Web Search: Adversarial Information Retrieval

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Adversarial IR

- One key difference between traditional IR and Web Search is that traditionally, works were published without IR systems in mind.
- An author writing a book or a newspaper, magazine or technical article would only concentrate on the work they were producing.
- The publishing of web pages is very different. An entire industry known as **Search Engine Optimisation (SEO)** has emerged.
- SEO involves taking steps to get your page high up in search engine rankings.
  - Motivation: Higher rankings → more clicks → more money.
- Case study on how Google's search algorithms can impact businesses: https://www.bbc.com/future/article/20240524-how-googles-new-algorithm-will-shape-your-internet
  - Housefresh.com went from thousands of daily visitors and 15 employees to hundreds of visitors and having to let go most of its staff because of a change in Google's ranking algorithms.

Search Engine Optimisation (SEO)

- To improve search engine ranking, many techniques are available, and can be divided into two categories:
  - "White Hat" SEO: operating ethically and following search engine rules and guidelines to get higher rankings.
    - Optimising the website for relevant keywords.
    - Creating mobile-friendly versions of websites.
    - Satisfying users' information needs.
    - Descriptive and relevant URLs, titles and meta descriptions.
    - Relevant, descriptive text in links.
    - etc.
  - "Black Hat" SEO: violating search engine guidelines and rules to try and get an artificial advantage: Adversarial IR.

Further reading: https://backlinko.com/white-hat-seo
https://support.google.com/webmasters/topic/6001971

Adversarial IR: Meta Tags

- HTML allows for page authors to describe the contents of their pages by using <meta> tags.
- These were designed to allow such information as a page's description, keywords and author to be specified in a standardised way.

```
<head>
  <title>Stamp Collecting World</title>
  <meta name="description" content="Everything you wanted to
     know about stamps, from prices to history." />
  <meta name="keywords" content="stamps, stamp
     collecting, stamp history, prices, stamps for sale" />
</head>
```

Adversarial IR: Meta Tags

- Some early search engines did not have the processing power to perform full-text search.
- Even for many that did, they initially gave a high weight to terms appearing in the description and keywords of a page.
- Theoretically, this would succinctly describe the key contents of their web sites.
- In reality, this was open to abuse by dishonest webmaster that would attempt to target certain search queries.
- A common example was people including the word “free” amongst their keywords, as many searchers would use this when they began searching for a particular service.
- Because of this behaviour, meta keyword tags are mostly ignored by search engines nowadays. Meta descriptions are still used to some degree.

