LING 120, Fall 2017
Language and Computers

Instructor: Sowmya Vajjala

Iowa State University, USA

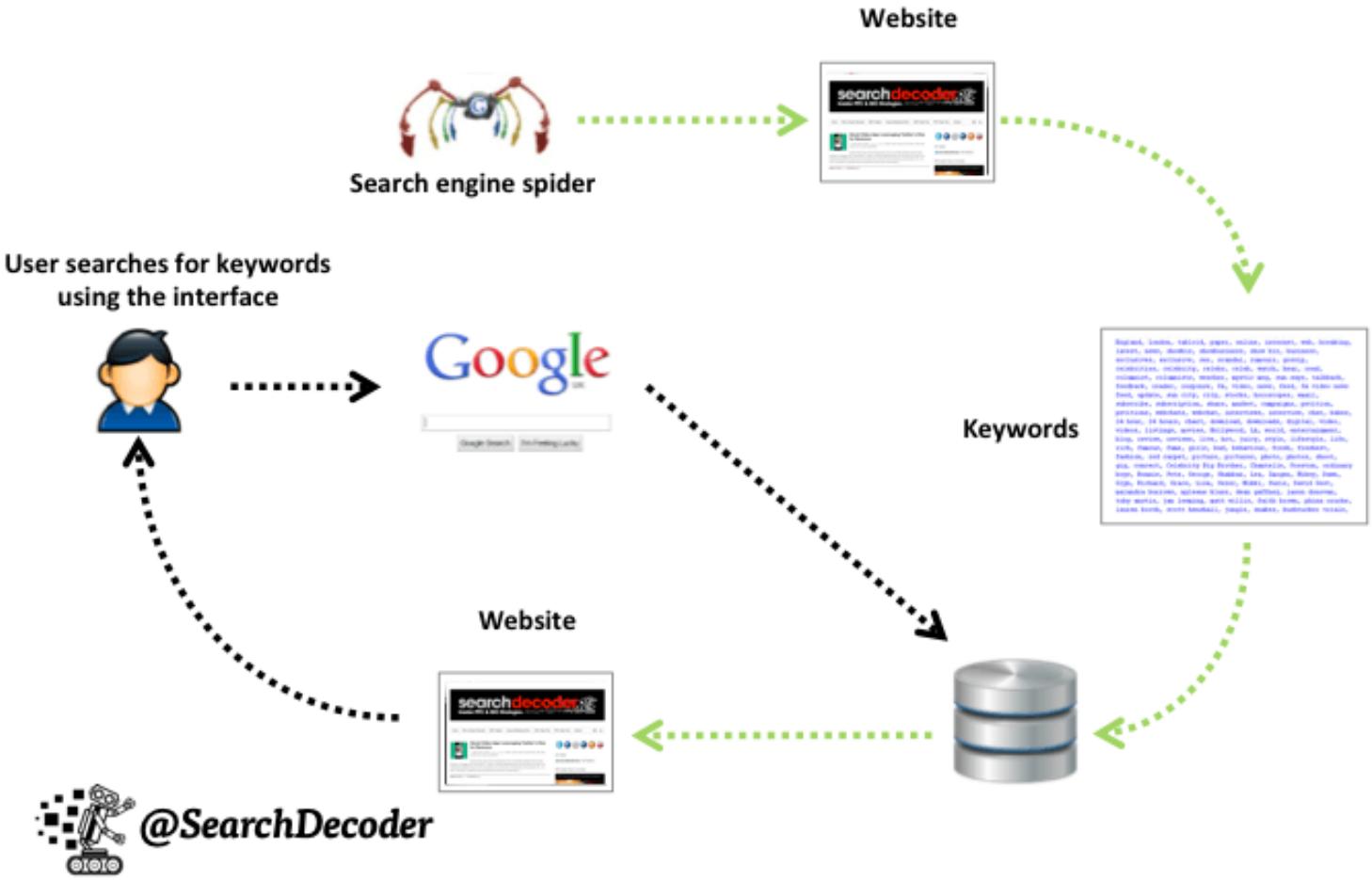
22 September 2017

Class outline

1. Recap of last class
2. Searching the web: Ranking search results
3. Announcement: Midterm - teams and topics
4. Group Exercise on understanding search results
5. Reminder: Submit Assignment 2 on time!

How does a search engine work?

