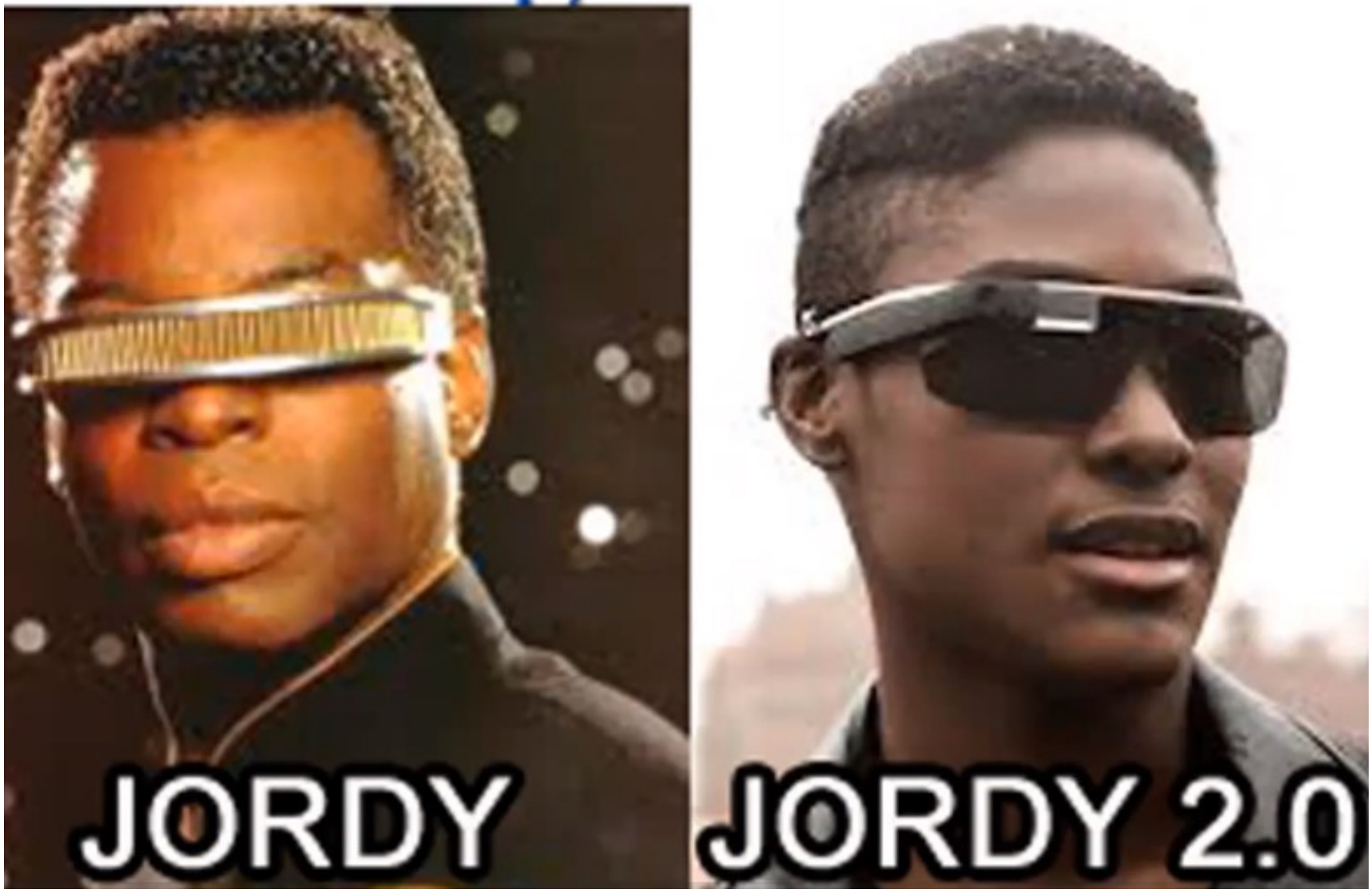
Other example (talking about view…!)

Google GLASS




GOOGLE GLASS
EVERY TIME I SEE SOMEONE WITH A GOOGLE GLASSES , I'M GOING TO GO UP TO THEM AND SCREAM "GOOGLE GLASSES : IMAGE SEARCH HORSE FUCK. SAFE SEARCH OFF. OPEN FIRST 50 RESULTS IN NEW TABS...
i WILL THEN RUN OFF INTO THE NIGHT..
memecenter.com
MemeCenter

Another essential feature...