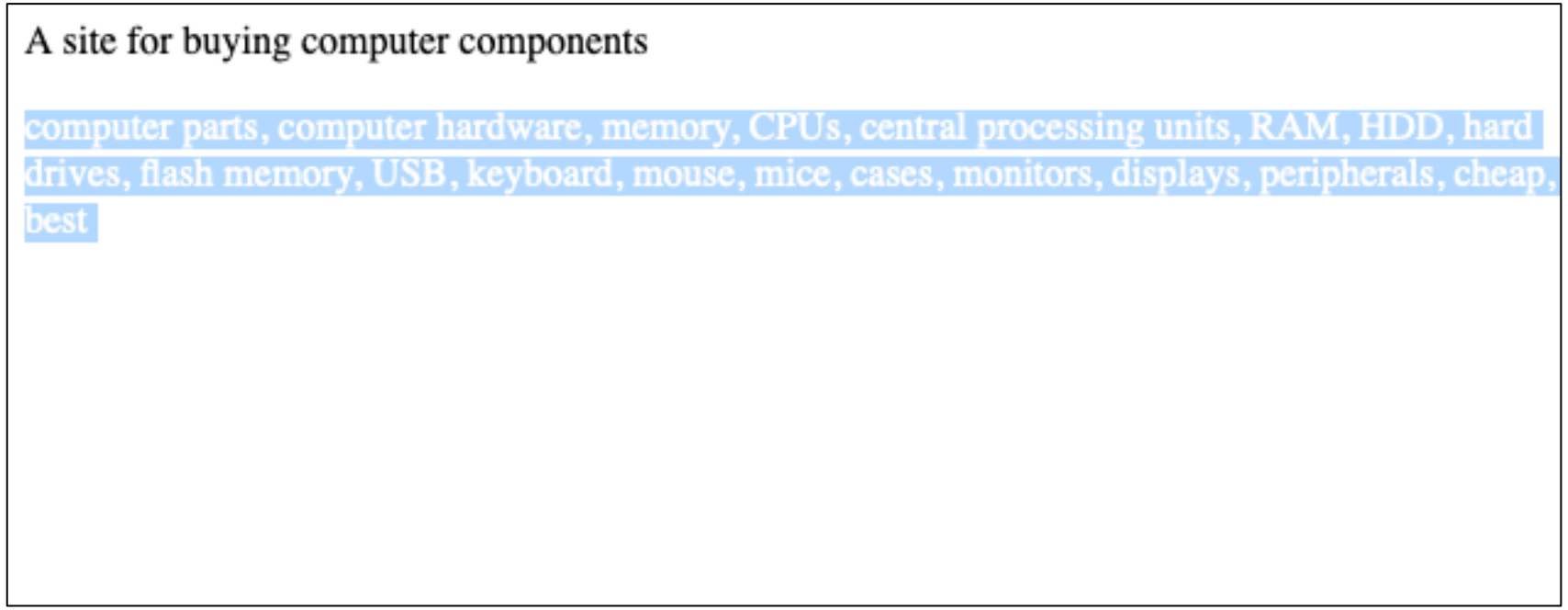
Adversarial IR: Hidden Text

- Consider this example:


```
<p>A site for buying computer components</p>
<p><font color="#ffffff"> computer parts,
computer hardware, memory </font></p>
```

Adversarial IR: Hidden Text

- Consider this example:




```
<p>A site for buying computer components</p>
<p><font color="#ffffff"> computer parts,
computer hardware, memory </font></p>
```

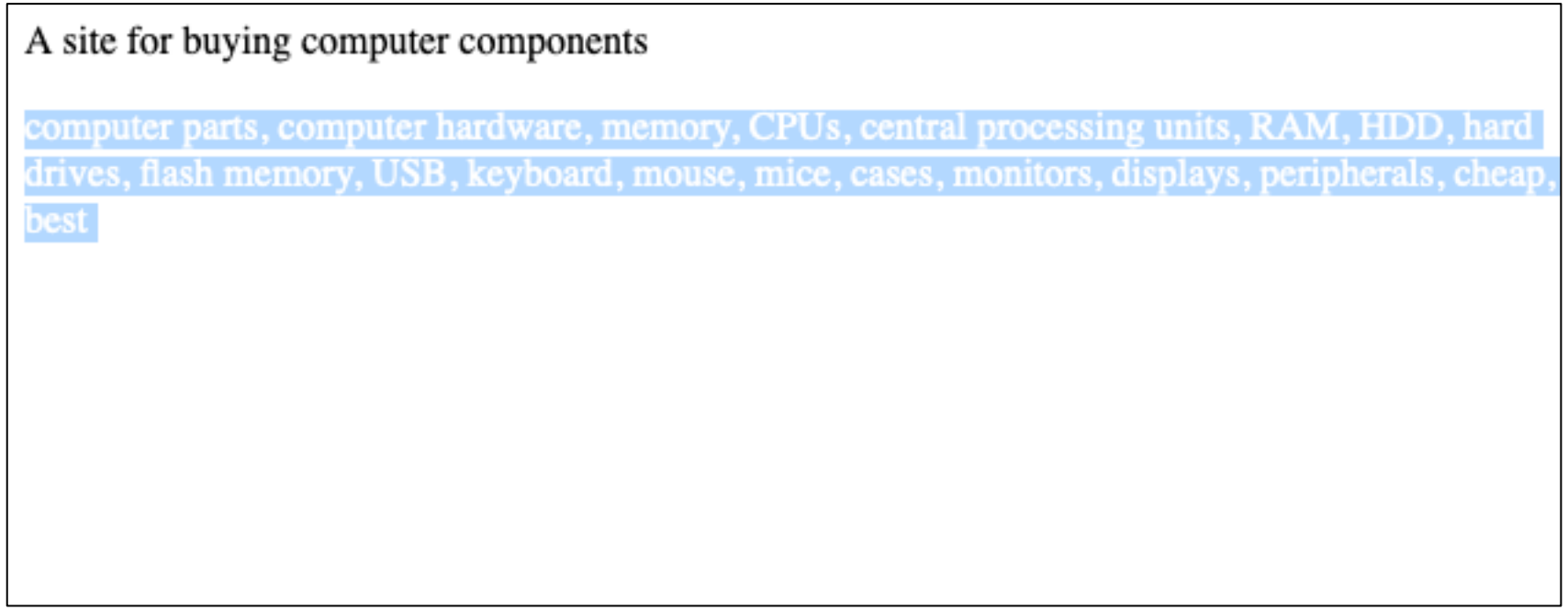
Adversarial IR: Hidden Text

- This is an example of manipulation using **hidden text**.
- The second paragraph in the previous slide will not be visible to the human eye, as the text is the same colour as the background.
- This is a method of changing the contents of the page without negatively affecting what a user sees.
- It has two main functions:

1. Add extra keywords for a search engine to pick up that aren't already contained in the content of the page (known as "keyword stuffing").
2. Increase the term frequency of important search terms.

