Web corpus construction

USING DATA FROM THE COMMON CRAWL
AGENDA

- Corpora - The web as corpora
- Inspecting the Common Crawl: Getting my feet WET
- How does one construct a web corpus?
- Presenting the English Common Crawl Corpus (ENC³)
From "traditional" corpora to web corpora

WHAT'S THE POINT?
Traditional corpora

- Typically manually constructed
- Sourced from newspapers or other published text (or transcribed spoken language)
- Used in many machine learning-based NLP areas, like machine translation, language modeling, distributional semantics etc.
- Two challenges: relatively small size and homogenous origin*

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown corpus (1960s)</td>
<td>1 million words</td>
</tr>
<tr>
<td>COBUILD (1980s)</td>
<td>8 million words</td>
</tr>
<tr>
<td>British National Corpus (1995)</td>
<td>100 million words</td>
</tr>
</tbody>
</table>
"We propose that a logical next step for the research community would be to direct efforts towards increasing the size of training collections, while deemphasizing the focus on comparing different learning techniques trained only on small training corpora."

From "Scaling to Very Very Large Corpora for Natural Language Disambiguation" (Banko and Brill, 2001)
More data is needed: The Web!

• Early efforts: Using search engine hits as counts for tokens or bigrams
• Today: Crawling web documents and extracting its text
• For my thesis, I do not crawl myself
• Rather...
Common Crawl

• Non-profit organisation
• Provides monthly crawls of the web
• Free data, through Amazon’s Public Dataset program
• Accessible through Amazon’s S3 protocol or as direct downloads
• Possible because of the US’ fair use laws
The data

- My project is based on **one** monthly crawl - the August 2016 crawl
- > 1.6 billion documents, 30TB of data

- A crawl is delivered in three formats
  - WARC - Web Archive format, the raw crawl data
  - WET - Web Extracted Text, extracted text data
  - WAT - Web Archive Transformation, computed metadata
## The data

<table>
<thead>
<tr>
<th></th>
<th>WARC</th>
<th>WET</th>
<th>WAT</th>
</tr>
</thead>
<tbody>
<tr>
<td># Files</td>
<td>29800</td>
<td>29800</td>
<td>29800</td>
</tr>
<tr>
<td>Avg. file size (compressed)</td>
<td>4515 MB (988)</td>
<td>405 MB (156)</td>
<td>1524 MB (353)</td>
</tr>
<tr>
<td># Entries per file</td>
<td>156000</td>
<td>52000</td>
<td>156000</td>
</tr>
<tr>
<td># Documents per file</td>
<td>52000</td>
<td>52000</td>
<td>52000</td>
</tr>
<tr>
<td>Avg. document size</td>
<td>122 MB</td>
<td>5 MB</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

1The WARC and WAT files have entries for the HTTP request, the response and the WARC headers. The WET files only contain the responses.
A note on scale

- 30 TB of data is a lot of data
- Each individual file is also large
- Every operation can take an excruciating amount of time, even simple ones
- Opening all the files, reading them, and closing them, with no additional operations takes a total of about 17.5 hours
- Any operation more advanced than opening the files increase this even more
- Parallel computing a necessity
- Interactive shell operations like moving files around or testing code edits much more troublesome

<table>
<thead>
<tr>
<th>Operation</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>17.5 hours</td>
</tr>
<tr>
<td>Decompressing</td>
<td>11.3 days</td>
</tr>
<tr>
<td>Downloading</td>
<td>14.5 days</td>
</tr>
<tr>
<td>Downloading with 12 threads</td>
<td>1.7 days</td>
</tr>
</tbody>
</table>
Getting my feet WET

• The WET files already contain extracted text
• If they are of sufficient quality, the task ahead is a lot easier

• Are they of sufficient quality?
### Remaining HTML tags

- Tags like `<div>`, `<a>`, `<html>` etc. still remain in the WET files.
- A test corpus made with the techniques I describe later perform a lot better.
- WET ∩ Corpus are WET documents that is processed by and included in the test corpus.

