Semi-Manual Annotation of Topics and Genres in Web Corpora, The Cheap and Fast Way

Including the Difference between Pearson's Chi-Square Test and Pyongyang's Nuclear Test

Vít Suchomel, Jan Kraus

Natural Language Processing Centre
Masaryk University, Brno, Czech Republic
https://nlp.fi.muni.cz/en/

Lexical Computing, Brno, Czech Republic
https://www.lexicalcomputing.com/
name.surname@sketchengine.eu

RASLAN
2022-12-10

Corpus Use

- Generally: data for studying natural language
- Linguists: analyses of language phenomena, language changes over time
- Lexicographers, teachers: dictionaries, word meanings, examples of a typical use
- Sociologists: style and theme, hot topics
- Marketing experts: brands/product evaluation, sentimet analysis
- Statistical NLP: language models for taggers, analysers, translation systems, predictive writing,...

Understanding the Content of Web Corpora

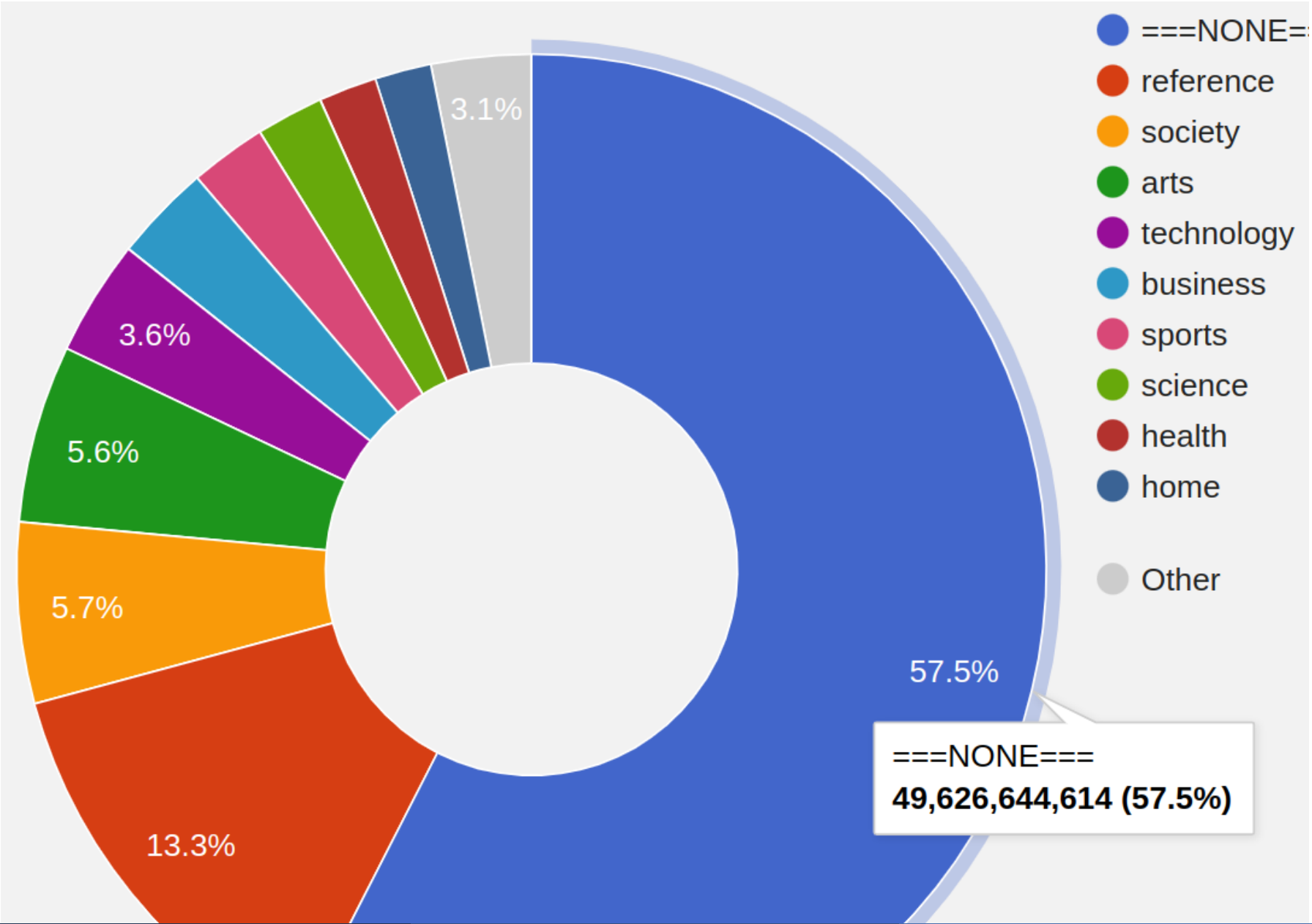
- Corpora from books, newspapers, magazines,. . . : rich metadata
- Web corpora?

