Searching: Needles and Haystacks

- Why it's important
- How it's done
- Technical difficulties
- Prospects, what's happening
- Social and ethical difficulties

Search and Hugh

- Involved in search here and there since about 1982, mainly EEC projects especially Celex

- Recent work for Vienna U on a multimedia database for stored manuscripts etc.

- Professionally in computing since about 1974. Actually small FORTRAN program to calculate π in about 1966!

- **Slides available at: http://www.slideshare.net/hughbar/**

hugh.barnard@gmail.com

As a Human Activity

- Looking for keys
- Remembering names and birthdays
- Looking up in a book
- And [the subject of this] making tools for the intertubes
- Getting a clue, from the above...

Why do/understand this at all?

- Since Google, Bing, Yahoo **already did it**, *for* you:
- Lots of interesting technical pieces
- Self education
- Fun and profit, do it 'better/different'
- Internal search engines, intranet search engines
- Domain specific engines
- School or research projects

How it's Done: 1

- Health warning: this explanation is simplified!
- Let's take Google/Bing for example
- How does it find one zillion documents/images with 'lolcat' in them, within a few seconds?

How it's Done:2

- It did it already
- A key concept: **Indexing**

Another key concept: **Inverted index [see wikipedia]: https://en.wikipedia.org/wiki/Inverted_index**

- lolcat in document x at position y

- **highlighted** cat in document x at position y

(we'll come back to this, for **rank**)

Some Indexing Problems

- Bandwidth and data transfer, 4.73 billion '*visible* pages' as of November

- Coverage, darkweb, dynamic creation of pages, javascript etc. etc.

- Speed of refresh, 'freshness' of indexed data

- Storage sizes

These don't apply or 'less' to domain specific projects

Parts of a Search Engine

- 1: Spidering, harvesting and directory processing (we'll come back to the differences)

- 2: Parser/Indexer

- 3: Index/data Storage

- 4; Retrieval [the bit of Google/Bing **that we see**!]

At any stage 'algorithms', for spam detection, for ranking etc. etc. **the secret sauce of search**

I'm going to go through these in order...

1: Spidering/Harvesting/Darkweb

- Spidering: start with a seed, follow links

- Directory based: index a load of things in a given directory

- Harvesting and metasearch: academic harvester interfaces, domain specific, I do this at present: https://en.wikipedia.org/wiki/Bioinformatic_Harvester

- Robots.txt: courtesy, the interactive web and the darkweb

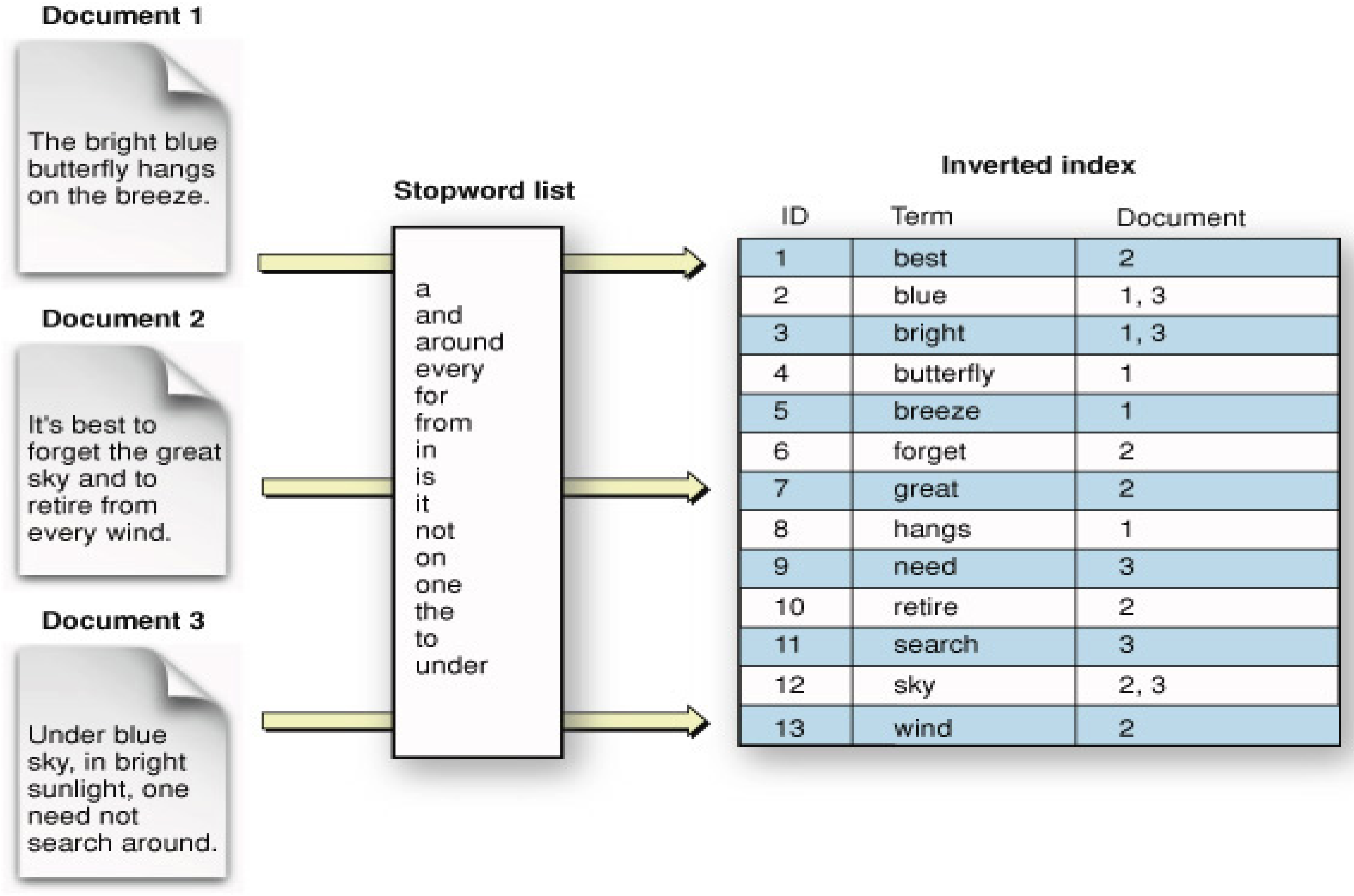
- Resource: https://commoncrawl.org/what-we-do/

