Towards automated creation and management of an evolving web corpus

Gareth Dwyer

Supervisor: James Connan

Goals

- Create Corpus of South African English that:
  - Contains accurate articles (not other WWW noise)
  - Contains no duplicate content
  - Automatically evolves
  - Allows for manual intervention

- Create tools for linguistic analysis
  - Keyword in Context
  - Collocates

Corpus builder and manager

Scrape WWW for South African English content:

- Watch RSS feeds (✓)
- Scrapy (✗) (memory issues, and too broad)

APIs give more accurate data

- WayBack Machine (✓)
- Disqus API (✓)

Remove Boilerplate, extract plain text:

Different algorithms for this -- Reporter (Python), Boilerpipe (Java)

WayBack Machine

- ➢ Takes periodic ‘snapshots’ of all (large) websites
- ➢ Provides API
- ➢ More accurate than general scrape

Backwards crawl:

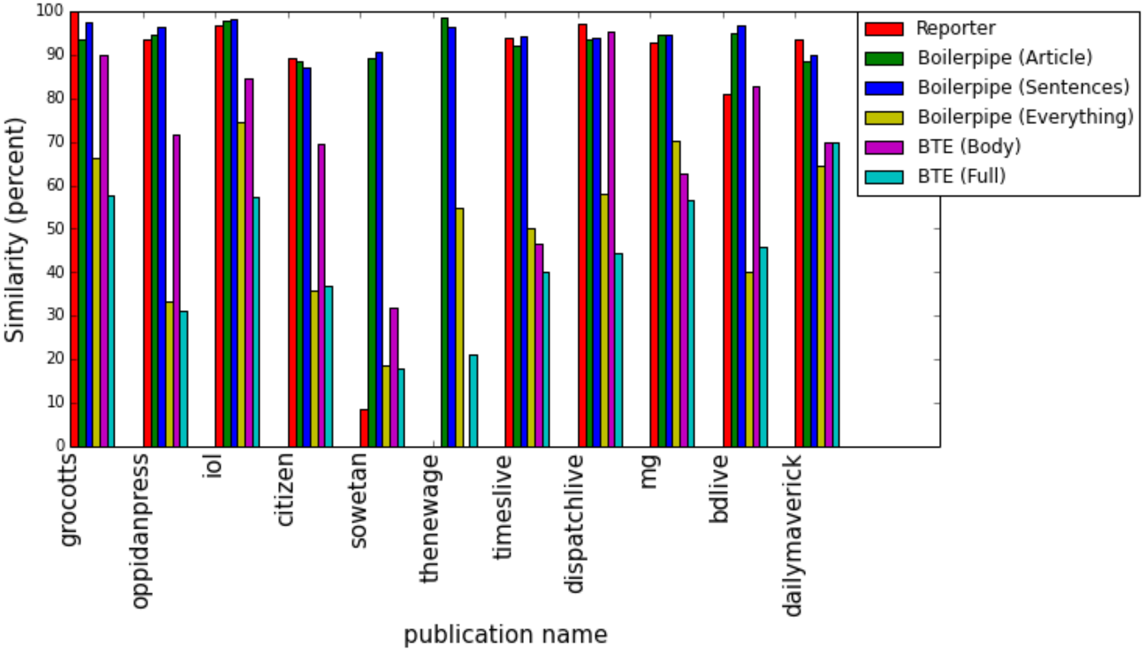
```
request most recent snapshot

crawl all links (depth 1)

go to previous snapshot
```

Cleaning

Actually 6 algorithms/variations

Manually cleaned dataset

Similarity measures for all options using TF-IDF/Cosin distance



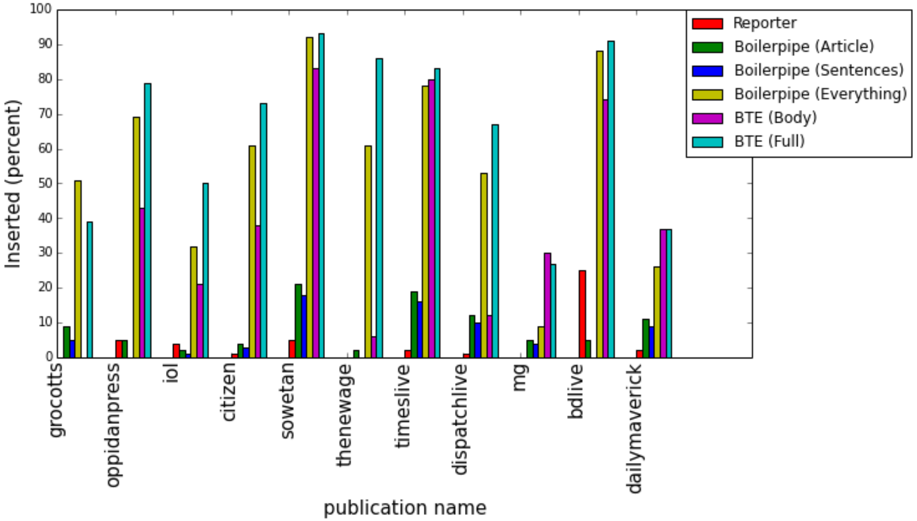
(BA student learnt to graph)