Understanding the Content of Web Corpora

- Corpora from books, newspapers, magazines,. . . : rich metadata
- Web corpora?

1. author
2. date of publishing
3. language variety
4. **text genre**
5. **text topic**

Tokens by Topic in enTenTen20



The Impossible Goals

- Cover a large part of a web corpus
- with only a small human effort
- $\Rightarrow$ spend human time efficiently

Website Homogeneity Assumption

- Assign the same topic/genre label for all pages from a website
- Holds for topics in 92 % of cases [Papčo 2022]

The rest of the corpus

- Train a classifier – every page is a separate instance
- and label the rest of the corpus

Topic and Genre Annotation of Whole Websites

1. Rank websites by token count in the corpus
2. Select top N sites
   - English: 3,000 $\Rightarrow$ 40 % of corpus tokens
   - other languages: 300 – 1,500 $\Rightarrow \geq$ 60 % and even up to 90 % of corpus tokens
3. Split site documents by frequent path prefixes, e.g. `/sports/`, `/culture/`
4. Spend time checking the website content proportionally to its rank
5. Generate table (a website per row) to record annotations
   - hostname (e.g. `bbc.com`)
   - link to the site landing page
   - link to concordance of random sentences from the site in Sketch Engine
6. Check site quality, topic, genre – all at the same time

Website Checks – Quality and Text Types Together

1. Hostname (e.g. `bbc.com`)
   - quality check: long phrases, language code, generic/foreign TLD are suspicious
2. live site checks in a browser
   - non-text, low quality text
   - hijacked/unrelated content
   - selectors with too many language mutations (high chance of MT)
   - MT scripts in the source code
   - a dead site (a high quality content does not get shut down often)
3. link to 100 random triples of consecutive sentences in context in Sketch Engine
   - 3 to 10 sentence triples are inspected
   - the rest is briefly seen and consulted more in the case of a doubtful content or suspicious site
   - each chunk of text can be tracked to the original web page
4. topic and genre
   - lexical or syntactic features typical for a recognized text type
   - unsure or multiple classes – don't label

Topics recongized in enTenTen21 (1/2)

Topic	Websites	Tokens
arts	12	169 655 242
beauty & women	6	45 899 006
cars & bikes	49	268 201 168
construction & real estate	1	4 610 212
culture & entertainment	123	695 609 769
economy, finance & business	62	387 271 125
education	15	79 155 574
food & drinks	2	9 774 572
gambling & casinos	1	7 839 308
games	52	324 004 431
health	59	426 786 724
history	24	176 510 675
hobbies	18	111 828 110

Topics recongized in enTenTen21 (2/2)

Topic	Websites	Tokens
home, family & children	7	47 126 547
lifestyle	0	0
nature & environment	6	64 495 602
pets & animals	9	33 432 198
politics & government	27	243 239 797
reference/encyclopedias	10	4 210 237 110
religion	71	424 919 420
science	51	594 461 579
sex	10	209 398 259
sports	103	647 268 352
technology & IT	138	887 566 212
travel & tourism	31	162 020 069
Total	**887**	**10 231 311 061**

Genres recongized in enTenTen21

Topic	Websites	Tokens
blog	99	748 208 188
discussion	194	1 327 118 539
fiction	55	1 009 319 746
legal	37	507 984 084
news	226	1 284 058 175
Total	**611**	**4 876 688 732**

Management of Annotation Resources

- Time spent with each website is proportional to the contribution of the site to the corpus
- several minutes to as less as 20 seconds with each item to inspect
- not assigning any labels is encouraged to reduce noise
- no expert linguistic or computer skills required
- no expert language skill required – live site and MT (GT/DeepL) of sentences is enough

Pearson's Chi-Square Test vs. Pyongyang's Nuclear Test