### Document counts:

<table>
<thead>
<tr>
<th></th>
<th>WET: 255 000</th>
<th>Test corpus: 90 500</th>
<th>Intersection: 90 500</th>
</tr>
</thead>
</table>
Stay out of the water

<table>
<thead>
<tr>
<th></th>
<th>WET files</th>
<th>Test corpus</th>
<th>WET ∩ Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td># entities</td>
<td>166176</td>
<td>8772</td>
<td>121349</td>
</tr>
<tr>
<td># docs w/ entities</td>
<td>5305</td>
<td>2284</td>
<td>2785</td>
</tr>
<tr>
<td>% docs w/ entities</td>
<td>2.1%</td>
<td>2.5%</td>
<td>3.1%</td>
</tr>
<tr>
<td># entities/doc</td>
<td>0.68</td>
<td>0.09</td>
<td>1.34</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>55.2</td>
<td>1.7</td>
<td>66.4</td>
</tr>
</tbody>
</table>

Remaining HTML entities

- HTML entities, like &gt; or &quot; are not cleaned sufficiently from WET either
- The test corpus performs a lot better
- But maybe not as well as we would have liked

- Document counts:
  - WET: 255 000
  - Test corpus: 90 500
  - Intersection: 90 500
Stay out of the water

<table>
<thead>
<tr>
<th>Language</th>
<th>WET files</th>
<th>Test corpus</th>
<th>WET ∩ Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>81.4%</td>
<td>75.3%</td>
<td>78.6%</td>
</tr>
<tr>
<td>Chinese</td>
<td>1.5%</td>
<td>3.3%</td>
<td>2.0%</td>
</tr>
<tr>
<td>German</td>
<td>2.0%</td>
<td>2.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.9%</td>
<td>2.5%</td>
<td>2.1%</td>
</tr>
<tr>
<td>Norwegian</td>
<td>0.12%</td>
<td>0.08%</td>
<td>0.08%</td>
</tr>
</tbody>
</table>

Language distribution in the different corpora

- The same language identifier was used for all three corpora
- Note the differences between the corpus and the intersection
If Jessica Hart was a dog she would be this thing.

Snooki and JWOWW dress to match their dogs.

We can’t really tell which is Ozzy! Incredible!

Coco and her pitbull. Identical!

Luke Wilson looks a lot like his stocky pup.

This is Amanda Byne’s dog... Blondes! Seann William Scott would most definitely have a yellow lab.

Brunettes! Anne Hathaway with her chocolate Labrador, Esmeralda.

Jon Hamm and his mutt of ten make similar facial expressions.
Text extraction from the Common Crawl
Constructing corpora from web crawls

• Turning web crawls into corpora have been worked on since about 2005
• Began with the "Web as Corpus kool ynitiative" (WaCky)
• Defined sub-tasks that are still the main challenges today

• In my project, I selected a tool chain called texrex
The sub-tasks

- **Crawling**
  - Selecting seed URLs
  - Reducing host bias
    - Crawling with a random walk
    - Maximizing the yield rate

- **Cleaning the data**
  - (Removing HTML tags)
  - Detecting connected text
  - Language identification
  - Boilerplate removal
  - Duplicate and near-duplicate removal
  - Handling encodings

- **Annotation/Post-processing**
  - Tokenisation, lemmatization, part-of-speech-tagging
  - Automatic metadata classification
Cleaning the data: The problem
Cleaning the data: The problem

Aleksej Navalnyj – ubehagelig urokråke eller reell utfordrer til Putin?

Den russiske opposisjonspolitikeren Aleksej Navalnyj må tilbringe de neste to ukene i fengsel. Det kommer neppe til å stoppe hans kritikk av dem som styrer i Russland. Spørsmålet er om han kan nå ut til bredere lag av det russiske folk, og bli en reell trussel for president Vladimir Putin.
Cleaning the data: The problem

Aleksej Navalnyj var nok kanskje selv overrasket over det som skjedde i Moskva søndag 26. mars.

Til tross for at det var en ulovlig demonstrasjon og myndighetene hadde gitt klar beskjed om at de ville slå hardt ned på demonstrantene, så tresset mer enn 10000 mennesker frykten og samlet seg på Tverskajagaten i sentrum av Moskva.