Can people think about problems with any of these?

2: Parser/Indexer

- Now the fun begins!
- Parsing: breaking down the stuff you get into tokens
- <b>lolcat</b> can be indexed as 'lolcat', for example
- Some tagging could be preserved as meta-information, **<h1>Cat</h1>** and **<p>Cat</p>** (see later)
- Parsing document types text, html, pdf etc
- Swish-e: http://www.swish-e.org/ has a small scale harvester/parser, for example
- Some problems/opportunities

3: Indexing: A 'plain vanilla' inverted index



Document 1

The bright blue butterfly hangs on the breeze.

Document 2

It's best to forget the great sky and to retire from every wind.

Document 3

Under blue sky, in bright sunlight, one need not search around.

Stopword list

a
and
around
every
for
from
in
is
it
not
on
one
the
to
under

Inverted index

ID	Term	Document
1	best	2
2	blue	1, 3
3	bright	1, 3
4	butterfly	1
5	breeze	1
6	forget	2
7	great	2
8	hangs	1
9	need	3
10	retire	2
11	search	3
12	sky	2, 3
13	wind	2

3:Indexing: 'chocolate' indexing

- Semantic aware/context storage:

For example: **cat** and **<b>cat</b>** and **CAT!** And **<h1>cat</h1>** and **cat, cat, cat** may have different 'values'

Pagerank is the most [in]famous: https://en.wikipedia.org/wiki/PageRank for estimating the 'value' of a resource

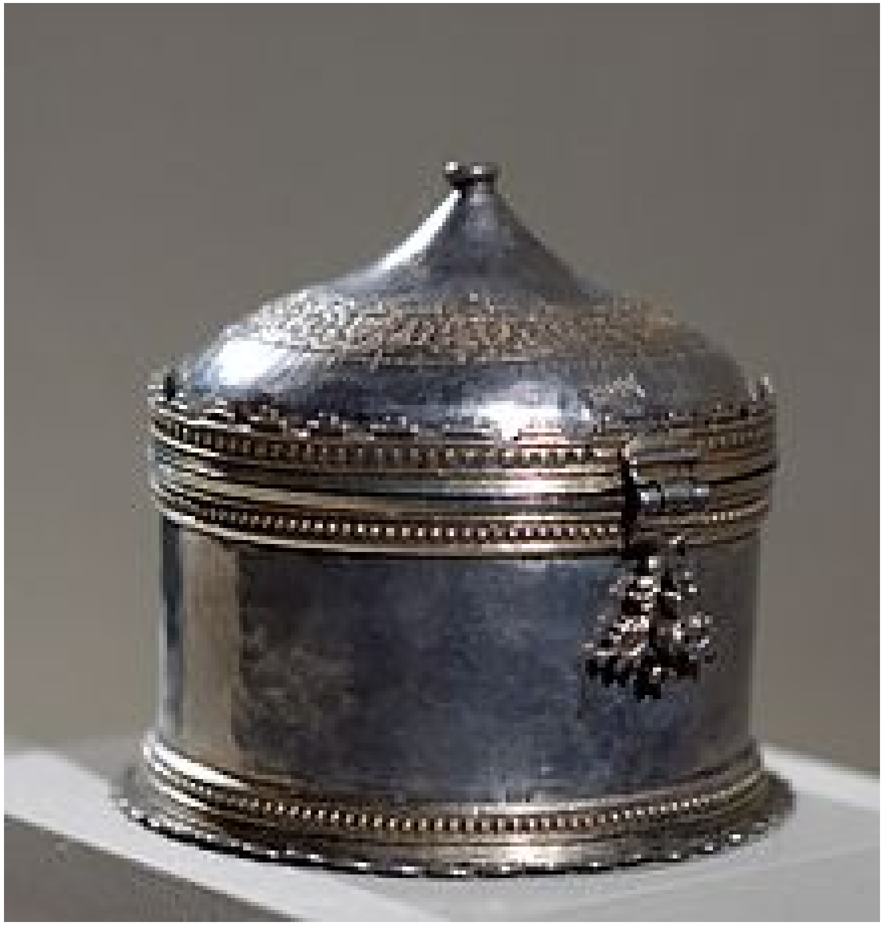
- Spam (**cat, cat, cat**) detection/SEO gaming detection (germ warfare)

- In general: 'algorithms', these are the bits that will probably give competitive advantage...

