LTI
Language, Technology and the Internet

Text, Meta-Text and Trust

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

Lecture 8

Text and Meta-text

- Revision of Web As Corpus
- Explicit Meta-data
  - Keywords and Categories
  - Rankings
  - Structural Markup
- Implicit Meta-data
  - Links and Citations
  - Tags
  - Tables
  - File Names
  - Translations

Internet Corpora Summary

- The web can be used as a corpus
  - Direct access
    - Fast and convenient
    - Huge amounts of data
    - ⊗ unreliable counts
  - Web sample
    - Control over the sample
    - Some setup costs (semi-automated)
    - ⊗ Less data
- Richer data than a compiled corpus
- ⊗ Less balanced, less markup

Explicit Metadata

- You can get information from metadata within documents
  - When they are accurate they are very good
  - They are often inaccurate
    - Sometimes deliberately deceitful
    - More often incomplete or out-of-date

Never attribute to malice that which is adequately explained by stupidity.

Hanlon's Razor

You have attributed conditions to villainy that simply result from stupidity

Robert A. Heinlein (1941) *Logic of Empire*

HTML Metadata

- Most document types contain metadata of some description:

```
<head>
  <title>CSLI LinGO Lab</title>
  <META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">
  <meta http-equiv="Content-Style-Type" content="text/css">
  <meta name="keywords" content="linguistic grammars online,
   LinGO, computational linguistics,
   head-driven phrase structure grammar, hpsg, natural language processing,
   parsing, generation, augmentative and alternative communication, aac,
   LinGO Redwoods, multiword expressions, MWE, grammar matrix">
  <meta name="description" content="This page provides information about
   the CSLI Linguistic Grammars Online (LinGO) Lab at Stanford
   University.">
```

- Should we also extract out this data, or is metadata too unreliable to consider using?

PDF Metadata

- Checkout this file (now look at earlier weeks)

```
InfoKey: Creator
InfoValue: xetex(k) 5.98 Copyright 2009 Radical Eye Software
InfoKey: Title
InfoValue: Lecture 11:Text and Meta-Text
InfoKey: Author
InfoValue: Francis Bond
InfoKey: Producer
InfoValue: GPL Ghostscript 8.71
InfoKey: Keywords
InfoValue: Language, Technology, Internet
InfoKey: Subject
InfoValue: HG2052/HG252: Language, Technology and the Internet
InfoKey: ModDate
InfoValue: D:20120315121040+08'00'
InfoKey: CreationDate
InfoValue: D:20120315121040+08'00'
PdfID0: 39bb293fa576c18e1ae64480cb8974
PdfID1: 39bb293fa576c18e1ae64480cb8974
```

```
NumberOfPages: 23
```

HTML Metadata

- HTML Metadata is generally considered unreliable
  - Authors don't see it, so they don't update it
  - As it is unseen, it is easy to lie in the MetaData

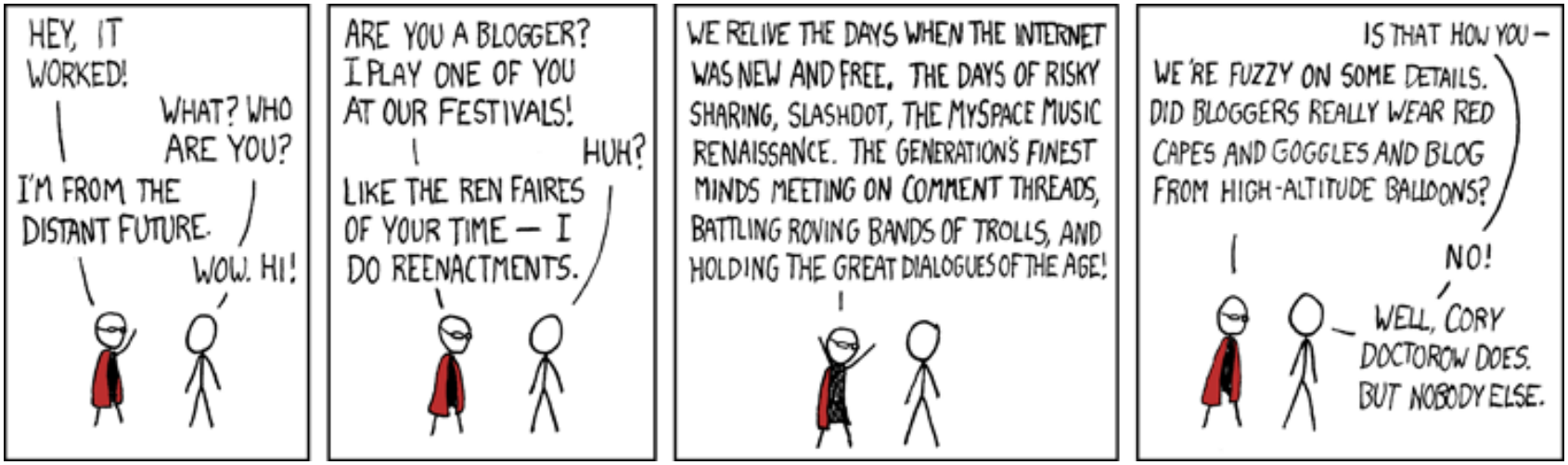
It wasn't long before webmasters with no scruples saw an opportunity to gain favour with the search engines by adding in keywords that did not pertain to the content of their pages. Various tactics were thought up to get ranked higher for certain keywords, and an entire industry sprang up to optimise search engine positioning. This was, in effect, cheating, and "keyword spamming" became a serious problem for search engines, who vainly attempted to add filters that would notice when a webmaster was loading up on the wrong keywords.

