Web corpus construction

USING DATA FROM THE COMMON CRAWL

Kjetil Bugge Kristoffersen

Master's thesis, Language technology

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 (ENC^3)

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*

Corpus	Size
Brown corpus (1960s)	1 million words
COBUILD (1980s)	8 million words
British National Corpus (1995)	100 million words

"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

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

	WARC	WET	WAT
# Files	29800	29800	29800
Avg. file size (compressed)	4515 MB (988)	405 MB (156)	1524 MB (353)
# Entries per file[1]	156000	52000	156000
# Documents per file	52000	52000	52000
Avg. document size	122 MB	5 MB	Not applicable

[1] The 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

Reading	17.5 hours
Decompressing	11.3 days
Downloading	14.5 days
Downloading with 12 threads	1.7 days

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?

Stay out of the water

	WET files	Test corpus	WET ∩ Corpus
# tags	37173	28	15114
# docs w/ tags	3055	28	1139
% docs w/ tags	1.2%	0.03%	1.26%
# tags/doc	0.2	0.0003	0.17
Std. Dev.	5.5	0.02	5.4

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:

WET:	255 000
Test corpus:	90 500
Intersection:	90 500

Stay out of the water

	WET files	Test corpus	WET ∩ Corpus
# entities	166176	8772	121349
# docs w/ entities	5305	2284	2785
% docs w/ entities	2.1%	2.5%	3.1%
# entities/doc	0.68	0.09	1.34
Std. Dev.	55.2	1.7	66.4

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

	WET files	Test corpus	WET ∩ Corpus
English	81.4%	75.3%	78.6%
Chinese	1.5%	3.3%	2.0%
German	2.0%	2.5%	2.0%
Spanish	1.9%	2.5%	2.1%
Norwegian	0.12%	0.08%	0.08%

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

```
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> I f J e s s i c a Hart was a dog she would be t h i s
thing .
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> Snooki and JWoww dres s to match t h e i r dogs .
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> We can ’ t r e a l l y t e l l which i s Ozzy !
Inc r edibl e !
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> Coco and her p i t b u l l . I d e n t i c a l !
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> Luke Wilson looks a l o t l i k e hi s s tocky pup .
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> This i s amanda Byne ’ s dog . . . <p><span style="
font −s i z e : 9px ; ">Photo : Splash news </span>
Blondes
! Seann William Scot t would most d e f i n i t e l y have a
yel low lab .
<p><span style=" font −s i z e : 9px ; ">Photo : Splash
news </
span> Brunet tes ! Anne Hathaway with her chocolat e
Labrador , Esmeralda . <p><span style=" font −s i z e :
9px
; ">Photo : Splash news </span> Jon Hamm and hi s mutt
of ten make s imi l a r f a c i a l expres s ions .
```

Example extract from a WET file

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

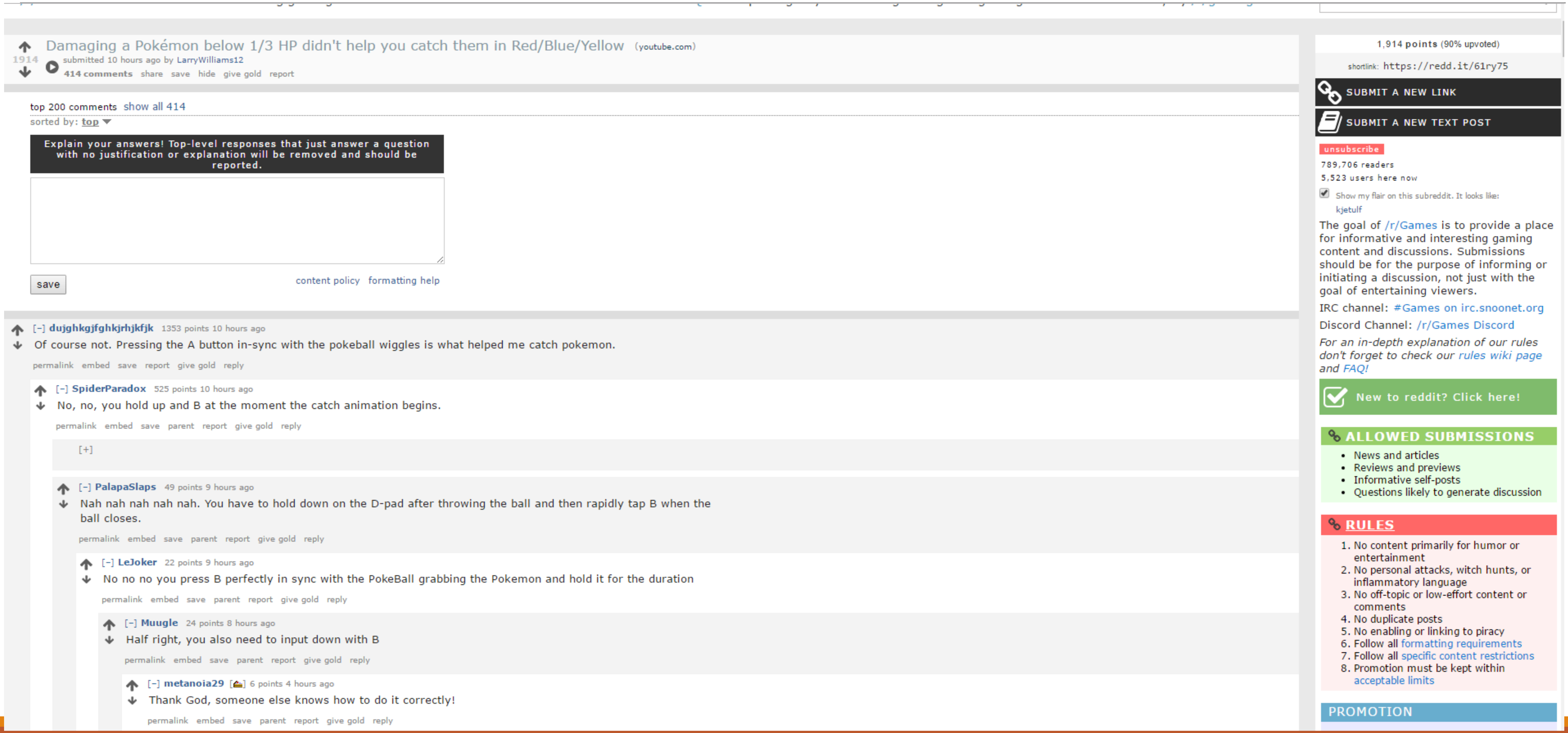
- In my project, I selected a tool chain called texrex

