#n#1, Lufki
n#n#1, McAllen#n#1, Midland#n#1, Odessa#n#2, Paris#n#3, Plano#n#1, San_Angelo#n#1, San_Antonio#n#1, Sherman#n#4, Texarkana#n#1, Tyler#n#2, Victoria#n#4, Waco#n#1, Wichi
ta_Falls#n#1, Brazos#n#1, Canadian#n#2, Colorado#n#2, Galveston_Bay#n#1, Guadalupe_Mountains#n#1, Llano_Estacado#n#1, Sabine#n#1

prm0080@csp01: perl applywordnet.pl texas dallas
texas dallas Texas Lone-Star_State TX Gulf_States Confederacy Dallas Texas
prm0080@csp01: 
```

Clustering

- Semantic Relatedness
- BOW model

Semantic Relatedness

- What is Semantic Relatedness?

  Semantic relatedness is a measure of how related two or more concepts are

  Example:"cat” and ”dog” are more related than

  ”cat” and ”bag”
- Exploit the semantic relatedness in wikipedia

  Wikipedia Miner

Method

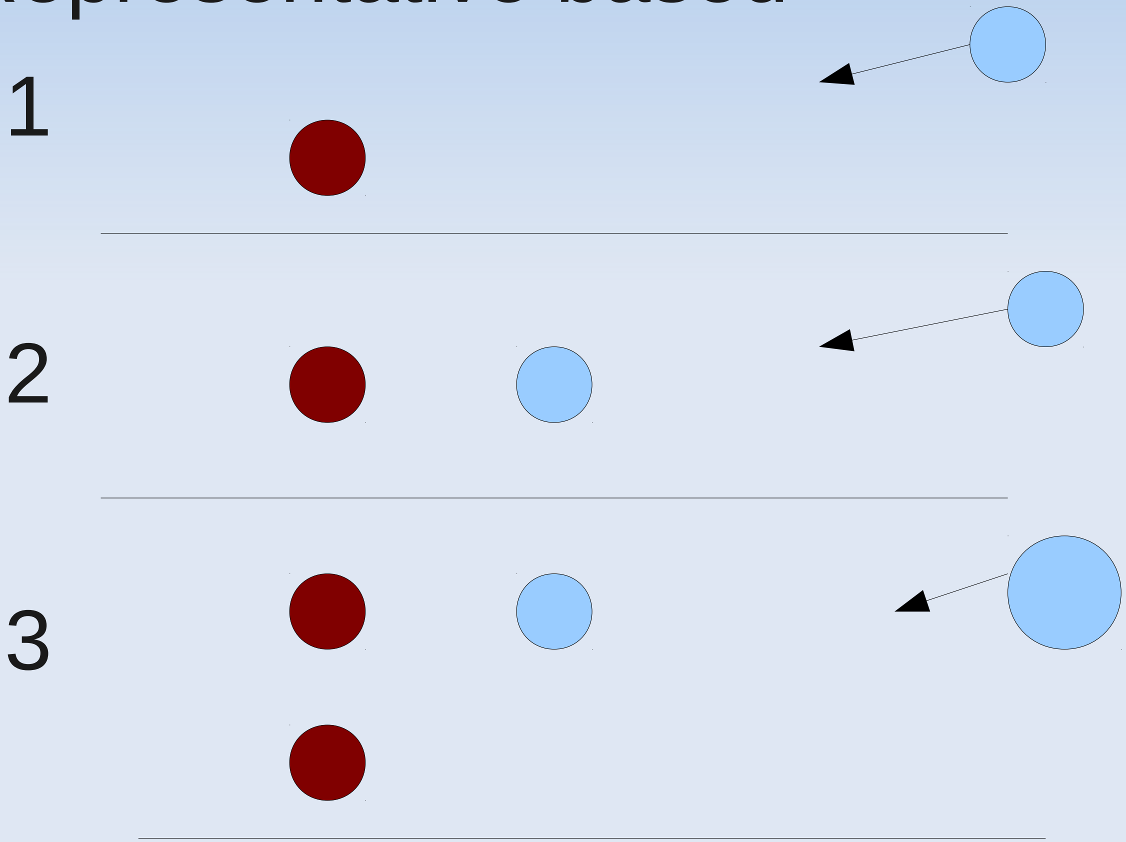
- Pos tagger (CRF)
- Extracting only noun and adjective phrases
- semantic relatedness using wikipedia miner
- Dbscan algorithm to get clustered concepts
- Mapping concept cluster against document (produces documents x concept_clust matrix)
- K-means algorithm to get clustered documents

Limitations

- Time consuming
- Practically imposible for online demo using CSP machine

BOW model

- Representative based

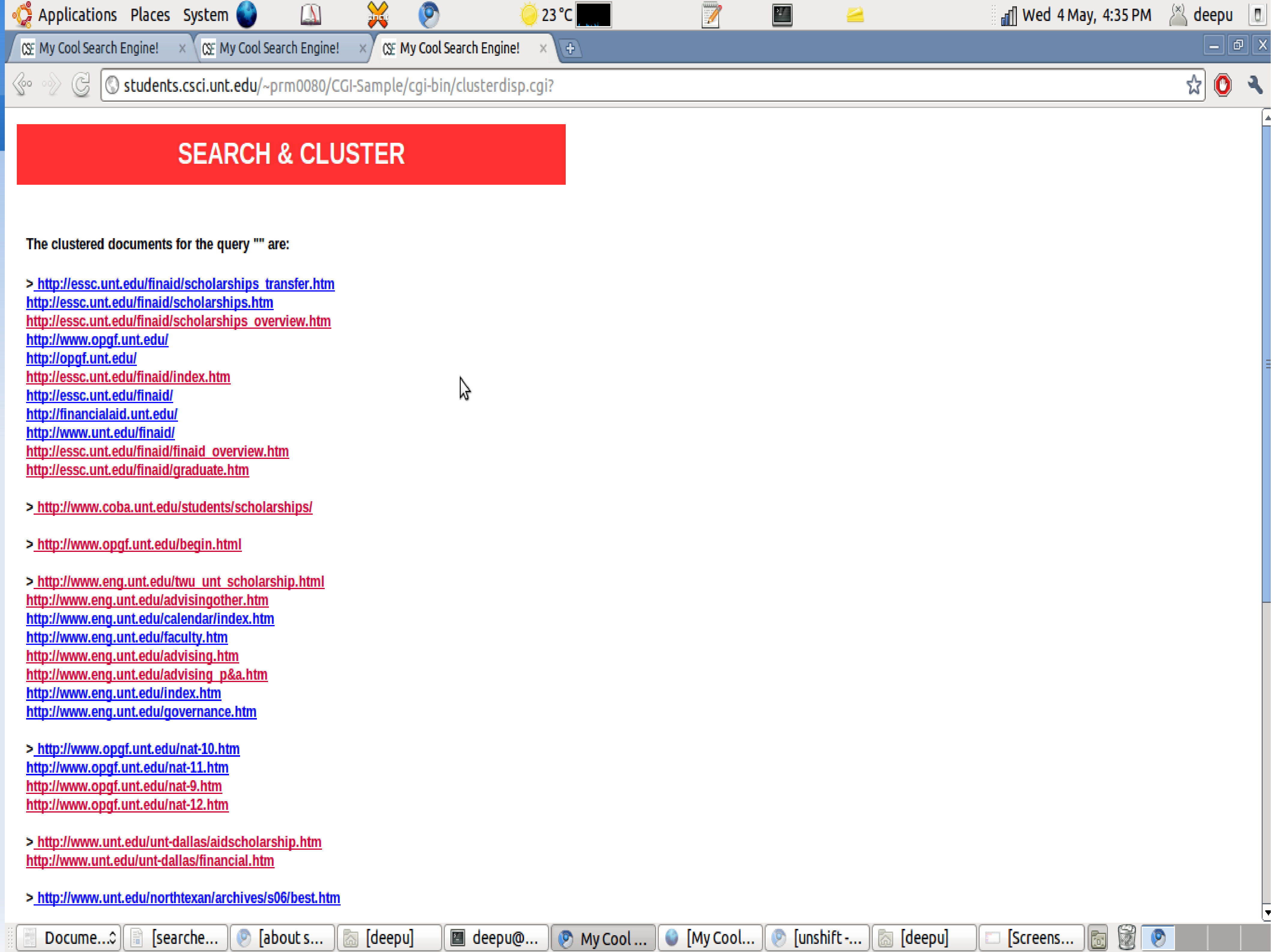


1

2

3

- Simple approach
- Induction Clustering (Idea)
- Clustering Based on Query
- Ranked Clusters to display
- Super Fast (Online demo using csp machine)



students.csci.unt.edu/~prm0080/CGI-Sample/cgi-bin/clusterdisp.cgi?