Criticism of the Semantic Web

Doctorow's seven insurmountable obstacles to reliable metadata are:

1. People lie
2. People are lazy
3. People are stupid
4. Mission Impossible: know thyself
5. Schemas aren't neutral
6. Metrics influence results
7. There's more than one way to describe something

Cory Doctorow



A Canadian-British blogger, journalist, and science fiction author who serves as co-editor of the blog Boing Boing. He is an activist in favour of liberalising copyright laws and a proponent of the Creative Commons organization, using some of their licences for his books. Some common themes of his work include digital rights management, file sharing, and post-scarcity economics. (Wikipedia)

People lie

Metadata exists in a competitive world. Suppliers compete to sell their goods, cranks compete to convey their crackpot theories (mea culpa), artists compete for audience.

Thus:

- A search for any commonly referenced term at a search-engine like Altavista will often turn up at least one porn link in the first ten results.
- Your mailbox is full of spam with subject lines like "Re: The information you requested."
- Publisher's Clearing House sends out advertisements that holler "You may already be a winner!"
- Press-releases have gargantuan lists of empty buzzwords attached

People are lazy

Here in the Info-Ivory-Tower, we understand the importance of creating and maintaining excellent metadata for our information.

But info-civilians are remarkably cavalier about their information. Your clueless aunt sends you email with no subject line, half the pages on Geocities are called "Please title this page" and your boss stores all of his files on his desktop with helpful titles like "UNTITLED.DOC."

People are stupid

Even when there's a positive benefit to creating good metadata, people steadfastly refuse to exercise care and diligence in their metadata creation.

Take eBay: every seller there has a damned good reason for double-checking their listings for typos and misspellings. Try searching for "plam" on eBay. Right now, that turns up nine typoed listings for "Plam Pilots." Misspelled listings don't show up in correctly-spelled searches and hence garner fewer bids and lower sale-prices. You can almost always get a bargain on a Plam Pilot at eBay.

The fine (and gross) points of literacy – spelling, punctuation, grammar – elude the vast majority of the Internet's users. To believe that J. Random Users will suddenly and en masse learn to spell and punctuate – let alone accurately categorize their information according to whatever hierarchy they're supposed to be using – is self-delusion of the first water.

Mission: Impossible – know thyself

In meta-utopia, everyone engaged in the heady business of describing stuff carefully weighs the stuff in the balance and accurately divines the stuff's properties, noting those results.

Simple observation demonstrates the fallacy of this assumption. When Nielsen used log-books to gather information on the viewing habits of their sample families, the results were heavily skewed to Masterpiece Theater and Sesame Street. Replacing the journals with set-top boxes that reported what the set was actually tuned to showed what the average American family was really watching: light entertainment.

People are lousy observers of their own behaviors. Entire religions are formed with the goal of helping people understand themselves better; therapists rake in billions working for this very end.

Schemas aren't neutral

In a given sub-domain, say, Washing Machines, experts agree on sub-hierarchies, with classes for reliability, energy consumption, color, size, etc.