Storage

- This used to be 'easy', now lots of options
- Sparse data, some entries 'lolcats' have lots of entries, **pyx** [wait for it!] won't have many
- It's a 'lot' of data, google came about by misspelling googolplex:
- Relationals are rather unsuitable, in general
- Sparse, roll-your-own: https://en.wikipedia.org/wiki/Bigtable
- Nosql and ready-mades: http://solr-vs-elasticsearch.com/ for example

This is a (pretty nice) Pyx!



4: Retrieval

- Here you get **results** of all this work
- Simple, one field, one button = Google
- Booleans and implied booleans [lots of works anded together]
- **Relevant** results, this is the main thing and links back to the storage and parsing
- Cookies/user awareness
- Let's look at a few problems with retrieval

Problems with Retrieval

- General relevancy, **the 'Paris Hilton' problem**
- Contextual relevancy, a zookeeper will want pythons, a programmer Python (cookies etc.)
- **Purpose**, buying, researching, looking (remember **you are the product**!)
- Diacritics, non-Roman (problem for both indexing and retrieval, actually)
- Non-textual, images, chemical structures
- Relevant synonyms

Etc.

Technical Difficulties: Examples

- Looking for André
- Looking for 中国 with UK keyboard [for example: zhong1 + guo2]
- Looking for cat [furry] and cat [computer command]
- Speed of index refresh [days, usually]
- Storage and computation, everything is 'big'
- Semantic search vs. search for 'words'

Social Difficulties

- Right to be forgotten, search tracking
- Security services and data mining (queries, history, for example)
- Privacy and doxing, see visual tagging too
- Linking 'unlinked' data, informal 'joins', generally
- Automatic visual tagging [facebook, ugh!]
- Automatic geolocation [most smartphones]
- Any more?

Opportunities

This is speculation, don't take too seriously:

- expansion of domain specific: https://www.shodan.io/

- above example was for IoT, clearly there'll be more

- 'honest' engines, non-profit etc. but how to finance?

- expanded and specialised metasearch

- improved semantics and synonyms

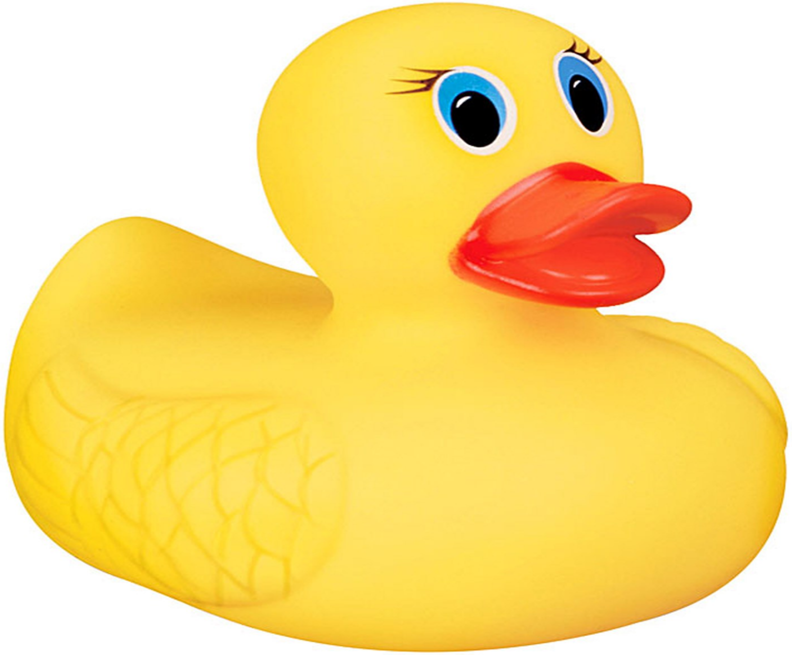
- better 'understanding' (Siri,see: http://sirius.clarity-lab.org/) , translation, thesauri (where I came in, in fact)

Rounding Up

- It's a central human activity
- It's a vital activity for the [intra|inter]tubes (web, IoT, internal applications)
- Very simple central idea(s), but lots of evolution possible
- There's a huge societal debate to go with the technical evolution
- Question and (possibly) some answers

Thanks!

Thanks for listening! As a reward, here is a nice picture of a bathtime duck:



Elastic Search Demo

Taken from:

http://joelabrahamsson.com/elasticsearch-101/

Questions

Also, I'm happy to give another talk this term:

- Perl?

- Raspberry Pi?

- Threats and opportunities in AI?

- Other talks that I am almost certainly unqualified to give...