CS412 Project

Group 11

Mei-Cheng Shih, Chen-Yu Li, Xinyang Liu

Course Information from Course Instruction Pages

Each school has several webpages listing the courses in the university together with their information

Course information:

Course ID,

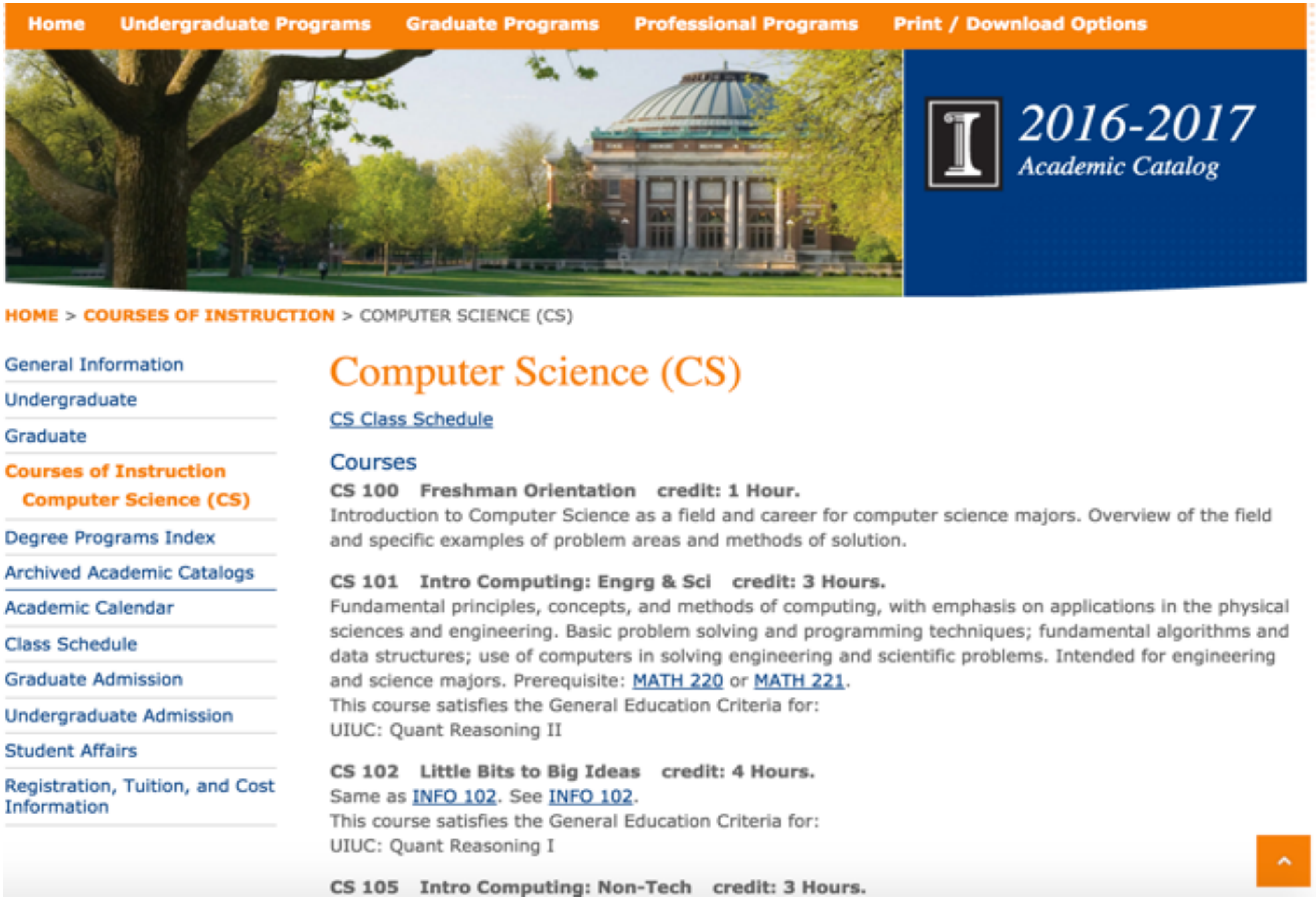
Course Name,

Credit,

Lecturer

Description,

Prerequisite.



