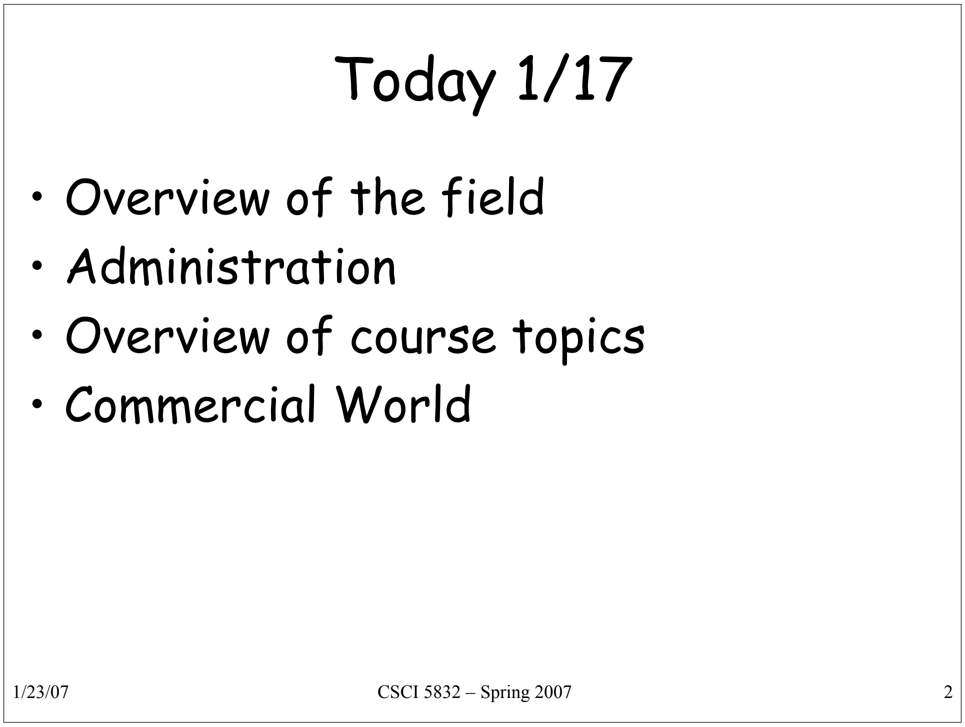
CSCI 5832
Natural Language Processing

Lecture 1

Jim Martin



Today 1/17

- Overview of the field
- Administration
- Overview of course topics
- Commercial World

Natural Language Processing

- What is it?
  - We're going to study what goes into getting computers to perform useful and interesting tasks involving human languages.
  - We will be secondarily concerned with the insights that such computational work gives us into human processing of language.

Why Should You Care?

Two trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication

Major Topics

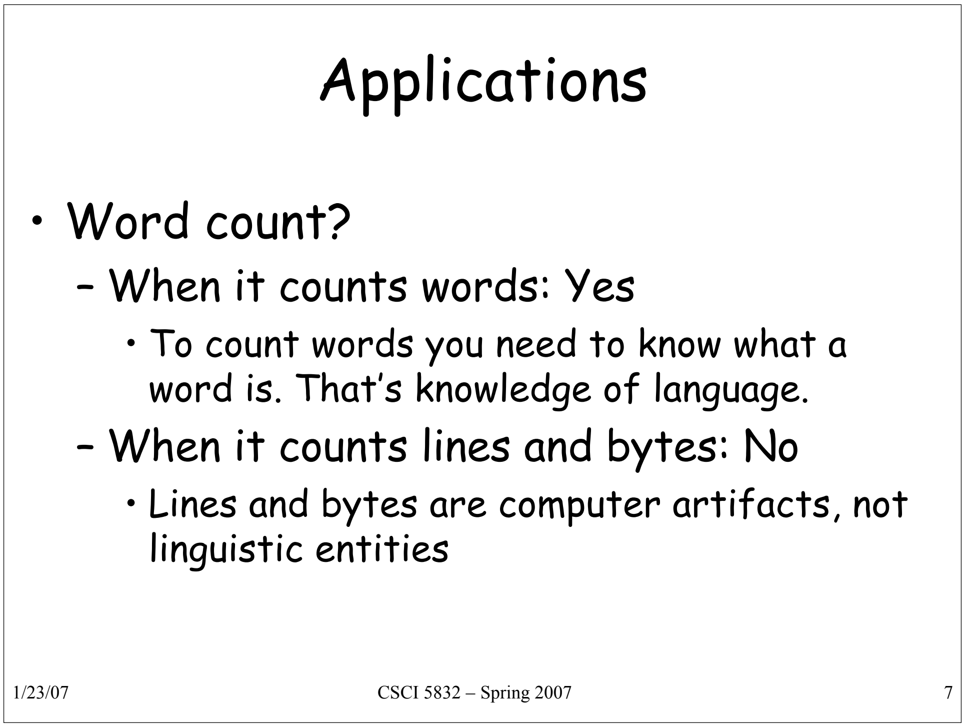
- Words
- Syntax
- Meaning
- Dialog and Discourse