Demonstrasjonen var en slags foreløpig kulminasjon på en kampanje som Aleksej Navalnyj og hans anti-korrupsjonsorganisasjon har hatt gående, rettet direkte mot den russiske statsministren og tidligere presidenten Dmitrij Medvedev.
Cleaning the data: The problem

Ulovlige piratsendinger på FM: Høres ut som noe som foregikk på 80-tallet

Ektepar i Singapore dømt for å ha sultet hushjelpen
(HTML tag-stripping)

- Remove the HTML tags, <a>, <html>, <div>
- Convert the entities, &gt; -&gt; >
- Second pass of removal (&lt; br &gt; becomes <br>)
- Insert paragraph breaks where paragraphy tags occur (div, p, article, etc.)
Connected text identification

CONNECTED TEXT

Hey everyone,
I am a math major currently taking a PhD in Natural Language Processing. My boyfriend is a history major but he wants to know more and understand what I do for a living, but while I try my best to explain to him the basics of Machine Learning/NLP it is very difficult for me to explain their inner workings in layman’s terms. It doesn’t help the fact that we are living in two different countries and our communication is done mostly through Skype, which is not the best way to explain more theoretical stuff.

I have looked for very basic introductory books in NLP but everything I find seems to be way too mathematical. Are there any books out there that explain the subject from a more “popular science” kind of perspective, i.e., directed at people who have no math background?

NOT CONNECTED TEXT

MODERATORS REMOVE

<table>
<thead>
<tr>
<th>(rule list)</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Personal information</td>
</tr>
<tr>
<td>▪ Excessive trolling</td>
</tr>
<tr>
<td>▪ Direct threats</td>
</tr>
<tr>
<td>▪ Blatant spam</td>
</tr>
<tr>
<td>▪ Deceptive links to shock sites, malware, etc</td>
</tr>
<tr>
<td>▪ Submissions irrelevant to StarCraft</td>
</tr>
<tr>
<td>▪ Submissions with vague or no context</td>
</tr>
<tr>
<td>▪ Uncorroborated accusations</td>
</tr>
<tr>
<td>▪ Promotional submissions that exceed “2 per 1 per 1”</td>
</tr>
<tr>
<td>▪ Duplicate results posts for the same individual match will be removed.</td>
</tr>
<tr>
<td>▪ Same- and similar-topic submissions that exceed 4 per top 25</td>
</tr>
</tbody>
</table>
Connected text identification

What about this one?

RULES

1. No content primarily for humor or entertainment
2. No personal attacks, witch hunts, or inflammatory language
3. No off-topic or low-effort content or comments
4. No duplicate posts
5. No enabling or linking to piracy
6. Follow all formatting requirements
7. Follow all specific content restrictions
8. Promotion must be kept within acceptable limits
Connected text identification

• Text is not considered connected or not connected
• Rather, its lack of "connectedness" contributes to a document's "Badness"
• "Badness" is calculated based on a lack of function words
  • Function words example: the, of, and, for etc.
  • A lack of these (which is the most used words in a language) point to the text being non-connected
  • If a document has too much non-connected text, it will be removed.
• If not, the document stays, and so does its non-connected text
• The next task is better suited for removing non-wanted content like this

• Bonus: As these function words are language-specific, this text quality assessment will consider sentences in other languages as "bad"
  • texrex use this assessment as their sole language identifier
Language identification

- Solved by the previous task
- Still important, as TLDs and meta tags in HTTP headers not reliable
Boilerplate removal

• The least trivial task
• Seeks to remove redundant content that is automatically inserted by a web page
  • Navigational elements
  • Buttons
  • Copyright notices
• Definition of boilerplate: "All that remains after markup stripping, and which does not belong to one of those blocks of content on the web page that contain coherent text"
  (Roland Schäfer, Accurate and efficient general-purpose boilerplate detection for crawled web corpora, 2016)
• Example:
Kunsten å lese baklengs

BEIJING (NRK): Hva er falske og hva er ekte nyheter i et land der alle medier er kontrollert av kommunistpartiet?

Russisk bank bekrefter kontakt med Trumps svigersønn

Senatorets etterretningskomité «forventer» at Jared Kushner vil kunne gi svar på «sentrale spørsmål» i forbindelse med etterforskningen av Russlands eventuelle inblanding i det amerikanske presidentvalget.