Which of those require some analysis of language?



source:

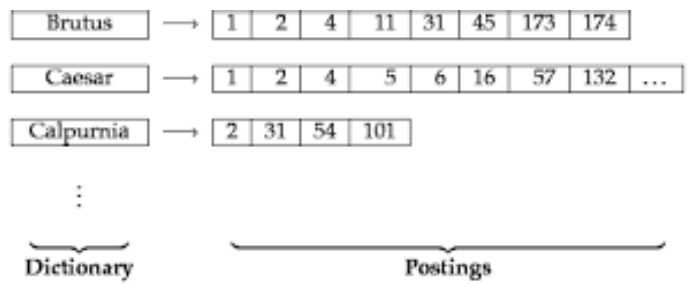
`https://www.searchdecoder.com/how-do-search-engines-work`

TDM and Inverted Index

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

► **Figure 1.1** A term-document incidence matrix. Matrix element (t, d) is 1 if the play in column d contains the word in row t, and is 0 otherwise.

(a) Term Document Matrix



► **Figure 1.2** The two parts of an inverted index. The dictionary is commonly kept in memory, with pointers to each postings list, which is stored on disk.

(b) Inverted Index

Questions to resolve Today

- Okay, indexing is cool. But for any query, I may still end up with 10000 results. How can I rank them?
- How do I ensure only good quality pages get ranked on top.

Ranking Search Results

- Search engines use different kinds of information to rank web-pages.
- We briefly saw some of them in the last class (e.g., are the query words in the title of the web page? etc)
- There can be several other factors such as: whether this page is "appropriate" for this user, whether this user previously spent long time on pages from this website etc.

Ranking Search Results

- Search engines use different kinds of information to rank web-pages.
- We briefly saw some of them in the last class (e.g., are the query words in the title of the web page? etc)
- There can be several other factors such as: whether this page is "appropriate" for this user, whether this user previously spent long time on pages from this website etc.
- Note: Search engines know a lot about us. Much more than you think. (e.g., `https://goo.gl/cwQoGL`)
- 200 factors google uses: `https://backlinko.com/google-ranking-factors` (NOTE: This is not validated by Google)

Ranking Search Results

- Search engines use different kinds of information to rank web-pages.
- We briefly saw some of them in the last class (e.g., are the query words in the title of the web page? etc)
- There can be several other factors such as: whether this page is "appropriate" for this user, whether this user previously spent long time on pages from this website etc.
- Note: Search engines know a lot about us. Much more than you think. (e.g., `https://goo.gl/cwQoGL`)
- 200 factors google uses: `https://backlinko.com/google-ranking-factors` (NOTE: This is not validated by Google)
- One popular ranking approach that was very influential in computer based search is: PageRank

Page Rank

- Intuition: If a page has a lot of other pages linking to it, it could mean the page is well-known, and is perhaps of good quality and is authoritaitve.
- So, the "rank" a page has is determined by pages linking to it
- ... and also the number of links into these pages as well.

Page Rank

- Intuition: If a page has a lot of other pages linking to it, it could mean the page is well-known, and is perhaps of good quality and is authoritaitve.
- So, the "rank" a page has is determined by pages linking to it
- ... and also the number of links into these pages as well.
- When the indexer returns two pages which are both relevant to a given query, one can then rank pages by their page rank and show to the user!

Other issues in search

- Let us say someone started searching for "I love Java" - ideally, a search engine should cluster results into multiple "senses" of meaning - Java coffee, Java - the place, Java - the programming language. (Kind of difficult. Open question).

Other issues in search

- Let us say someone started searching for "I love Java" - ideally, a search engine should cluster results into multiple "senses" of meaning - Java coffee, Java - the place, Java - the programming language. (Kind of difficult. Open question).
- Detecting pages with duplicate content

Other issues in search

- Let us say someone started searching for "I love Java" - ideally, a search engine should cluster results into multiple "senses" of meaning - Java coffee, Java - the place, Java - the programming language. (Kind of difficult. Open question).
- Detecting pages with duplicate content
- Working through multiple forms of data (if I give a query, how about getting back mp3 files containing that query as results??)

... and so on.