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

→ Common Crawl data

Cleaning the data: The problem



1914 Damaging a Pokémon below 1/3 HP didn't help you catch them in Red/Blue/Yellow (youtube.com)
submitted 10 hours ago by LarryWilliams12
414 comments share save hide give gold report

top 200 comments show all 414
sorted by: top

Explain your answers! Top-level responses that just answer a question with no justification or explanation will be removed and should be reported.

save content policy formatting help

[–] dujghkgjfghkjrhjkfjk 1353 points 10 hours ago
Of course not. Pressing the A button in-sync with the pokeball wiggles is what helped me catch pokemon.
permalink embed save report give gold reply

[–] SpiderParadox 525 points 10 hours ago
No, no, you hold up and B at the moment the catch animation begins.
permalink embed save parent report give gold reply

[+]

[–] PalapaSlaps 49 points 9 hours ago
Nah nah nah nah nah. You have to hold down on the D-pad after throwing the ball and then rapidly tap B when the ball closes.
permalink embed save parent report give gold reply

[–] LeJoker 22 points 9 hours ago
No no no you press B perfectly in sync with the PokeBall grabbing the Pokemon and hold it for the duration
permalink embed save parent report give gold reply

[–] Muugle 24 points 8 hours ago
Half right, you also need to input down with B
permalink embed save parent report give gold reply

[–] metanoia29 [] 6 points 4 hours ago
Thank God, someone else knows how to do it correctly!
permalink embed save parent report give gold reply

1,914 points (90% upvoted)

shortlink: https://redd.it/61ry75

SUBMIT A NEW LINK

SUBMIT A NEW TEXT POST

unsubscribe

789,706 readers

5,523 users here now

☑ Show my flair on this subreddit. It looks like:
kjetulf

The goal of /r/Games is to provide a place for informative and interesting gaming content and discussions. Submissions should be for the purpose of informing or initiating a discussion, not just with the goal of entertaining viewers.

IRC channel: #Games on irc.snoonet.org

Discord Channel: /r/Games Discord

For an in-depth explanation of our rules don't forget to check our rules wiki page and FAQ!

New to reddit? Click here!

ALLOWED SUBMISSIONS

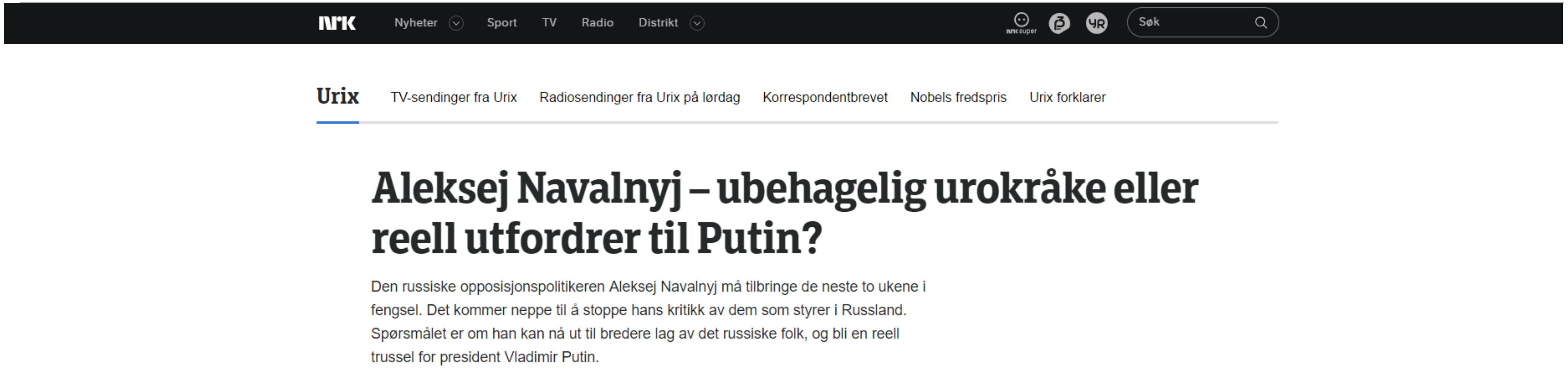
- News and articles
- Reviews and previews
- Informative self-posts
- Questions likely to generate discussion