Lik funnet i koffert ved havna i Rimini

En koffert med liket av en asiatisk kvinne er funnet ved havna i den italienske turistbyen Rimini. Politiets teori er at kvinner er drept av sin tyske ektemann på cruisetogene.
But also the obvious ones:
Boilerplate removal

• A multilayered perceptron classifies boilerplate
• Trained on manually classified paragraphs (15 000 paragraphs used for English)
• 37 features used
• Language-dependent*
• Results in an evaluated F1-score of:
  • 0.99 for English
  • 0.977 for German
  • 0.983 for Swedish
  • 0.994 for French
• (If anyone wants to do this for Norwegian: the tool chain supports training mode, which extracts all the paragraphs with features for you)

Example features:
• Length of paragraph
• Proportion of HTML markup to all text in the non-stripped document
• Number of sentences
• Does the paragraph end with punctuation?
• Average sentence length
• Number of sentences ended in punctuation
• The proportion of HTML markup to text in the neighboring paragraphs
• The proportion of the number of paragraph characters to the whole page
Duplicate- and near-duplicate removal

• Perfect duplicate removal:
  • Populate an array of, 64/128 characters, which are evenly distributed across the document
  • If two documents match, they are duplicates

• Near-duplicate removal
  • Retrieve the token n-grams of the documents (n typically equals 5)
  • Hash them - these are now called shingles
  • Calculate the number of shingles shared between each document
  • If the number exceed a controllable threshold; remove the shortest
Duplicate- and near-duplicate removal

**DOCUMENT A**

"a rose is a rose is a rose"

4-gram:

{"a rose is a", "rose is a rose", "is a rose is"}

Overlap of 3 shingles

**DOCUMENT B**

"a tulip is a rose is a rose"

4-gram:

{"a tulip is a", "tulip is a rose", "is a rose is",
"a rose is a", "rose is a rose"}

Overlap of 3 shingles
**Encodings**

- The web uses a lot of different encodings
- They need to be normalized
- Challenge: The announced encoding in the HTTP headers might be wrong
- Or: Several encodings may be used on the same web page
- IBM’s International Components for Unicode can deal with it
Almost all of the subtasks mentioned take place in the tool called texrex
Tender handles sorting and counting shingles to identify near-duplicated documents
Tecl performs the destructive removal of the duplicates
Running it on Abel

• I won’t bother you with the technical details
• But: Running the tools on abel is a lot of tinkering
  • Splitting up jobs
  • Trial and error to see what works, what is efficient etc.
  • Rerunning failed jobs
**Job setup – a breakdown**

<table>
<thead>
<tr>
<th></th>
<th>texrex</th>
<th>tender</th>
<th>tecl</th>
<th>CoreNLP tokenisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of jobs</td>
<td>10000</td>
<td>995</td>
<td>498</td>
<td>49</td>
</tr>
<tr>
<td>Threads / job</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Allocated memory / job</td>
<td>12 GB</td>
<td>16 GB</td>
<td>12 GB</td>
<td>24</td>
</tr>
<tr>
<td>Hours / job</td>
<td>1 - 4</td>
<td>0.5 - 1 (final one: 72)</td>
<td>3</td>
<td>1 - 3</td>
</tr>
<tr>
<td>Hours total</td>
<td>200</td>
<td>100</td>
<td>48</td>
<td>24</td>
</tr>
</tbody>
</table>

Effective run time to produce corpus from start to end: 16 days
ENC$^3$

- 86 million documents
- 6 billion sentences
- 135 billion tokens
- Delivered in three formats: XML, text, ConLL
ENC³ XML

• The primary source for the corpus
• Contains all the paragraphs considered boilerplate
• Instead classified with a number from 0-1 so user can select the threshold
• Other meta data also kept
• Because it's XML, some characters must be escaped
The Allens knew from the beginning that they wanted a T-top coupe, and spent two years searching before finally coming across one in Oxford White for sale in 2007. It had been posted online by a man in Connecticut, who apparently wanted the car for himself, but upon further inspection, realized he had bitten off more than he could chew.

