CS 6740 / INFO 6300
Advanced Language Technologies

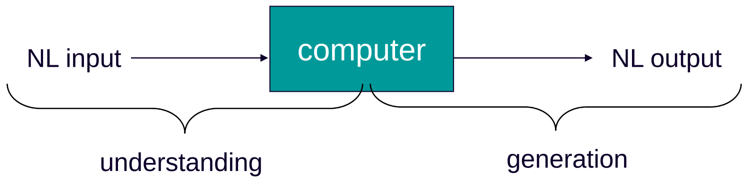
Graduate-level introduction to technologies for the computational treatment of information in human-language form,

covering **natural-language processing** (NLP) and/or **information retrieval** (IR).

Possible topics include text categorization and clustering, information extraction, latent semantic analysis (LSI), click-through data for web search, language modeling, computational syntactic and semantic formalisms, grammar induction, and machine translation.

Natural Language Processing (NLP)

- "Natural" language
  - Languages that people use to communicate
- Ultimate goal
  - To build computer systems that perform as well at using natural language as humans do
- Immediate goal
  - To build computer systems that can process text and speech more intelligently

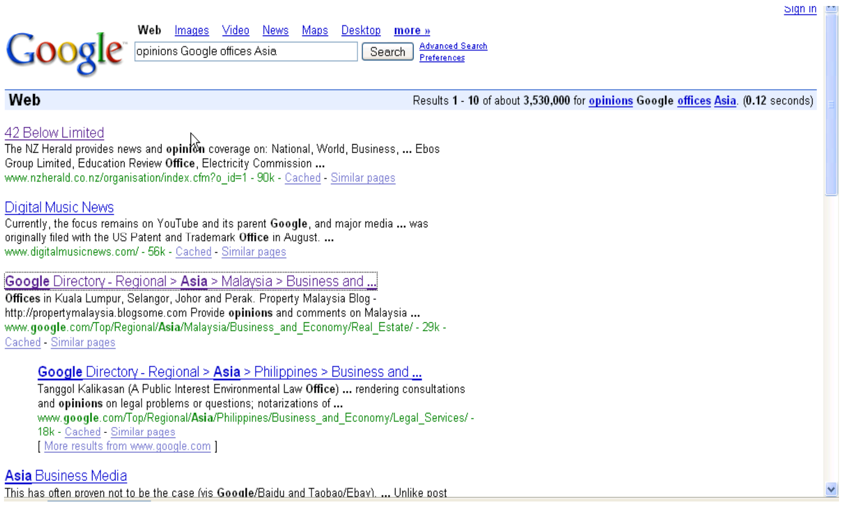


NL input → computer → NL output

understanding

generation

Information retrieval (IR)



opinions Google offices Asia

Web Results **1** - **10** of about **3,530,000** for **opinions Google offices Asia**. (**0.12** seconds)

42 Below Limited
The NZ Herald provides news and **opinion** coverage on: National, World, Business, **...** Ebos Group Limited, Education Review **Office**, Electricity Commission **...**
www.nzherald.co.nz/organisation/index.cfm?o_id=1 - 90k - Cached - Similar pages

Digital Music News
Currently, the focus remains on YouTube and its parent **Google**, and major media **...** was originally filed with the US Patent and Trademark **Office** in August. **...**
www.digitalmusicnews.com/ - 56k - Cached - Similar pages

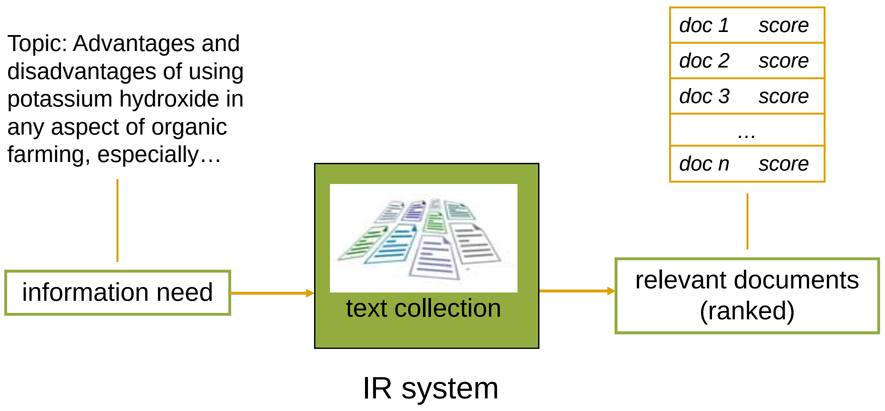
Google Directory - Regional > **Asia** > Malaysia > Business and **...**
Offices in Kuala Lumpur, Selangor, Johor and Perak. Property Malaysia Blog - http://propertymalaysia.blogsome.com Provide **opinions** and comments on Malaysia **...**
www.**google**.com/Top/Regional/**Asia**/Malaysia/Business_and_Economy/Real_Estate/ - 29k - Cached - Similar pages