Computer Science (CS)

CS Class Schedule

Courses

CS 100 Freshman Orientation credit: 1 Hour.
Introduction to Computer Science as a field and career for computer science majors. Overview of the field and specific examples of problem areas and methods of solution.

CS 101 Intro Computing: Engrg & Sci credit: 3 Hours.
Fundamental principles, concepts, and methods of computing, with emphasis on applications in the physical sciences and engineering. Basic problem solving and programming techniques; fundamental algorithms and data structures; use of computers in solving engineering and scientific problems. Intended for engineering and science majors. Prerequisite: MATH 220 or MATH 221.
This course satisfies the General Education Criteria for:
UIUC: Quant Reasoning II

CS 102 Little Bits to Big Ideas credit: 4 Hours.
Same as INFO 102. See INFO 102.
This course satisfies the General Education Criteria for:
UIUC: Quant Reasoning I

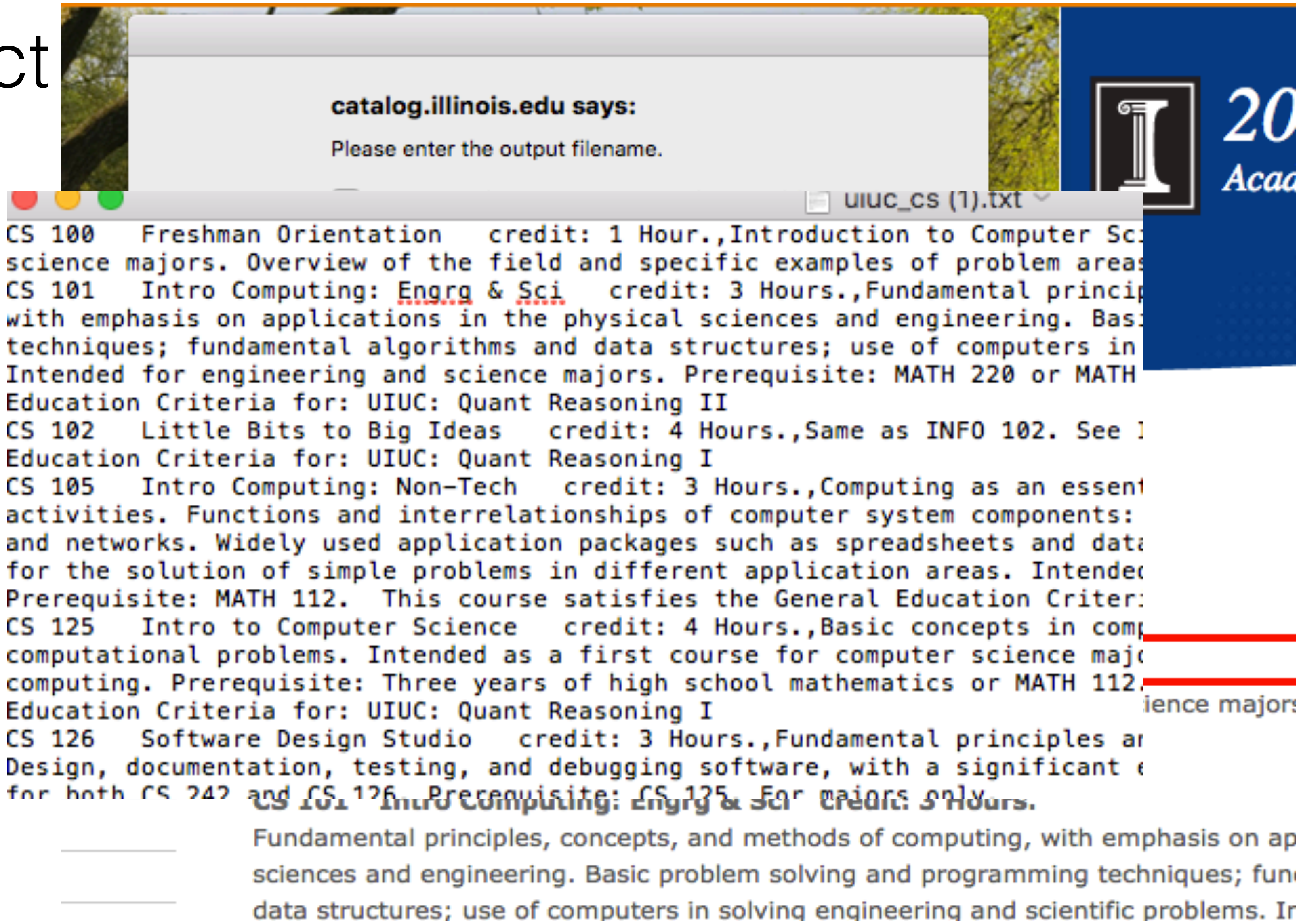
CS 105 Intro Computing: Non-Tech credit: 3 Hours.

