FINAL PROJECT

COLLEGE SEARCH INTERFACE

Xinnan Chen, 315, xchen100@jhu.edu
Xueshan Bai, 315, xbai10@jhu.edu

http://ugrad.cs.jhu.edu/~xchen100/home.html | **AN ON-LINE INTERFACE FOR COLLEGE SEARCH WITH TWO OPTIONS: 1) SPECIFIC CRTIERIA SEARCH 2) NATURAL LANGUAGE QUIZ SEARCH**

PHASE 2

(1) We obtained data on colleges from the U.S. Department of Education in the form of csv files (https://www.ed.gov). We transformed the csv files into SQL DDL, created tables of distinct categories(e.g. Diversity, Costs, Standard Test Scores etc.) and then inserted the corresponding tuples into each table.

(2) The stored procedure Search(IN SchName VARCHAR(120), SAT_MIN INTEGER, SAT_MAX INTEGER, ACT_MIN INTEGER, ACT_MAX INTEGER, State VARCHAR(2), City VARCHAR(30)) is used to implement a search engine that would output schools based on input from users for certain criteria. These criteria, i.e. the parameters for the Search procedure, are School Name, Minimum SAT Median Score, Maximum SAT Median Score, Minimum ACT Median Score, Maximum ACT Median Score, State of the schools, and the City of the schools.

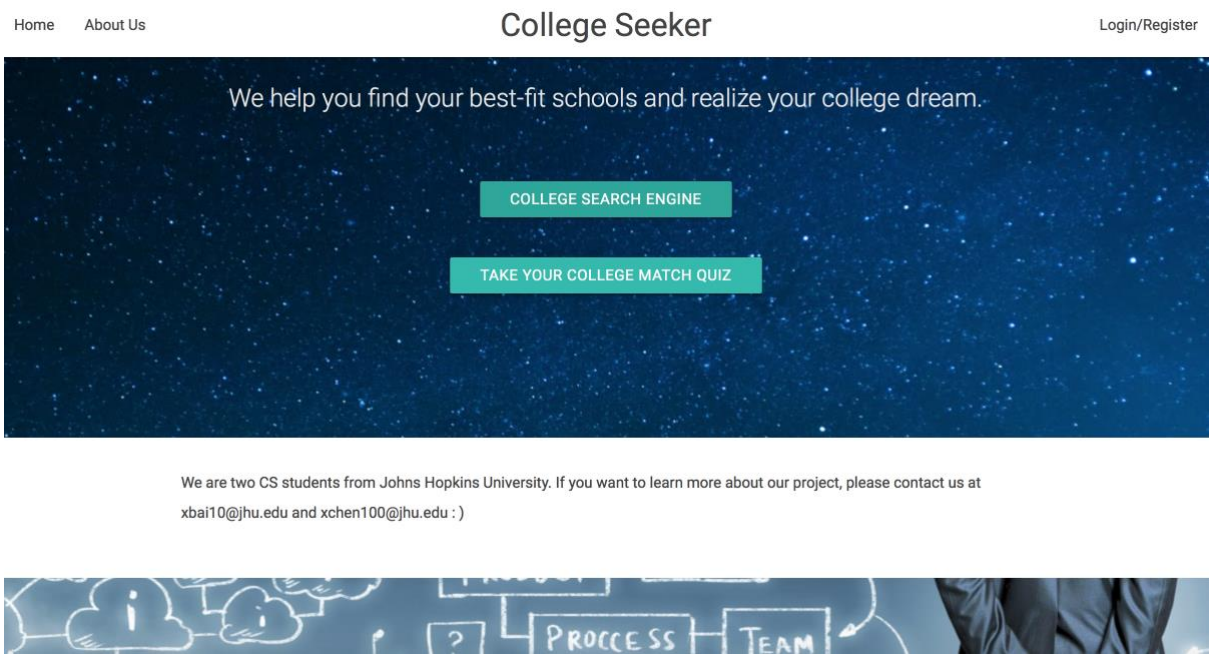
Users can choose to give any criterion an input or not. The search will be limited to schools that meet the entered criteria, and the criteria fields without an input have no constraint on them. For example, if a user wants to search for schools in Maryland that has a minimum SAT median score of 2130 and a maximum ACT score of 30, the interface would call the stored procedure as Search(NULL, 2130, NULL, NULL, 30, 'MD', NULL). The result is the list of schools along with their names, urls, school types, SAT/ACT median scores, locations, and tuition.