}

Applications

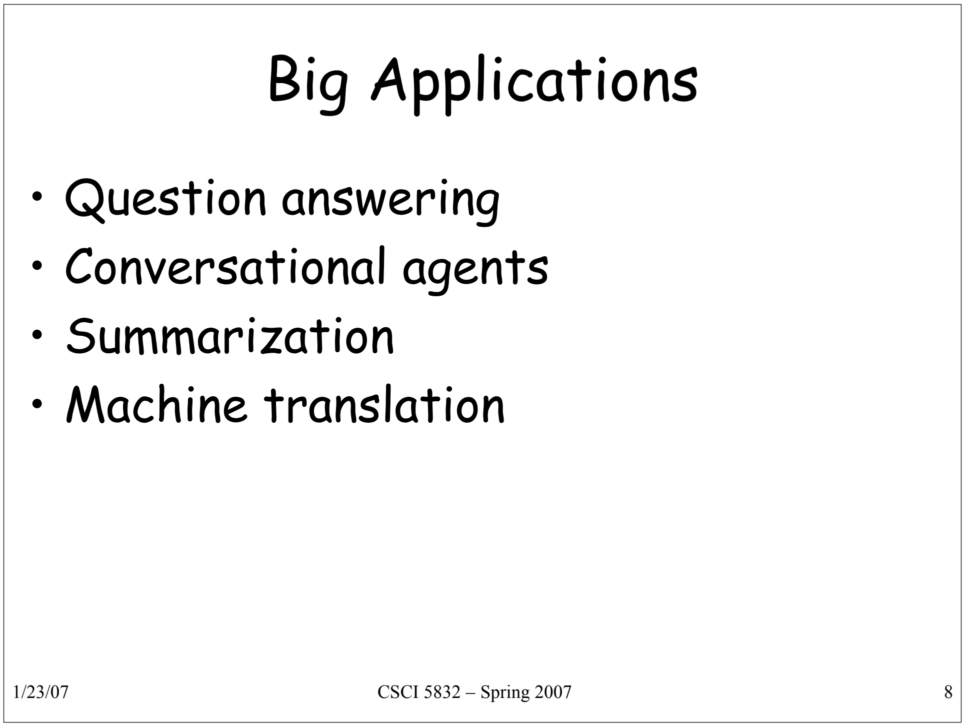
Applications

- First, what makes an application a *language processing application* (as opposed to any other piece of software)?
  - An application that requires the use of knowledge about human languages
    - Example: Is Unix wc (word count) a language processing application?



Applications

- Word count?
  - When it counts words: Yes
    - To count words you need to know what a word is. That's knowledge of language.
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities



Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Big Applications

- These kinds of applications require a tremendous amount of knowledge of language.
- Consider the following interaction with HAL the computer from 2001: A Space Odyssey

HAL

- Dave: *Open the pod bay doors, Hal.*
- HAL: *I'm sorry Dave, I'm afraid I can't do that.*

What’s needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
  - What they mean
  - How they combine (bay, vs. pod bay)
- How groups of words clump
  - What the clumps mean

What’s needed?

- Dialog
  - It is polite to respond, even if you’re planning to kill someone.
  - It is polite to pretend to want to be cooperative (I’m afraid, I can’t…)

Real Example

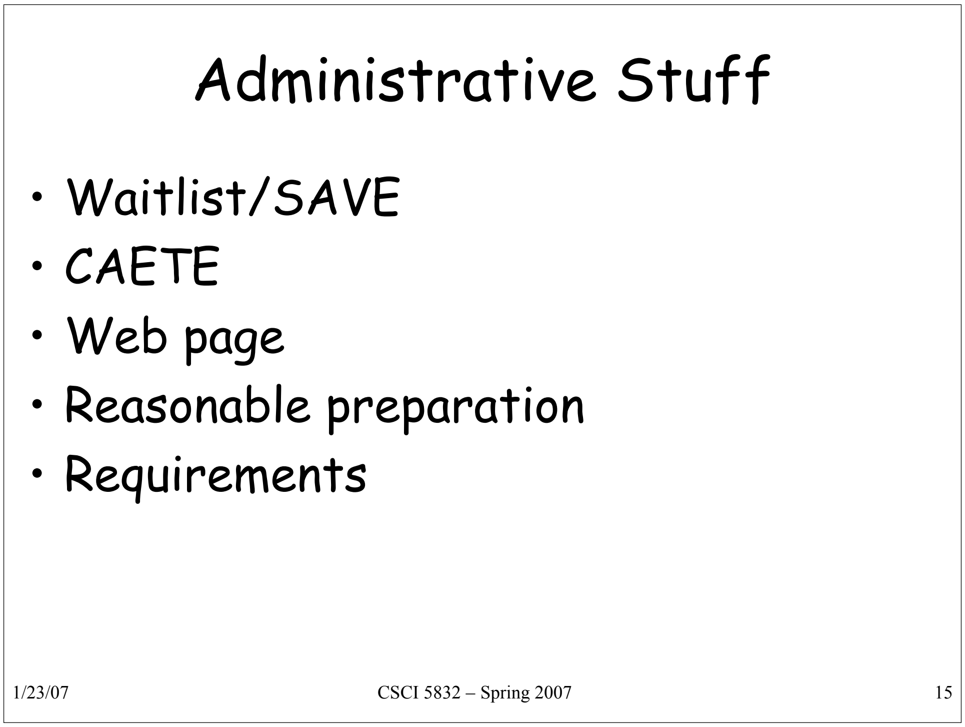
What is the Fed's current position on interest rates?

- What or who is the "Fed"?
- What does it mean for it to to have a position?
- How does "current" modify that?