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

PROMOTION

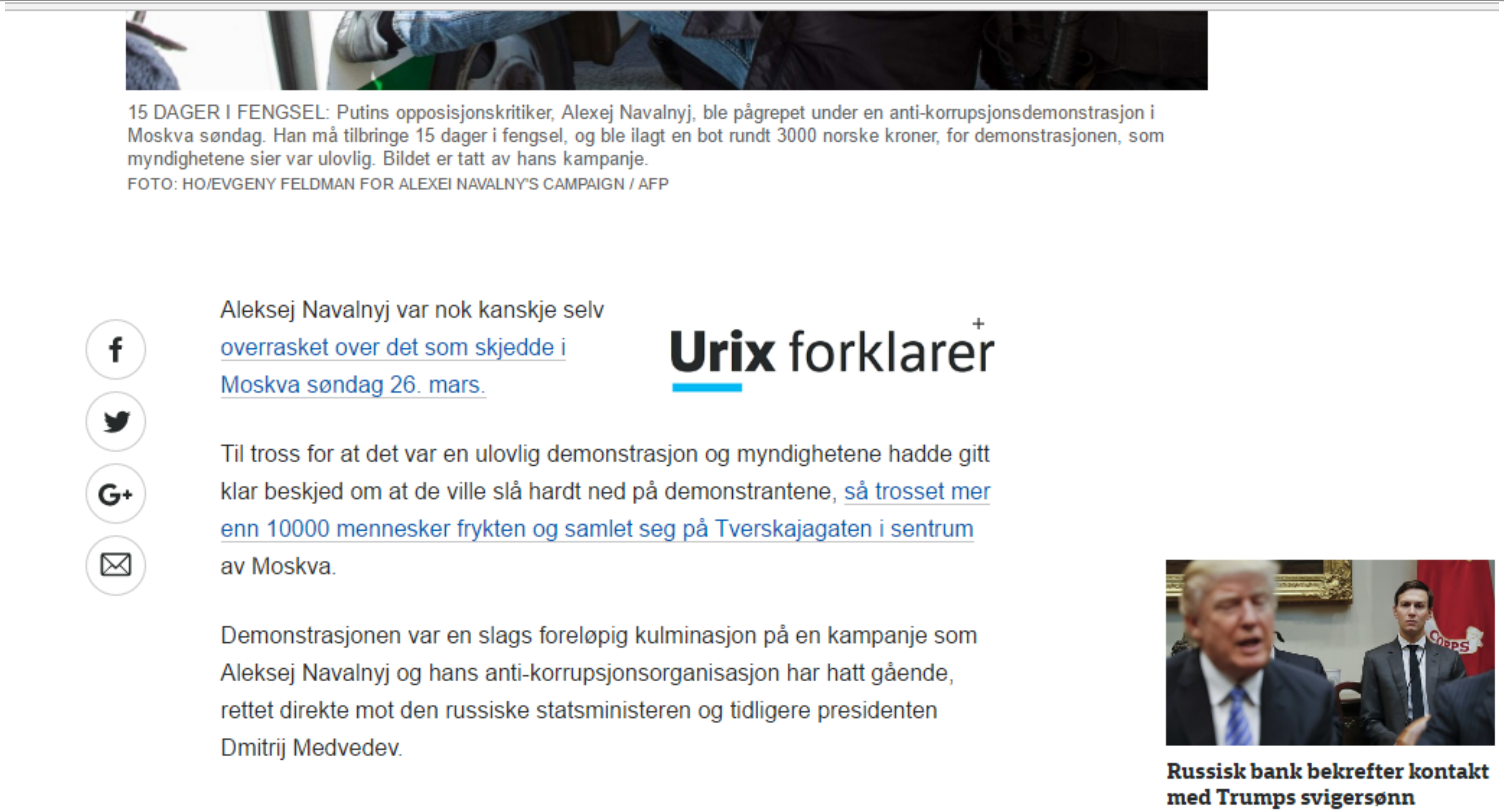
Cleaning the data: The problem



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



15 DAGER I FENGSEL: Putins opposisjonskritiker, Alexej Navalnyj, ble pågrepet under en anti-korrupsjonsdemonstrasjon i Moskva søndag. Han må tilbringe 15 dager i fengsel, og ble ilagt en bot rundt 3000 norske kroner, for demonstrasjonen, som myndighetene sier var ulovlig. Bildet er tatt av hans kampanje.