Nothing could be farther from the truth. Any hierarchy of ideas necessarily implies the importance of some axes over others. A manufacturer of small, environmentally conscious washing machines would draw a hierarchy that looks like this:

```
Energy consumption:
    Water consumption:
        Size:
            Capacity:
                Reliability:
```

While a manufacturer of glitzy, feature-laden washing machines would want something like this:

```
Color:
    Size:
        Programmability:
            Reliability:
```

The conceit that competing interests can come to easy accord on a common vocabulary totally ignores the power of organizing principles in a marketplace.

Metrics influence results

Ranking axes are mutually exclusive: software that scores high for security scores low for convenience, desserts that score high for decadence score low for healthiness. Every player in a metadata standards body wants to emphasize their high-scoring axes and de-emphasize (or, if possible, ignore altogether) their low-scoring axes.

It's wishful thinking to believe that a group of people competing to advance their agendas will be universally pleased with any hierarchy of knowledge. The best that we can hope for is a detente in which everyone is equally miserable.

There's more than one way to describe something

"No, I'm not watching cartoons! It's cultural anthropology."

"This isn't smut, it's art."

"It's not plagiarism, it's borrowing!"

Reasonable people can disagree forever on how to describe something. Arguably, your Self is the collection of associations and descriptors you ascribe to ideas. Requiring everyone to use the same vocabulary to describe their material denudes the cognitive landscape, enforces homogeneity in ideas.

And that's just not right.

So how can we get metadata?

- Look for visible metadata (you can check)
- Look for structural metadata (from the system)
- Look for implicit metadata
  - Finding a new source of metadata opens up a new world of knowledge

Keywords and Categories

- Sites with visible tags are more trustworthy/reliable
- Tags within blogs/photo cites
- Keywords in journals and conferences

Example tags from Science Professor

academia (109)
academic novels (9)
accounting nightmares (10)
administrative assistants (7)
adviser-student (69)
attempt at humor (18)
awards (7)
bizarre (56)
blogging (22)
books (23)
broader impacts (8)
career issues (27)
cats (19)
citations and citation index (19)

Rankings

- Another good source of meta-data is rankings/forums

  HG251
  Q A Solved

- Sentiment Analysis tries to judge whether text is favorable or unfavorable
  - Link text to rankings for data
  - Link posts to tags for usefulness in QA

hungry go where

Overall: 7 Recommend.
I spent about S$10 Per Person

Food/Beverage: 6
Ambience: 5
Value: 9
Service: 5

Cheap but not very cheerful 10 June, 2010

Absolutely love this neighbourhood eatery, mainly because I have been eating here since I was a child, so it brings back many happy memories. Granted, service is kind of lacking but a cheap and yummy home-style meal can always be had. Must-haves for me are the Honey Pork (love the 3 or 4 little green peas they garnish it with), Ayam Buah Keluak, Bakwan Kepeting (meatball soup) and Sayur Lodeh. The Otak and Ngor Hiang are not bad too.

Links and Citations

- Citation frequency can be used to measure the impact of an article.
  - Simplest measure: Each article gets one vote – not very accurate.
- On the web: citation frequency = inlink count
  - A high inlink count does not necessarily mean high quality ...
  - ...mainly because of link spam.
- Better measure: weighted citation frequency or citation rank
  - An article's vote is weighted according to its citation impact.
  - This can be formalized in a well-defined way and calculated. PageRank!

Structure

- Structural Markup gives useful cues
  - Words in headers are often good keywords
  - TableOfContents!

```
1 Review
1.1 Language Identification
1.2 Normalization
2 Text and Meta-text
2.1 Implicit Tags
2.1 Explicit Tags
```

Implicit Metadata

- You can get clues from metadata within documents
  - as they are non-intended, they tend to be noisy
  - but they are rarely deceitful

Tags

- Hypertext Anchors and other formatting gives phrase boundaries

```
whereas McCain is secure on the topic, Obama
<a>[VP worries about winning the pro-Israel vote]</a>

[NP [NP Libyan ruler]
<a>[NP Mu ʻammar al-Qaddafi]</a>] referred to
```

Mainly NPs

- This can be very useful in restricting parser possibilities

Valentin I. Spitkovsky, Daniel Jurafsky, Hiyan Alshawi (2010) *Profiting from Mark-Up: Hyper-Text Annotations for Guided Parsing* ACL