- ... For the success of a site, is to be visible ***outside***
- Typically, via ***search engines***

So, it is critical…

- To be ranked highly from a search engine
- → in general, in the **SERP (Search Engine Results Page)**

For instance…

- How much is it important to appear in the top ten?

Top ten or not top ten?

- La top ten absorbs…
- … more than **95%** of all clicks!!

First position?

- 1) **51%** of clicks (!!!!!!!!)
- So the first positions already attracts ***more clicks than all the others combined!!***

Go on…

- 2) 16% of clicks (!)
- 3) 6%
- 4) 6%
- 5) 5%
- 6) 4%
- 7) 2%
- 8) 1%
- 9) 1%

THE CHINESE - Malabrocca, born in Tortona on June 22, 1920, also called "the Chinese" because of his almond-shaped eyes, was a runner's finest. In his career he has won 138 races, 15 of which as a professional (Paris-Nantes 1947, Coppa Agostoni 1948, Tour of Croatia and Slovenia 1949), and was twice Italian cyclocross champion (1951 and 1953).

He was running away behind the group. He went into

So: “Malabrocca effect”

- At the last place in the top teb, the click-rate ***doubles*** compared to the previous positions: **2%** !!

And the mix testo-immagini?

Consequence

- Mixing images to text doesn't alter the proprieties of the «textual» top ten

So how do we climb the ranking?

- **SPAMDEX** = SPAM INDEX
- Also called
  SEO (Search Engine Optimization), or
  SEP (Search Engine Persuasion)

File not found

Coming Soon.....

Groovymovies.com

Reviews!
Video!
DVD!
Current Films!
More!

How is rank calculated?

- Currently, information is given by the ***textual*** component of a page, plus its ***hypertextual*** component

On hypertextual...

- You probably heard already something about it (pagerank etc): we will go back to this soon

Let's focus before…

- On the «less famous» part, the textual one
- And let's try to see things from the search engine perspective…

The textual part

- All search engines use at base level variants of the same technique
- Called TFIDF (or TF-IDF)

TFIDF

- Stands for
  Term **F**requency
  Inverse **D**ocument **F**requency
- Gives a measure of how important a word is for the page
- **TFIDF = TF * IDF**

The problem with TF

- IF we were to use TF only, many cases just would not work
- Think for instance words like «the»: they would be the most important words in a page!
- To solve this, the second component (IDF) comes to help

IDF

- Inverse Document Frequency → the inverse of the frequency of the word **within** the set of documents (web site, www etc)…
- scaled logarithmically

Example

- Web site of 1000 pages, «the» appears in 980 pages → 98% frequency (0.98) → IDF is log(1/0.98) = **0.008**
- Web site of 1000 pages, "bike" appears in 100 pages → 10% frequency (0.1) → IDF is log(1/.1) = **1**
- Web site of 1000 pages, "Schopenauer" appears in 10 pages → 1% frequency (0.01) → IDF is log (1/0.01) = **10**

Note…

- The "dilemma" (tension)…:
- If we want to raise the textual score of a page for a word w, we have to be careful because raising too much its TFIDF automatically causes lowering the TFIDF of the other words!

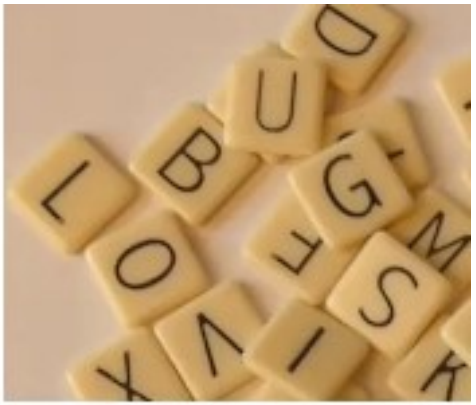
The strategy

- Focus on a set of words (the «champions»), and raise their TFIDF (and textual score), lowering the others