FOTO: HO/EVGENY FELDMAN FOR ALEXEI NAVALNY'S CAMPAIGN / AFP

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

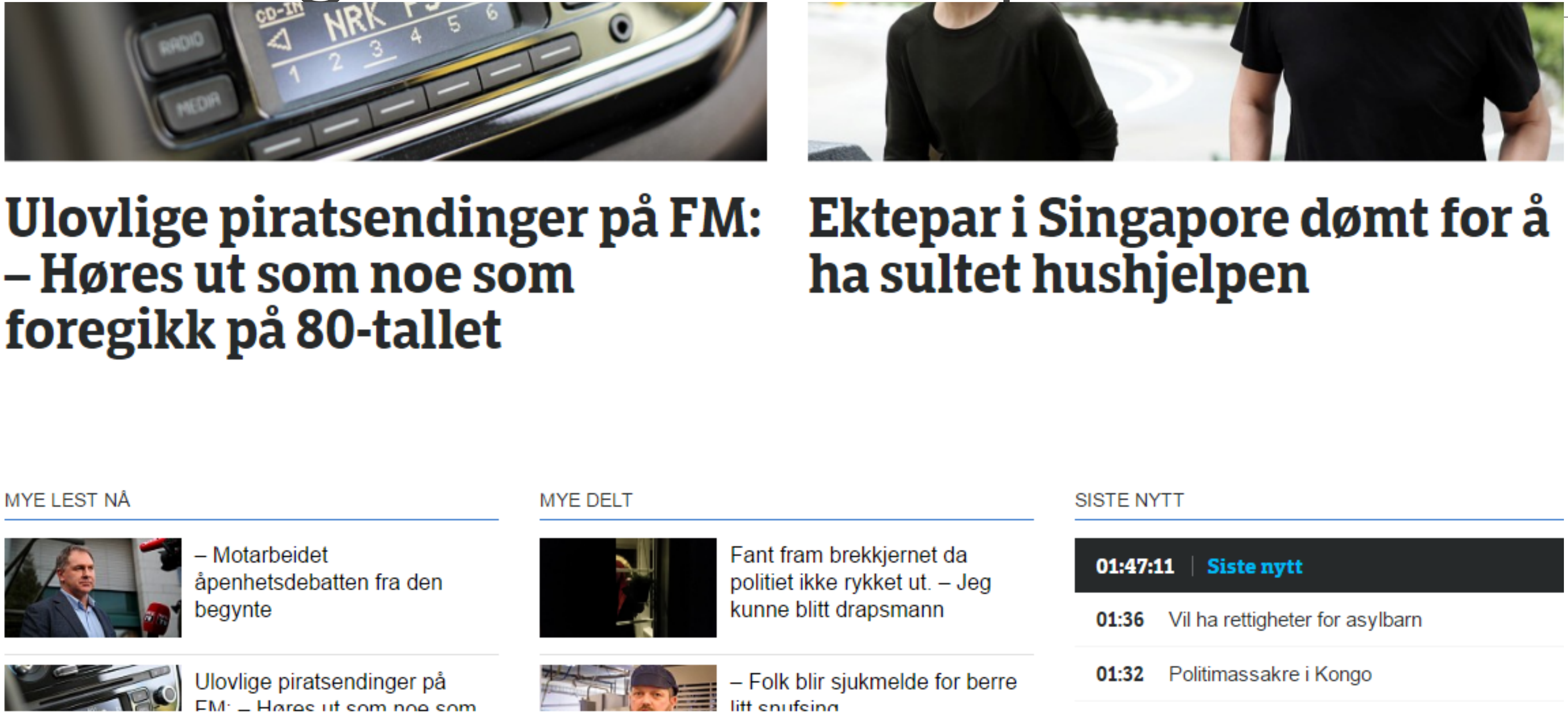
Urix forklarer

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 trosset 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 statsministeren og tidligere presidenten Dmitrij Medvedev.

Russisk bank bekrefter kontakt med Trumps svigersønn

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

MYE LEST NÅ

– Motarbeidet åpenhetsdebatten fra den begynte

Ulovlige piratsendinger på FM: – Høres ut som noe som

MYE DELT

Fant fram brekkjernet da politiet ikke rykket ut. – Jeg kunne blitt drapsmann

– Folk blir sjukmelde for berre litt snufsing

SISTE NYTT

01:47:11 | Siste nytt

01:36 Vil ha rettigheter for asylbarn

01:32 Politimassakre i Kongo

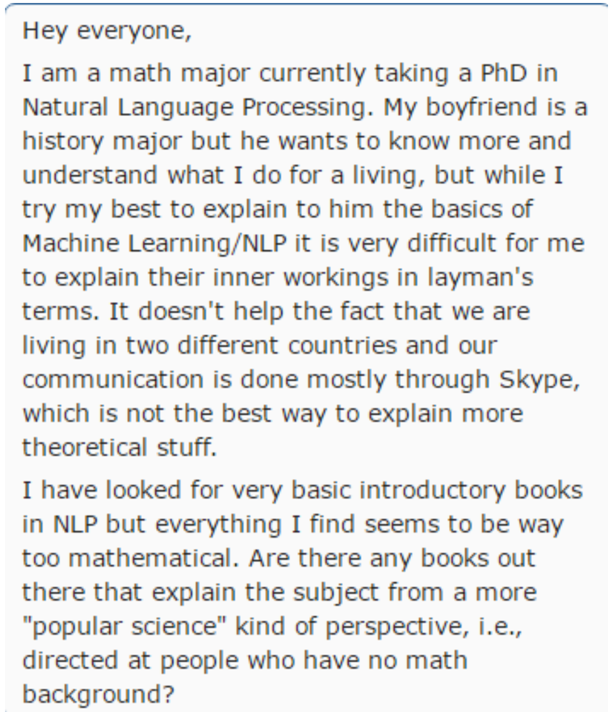
(HTML tag-stripping)

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

Also called "text quality assessment"

Connected text identification

CONNECTED TEXT

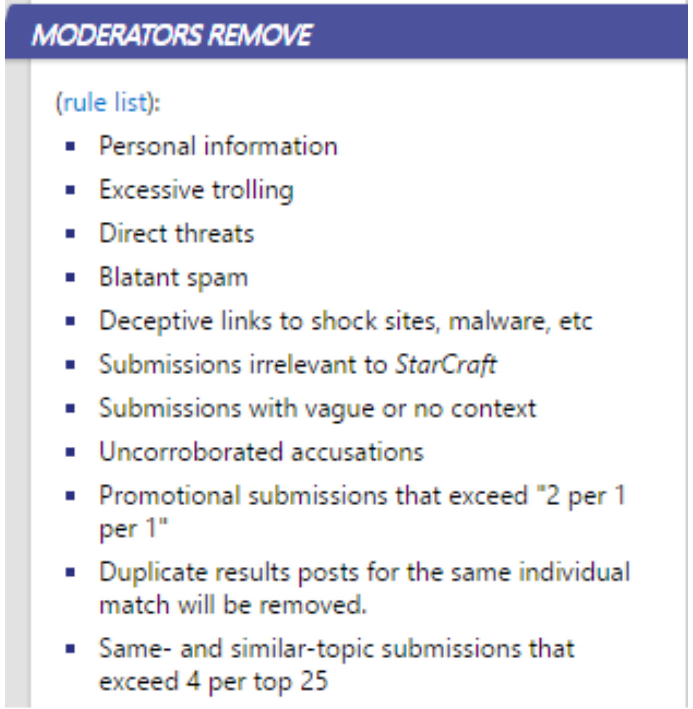


Hey everyone,

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



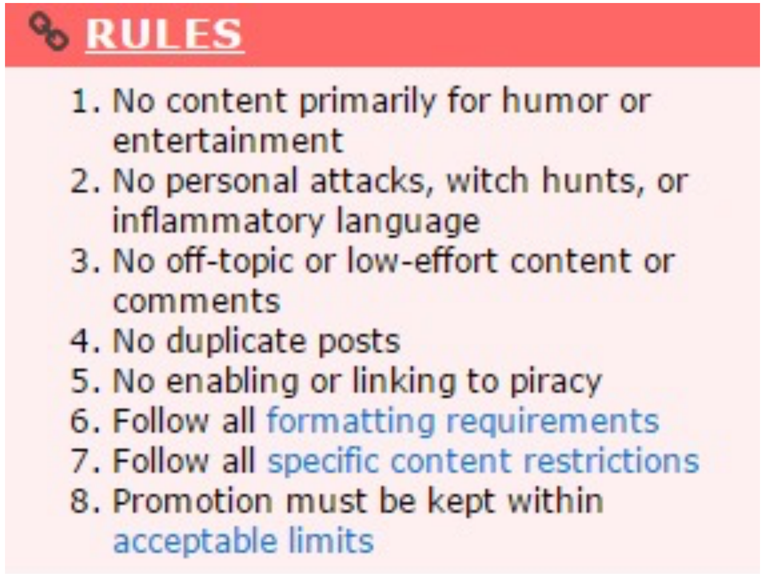
MODERATORS REMOVE

(rule list):