Adversarial IR: Hidden Text

- Consider this example:




```
<p>A site for buying computer components</p>
<p><font color="#fffffe"> computer parts,
computer hardware, memory </font></p>
```

Another example.

```
<p class="class1">A site for buying computer components</p>

<p class="class2">computer parts, computer hardware, memory</p>
```

... with the following CSS:

```
.class1 {
   font-family: century gothic;
}

.class2 {
   color: white;
}
```

Adversarial IR: Hidden Text

- Here, the manipulation is not as obvious, as you'd have to check the Cascading Style Sheet class "class2" to see if there are any color changes defined there.
- You would also have to check any other elements that the second paragraph is contained in.
- This is not a trivial task for a crawler to carry out (remember that a crawler needs to operate very quickly to build up a sufficiently large index).
- Many search engines will remove your site from its index if this type of activity is found.

Adversarial IR: Cloaking

- **Cloaking** is the term given to the practice of serving different content to web crawlers than to users.
- A well-behaved crawler will identify itself to a web server when it visits, by way of a "user agent string".
- A server (or a script running on it) can easily be configured to give different content based on this user agent.
  - This technique is often used legitimately to show a mobile-optimised site for phone users.
- The idea is to serve highly-targeted text to the crawler, so as to gain a high ranking on particular searches.
  - Regular users would be served the normal content.
- This can also be done using redirect (redirecting users and crawlers to different URLs).

Adversarial IR: Sneaky JavaScript

- Web crawlers (and old browsers) are normally not capable of executing JavaScript code.
- In this situation, there is a <noscript> tag that allows you to specify text content for users who are unable to run JavaScript (this is not so common anymore) or those who are unwilling to run it.
- This text is what the web crawler will see as the content of the page.
- As a result, this can be exploited in the same way as cloaking: web crawlers index the contents of the "noscript" tag, whereas real users will get redirected (via JavaScript) to a different page without them noticing.

Adversarial IR: Exploiting PageRank

- Since a high PageRank (or similar) results in high rankings in search results, there is a commercial benefit to having a high PageRank score.
- This has led to the creation of "link farms": sites (or networks of sites) with the sole purpose of linking to other sites so as to increase their PageRank.
- Each page has a small contribution to overall PageRank, so it is easy to create huge numbers of pages consisting solely of links.
- Additionally, pages that benefited from this PageRank would link back to the farm so as to boost the PageRank of the farm.
- Often, this link would be hidden by being the same colour as the rest of the page.
- Some owners of sites with high PageRank built up in this way would charge a fee from anybody who wanted to be linked to from the farm.

Adversarial IR: Exploiting PageRank

- Another approach is "comment spam", where pages that allow third-party commenting are abused to try and boost PageRank.
  - Advertising links are posted as comments in hundreds or thousands of places, to gain PageRank from each.
- This is why the "nofollow" attribute for links was created. This would means that search engines do not take the link into account when calculating PageRank. This allows site owners to have more control over which links from their site add to the PageRank of another page.
- This still needs to be implemented correctly. Bots will search for sites that do not use these attributes and continue to try and exploit this where possible.

Adversarial IR: Doorway Pages

- Rather than attempting to include all of the desired search queries on a single page (which will affect the term frequencies of the individual terms), sometimes individual "doorway pages" are created.
- These are pages that are optimised for a particular search query, each of which then funnels a user to the same destination.
- Again, this is a method of creating content that is solely for the benefit of search engines so as to gain a high ranking.

Adversarial IR: Summary

- Some website owners can violate search engine's rules and guidelines to get higher search engine rankings ("Black hat SEO").
- This can be difficult to detect, but search engines will penalise or ban sites who are found to break the rules.
  - A guide to Google Penalties (https://www.searchenginejournal.com/the-complete-list-of-google-penalties-and-how-to-recover/201510/)
  - 10 Big Brands that were Penalized by Google (https://marketingland.com/10-big-brands-that-were-penalized-by-google-69646).
  - 5 Times Google Penalized Itself for Breaking Its Own SEO Rules (https://searchengineland.com/google-penalized-breaking-seo-rules-184098).
- The job of the search engine creator is to try to make sure that web pages using these tactics are not included in the index, because they are unlikely to satisfy users' information needs.