Google Directory - Regional > **Asia** > Philippines > Business and **...**
Tanggol Kalikasan (A Public Interest Environmental Law **Office**) **...** rendering consultations and **opinions** on legal problems or questions; notarizations of **...**
www.**google**.com/Top/Regional/**Asia**/Philippines/Business_and_Economy/Legal_Services/ - 18k - Cached - Similar pages
[More results from www.google.com]

Asia Business Media
This has often proven not to be the case (vis **Google**/Baidu and Taobao/Ebay). **...** Unlike post

Information retrieval

- Ad-hoc IR



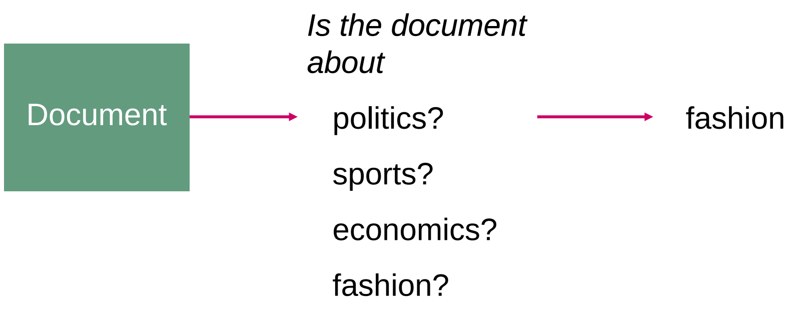
Topic: Advantages and disadvantages of using potassium hydroxide in any aspect of organic farming, especially…

information need → text collection → relevant documents (ranked)

doc 1	*score*
doc 2	*score*
doc 3	*score*
…	
doc n	*score*

IR system

Text categorization



Document

Is the document about

politics?

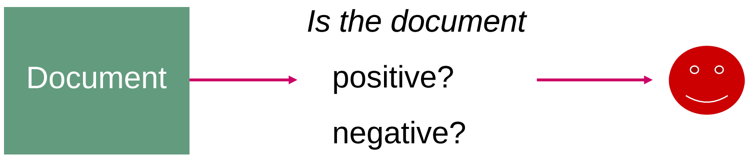
sports?

economics?

fashion?

fashion

Sentiment categorization



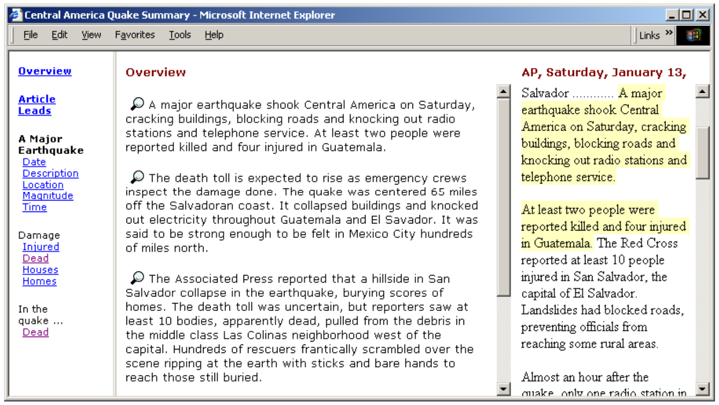
Document

Is the document

positive?

negative?

Summarization



Central America Quake Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links »

Overview

Article Leads

A Major Earthquake
Date
Description
Location
Magnitude
Time

Damage
Injured
Dead
Houses
Homes

In the quake ...
Dead

Overview

A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service. At least two people were reported killed and four injured in Guatemala.

The death toll is expected to rise as emergency crews inspect the damage done. The quake was centered 65 miles off the Salvadoran coast. It collapsed buildings and knocked out electricity throughout Guatemala and El Savador. It was said to be strong enough to be felt in Mexico City hundreds of miles north.

The Associated Press reported that a hillside in San Salvador collapse in the earthquake, burying scores of homes. The death toll was uncertain, but reporters saw at least 10 bodies, apparently dead, pulled from the debris in the middle class Las Colinas neighborhood west of the capital. Hundreds of rescuers frantically scrambled over the scene ripping at the earth with sticks and bare hands to reach those still buried.

AP, Saturday, January 13,

Salvador .... A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service.