Caveat

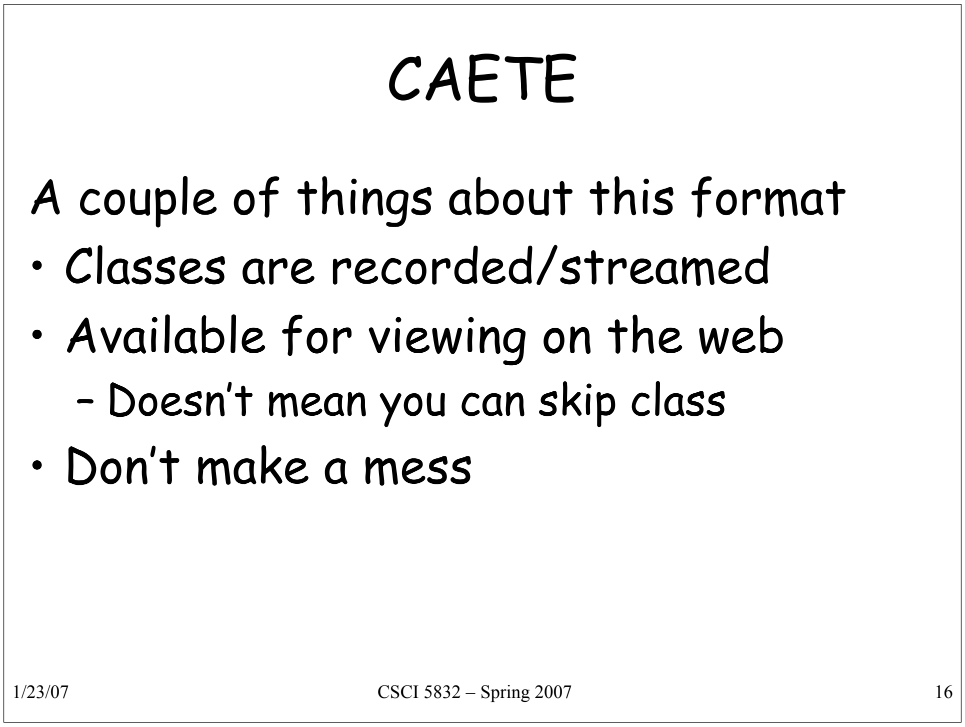
NLP has an AI aspect to it.

- We're often dealing with ill-defined problems
- We don't often come up with perfect solutions/algorithms
- We can't let either of those facts get in our way



Administrative Stuff

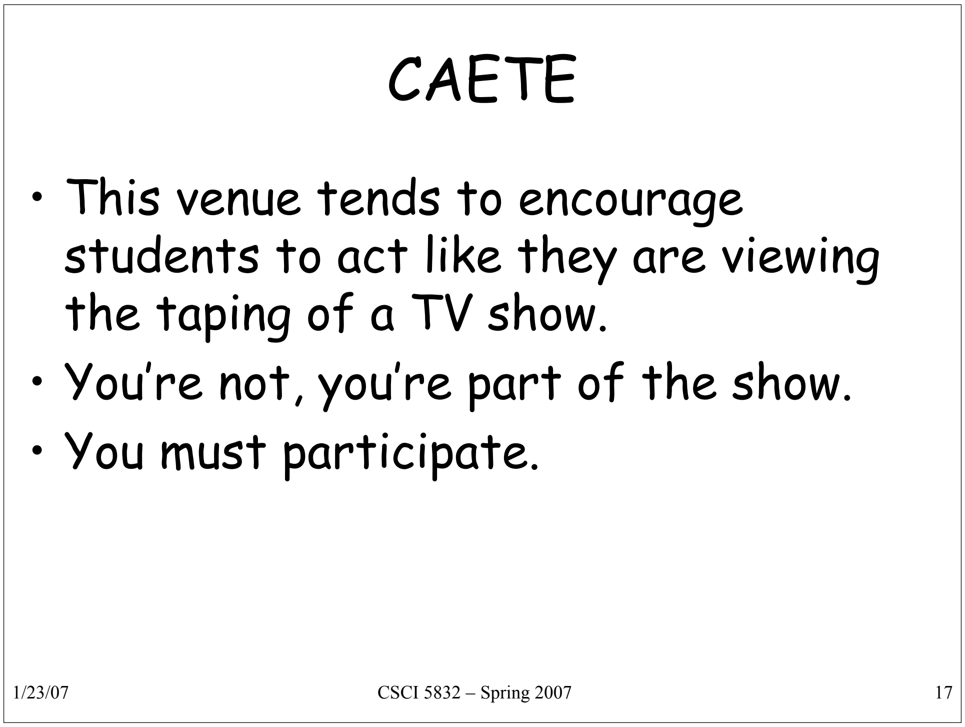
- Waitlist/SAVE
- CAETE
- Web page
- Reasonable preparation
- Requirements