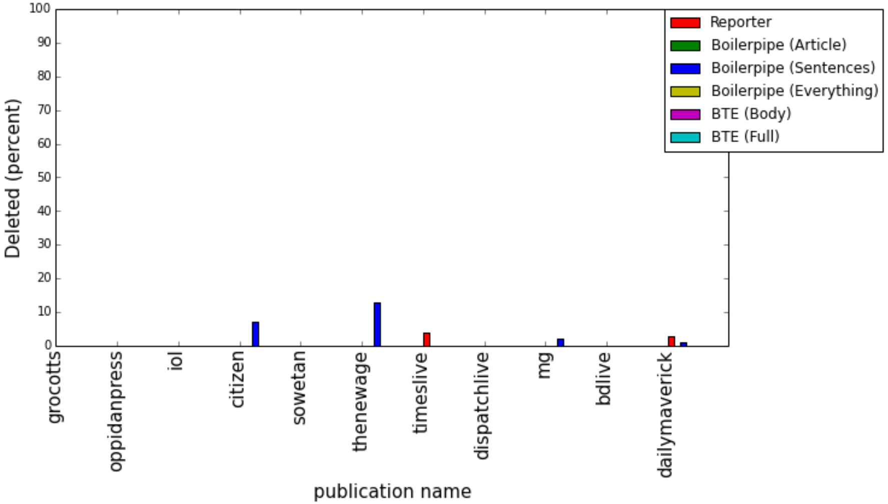
Cleaning (2)

Wdiff for more detailed insights: inserted and deleted text

Boilerpipe Sentences (blue) looks good, but deletes too much

Final choice Reporter *and* Boilerpipe (all variants)





Cleaning (manual)

- Automated system is not perfect
- Even with machine learning, 100% accuracy isn’t feasible
  - Sneaky adverts, (long) comments, etc

- Manual cleaning
  - Filter (tag type, attribute name, attribute value)
  - demo http://www.iol.co.za/news/politics/closing-arguments-in-eff-disciplinary-1.1771855#.VE-cCHVSykA
  - http://146.231.133.148/sandbox

Deduplication

- Check every article against every other - O(n^2)
- Or load all docs into memory, create matrix - SegFault
- Or use hash tables:
  - Store md5(sentence):[article_ids] in DB
  - To check for duplicates of Article*:*

```
Hash all sentences

Get possible matches for each sentence

Do pairwise comparison of each possible match
```

But how unique is a sentence?

Deduplication (2)

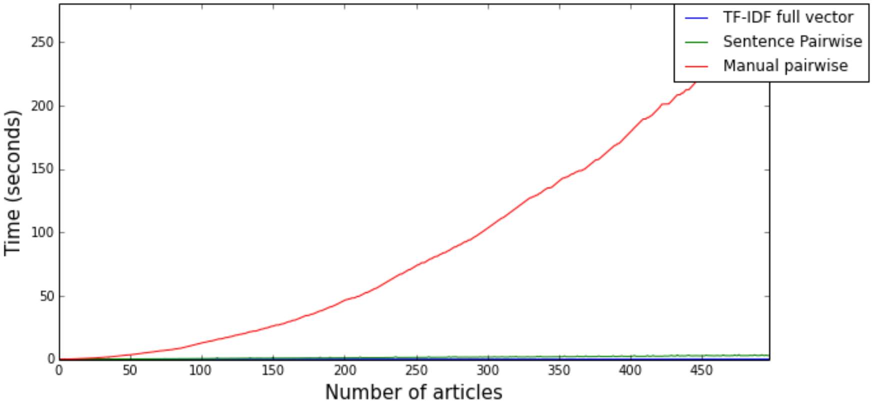
Unique enough! (Min length, max matches)