Tables

- You can learn many things from Tables
  - for example, categories

Vehicle	**Price**	**Manufacturer**	**Type**	**Rating**
Raum	XXX	Toyota	Hatch-back	Solid
icw30	XXX	Hyundae	Station Wagon	Exciting
Corolla	XXX	Toyota	Station Wagon	Solid
Camry	XXX	Toyota	Sedan	Bland

Toyota $\subset$ manufacturer

File Names

- How to find definitions?
  - Look for files called `glossary`, `dictionary`, ...
- Is there an English version of this?
  - `http://nlpwww.nict.go.jp/wn-ja/index.ja.html`
  - `http://nlpwww.nict.go.jp/wn-ja/index.en.html`

Translations

- A translation into another language can be seen as markup

Bracketed Glosses

EN:SomeThoughtsConcerningEducation
教AZH: 育漫话

笔者在认真阅读洛克的教育著作《教育漫话》（ＳｏｍｅＴｈｏｕｇｈｔｓＣｏｎｃｅｒｎｉｎｇＥｄｕｃａｔｉｏｎ）、《关于理解的指导》以及《贫穷儿童劳动学校计划》（ＰｌａｎｏｆＷｏｒｋｉｎｇＳｃｈｏｏｌｆｏｒＰｏｏｒ ...
{25094:corpus0.txt}

EN: GPS
ZH: 通用 回解诀者

２８附录：（通用 回解诀者》（ＧＰＳ）计算机程序解决・・河内塔。
{2069:corpus0.txt}

Extracted using regular expressions from the Chinese Gigaword Corpus.

Cross-lingual Disambiguation

(1) *I*$_1$ *saw*$_2$ *the kid*$_3$ *with a telescope*$_4$

(2) ϕ_1 望遠鏡 $_4$ で 子供 $_1$ を 見た $_2$
bouenkyou de kodomo wo mita
NULL telescope with child ACC see-past
With the telescope, I saw the kid.

- We can disambiguate the PP attachment: ***de*** only modifies verbs
- We can disambiguate the verb ***see/saw***: ***mita*** is only "see"
- We can resolve the zero pronoun: It must be the speaker.

Query Data as Meta-data

AOL user 2708:

- revenge tactics
- the woman's book of revenge
- dirty tricks for chicks
- ...
- locatecell.com
- what can i do to an old lover for revenge
- mean revenge tactics
- death records in hampstead

Wikipedia Redirections

- Alternative names (*Edison Arantes do Nascimento* → *Pelé*).
- Abbreviations (*DSM-IV* → *Diagnostic and Statistical Manual of Mental Disorders*).
- Alternative spellings or punctuation. (*Colour* → *Color*; *Al-Jazeera* → *Al Jazeera*).
- Likely misspellings (*Condoleeza Rice* → *Condoleezza Rice*).
- Plurals (*Greenhouse gases* → *Greenhouse gas*).
- Related words (*Symbiont* → *Symbiosis*).
- Representations using ASCII characters (*Kurt Goedel* and *Kurt Godel* → *Kurt Gödel*).

Unfortunately redirects are rarely typed (so we don't know the relation, but have to infer it).

type: identify as belonging to a certain type — *Such people can practically be typed* PWN3.0

Cross Wikipedia Links

en Forensic linguistics

ca Lingüística forense

cs Forenzní lingvistika

de Forensische Linguistik

es Lingüística forense

nl Forensische taalkunde

no Forensisk lingvistikk

tr Adlî dil bilimi

zh 司法语言学

Why is metadata important?

- The 1990s started a revolution in empirical linguistics
  - New insights come from Data Mining large text collections
    - Corpus Linguistics
    - You can do with a computer what you can't do with paper
  - New tools come from supervised Machine Learning
- Annotation is expensive and tedious to do
- We want to get annotation for free
- People appreciate clever ideas

What about bad actors?

- Some people are deliberately trying to deceive you
- To steal from (scam)
- To persuade you (fake news)
- To trick you (troll)

Deliberate Deceit: Phishing

- Phishing is a way of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication.
- Communications typically pretend to be from social web sites, auction sites, online payment processors or IT administrators.
- Phishing is typically carried out by e-mail spoofing, linking users to a fake website whose look and feel are almost identical to the legitimate one.
- Phishing is an example of social engineering.
- A phishing technique was described in detail in 1987, and (according to its creator) the first recorded use of the term *phishing* was made in 1995.