(3) The quiz we designed is related to natural language surfaces. In asking the questions, the quiz communicates with the users and translates the answers into queries for data. Given more time, we'd like for the quiz to ask more meaningful questions.

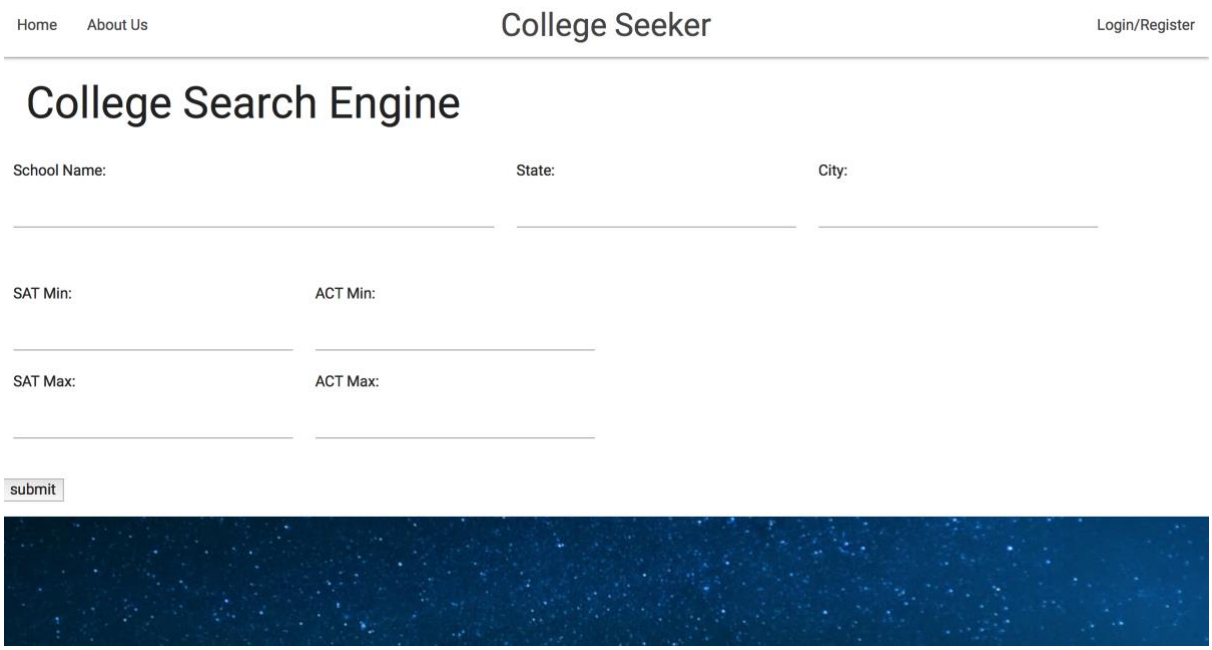
(4) In addition to a school search engine for specific criterion, we also provide a quiz that can help a user find out what kind of schools is a right fit! This quiz asks the user's preference for locations, diversity, studying habits etc., and computes a list of schools that best accommodates the user.

(5) Some limitations exist in the breadth of the search. If we had additional time, we would like to acquire more data and answer some more varied and worthwhile/interesting questions such as "Which schools have the highest number of student reporting the greatest satisfaction?" "Which campuses are the safest in the US?" and "Which school provides the best cafeteria?"