SEARCH & CLUSTER

The clustered documents for the query "" are:

> http://essc.unt.edu/finaid/scholarships_transfer.htm
http://essc.unt.edu/finaid/scholarships.htm
http://essc.unt.edu/finaid/scholarships_overview.htm
http://www.opgf.unt.edu/
http://opgf.unt.edu/
http://essc.unt.edu/finaid/index.htm
http://essc.unt.edu/finaid/
http://financialaid.unt.edu/
http://www.unt.edu/finaid/
http://essc.unt.edu/finaid/finaid_overview.htm
http://essc.unt.edu/finaid/graduate.htm

> http://www.coba.unt.edu/students/scholarships/

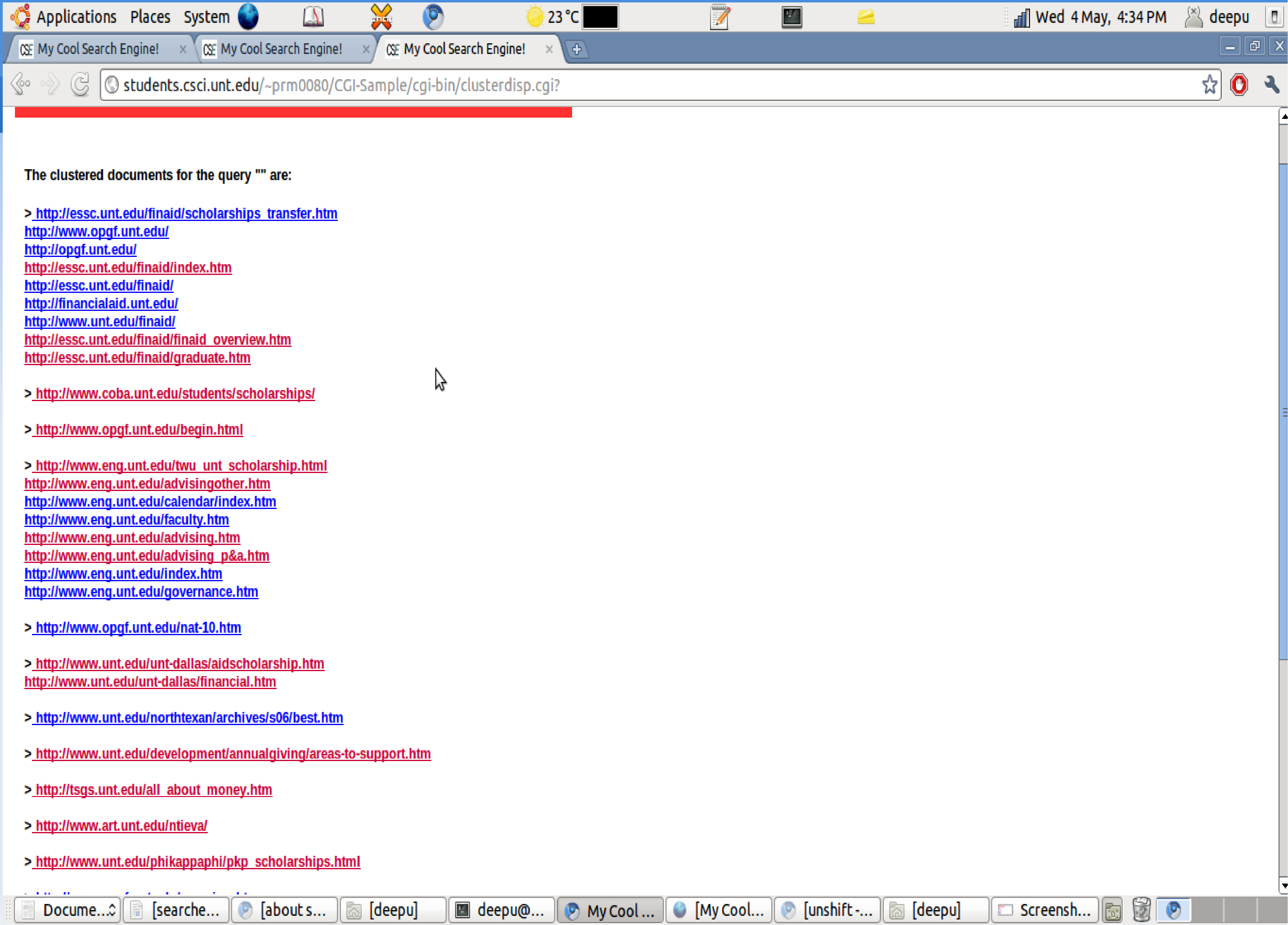
> http://www.opgf.unt.edu/begin.html

> http://www.eng.unt.edu/twu_unt_scholarship.html
http://www.eng.unt.edu/advisingother.htm
http://www.eng.unt.edu/calendar/index.htm
http://www.eng.unt.edu/faculty.htm
http://www.eng.unt.edu/advising.htm
http://www.eng.unt.edu/advising_p&a.htm
http://www.eng.unt.edu/index.htm
http://www.eng.unt.edu/governance.htm

> http://www.opgf.unt.edu/nat-10.htm
http://www.opgf.unt.edu/nat-11.htm
http://www.opgf.unt.edu/nat-9.htm
http://www.opgf.unt.edu/nat-12.htm

> http://www.unt.edu/unt-dallas/aidscholarship.htm
http://www.unt.edu/unt-dallas/financial.htm

> http://www.unt.edu/northtexan/archives/s06/best.htm



The clustered documents for the query "" are:

> http://essc.unt.edu/finaid/scholarships_transfer.htm
http://www.opgf.unt.edu/
http://opgf.unt.edu/
http://essc.unt.edu/finaid/index.htm