Easy Tool for Course Information Extraction!

Course Extraction - Rule Based Method

Users need to select an interested block first

Find similar blocks based on DOM features in the vipsTree



catalog.illinois.edu says:

Please enter the output filename.

```
CS 100   Freshman Orientation   credit: 1 Hour.,Introduction to Computer Sci
science majors. Overview of the field and specific examples of problem areas
CS 101   Intro Computing: Engrg & Sci   credit: 3 Hours.,Fundamental princip
with emphasis on applications in the physical sciences and engineering. Basi
techniques; fundamental algorithms and data structures; use of computers in
Intended for engineering and science majors. Prerequisite: MATH 220 or MATH
Education Criteria for: UIUC: Quant Reasoning II
CS 102   Little Bits to Big Ideas   credit: 4 Hours.,Same as INFO 102. See I
Education Criteria for: UIUC: Quant Reasoning I
CS 105   Intro Computing: Non-Tech   credit: 3 Hours.,Computing as an essent
activities. Functions and interrelationships of computer system components:
and networks. Widely used application packages such as spreadsheets and data
for the solution of simple problems in different application areas. Intended
Prerequisite: MATH 112.  This course satisfies the General Education Criter:
CS 125   Intro to Computer Science   credit: 4 Hours.,Basic concepts in comp
computational problems. Intended as a first course for computer science majo
computing. Prerequisite: Three years of high school mathematics or MATH 112.
Education Criteria for: UIUC: Quant Reasoning I
CS 126   Software Design Studio   credit: 3 Hours.,Fundamental principles ar
Design, documentation, testing, and debugging software, with a significant e
for both CS 242 and CS 126. Prerequisite: CS 125. For majors only.
```

ience majors

CS 101 Intro Computing: Engrg & Sci Credit: 3 Hours.

Fundamental principles, concepts, and methods of computing, with emphasis on ap
sciences and engineering. Basic problem solving and programming techniques; fun
data structures; use of computers in solving engineering and scientific problems. Ir

Features: block height, subblock number, vips-id, left alignment, block height, font size, font weight, text color, text content