(7) Screenshots of our interface:



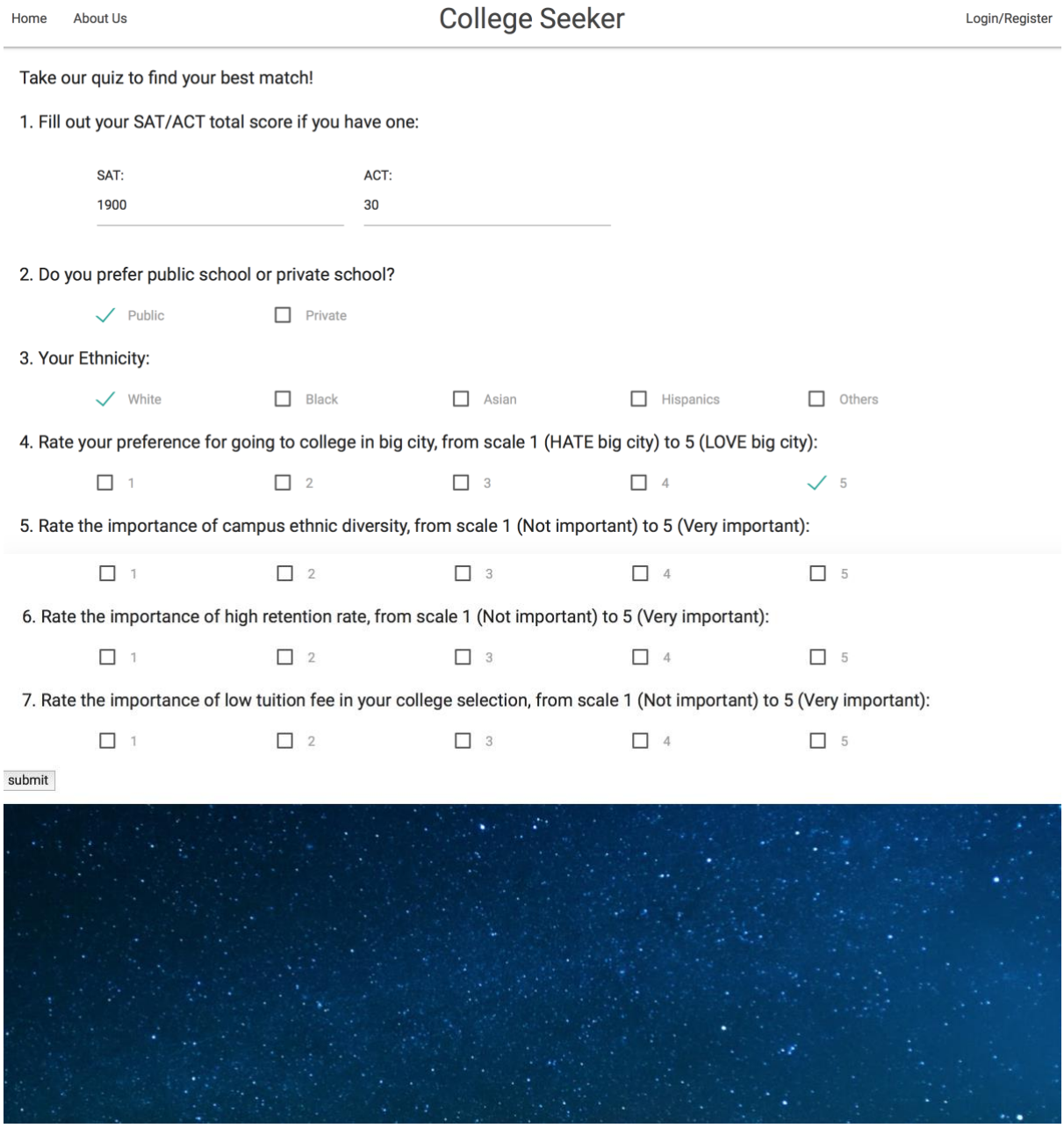
Our project features a clean yet stylish interface. The home page lists the two options for users to search for schools. The first one is a search engine with useful input fields -- some of the most frequently asked aspects about colleges. The search engine is shown below:



This search engine option is particularly efficient for users who know exactly what kind of schools they are looking for. When the users already have a certain set of qualifications in mind or when they are interested in one specific school, the engine does a fast and refined search for them.