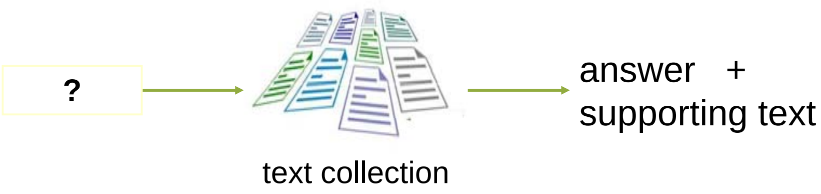
At least two people were reported killed and four injured in Guatemala. The Red Cross reported at least 10 people injured in San Salvador, the capital of El Salvador. Landslides had blocked roads, preventing officials from reaching some rural areas.

Almost an hour after the quake, only one radio station in

[White et al., 2002]

Question answering (QA)

- Task
  - » How many calories are there in a Big Mac?
  - » Who is the voice of Miss Piggy?
  - » Who was the first American in space?
  - – Retrieve not just relevant documents, but return the answer



?

answer + supporting text

text collection

Machine translation

- MT systems would clearly facilitate human-human communication
  - Certainly see a need for it…
    - u The extension of the coverage of the health services to the underserved or not served population of the countries of the region was the central goal of the Ten-Year Plan and probably that of greater scope and transcendence.
    - u Welcome to Chinese Restaurant. Please try your Nice chinese Food With chopsticks. the traditional and typical of Chinese glorious history and cultual. PRODUCT OF CHINA

Bill Gates, 1997 "…now we're betting the company on these natural interface technologies"

Dialogue systems

- Require both understanding and generation
  - Dave: Open the pod bay doors, HAL.
  - HAL: I'm sorry Dave, I'm afraid I can't do that.
  - Dave: What's the problem?
  - HAL: I think you know what the problem is just as well as I do.

…cks Guides > 2001: A Space Odyssey
Created with HyperSnap-DX 4
To avoid this stamp, buy a license at
http://www.hyperionics.com
Picture Gallery
Next >>

Pictures courtesy of 2001 Internet Resource Archive, HAL 9000, Kubrick Multimedia Film Guide and MGM/UA.

Umm…there WILL be complications

- NLP
  - AI-complete
    - » To "solve" NLP, you'd need to solve all of the problems in AI
  - Turing test
    - » Posits that engaging effectively in linguistic behavior is a sufficient condition for having achieved intelligence.

…But little kids can "do" NLP…

- Why is understanding language hard?

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- Syntax
  - Concerns sentence structure
  - Different syntactic structure implies different interpretation
    - » Squad helps dog bite victim.
      - u [$_{np}$ squad] [$_{vp}$ helps [$_{np}$ dog bite victim]]
      - u [$_{np}$ squad] [$_{vp}$ helps [$_{np}$ dog] [$_{inf\text{-}clause}$ bite victim]]
    - » Visiting relatives can be trying.

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- Semantics
  - Concerns what words mean and how these meanings combine to form sentence meanings.
    - » We steered the sailboat away from the bank.
      - u River bank vs. $$$ bank?
    - » Visiting relatives can be trying.
    - » Visiting museums can be trying.
      - u Same set of possible syntactic structures for this sentence
      - u But the meaning of **museums** makes only one of them plausible

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- Discourse
  - Concerns how the immediately preceding sentences affect the interpretation of the next sentence
    - » Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** will be called Prodome Ltd.
    - » Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** will own 50% of the new company to be called Prodome Ltd.
    - » Merck & Co. formed a joint venture with Ache Group, of Brazil. **It** had previously teamed up with Merck in two unsuccessful pharmaceutical ventures.

Why is dealing with NL hard?

Ambiguity!!!! …at **all** levels of analysis ☹

- Pragmatics
  - Concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

    "I just came from New York."

    - » Would you like to go to New York today?
    - » Would you like to go to Boston today?
    - » Why do you seem so out of it?
    - » Boy, you look tired.

What topics will we cover?

Language modeling
Lexical semantics and word-sense disambiguation
Part-of-speech tagging and HMMs
Parsing
Semantic analysis
Discourse processing
Coreference analysis
NL Generation
Machine translation

Information extraction
Information retrieval models
Sentiment analysis
Text categorization
Question answering
Summarization

GOAL: learn to evaluate and execute research in NLP/IR.

Reference Material

- Required text book:
  - Jurafsky and Martin, *Speech and Language Processing*, Prentice-Hall, **2nd edition**.
- Other useful references:
  - Manning and Schutze. *Foundations of Statistical NLP*, MIT Press, 1999.
  - Frederick Jelinek. *Statistical Methods for Speech Recognition*, MIT Press, 1998.
  - Others listed on course web page…

Prereqs, Coursework, & Grading

- Prerequisites
  - Permission of instructor. Neither INFO/CS4300 nor CS4740 are prerequisites.
- Grading
  - 25%: participation
    You'll be expected to participate in class discussion and class exercises.
  - 40%: semester project
  - 33%: presentation of readings and research papers
  - 2%: course evaluation completion