- Personal information
- Excessive trolling
- Direct threats
- Blatant spam
- Deceptive links to shock sites, malware, etc
- Submissions irrelevant to *StarCraft*
- Submissions with vague or no context
- Uncorroborated accusations
- Promotional submissions that exceed "2 per 1 per 1"
- Duplicate results posts for the same individual match will be removed.
- Same- and similar-topic submissions that exceed 4 per top 25

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

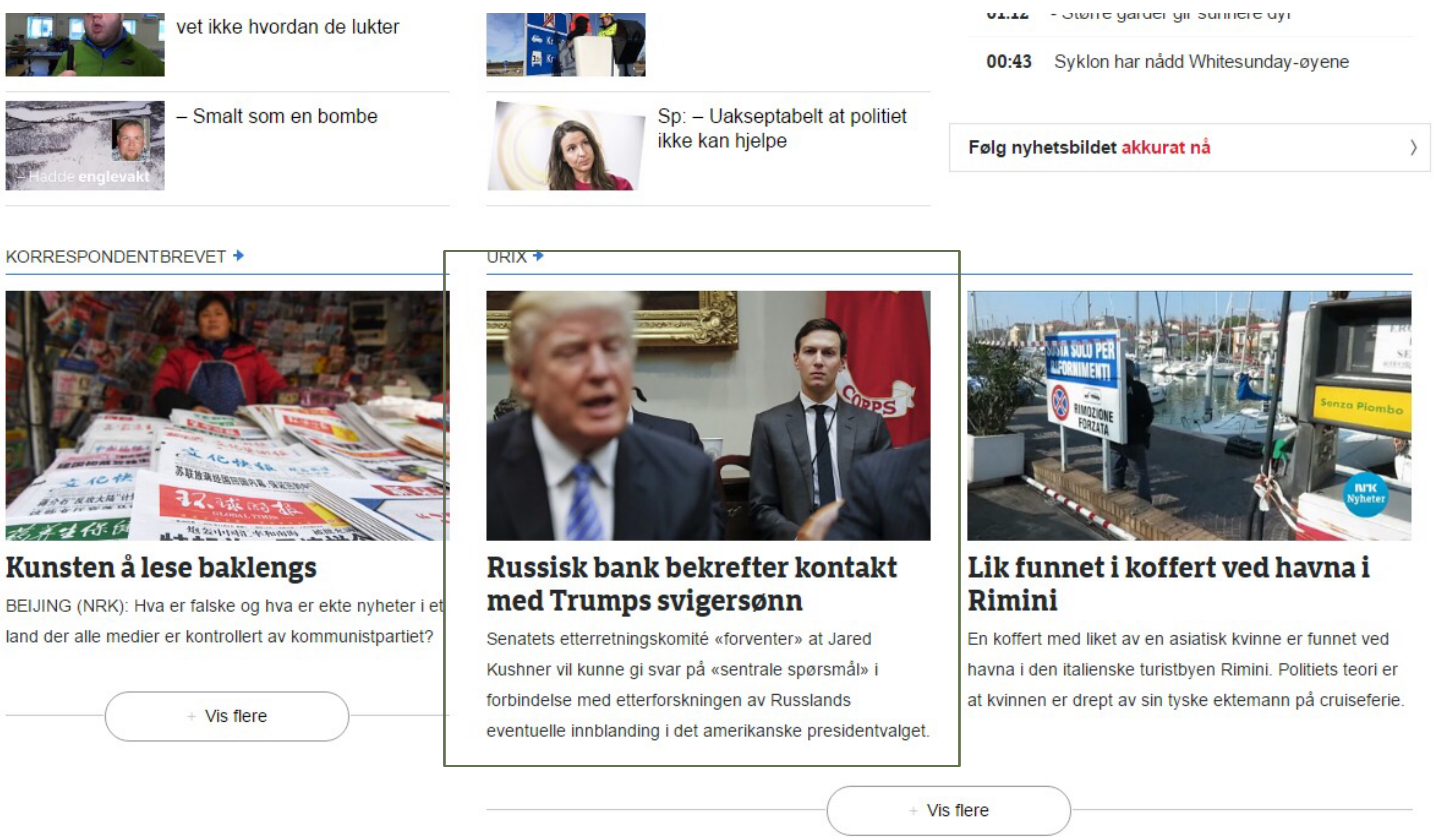
- Bonus: As these function words are language-specific, this text quality assessment will consider sentences in other languages as "bad"
- texrex use this assessment as their sole language identifier