From my own spam box

System Administrator s28407548@tuks.co.za via srs.ieee.org
to undisclosed recipients

You have exceeded the storage limit on your mailbox.

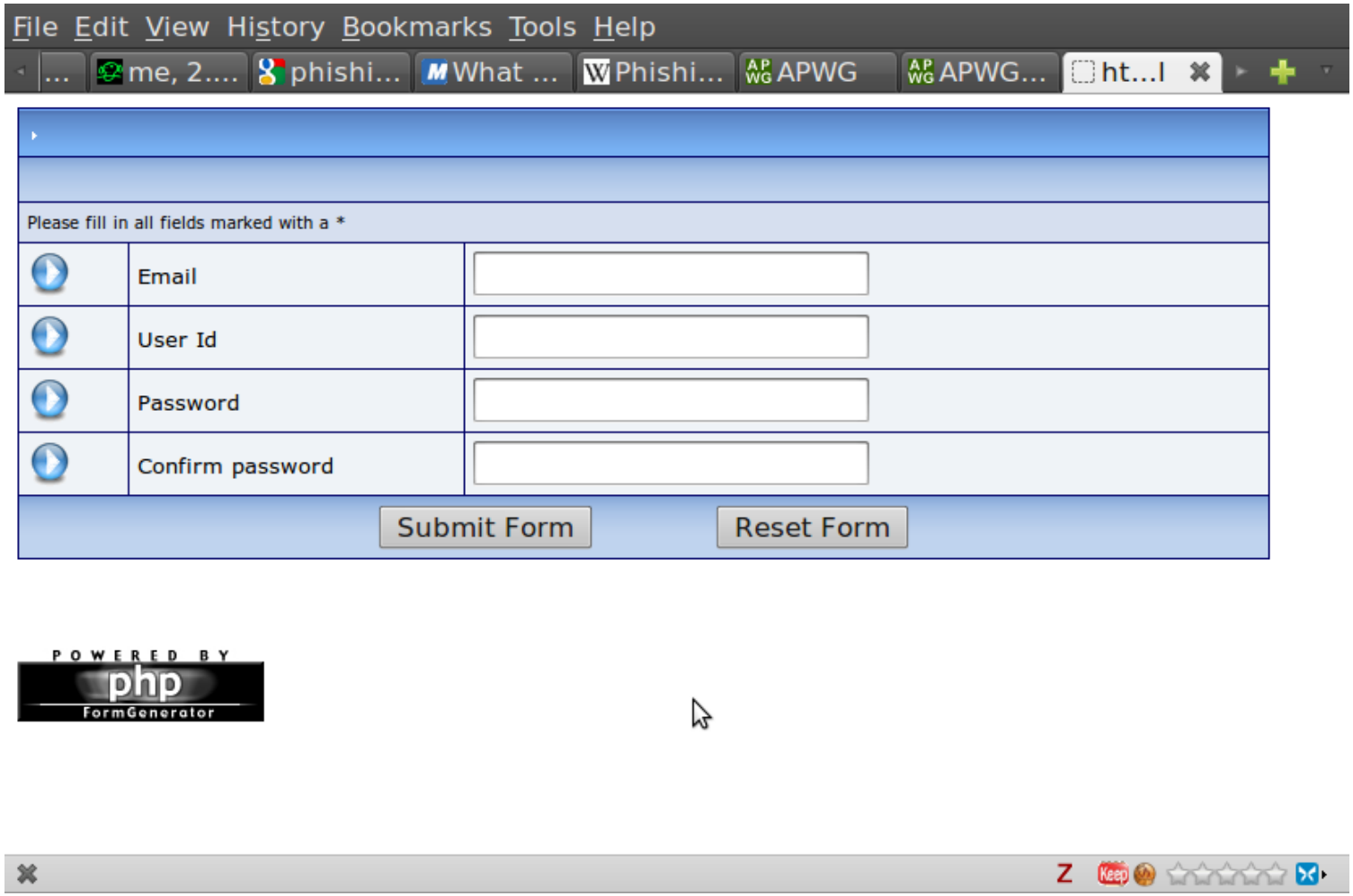
You will not be able to send or receive new mail until
you upgrade your email quota.