On the other hand, for people who are not exactly sure what kind of criteria they are looking for, the second option comes in handy. It is a quiz that helps the users find a range of

possible schools, by assessing their preferences and looking up schools that meet their interests.



PHASE 1

(1). Xinnan Chen(xchen100), Xueshan Bai(xbai10)
(2). A college education database
(3).
(i) Compute the names, SAT mid reading scores, ACT mid composite scores of colleges with admission rate below 50%, four year full-time retention rate above 60% and more than 15% Black Undergraduates.
(ii) List all the colleges that are flagged as Tribal Colleges and Universities and have a 75th percentile of SAT reading that is greater than 600.
(iii) List all the women-only public schools in Maryland.

(iv) Find the college with the lowest admission rate.
(v) List the name and tuition (for both in-state and out-of-state) of all the schools that are not minority-serving institutions.
(vi) Find the average number of undergraduate students in distance-only schools.
(vii) List the name of all private schools with a retention rate greater than 80% and a completion rate lower than 70%.
(viii) Find the maximum repayment rate for students who completed school.
(ix) List the name and state of all schools where the low income debt median is lower than the high income debt median.
(x) Find the percentage of Hispanic undergraduate students at Johns Hopkins University.
(xi) List the url of all the schools that are not currently operating.
(xii) Find the minimum cost for students with family income lower than $75,000 at schools with zip code 90010.
(xiii) List all schools accredited by Southern Association of Colleges and Schools Commission on Colleges
(xiv) List the school name, retention rate, and admission rate of schools with an age entry greater than 30.
(xv) List the debt median for students who have withdrawn at schools with a first generation debt median greater than $15,000.

(4). Relational model:

College_info	SchID	SchName	URL	Distance_only	Currently_operating
	104179	University of Arizona	www.arizona.edu	no	yes

Location	SchID	City	State	Zip_Code
	100654	Normal	AL	35762

Has_Special_Mission	SchID	TypeID
	100654	1

Mission_Type	TypeID	TypeName
	1	HBCU

Same_Sex	SchID	GenderType
	100663	coed

SAT	SchID	SATVR25	SATVR75	SATMT25	SATMT75	SATWR25	SATWR75	SATVRMID	SATMTMID	SATWRMID
	100663	520	630	520	668	NULL	NULL	575	594	NULL

Costs	SchID	Avg_annual_cost	income_$0_to_$30,000	income_$30,001_to_$48,000	income_$48,001_to_$75,000	income_$75,001_to_$110,000	income_$110,001+
	100663	16023	13614	14746	17601	18873	18482

Tuition	SchID	in_state	out_of_state	program_year
	100663	7766	17654	NULL

ACT	SchID	ACTCM25	ACTCM75	ACTEN25	ACTEN75	ACTMT25	ACTMT75	ACTWR25	ACTWR75	ACTCMMID	ACTENMID	ACTMTMID	ACTWRMID
	100663	22	28	22	30	19	26	NULL	NULL	25	26	23	NULL

Undergraduates	SchID	Number_of_students	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MORE	UGDS_NRA	UGDS_Unknown	Part_time_proportion
	100063	11269	0.5863	0.2541	0.0317	0.0595	0.0023	0.0006	0.0389	0.0181	0.0085	0.2671

Statistics	SchID	Adm_rate	Retention_rate_FT4	Retention_rate_FTL4	Retention_rate_PT4	Retention_rate_PTL4	Completion_rate_4	Completion_rate_L4
	100663	0.6538	0.7864	NULL	0.6071	NULL	0.5444	NULL

Repayment_rate	SchID	Completed_RPY	Withdrawn_RPY
	100063	0.624601367	0.436409884

Debt	SchID	All_debt_MDN	Graduated_debt_MDN	Withdrawn_debt_MDN	LO_INC_debt_MDN	MD_INC_debt_MDN	HI_INC_debt_MDN	Dependent_debt_MDN	Independent_debt_MDN	Pell_debt_MDN	NoPell_debt_MDN	Female_debt_MDN	Male_debt_MDN	First_Gen_MDN	No_First_Gen_MDN
	100063	14250	21500	9500	14739	14250	14000	14818.5	13250	17000	11907	14750	13750	14100	14500

