You would like to...

- Know whether someone could get the **flu virus** by handling money.
- Know who **invented** the television and when.
- Decide whether having an **abortion** should be considered moral.
- Figure out whether **technology** could potentially help reduce poverty in Africa.
- Decide who is a better **candidate** for the upcoming congressional elections.

**Who do you consult for answers?**

- Myself and my experiences
- A friend or family member
- A religious authority
- A professor
- Someone else
- A book or librarian
- A newspaper
- An ad that I like
- The Internet
- Social Media (Facebook,...)

**How did you decide who to ask:**

- Would they know?
- Can I trust them?
- Is it convenient to ask?
In the past, the greatest problem we were facing was finding information; now, with the advent of the Internet, the greatest problem is evaluating the vast wealth of information available. Students [...]

Students placed greater emphasis on the process of finding an answer than on analyzing the actual information.

In Google We Trust...

“OF COURSE IT’S TRUE; I SAW IT ON THE INTERNET!”

Google

In Google We Trust...

Reliability of Information on the Web

Anyone can be an author on the web!

JUMP START by Robb Armstrong

crawl the web
create inverted index

Document IDs

Search engine servers
Inverted index

How Google Search Works?

1st Gen. Ranking = Content Similarity...

**Content Similarity Ranking:**
The more rare words two documents share, the more similar they are

<table>
<thead>
<tr>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>157</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brutus</td>
<td>4</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>232</td>
<td>227</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>worser</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Documents are treated as “bags of words” (no effort to “understand” the contents)
- Term frequencies are computed (Above: a few rare words in W.S. books)
- Very common words (“a”, “of”, “the”, “in”, ...) are ignored
PageRank algorithm = Page Reputation

- The reputation “PageRank” of a page $P_i$ is the sum of a fraction of the reputations of all pages $P_j$ that point to $P_i$.

- Idea similar to academic reputation

- Beautiful Math behind it

- The PageRank of a page can be calculated by a sequence of matrix multiplications.

- It is equivalent to the probability of randomly surfing to a page.

```
\[ P_i = (1-d) + d \cdot \sum_{j=1}^{N} \frac{P_j}{C_j} \]
```

How To Spam?

“Fake News”

- Lies, formatted and circulated online in such a way that a reader might mistake them for legitimate news.

FBI Insider: Clinton Emails Linked To Political Pedophile Sex Ring

Asking Google for Pizzagate evidence...

Search results for pizzagate evidence 30 April 2017

(Mis) Information in Social Media

Anti-Trump protestors in Austin today are not as organic as they seem. Here are the busses they came in. #fakeprotests #trump2016 #austin
Handling Misinformation

- **Read the article about Facebook:**
  - [https://goo.gl/fE3XDB](https://goo.gl/fE3XDB)
  - What approach did Facebook try?
  - What were the main factors leading to changing this approach?
  - What alternative approach did Facebook implement?

- **Read this article about Google:**
  - [https://goo.gl/WJHuyd](https://goo.gl/WJHuyd)
  - What approach is Google applying?
  - Compare and contrast to Facebook’s approach
  - What other approaches can you think of?