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:



vet ikke hvordan de lukter

01:12 - Større gårder gir sunnere dyr

00:43 Syklon har nådd Whitesunday-øyene

– Smalt som en bombe

Sp: – Uakseptabelt at politiet ikke kan hjelpe

Følg nyhetsbildet akkurat nå 〉

KORRESPONDENTBREVET ➔

URIX ➔

Kunsten å lese baklengs

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

+ Vis flere

Russisk bank bekrefter kontakt med Trumps svigersønn

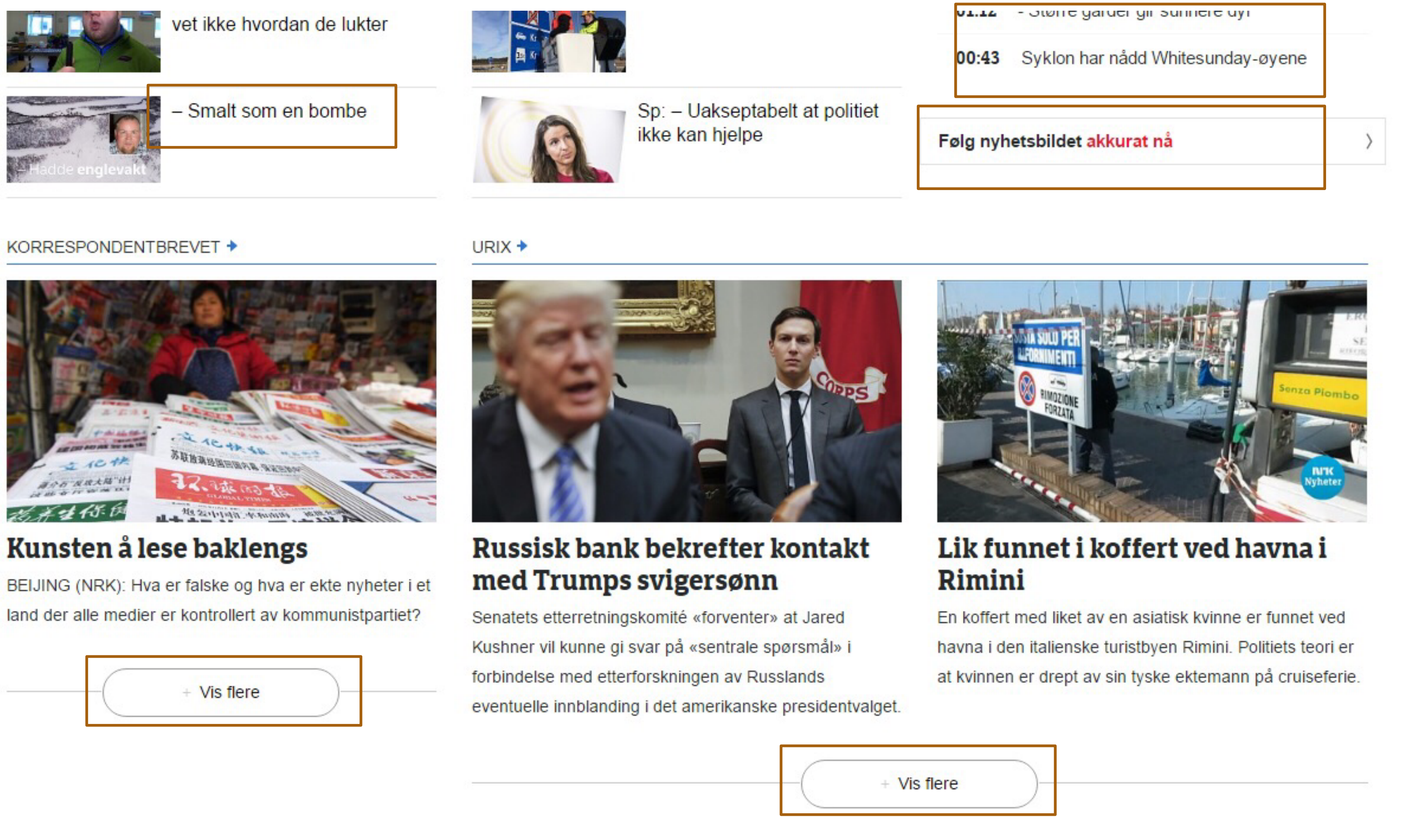
Senatets etterretningskomité «forventer» at Jared Kushner vil kunne gi svar på «sentrale spørsmål» i forbindelse med etterforskningen av Russlands eventuelle innblanding i det amerikanske presidentvalget.

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 kvinnen er drept av sin tyske ektemann på cruiseferie.

+ Vis flere

But also the obvious ones:



vet ikke hvordan de lukter

– Hadde englevakt

– Smalt som en bombe

Sp: – Uakseptabelt at politiet ikke kan hjelpe

01:12 - Større gårder gir sunnere dyr

00:43 Syklon har nådd Whitesunday-øyene

Følg nyhetsbildet akkurat nå

KORRESPONDENTBREVET

URIX

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

Senatets etterretningskomité «forventer» at Jared Kushner vil kunne gi svar på «sentrale spørsmål» i forbindelse med etterforskningen av Russlands eventuelle innblanding 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 kvinnen er drept av sin tyske ektemann på cruiseferie.