SQL DDL Implementation:

```
CREATE TABLE College_info (
        SchID INTEGER NOT NULL primary key,
        SchName VARCHAR(35),
        SchType VARCHAR (20), #public, private nonprofit, or private for-profit
        URL VARCHAR(100),
        Distance_only VARCHAR(3), #yes or no
        Currently_operating VARCHAR(3)
);

CREATE TABLE Location (
        SchID INTEGER NOT NULL primary key,
        City VARCHAR(20),
        State VARCHAR(2),
        Zip_Code VARCHAR(15)
);

CREATE TABLE Has_Special_Mission (
        SchID INTEGER NOT NULL primary key,
        TypeID Integer
);

CREATE TABLE Mission_Type (
        TypeID INTEGER NOT NULL primary key,
        TypeName VARCHAR (80) #HBCU, PBI etc.
);

CREATE TABLE Same_Sex (
        SchID INTEGER primary key,
        GenderType VARCHAR(10)
);

CREATE TABLE SAT (
        SchID INTEGER NOT NULL primary key,
        SATVR25 INTEGER,
        SATVR75 INTEGER,
        SATMT25 INTEGER,
        SATMT75 INTEGER,
        SATWR25 INTEGER,
        SATWR75 INTEGER,
        SATVRMID INTEGER,
        SATMTMID INTEGER,
        SATWRMID INTEGER
);

CREATE TABLE ACT (
        SchID INTEGER NOT NULL primary key,
```

```
        ACTCM25 INTEGER,
        ACTCM75 INTEGER,
        ACTEN25 INTEGER,
        ACTEN75 INTEGER,
        ACTMT25 INTEGER,
        ACTMT75 INTEGER,
        ACTWR25 INTEGER,
        ACTWR75 INTEGER,
        ACTCMMID INTEGER,
        ACTENMID INTEGER,
        ACTMTMID INTEGER,
        ACTWRMID INTEGER
);

CREATE TABLE Undergraduates (
        SchID INTEGER NOT NULL primary key,
        Number_of_students INTEGER,
        UGDS_WHITE DECIMAL(5, 4),
        UGDS_BLACK DECIMAL(5, 4),
        UGDS_HISP DECIMAL(5, 4),
        UGDS_ASIAN DECIMAL(5, 4),
        UGDS_AIAN DECIMAL(5, 4),
        UGDS_NHPI DECIMAL(5, 4),
        UGDS_2MORE DECIMAL(5, 4),
        UGDS_NRA DECIMAL(5, 4),
        UGDS_Unknown DECIMAL(5, 4),
        Part_time_proportion DECIMAL(5, 4)
);

CREATE TABLE Costs (
        SchID INTEGER NOT NULL primary key,
        Avg_annual_cost INTEGER,
        income_$0_to_$30,000 INTEGER,
        income_$30,001_to_$48,000 INTEGER,
        income_$48,001_to_$75,000 INTEGER,
        income_$75,001_to_$110,000 INTEGER,
        income_$110,001+ INTEGER
);

CREATE TABLE Tuition (
        SchID INTEGER NOT NULL primary key,
        in_state INTEGER,
        out_of_state INTEGER,
        program_year INTEGER
);


CREATE TABLE Statistics (
```

```
        SchID INTEGER NOT NULL primary key,
        Adm_rate DECIMAL(5, 4),
        Retention_rate_FT4 DECIMAL(5, 4),
        Retention_rate_FTL4 DECIMAL(5, 4),
        Retention_rate_PT4 DECIMAL(5, 4),
        Retention_rate_PTL4 DECIMAL(5, 4),
        Completion_rate_4 DECIMAL(5, 4),
        Completion_rate_L4 DECIMAL(5, 4),
);

CREATE TABLE Repayment_rate (
        SchID INTEGER NOT NULL primary key,
        Completed_RPY DECIMAL(5, 4),
        Withdrawn_RPY DECIMAL(5, 4)
);

CREATE TABLE Debt (
        SchID INTEGER NOT NULL primary key,
        All_debt_MDN INTEGER,
        Graduated_debt_MDN INTEGER,
        Withdrawn_debt_MDN INTEGER,
        LO_INC_debt_MDN INTEGER,
        MD_INC_debt_MDN INTEGER,
        HI_INC_debt_MDN INTEGER,
        Dependent_debt_MDN INTEGER,
        Independent_debt_MDN INTEGER,
        Pell_debt_MDN INTEGER,
        NoPell_debt_MDN INTEGER,
        Female_debt_MDN INTEGER,
        Male_debt_MDN INTEGER,
        First_Gen_MDN INTEGER,
        Not_First_Gen_MDN INTEGER
);
```