How google works: `https://www.google.com/search/howsearchworks/algorithms/`

Mid-term: Dates and Expectations

- Dates:
- Teams (next slide) 1–3 on Monday (9th October); 4–6 on Wednesday (11th October)

Mid-term Groups

1. Group 1: Dallas Evers, Krishna Chaitanya Gummalla, Nellie Waidelich, Congyu Zhao
2. Group 2: Marcelanne Gaebler, Gabriel Muhammed, Trevor Holbrook, Yangyifan Lu
3. Group 3: Nathan Friedrichsen, Linghan Zhang, Sydney Steinauer, Jordan Yardley
4. Group 4: Joseph Hudson, Brandon Huegli, Miles Lucas
5. Group 5: Maceo Jackson, Matthew Schaffer, Joshua Slagle
6. Group 6: Emily Kurimski, Dylan Mrzlak, Jonah McDermed, Ethan McGill

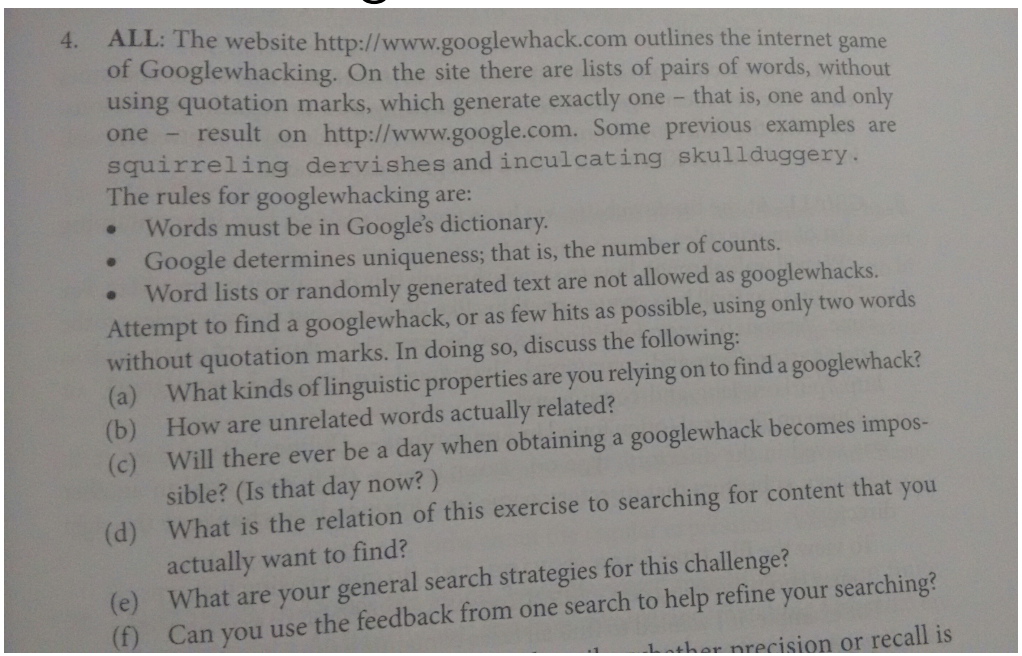
Mid-term Topics

Uploaded on Canvas. 11 topics in total - choose one of them. (Descriptions)

Last class' Exercise

Think about these problems and submit your thoughts online on Canvas forum for today.

- Is "popularity" a good heuristic to rank a page? Provide one example where your search query results in a popular page as the top result, but is incorrect for your need. What is the rank of the page that met your need?
- Question on Googlewhack. That website in the picture does not work though. You should use other means to know more.



4. **ALL**: The website http://www.googlewhack.com outlines the internet game of Googlewhacking. On the site there are lists of pairs of words, without using quotation marks, which generate exactly one – that is, one and only one – result on http://www.google.com. Some previous examples are `squirreling dervishes` and `inculcating skullduggery`.
The rules for googlewhacking are:
   - Words must be in Google's dictionary.
   - Google determines uniqueness; that is, the number of counts.
   - Word lists or randomly generated text are not allowed as googlewhacks.

Attempt to find a googlewhack, or as few hits as possible, using only two words without quotation marks. In doing so, discuss the following:
(a) What kinds of linguistic properties are you relying on to find a googlewhack?
(b) How are unrelated words actually related?
(c) Will there ever be a day when obtaining a googlewhack becomes impossible? (Is that day now?)
(d) What is the relation of this exercise to searching for content that you actually want to find?
(e) What are your general search strategies for this challenge?
(f) Can you use the feedback from one search to help refine your searching?

Attendance Exercise

- Work in groups of 3 and submit a solution to the problem in the handout.
- You can also submit this online on Canvas.
- question url: `http://nacloweb.org/resources/problems/2007/N2007-B.pdf`

Next Week

- Semi-structured search, Regular Expressions
- Assignment 3 Description and Assignment 2 Discussion (Monday)
- To do: Read Chapter 4, Finish Assignment 2
- Could do: Know your mid-term team members. Start thinking about a topic.