Course Extraction – Function Develoepd

Two Functions to Extract Course Information

- Feature Based Analysis
- Kmean Clustering Based

Two Functions to Extract Information of Similar Courses

- Kmean Clustering Based
- Hierarchical Clustering Based

Course Extraction –Tested Websites and Output Example

Working

Stanford

Georgia Tech

UC Berkeley

UIUC

CMU

UIC

Purdue

Princeton

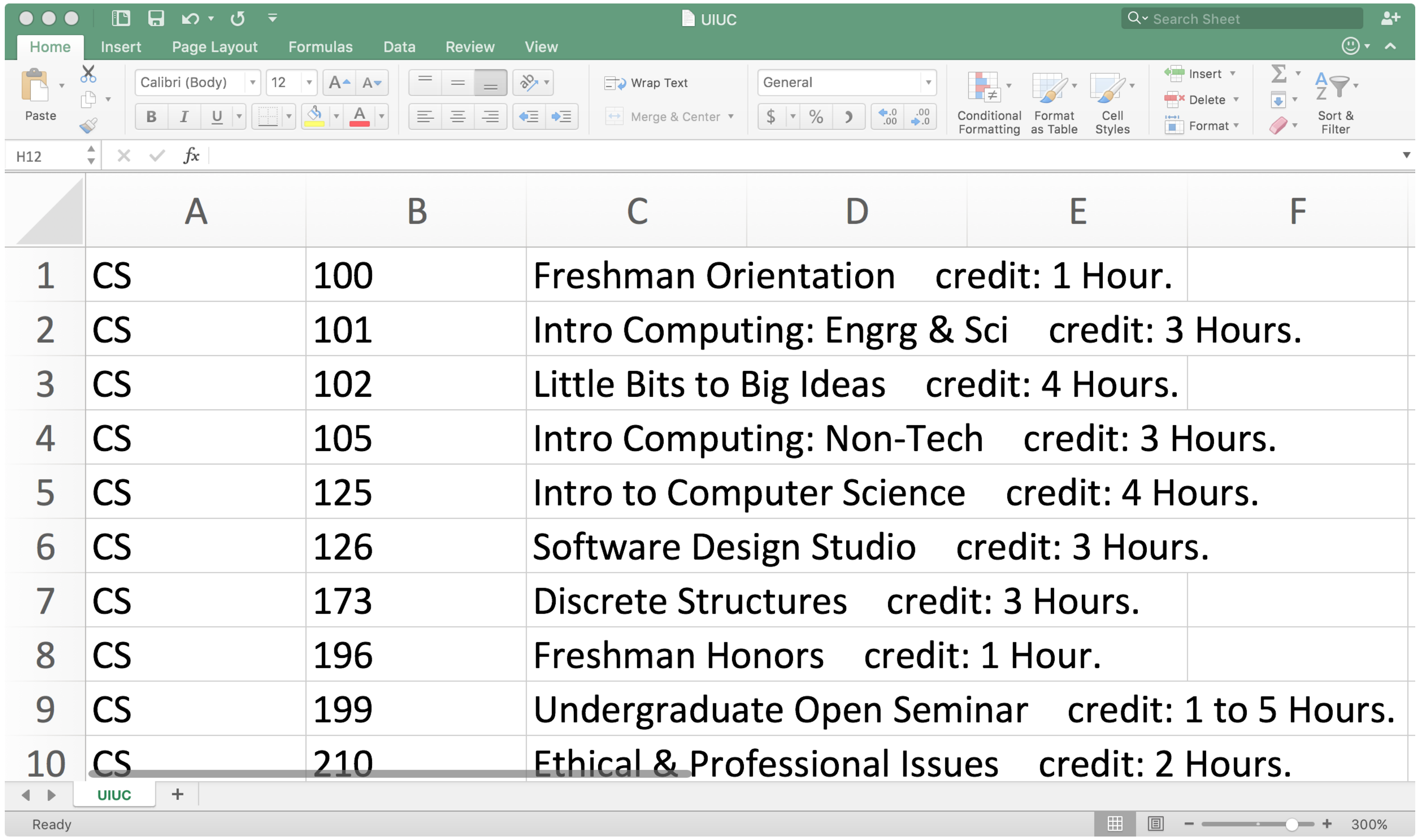
UT Austin

Cornell

MIT

U Washington

Not Working



	A	B	C	D	E	F
1	CS	100	Freshman Orientation credit: 1 Hour.			
2	CS	101	Intro Computing: Engrg & Sci credit: 3 Hours.			
3	CS	102	Little Bits to Big Ideas credit: 4 Hours.			
4	CS	105	Intro Computing: Non-Tech credit: 3 Hours.			
5	CS	125	Intro to Computer Science credit: 4 Hours.			
6	CS	126	Software Design Studio credit: 3 Hours.			
7	CS	173	Discrete Structures credit: 3 Hours.			
8	CS	196	Freshman Honors credit: 1 Hour.			
9	CS	199	Undergraduate Open Seminar credit: 1 to 5 Hours.			
10	CS	210	Ethical & Professional Issues credit: 2 Hours.			