(5). SQL statements:

(i)
```
SELECT c.SchName, s.SATVRMID, a.ACTCMMID
FROM College_info as c, SAT as s, ACT as a, Statistics as t, Undergraduates as u
WHERE c.SchID=s.SchID
AND c.SchID=a.SchID
AND c.SchID=t.SchID
AND c.SchID=u.SchID
AND s.Adm_rate < 0.5
AND s.Retention_rate_FT4 > 0.6
AND u.UGDS_BLACK > 0.15
```

(ii)
```
SELECT c.SchName
FROM College_info as c, Same_Sex as s, Location as l
```

```
    WHERE c.SchID=s.SchID
    AND c.SchID=l.SchID
    AND c.SchType='public'
    AND l.State='MD'
    AND s.GenderType='womenonly'
```

(iii)
```
SELECT c.SchName
 FROM College_info as c, Statistics as s, Tuition as t
 WHERE c.SchID=s.SchID
 AND c.SchID=t.SchID
 AND s.Adm_rate = (
     SELECT min(s2.Adm_rate)
     FROM Statistics as s2
     GROUP BY s2.SchID
 )
```

(iv)
```
SELECT c.SchName, t.in_state, t.out_of_state
 FROM College_info as c, Tuition as t, Has_Special_Mission as h
 WHERE c.SchID=t.SchID
 AND c.SchID=h.SchID
 AND h.TypeID=0
```

(v)
```
SELECT max(r.Completed_RPY)
FROM Repayment_rate as r
GROUP BY r.SchID
```

(vi)
```
SELECT c.SchName, l.State
 FROM College_info as c, Location as l, Debt as d
 WHERE c.SchID=l.SchID
 AND c.SchID=d.SchID
 AND d.LO_INC_debt_MDN < d.HI_INC_debt_MDN
```

(vii)
```
SELECT min(c.income_$0_to_$30,000)
 FROM Costs as c, Location as l, Statistics as s
 WHERE c.SchID=l.SchID
 AND c.SchID=s.SchID
 AND l.Zip_code=90010
 AND s.Retention_rate > 50%
 GROUP BY c.SchID
```

(viii)
```
SELECT  min(d.Withdrawn_debt_MDN)
 FROM Debt as d
 WHERE d.First_Gen_MDN > 15000
 GROUP BY d.SchID
```

(6).

- Data.gov/Education/College ScoreCrad

  URL: https://catalog.data.gov/dataset/college-scorecard

Conversion issues:

1. Attribute name abbreviation in the original tables can be hard to understand.
2. Delete unimportant attributes that have null values or uniform values for most of the instances(e.g. NPT4_PROG, WOMENONLY).
3. Delete unnecessary/unrelated attributes.
4. The single College Scorecard table is too large and hard to manipulate--split the largest table into several small tables with more concentrated attributes.
5. College name variation in different tables(e.g. Johns Hopkins University v.s. The Johns Hopkins University).

- Web interface: Dropdown selection for value/range for each attribute so users can search for colleges information based on SAT/ACT range, admission rate, retention rate, annual cost and etc.

(7). We are going to implement a college search engine.

user input: customized value/range for each attribute.

output: all the critical information related to the colleges that satisfy the constraints.

e.g. UNITID, Institute Name, City, Website, Admission Rate, SAT_AVG, ACT_AVG, UGDS, UGDS_White/Black/Hisp/Asian...., Cost_AVG, Retention Rate, Completion Rate, Entry Age.

(8). data mining

natural language interfaces