So…

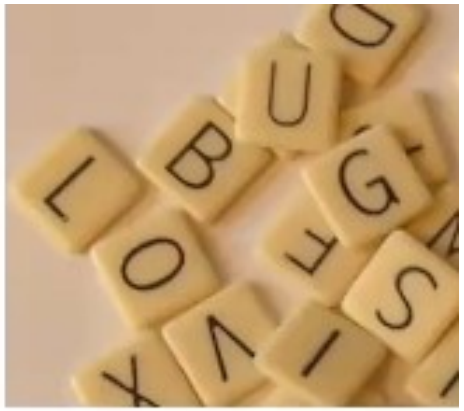
- … we will have to carefully ***choose a set of keywords***, and appropriately ***manage*** them in the web site

Body spam



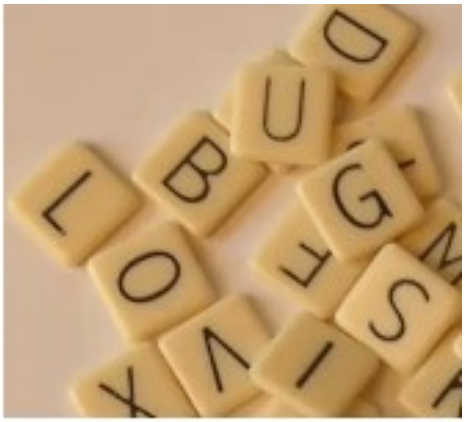
- The easiest way: insert words into the BODY of an HTML page
- Simple and effective (apart from the compromises with the TFIDF…)
- Disadvantage: we are touching the actual content of the page

Meta tag spam



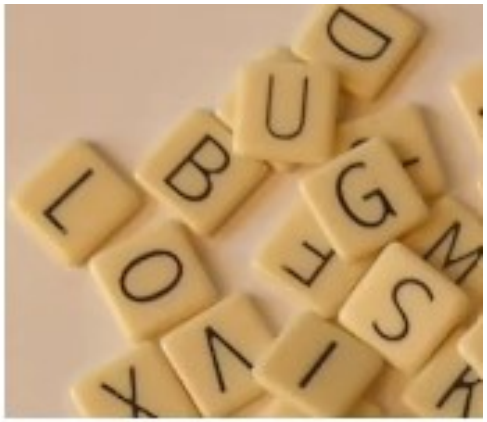
- <meta name="keywords"
  content="bike, football, sport">
- Advantages: no (user side) visible content is touched
- Disadvantages: abused, very low score by current search engines

Anchor text spam



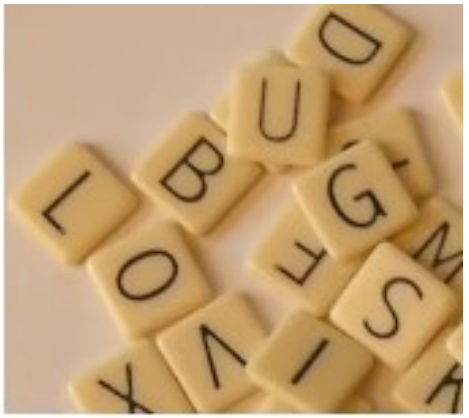
- Technically, a part of body spam, but it is usually considered apart
- Insert the words in the anchor text (<A>…</A>)
- Special scores, and a peculiarity…

Anchor text spam (cont.)



- Breaking the (usually purely textual) model, keywords are typically also added by the search engine to the ***target*** page of the link
- And as added bonus, so with *less limitations* with respect to TFIDF!

URL spam



- Finishing term spamming, URL spam is the technique to insert keywords directly into the web address of the page (!)
- This because search engines also analyze the addresses, giving ***bonuses*** similar to the anchor text spam

The other side of the coin

- So far, we saw positioning in term spamming, so to say the «***form***»
- But obviously also the «***content***» matters: what keywords do we use

The "Starter Kit"

- Repetition
- Dumping
- Weaving
- Stitching
- Broadening

Repetition

- Repeat the same keyword, of course paying attention…
- … to the TFIDF (balancing), and also to
- the **countermeasures** (!), given that repetition spam is easy to identify (and penalize…!)

Dumping

- Insert many terms that are seldom used, even if not related to the page (!)
- → being rare keywords their score will be relative high!

Weaving

- Take pieces of other web sites, and modify them by inserting our keywords (usually in a random way)
- So, automatic way to create interesting content, and power it up in attraction with our specific keywords

Stitching

- Paste & copy of fragments of other web pages, uniting them into a single page
- → automatic way to create «interesting» content to populate a site (various search engine also use ***global bonuses*** to measure how much ***information*** does a site offer)