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

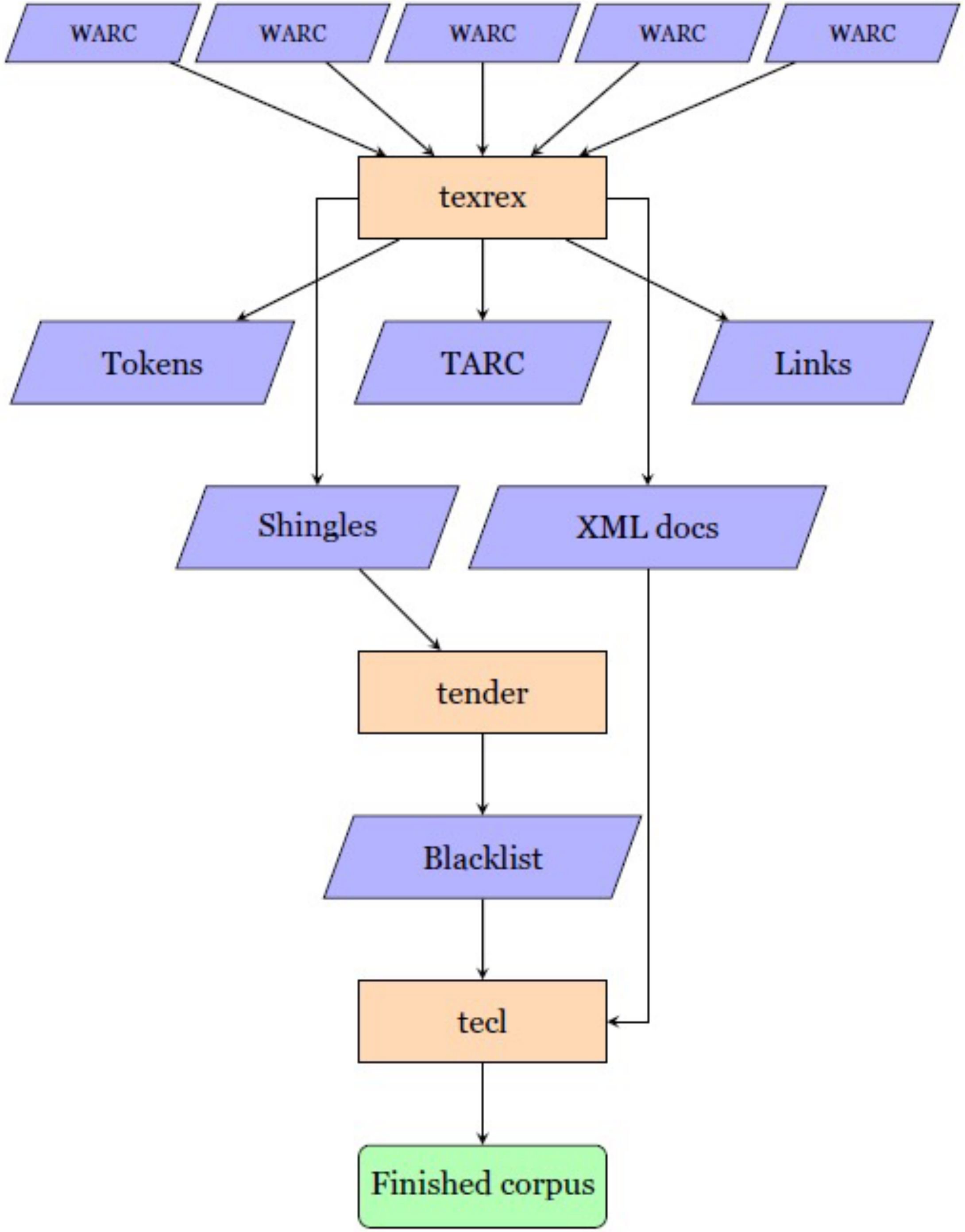
Encodings

- 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

texrex / tender / tecl

- 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

texrex / tender / tecl



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

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

Effective run time to produce corpus from start to end:

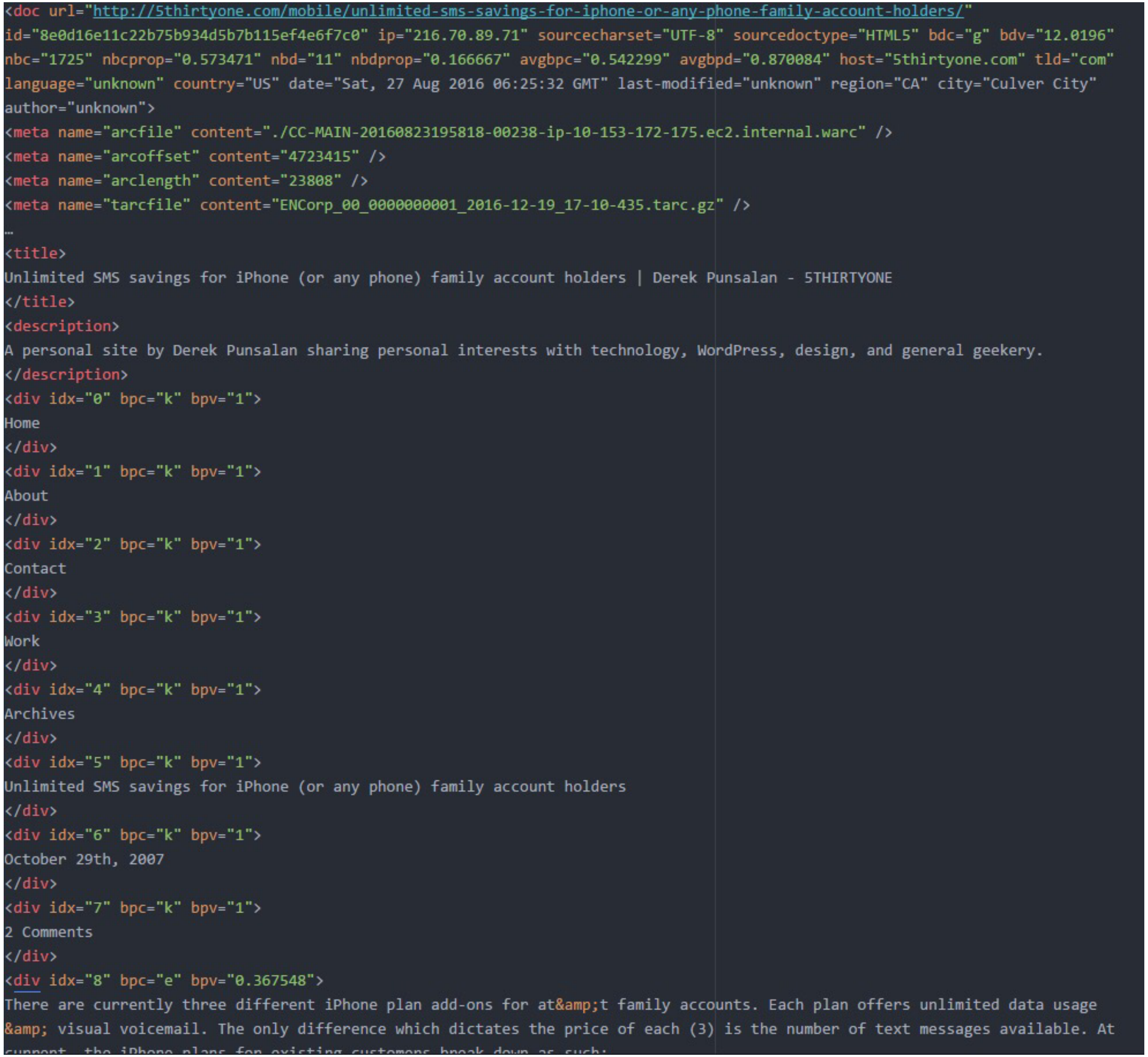
16 days

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



```
<doc url="http://5thirtyone.com/mobile/unlimited-sms-savings-for-iphone-or-any-phone-family-account-holders/"
id="8e0d16e11c22b75b934d5b7b115ef4e6f7c0" ip="216.70.89.71" sourcecharset="UTF-8" sourcedoctype="HTML5" bdc="g" bdv="12.0196"
nbc="1725" nbcprop="0.573471" nbd="11" nbdprop="0.166667" avgbpc="0.542299" avgbpd="0.870084" host="5thirtyone.com" tld="com"
language="unknown" country="US" date="Sat, 27 Aug 2016 06:25:32 GMT" last-modified="unknown" region="CA" city="Culver City"
author="unknown">
<meta name="arcfile" content="./CC-MAIN-20160823195818-00238-ip-10-153-172-175.ec2.internal.warc" />
<meta name="arcoffset" content="4723415" />
<meta name="arclength" content="23808" />
<meta name="tarcfile" content="ENCorp_00_0000000001_2016-12-19_17-10-435.tarc.gz" />
…
<title>
Unlimited SMS savings for iPhone (or any phone) family account holders | Derek Punsalan - 5THIRTYONE
</title>
<description>
A personal site by Derek Punsalan sharing personal interests with technology, WordPress, design, and general geekery.
</description>
<div idx="0" bpc="k" bpv="1">
Home
</div>
<div idx="1" bpc="k" bpv="1">
About
</div>
<div idx="2" bpc="k" bpv="1">
Contact
</div>
<div idx="3" bpc="k" bpv="1">
Work
</div>
<div idx="4" bpc="k" bpv="1">
Archives
</div>
<div idx="5" bpc="k" bpv="1">
Unlimited SMS savings for iPhone (or any phone) family account holders
</div>
<div idx="6" bpc="k" bpv="1">
October 29th, 2007
</div>
<div idx="7" bpc="k" bpv="1">
2 Comments
</div>
<div idx="8" bpc="e" bpv="0.367548">
There are currently three different iPhone plan add-ons for at&amp;t family accounts. Each plan offers unlimited data usage
&amp; visual voicemail. The only difference which dictates the price of each (3) is the number of text messages available. At
current, the iPhone plans for existing customers break down as such:
```

ENC^3 Text

- Extracted text from the XMLs
- Boilerplate removed
- Picked a threshold at 0.5 for boilerplate removal
- 116 bln white-space separated tokens
- No meta data in text files, but…
- Comes with linker files to link each document back to the XML files

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 wordùgone. 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.

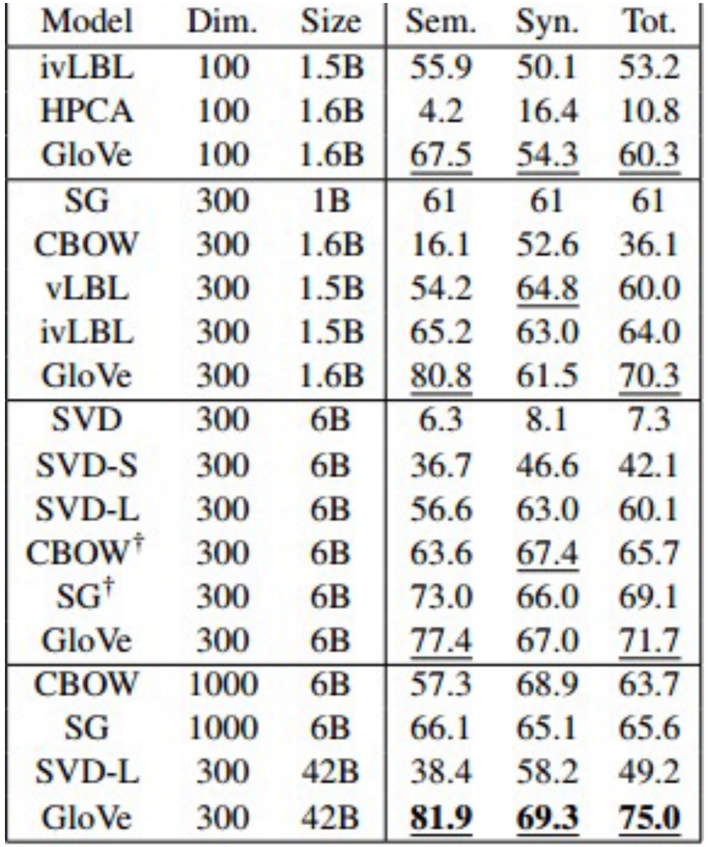
ENC^3 ConLL

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

```
1    As      _    _    _    _    _
2    I       _    _    _    _    _
3    'm      _    _    _    _    _
4    sure    _    _    _    _    _
5    you     _    _    _    _    _
6    are     _    _    _    _    _
7    aware   _    _    _    _    _
8    ,       _    _    _    _    _
9    a       _    _    _    _    _
10   soccer  _    _    _    _    _
11   coach   _    _    _    _    _
12   does    _    _    _    _    _
13   more    _    _    _    _    _
14   than    _    _    _    _    _
15   just    _    _    _    _    _
16   coach   _    _    _    _    _
17   .       _    _    _    _    _
```

Evaluation with word embeddings

- I wish to evaluate the corpus in a down-stream 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



Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

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