CAETE

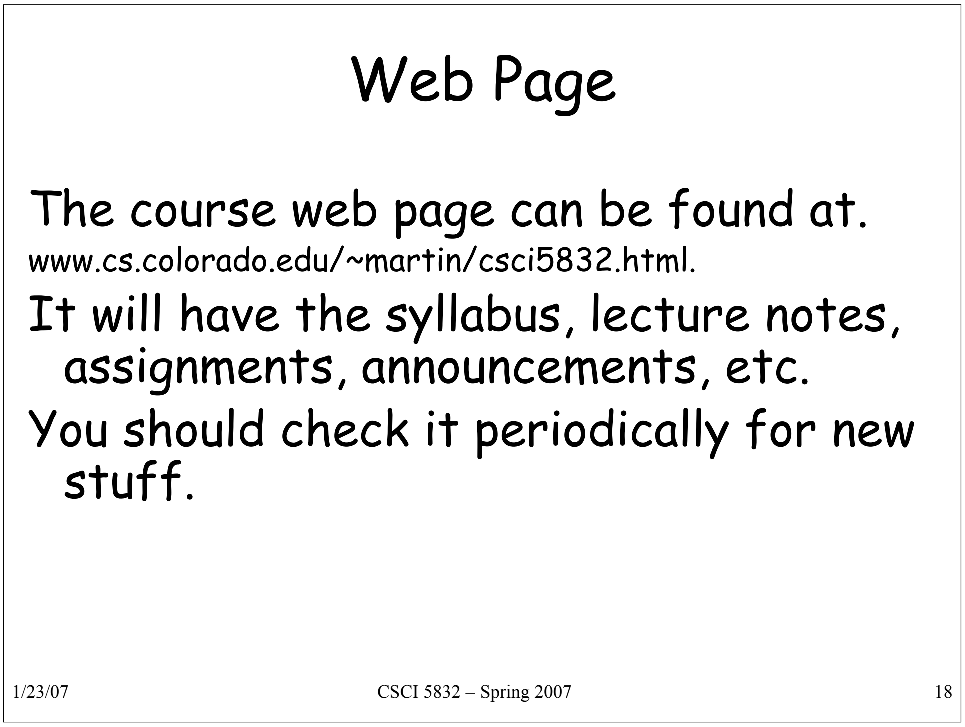
A couple of things about this format

- Classes are recorded/streamed
- Available for viewing on the web
  - Doesn’t mean you can skip class
- Don’t make a mess



CAETE

- This venue tends to encourage students to act like they are viewing the taping of a TV show.
- You’re not, you’re part of the show.
- You must participate.

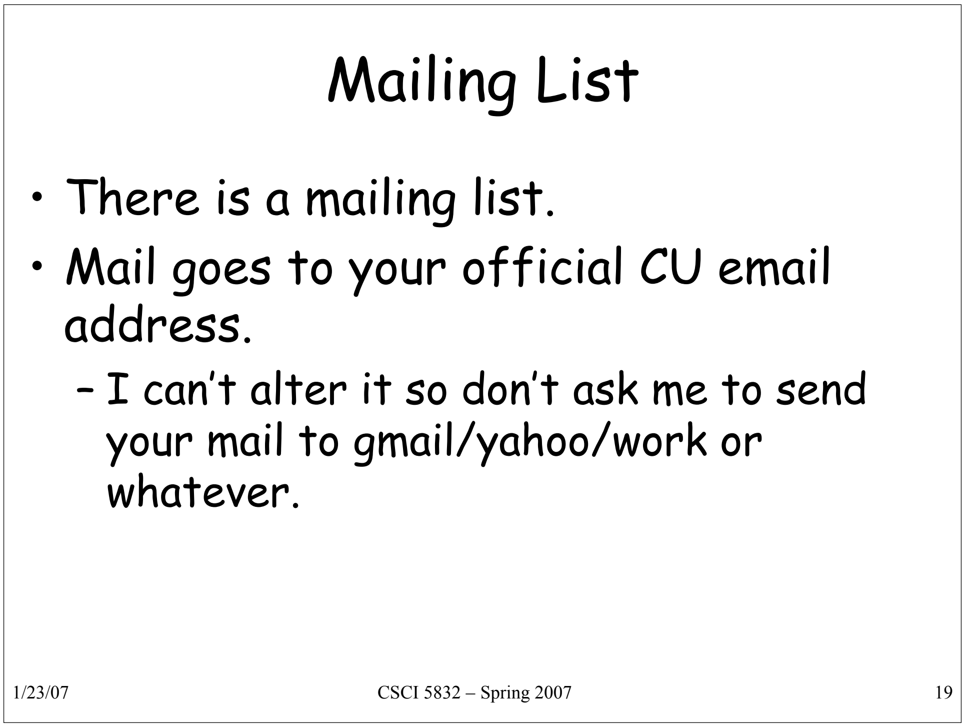


Web Page

The course web page can be found at.
www.cs.colorado.edu/~martin/csci5832.html.

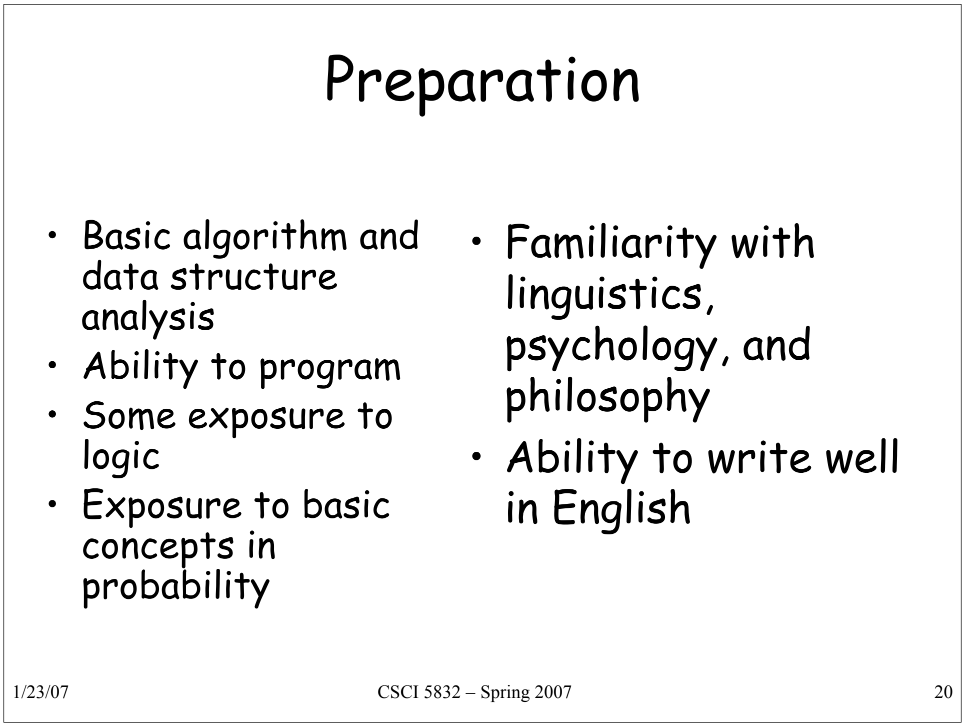
It will have the syllabus, lecture notes, assignments, announcements, etc.