Click the below link and fill the form to upgrade your
account.

http://millerofficetrailers.com/forms/use/hepldesk/form1.html

System Administrator
192.168.0.1

The fake form



Some Distinguishing Features

- Surprisingly many grammatical mistakes
- Spoofed URLs
- Ultimatums
- Weird misspellings: `NTU.edu.org`
- Link to a strange web site

- Most phishing attempts fail
- But they are cheap to construct, so even one in a million response rates are enough

Fake News

- Fake news is false news
  - Malicously False News
  - Satire
  - Disinformation
  - Misinformation
  - Rumour
- Fake news is intentionally and verifiably false news published by a news outlet

Analyzing Language in Fake News

- Verifying facts is difficult, but we can analyze language (Rashkin et al., 2017)

Phenomenon	Ratio	Example	Type
Swear	7.00	Ms. Rand, who has been damned to eternal torment ...	S
2p (You)	6.73	You would instinctively justify ...	P
Modal Adv	2.63	... investigation of Clinton was inevitably linked ...	S
Negation	1.51	There is nothing that outrages liberals more than ...	H
Superlatives	1.17	Fresh water is the single most important natural resource	P
Comparitives	0.86	... from fossil fuels to greener sources of energy	P
Hear	0.50	The prime minister also spoke about the commission ...	S
Number	0.43	... 7 million foreign tourists coming to the country in 2010	S

  - Satire, Hoax, Propoganda vs real news
  - Ratio of appearance in fake to real news

- We can identify things likely to be fake news just from the language

- But the single most useful piece of evidence (feature) is the source — where it comes from

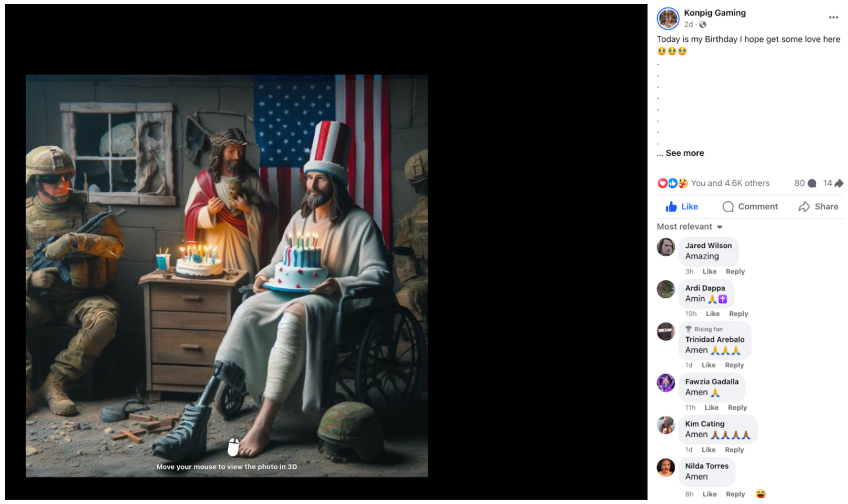
A new problem: AI slop

- **slop** is the new name for unwanted AI-generated content (like **spam** for email)
- Mindlessly generated content presented as real content
- We used to have content farms, where poorly paid people create fake accounts, web pages and news
- Now AI can do it faster and cheaper!

https://www.theguardian.com/technology/article/2024/may/19/spam-junk-slop-the-latest-wave-of-ai-behind-the-zombie-internet

Facebook Is the 'Zombie Internet'

- Jason Koebler looks at AI generated pages on facebook, which are then often full of AI bots commenting on them, ...
- For example:



  - Amen
  - Amin
  - Happy birthday
  - Thank you Lord Amen happy birthday

- But sometimes pictures fool people, and real people waste time discussing them

FungiFriend

- An AI chatbot called "FungiFriend" joined mushroom identification Facebook group
- a user asked "how do you cook Sarcosphaera coronaria," a type of mushroom that was once thought edible but is now known to hyperaccumulate arsenic and has caused a documented death
- FungiFriend told the member that it is "edible but rare," and said "cooking methods mentioned by some enthusiasts include sautéing in butter, adding to soups or stews, and pickling."
- This was a bot created and run by META, Facebook's parent company!

You have to be careful all the time!

https://www.404media.co/ai-chatbot-added-to-mushroom-foraging-facebook-group-immediately-gives-tips-for-cooking-dangerous-mu

What to Do?

- be vigilant
- look at metadata
- look at implicit metadata
- be ethical
  - don't create slop
  - don't exploit slop
- push for others to do the same
- look for human-centric media
  - Blue Sky
  - Blogs
  - Substack with authors you can trust
  - Newspapers!

References

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.