"It was actually in a salvage yard, going to be destroyed," explains Julie Allen. "Luckily, this young guy saw it and knew what it was." That young guy sent the Allens some pictures and information, and they instantly made up their minds. When they arrived to see the car the following day, they realized they had their work cut out for them. Julie described it in one word—gone. With most of the car plagued by rust and decay, the Mustang’s only salvageable parts were a quarter-panel and the T-top roof.
ENC³ ConLL

- Tokenized with CoreNLP
- Delivered in popular ConLL format
- Hopefully I get time to PoS-tag and lemmatise as well
Evaluation with word embeddings

- I wish to evaluate the corpus in a downstream task
- Making dense word vectors is my chosen task
  - Current
  - Other intriguing results
- Will be using GloVe
- Evaluation will be done with the "typical" evaluation techniques
  - Google analogy: “Stockholm is to Sweden as Oslo is to _____?” or “Run is to running as dance is to _____?”
  - Word similarity: WordSim353, Simlex-999, etc.
- Currently running the co-occurrence counter
  - Processed 110 bln of 135 bln now

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ivLBD</td>
<td>100</td>
<td>1.5B</td>
<td>55.9</td>
<td>50.1</td>
<td>53.2</td>
</tr>
<tr>
<td>HPCA</td>
<td>100</td>
<td>1.6B</td>
<td>4.2</td>
<td>16.4</td>
<td>10.8</td>
</tr>
<tr>
<td>GloVe</td>
<td>100</td>
<td>1.6B</td>
<td>67.5</td>
<td>54.3</td>
<td>60.3</td>
</tr>
<tr>
<td>SG</td>
<td>300</td>
<td>1B</td>
<td>61.0</td>
<td>61.0</td>
<td>61.0</td>
</tr>
<tr>
<td>CBOW</td>
<td>300</td>
<td>1.6B</td>
<td>16.1</td>
<td>52.6</td>
<td>36.1</td>
</tr>
<tr>
<td>vLBD</td>
<td>300</td>
<td>1.5B</td>
<td>54.2</td>
<td>64.8</td>
<td>60.0</td>
</tr>
<tr>
<td>ivLBD</td>
<td>300</td>
<td>1.5B</td>
<td>65.2</td>
<td>63.0</td>
<td>64.0</td>
</tr>
<tr>
<td>GloVe</td>
<td>300</td>
<td>1.6B</td>
<td>80.8</td>
<td>61.5</td>
<td>70.3</td>
</tr>
<tr>
<td>SVD</td>
<td>300</td>
<td>6B</td>
<td>6.3</td>
<td>8.1</td>
<td>7.3</td>
</tr>
<tr>
<td>SVD-S</td>
<td>300</td>
<td>6B</td>
<td>36.7</td>
<td>46.6</td>
<td>42.1</td>
</tr>
<tr>
<td>SVD-L</td>
<td>300</td>
<td>6B</td>
<td>56.6</td>
<td>63.0</td>
<td>60.1</td>
</tr>
<tr>
<td>CBOW†</td>
<td>300</td>
<td>6B</td>
<td>63.6</td>
<td>67.4</td>
<td>65.7</td>
</tr>
<tr>
<td>SG†</td>
<td>300</td>
<td>6B</td>
<td>73.0</td>
<td>66.0</td>
<td>69.1</td>
</tr>
<tr>
<td>GloVe†</td>
<td>300</td>
<td>6B</td>
<td>77.4</td>
<td>67.0</td>
<td>71.7</td>
</tr>
<tr>
<td>CBOW</td>
<td>1000</td>
<td>6B</td>
<td>57.3</td>
<td>68.9</td>
<td>63.7</td>
</tr>
<tr>
<td>SG</td>
<td>1000</td>
<td>6B</td>
<td>66.1</td>
<td>65.1</td>
<td>65.6</td>
</tr>
<tr>
<td>SVD-L</td>
<td>300</td>
<td>42B</td>
<td>38.4</td>
<td>58.2</td>
<td>49.2</td>
</tr>
<tr>
<td>GloVe</td>
<td>300</td>
<td>42B</td>
<td>81.9</td>
<td>69.3</td>
<td>75.0</td>
</tr>
</tbody>
</table>

Evaluation results for GloVe word vectors
Outlook

- **Non-destructive subtasks** makes it possible to test a much larger, but more boilerplaty corpus
  - And testing the significance of the different subtasks with respect to downstream tasks
- **Web corpus for Norwegian?**
  - Common Crawl has since I started my project released more international data
  - Need to train the neural network with Norwegian data: manual annotation needed
  - Trivial: Need to construct a language profile for the text quality assessment
- **Testing the significance of corpus size more thoroughly**
  - With this corpus and this technique, evaluating different corpus sizes with respect to different types of downstream tasks is possible