You should check it periodically for new stuff.



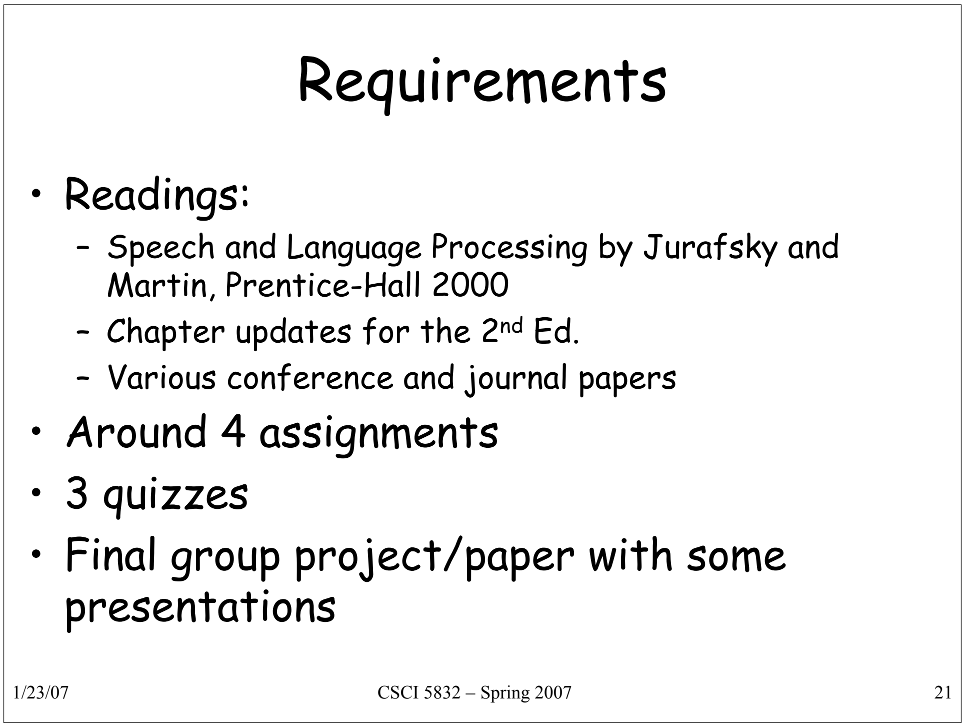
Mailing List

- There is a mailing list.
- Mail goes to your official CU email address.
  - I can’t alter it so don’t ask me to send your mail to gmail/yahoo/work or whatever.



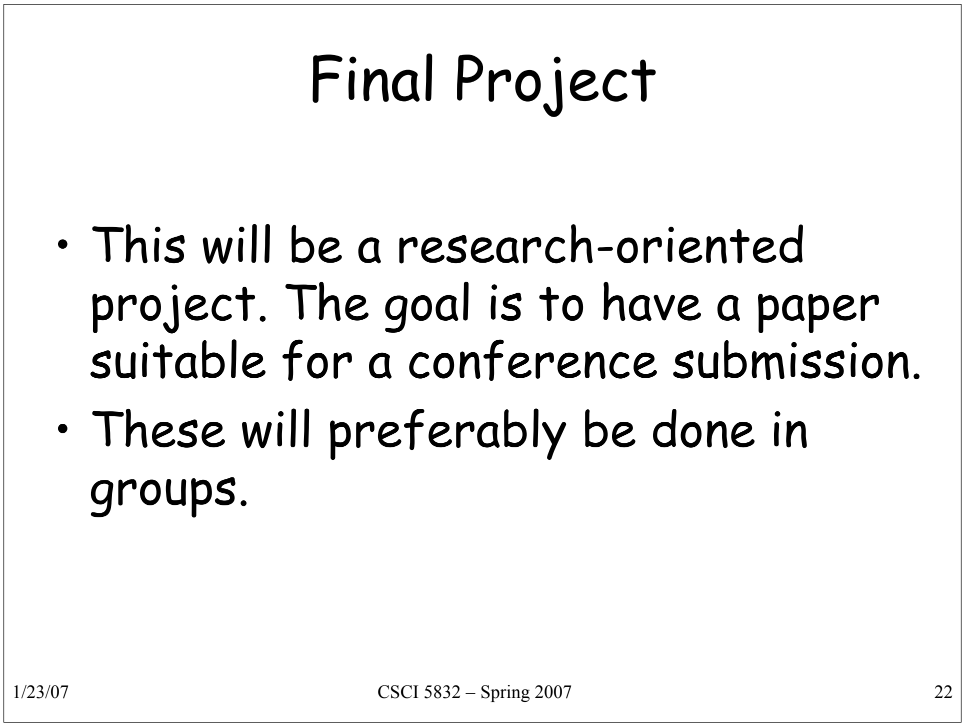
Preparation

- Basic algorithm and data structure analysis
- Ability to program
- Some exposure to logic
- Exposure to basic concepts in probability
- Familiarity with linguistics, psychology, and philosophy
- Ability to write well in English



Requirements

- Readings:
  - Speech and Language Processing by Jurafsky and Martin, Prentice-Hall 2000
  - Chapter updates for the 2nd Ed.
  - Various conference and journal papers
- Around 4 assignments
- 3 quizzes
- Final group project/paper with some presentations