http://essc.unt.edu/finaid/
http://financialaid.unt.edu/
http://www.unt.edu/finaid/
http://essc.unt.edu/finaid/finaid_overview.htm
http://essc.unt.edu/finaid/graduate.htm

> http://www.coba.unt.edu/students/scholarships/

> http://www.opgf.unt.edu/begin.html

> http://www.eng.unt.edu/twu_unt_scholarship.html
http://www.eng.unt.edu/advisingother.htm
http://www.eng.unt.edu/calendar/index.htm
http://www.eng.unt.edu/faculty.htm
http://www.eng.unt.edu/advising.htm
http://www.eng.unt.edu/advising_p&a.htm
http://www.eng.unt.edu/index.htm
http://www.eng.unt.edu/governance.htm

> http://www.opgf.unt.edu/nat-10.htm

> http://www.unt.edu/unt-dallas/aidscholarship.htm
http://www.unt.edu/unt-dallas/financial.htm

> http://www.unt.edu/northtexan/archives/s06/best.htm

> http://www.unt.edu/development/annualgiving/areas-to-support.htm

> http://tsgs.unt.edu/all_about_money.htm

> http://www.art.unt.edu/ntieva/

> http://www.unt.edu/phikappaphi/pkp_scholarships.html

Future Work

- Using combination of single and complete linkage clustering algorithm's to cluster large data
- Prof. Richard Goodrum had an idea of using message passing concept in c++ to perform clustering using multiple process.
- Incorporate user's (like/dislike) based on Measuring proximity on graphs with side information

Acknowldegement

- Dr. Qunfeng Dong
- Prof.Richard Goodrum
- Dr. Rada Mihalcea
- Bharath dhandala

Questions ??

Thanks Everyone...