Main reason:
VIPS cannot recognize the visual blocks.

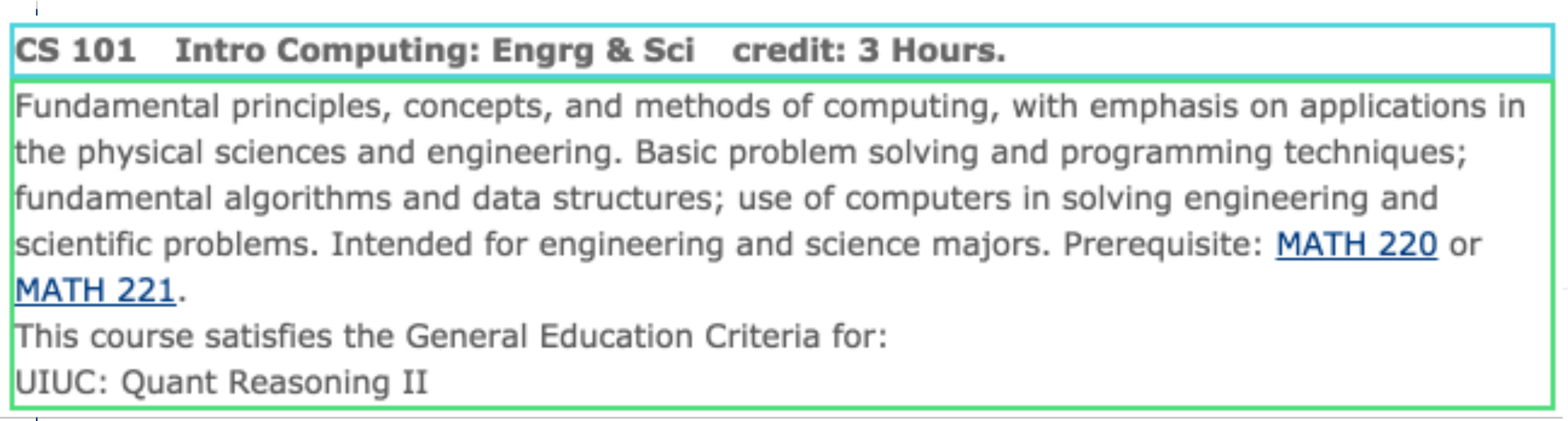
Sample from top 10 CS grad school (US News):
http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-science-schools/computer-science-rankings

Q
&
A

Course Extraction - Clustering Based Method

Highlight:

1. Allow using large or small visual blocks
2. K-means clustering, user can decide K (range: 2~8)

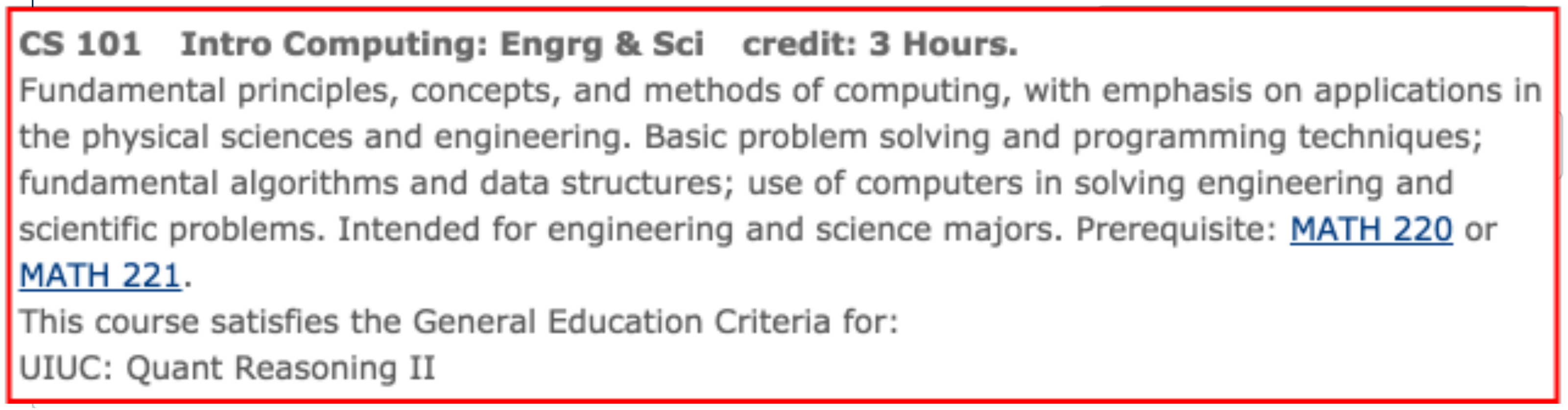


CS 101 Intro Computing: Engrg & Sci credit: 3 Hours.

Fundamental principles, concepts, and methods of computing, with emphasis on applications in the physical sciences and engineering. Basic problem solving and programming techniques; fundamental algorithms and data structures; use of computers in solving engineering and scientific problems. Intended for engineering and science majors. Prerequisite: MATH 220 or MATH 221.

This course satisfies the General Education Criteria for:
UIUC: Quant Reasoning II

OR



CS 101 Intro Computing: Engrg & Sci credit: 3 Hours.

Fundamental principles, concepts, and methods of computing, with emphasis on applications in the physical sciences and engineering. Basic problem solving and programming techniques; fundamental algorithms and data structures; use of computers in solving engineering and scientific problems. Intended for engineering and science majors. Prerequisite: MATH 220 or MATH 221.

This course satisfies the General Education Criteria for:
UIUC: Quant Reasoning II

Features: (large block) Word count, block height, subblock number (small block) subblock index, left alignment, word count, block height, font size, font weight

Course Extraction - Rule Based Method

Works for about 60% of the course catalog we tested

Working	Not Working
UIUC, UCLA, STANFORD, CMU PRINCETON, UIC, YALE, UMICH, UCDAVIS, GATECH, U Washington	UC BERKELEY, PURDUE, CORNELL, MIT, UTEAXAS

This methods lacks of flexibility, e.g. course catalog of UCB is not left aligned