Final Project

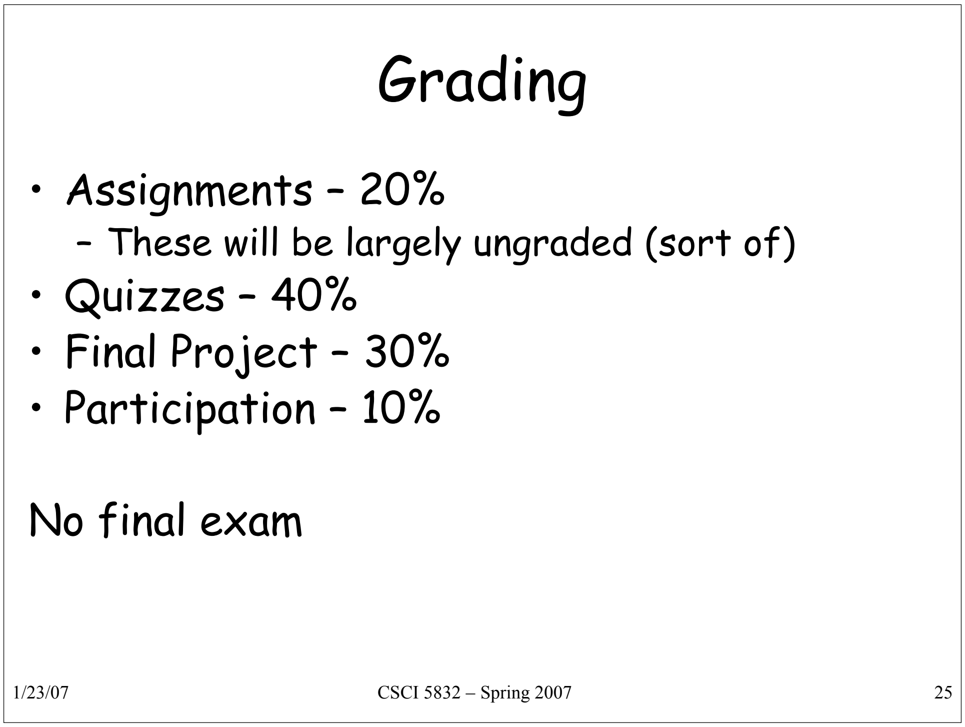
- This will be a research-oriented project. The goal is to have a paper suitable for a conference submission.
- These will preferably be done in groups.

Programming

- All the programming will be done in Python.
  - It’s free and works on Windows, Macs, and Linux
  - It’s easy to install
  - Easy to learn

Programming

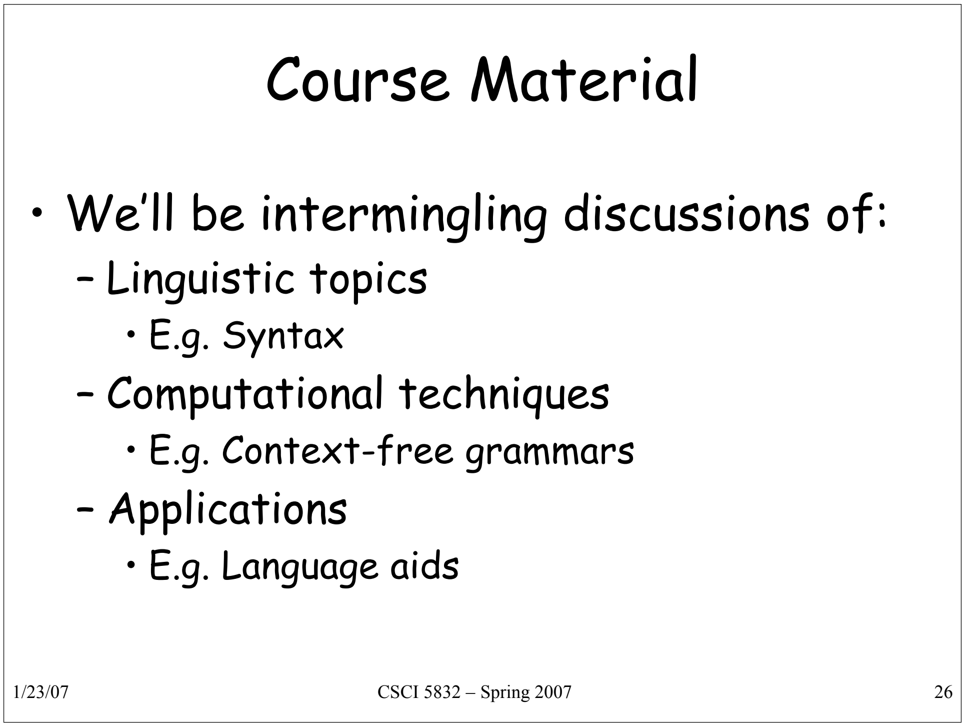
- Go to www.python.org to get started.
- The default installation comes with an editor called IDLE. It’s a serviceable development environment.
- Python mode in emacs is pretty good. It’s what I use but I’m a dinosaur.
- If you like eclipse, there is a python plug-in for it.



Grading

- Assignments – 20%
  - These will be largely ungraded (sort of)
- Quizzes – 40%
- Final Project – 30%
- Participation – 10%

No final exam



Course Material

- We'll be intermingling discussions of:
  - Linguistic topics
    - E.g. Syntax
  - Computational techniques
    - E.g. Context-free grammars
  - Applications
    - E.g. Language aids

Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics
- Discourse and dialog processing

My biases…

- I'm not terribly into phonology or speech
- I care about meaning in general, and word meanings in particular

Topics: Techniques

- Finite-state methods
- Context-free methods
- Augmented grammars
  - Unification
  - Logic

- Probabilistic versions
- Supervised machine learning

Topics: Applications

- Small
  - Spelling correction
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation

- Often stand-alone
- Enabling applications
- Funding/Business plans

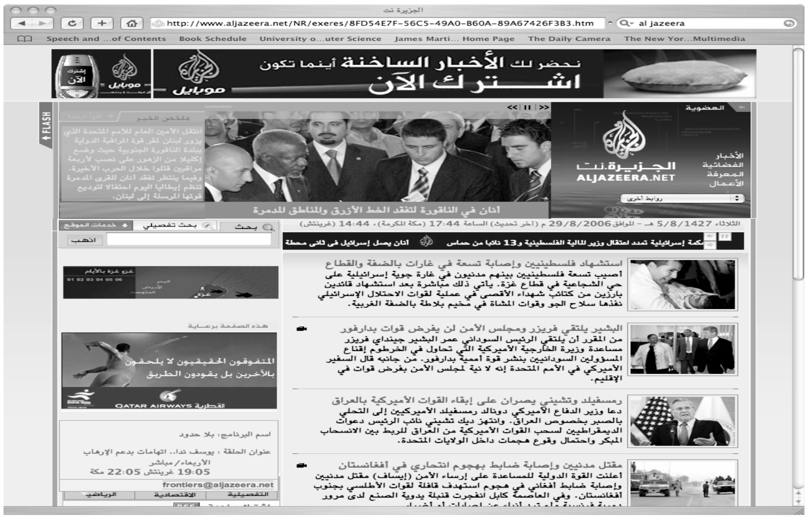
Just English?

- The examples in this class will for the most part be English.
  - Only because it happens to be what I know.
- Projects on other languages are welcome.
- We'll cover other languages primarily in the context of machine translation.

Commercial World

- Lot's of exciting stuff going on…
- Some samples…
  - Machine translation
  - Question answering
  - Buzz analysis

Google/Arabic



الجزيرة نت

http://www.aljazeera.net/NR/exeres/8FD54E7F-56C5-49A0-B60A-89A67426F3B3.htm

al jazeera

Speech and ...of Contents Book Schedule University o...uter Science James Marti... Home Page The Daily Camera The New Yor...Multimedia

نحضر لك الأخبار الساخنة أينما تكون
اشـــــترك الآن

ملخص الخبر

انتقل الأمين العام للأمم المتحدة الذي يزور لبنان لمقر قوة المراقبة الدولية ببلدة الناقورة الجنوبية حيث وضع إكليلا من الزهور على نصب لأربعة مراقبين قتلوا خلال الحرب الأخيرة. وفيما ينتظر تفقد أنان للقرى المدمرة تنظم إيطاليا اليوم احتفالا لتوديع قوتها المرسلة إلى لبنان.

أنان في الناقورة لتفقد الخط الأزرق والمناطق المدمرة

العضوية

الجزيرة.نت
ALJAZEERA.NET

الأخبار
الفضائية
المعرفة
الأعمال

روابط أخرى

الثلاثاء 5/8/1427 هـ - الموافق 29/8/2006 م (آخر تحديث) الساعة 17:44 (مكة المكرمة)، 14:44 (غرينتش)

حكمة إسرائيلية تمدد اعتقال وزير المالية الفلسطينية و13 نائبا من حماس

أنان يصل إسرائيل في ثاني محطة

بحث | بحث تفصيلي | خدمات الموقع

اذهب

استشهاد فلسطينيين وإصابة تسعة في غارات بالضفة والقطاع

أصيب تسعة فلسطينيين بينهم مدنيون في غارة جوية إسرائيلية على حي الشجاعية في قطاع غزة. يأتي ذلك مباشرة بعد استشهاد قائدين بارزين من كتائب شهداء الأقصى في عملية لقوات الاحتلال الإسرائيلي نفذها سلاح الجو وقوات المشاة في مخيم بلاطة بالضفة الغربية.

البشير يلتقي فريزر ومجلس الأمن لن يفرض قوات بدارفور

من المقرر أن يلتقي الرئيس السوداني عمر البشير جينداي فريزر مساعدة وزيرة الخارجية الأميركية التي تحاول في الخرطوم إقناع المسؤولين السودانيين بنشر قوة أممية بدارفور. من جانبه قال السفير الأميركي في الأمم المتحدة إنه لا نية لمجلس الأمن بفرض قوات في الإقليم.

رمسفيلد وتشيني يصران على إبقاء القوات الأميركية بالعراق

دعا وزير الدفاع الأميركي دونالد رمسفيلد الأميركيين إلى التحلي بالصبر بخصوص العراق. وانتهز ديك تشيني نائب الرئيس دعوات الديمقراطيين لسحب القوات الأميركية من العراق للربط بين الانسحاب المبكر واحتمال وقوع هجمات داخل الولايات المتحدة.

مقتل مدنيين وإصابة ضابط بهجوم انتحاري في أفغانستان

أعلنت القوة الدولية للمساعدة على إرساء الأمن (إيساف) مقتل مدنيين وإصابة ضابط أفغاني في هجوم استهدف قافلة لقوات الأطلسي بجنوب أفغانستان. وفي العاصمة كابل انفجرت قنبلة يدوية الصنع لدى مرور

هذه الصفحة برعاية

المتفوقون الحقيقيون لا يلحقون بالآخرين بل يقودون الطريق

QATAR AIRWAYS القطرية

اسم البرنامج: بلا حدود

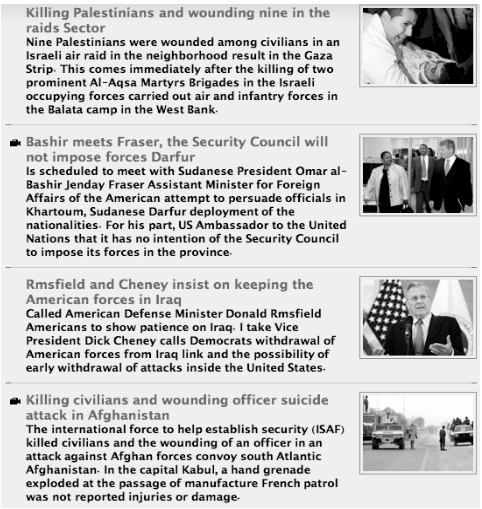
عنوان الحلقة : يوسف ندا.. اتهامات بدعم الإرهاب