Course Extraction - Clustering Based Method

Working

Stanford

Georgia Tech

UC Berkeley

UIUC

CMU

UIC

Purdue

Princeton

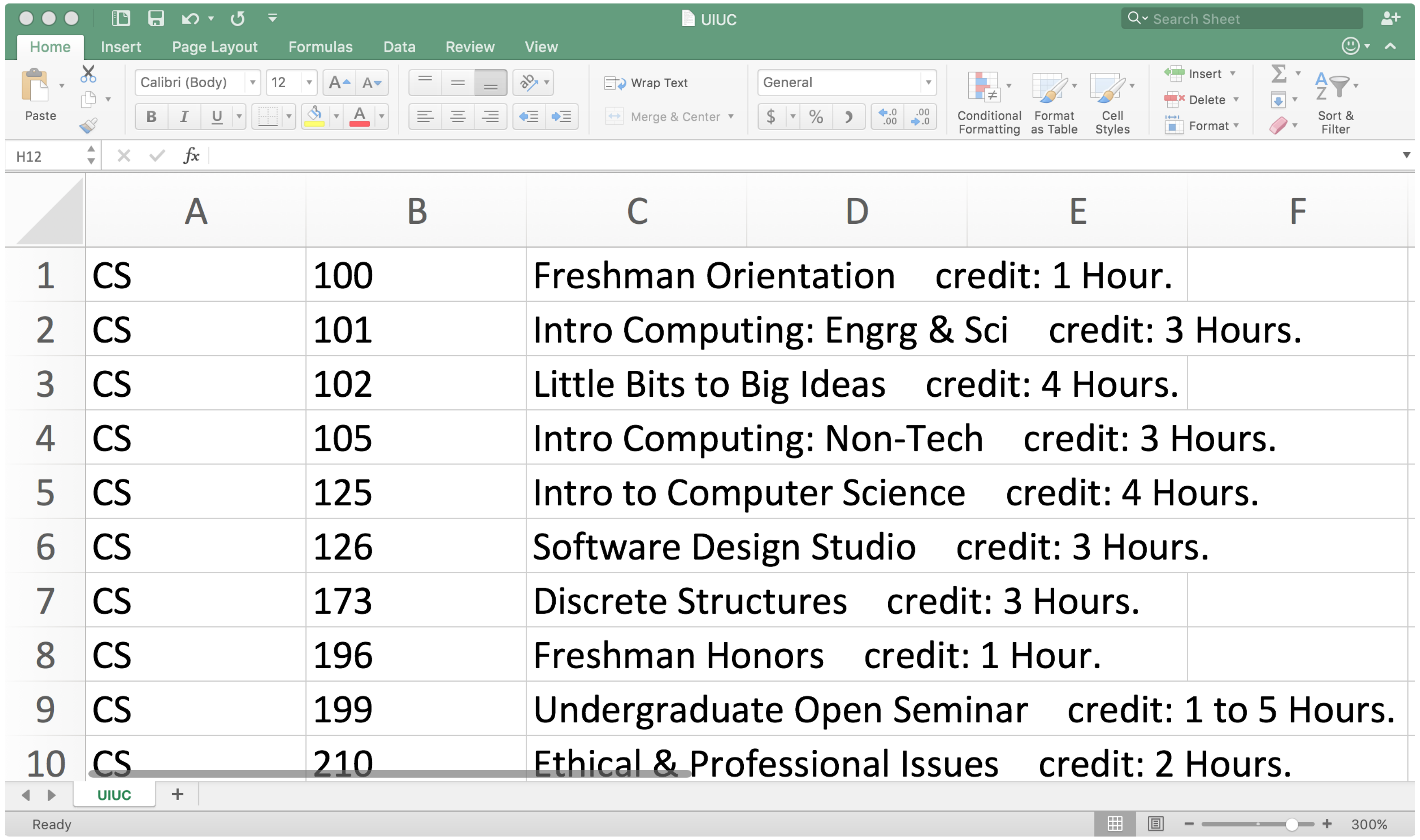
UT Austin

Cornell

MIT

U Washington

Not Working

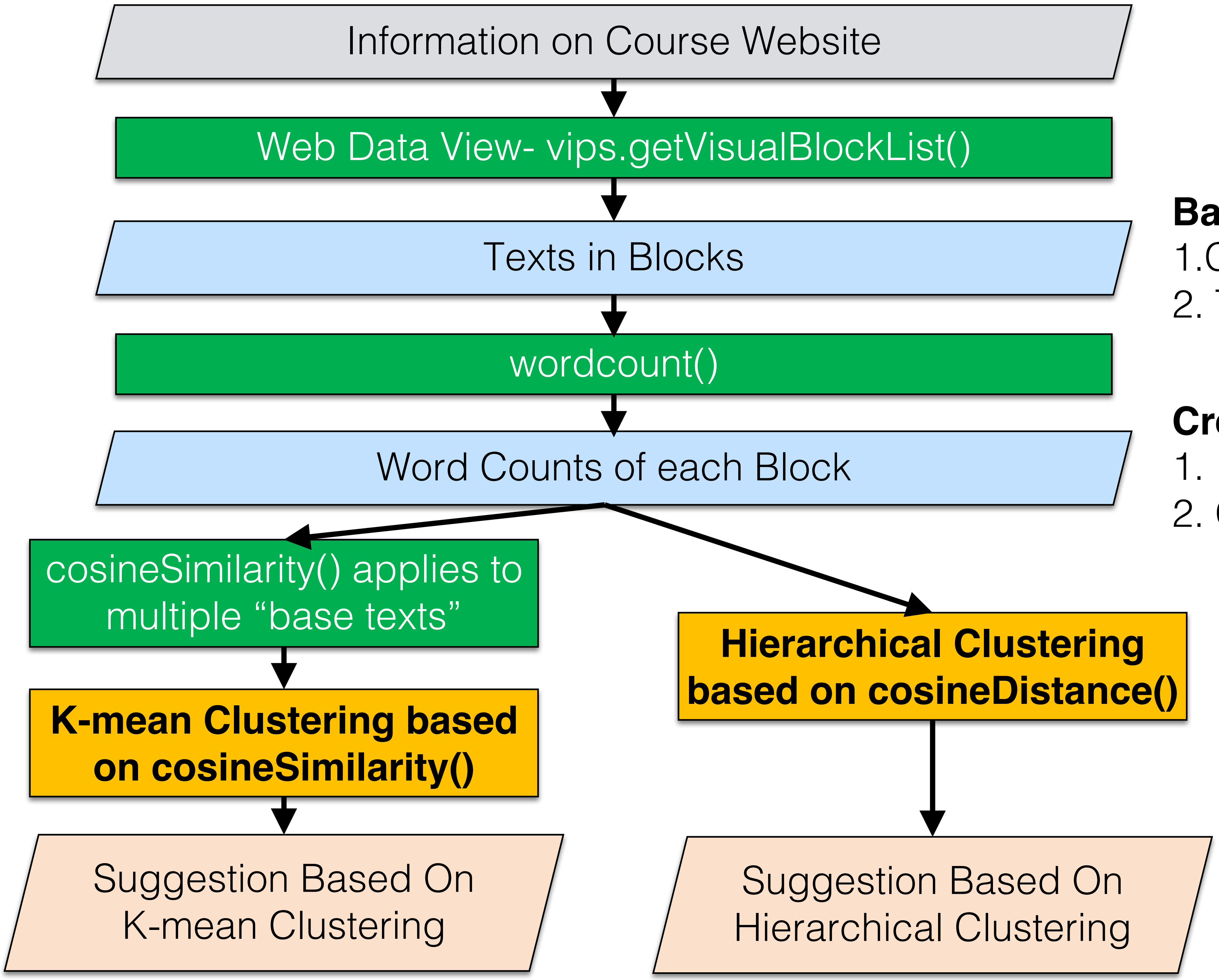


	A	B	C	D	E	F
1	CS	100	Freshman Orientation credit: 1 Hour.			
2	CS	101	Intro Computing: Engrg & Sci credit: 3 Hours.			
3	CS	102	Little Bits to Big Ideas credit: 4 Hours.			
4	CS	105	Intro Computing: Non-Tech credit: 3 Hours.			
5	CS	125	Intro to Computer Science credit: 4 Hours.			
6	CS	126	Software Design Studio credit: 3 Hours.			
7	CS	173	Discrete Structures credit: 3 Hours.			
8	CS	196	Freshman Honors credit: 1 Hour.			
9	CS	199	Undergraduate Open Seminar credit: 1 to 5 Hours.			
10	CS	210	Ethical & Professional Issues credit: 2 Hours.			

Main reason:
VIPS cannot recognize the visual blocks.

Sample from top 10 CS grad school (US News):
http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-science-schools/computer-science-rankings

Similar Course Suggestion


Information on Course Website

Web Data View- vips.getVisualBlockList()

Texts in Blocks

wordcount()

Word Counts of each Block

cosineSimilarity() applies to multiple “base texts”

K-mean Clustering based on cosineSimilarity()

Suggestion Based On K-mean Clustering

Hierarchical Clustering based on cosineDistance()

Suggestion Based On Hierarchical Clustering

Basic Text Cleaning:

1.Clear punctuations

2. Transform to lowercase

Create Word List:

1. Identify word elements

2. Count Frequency