الأربعاء/مباشر

19:05 غرينتش 22:05 مكة

frontiers@aljazeera.net

Google/Arabic Translation



Killing Palestinians and wounding nine in the raids Sector

Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

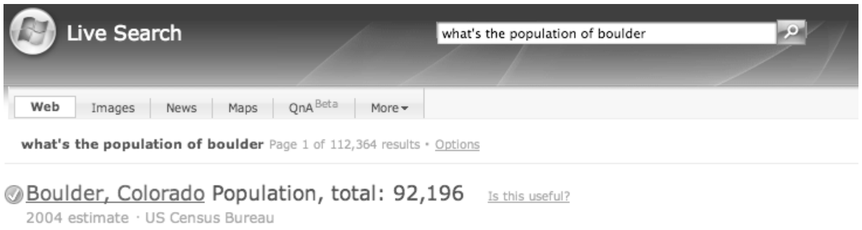
Rmsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan

The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Web Q/A

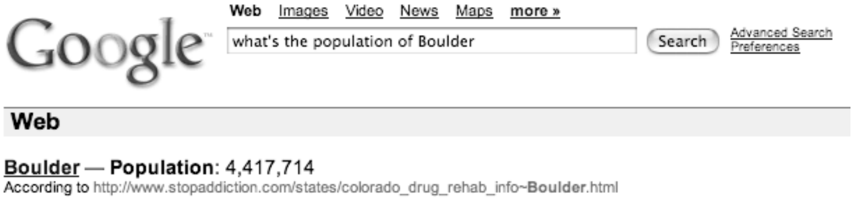


what's the population of boulder

what's the population of boulder Page 1 of 112,364 results • Options

Boulder, Colorado Population, total: 92,196 Is this useful?

2004 estimate · US Census Bureau



what's the population of Boulder

Web

Boulder — **Population**: 4,417,714

According to http://www.stopaddiction.com/states/colorado_drug_rehab_info~**Boulder**.html

Summarization

- Current web-based Q/A is limited to returning simple fact-like (factoid) answers (names, dates, places, etc).
- Multi-document summarization can be used to address more complex kinds of questions.

  Circa 2002:

  What's going on with the Hubble?

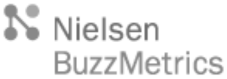
NewsBlaster Example

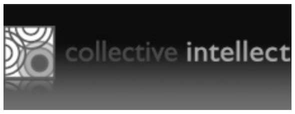
The U.S. orbiter Columbia has touched down at the Kennedy Space Center after an 11-day mission to upgrade the Hubble observatory. The astronauts on Columbia gave the space telescope new solar wings, a better central power unit and the most advanced optical camera. The astronauts added an experimental refrigeration system that will revive a disabled infrared camera. "Unbelievable that we got everything we set out to do accomplished," shuttle commander Scott Altman said. Hubble is scheduled for one more servicing mission in 2004.

Weblog Analytics

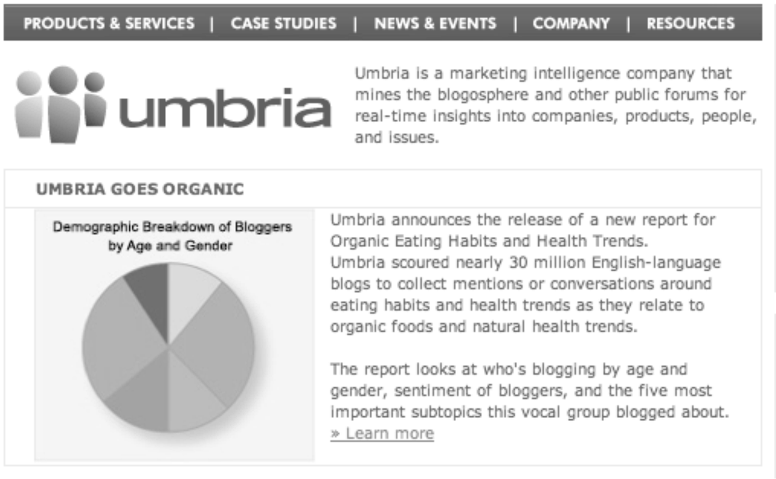
- Textmining weblogs, discussion forums, user groups, and other forms of user generated media.
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what's hot, what topics are people talking about right now).

Web Analytics





Umbria


PRODUCTS & SERVICES | CASE STUDIES | NEWS & EVENTS | COMPANY | RESOURCES
umbria

Umbria is a marketing intelligence company that mines the blogosphere and other public forums for real-time insights into companies, products, people, and issues.

UMBRIA GOES ORGANIC

Demographic Breakdown of Bloggers by Age and Gender

Umbria announces the release of a new report for Organic Eating Habits and Health Trends.
Umbria scoured nearly 30 million English-language blogs to collect mentions or conversations around eating habits and health trends as they relate to organic foods and natural health trends.

The report looks at who's blogging by age and gender, sentiment of bloggers, and the five most important subtopics this vocal group blogged about.
» Learn more

Next Time

- Read Chapter 1, start on Chapter 2
- Download, install and learn Python. The first assignment will be given out next time.