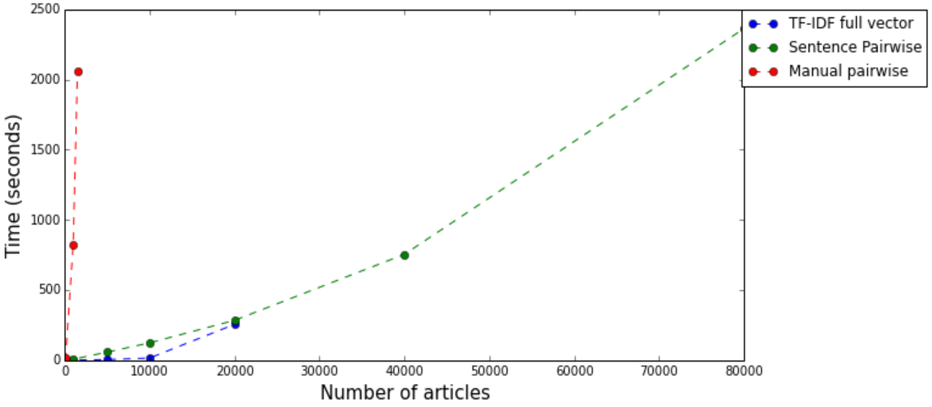
$$\begin{aligned} Sim(\mathbf{q}, \mathbf{d}) &= \cos(\mathbf{q} \angle \mathbf{d}) \\ &= \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{d}\| \ \|\mathbf{q}\|} \\ &= \frac{\sum_{k \in (q \cap d)} w_{kd} \cdot w_{kq}}{\|\mathbf{d}\| \ \|\mathbf{q}\|} \\ &= \frac{\sum_{k \in (q \cap d)} w_{kd} \cdot w_{kq}}{\sqrt{\sum_{k \in d} (w_{kd})^2} \sqrt{\sum_{k \in q} (w_{kq})^2}} \end{aligned}$$

Pairwise is good for checking accuracy

TF-IDF matrix needs too much RAM (and Maths)

Sentence Algorithm works well, nearly linear, small memory footprint, accurate enough.

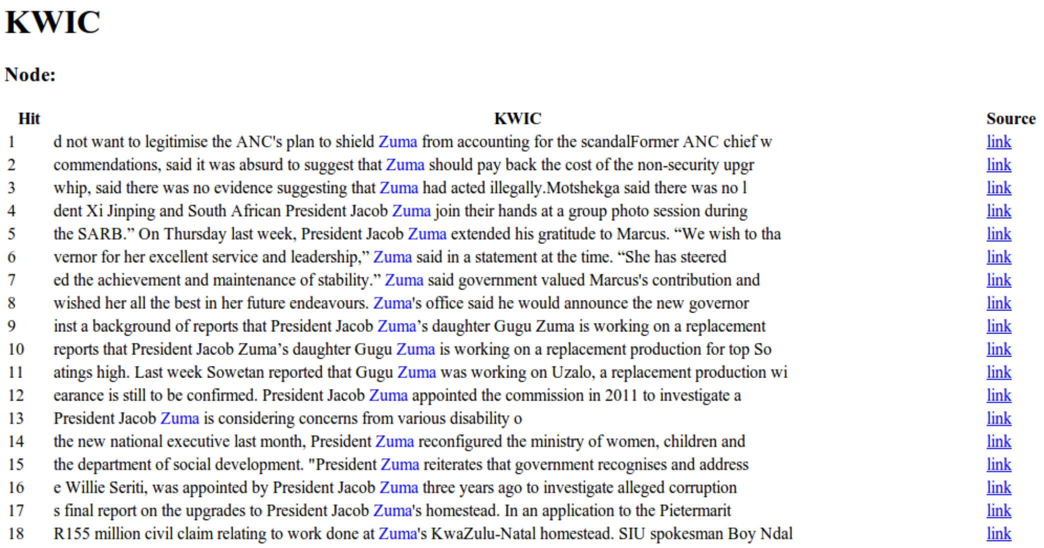




Language Analysis Tools

1) Keyword in Context

Sorted by date, show every match in context (customizable)



KWIC

Node:

Hit	KWIC	Source
1	d not want to legitimise the ANC's plan to shield Zuma from accounting for the scandalFormer ANC chief w	link
2	commendations, said it was absurd to suggest that Zuma should pay back the cost of the non-security upgr	link
3	whip, said there was no evidence suggesting that Zuma had acted illegally.Motshekga said there was no l	link
4	dent Xi Jinping and South African President Jacob Zuma join their hands at a group photo session during	link
5	the SARB.” On Thursday last week, President Jacob Zuma extended his gratitude to Marcus. “We wish to tha	link
6	vernor for her excellent service and leadership,” Zuma said in a statement at the time. “She has steered	link
7	ed the achievement and maintenance of stability.” Zuma said government valued Marcus's contribution and	link
8	wished her all the best in her future endeavours. Zuma's office said he would announce the new governor	link
9	inst a background of reports that President Jacob Zuma’s daughter Gugu Zuma is working on a replacement	link
10	reports that President Jacob Zuma’s daughter Gugu Zuma is working on a replacement production for top So	link
11	atings high. Last week Sowetan reported that Gugu Zuma was working on Uzalo, a replacement production wi	link
12	earance is still to be confirmed. President Jacob Zuma appointed the commission in 2011 to investigate a	link
13	President Jacob Zuma is considering concerns from various disability o	link
14	the new national executive last month, President Zuma reconfigured the ministry of women, children and	link
15	the department of social development. "President Zuma reiterates that government recognises and address	link
16	e Willie Seriti, was appointed by President Jacob Zuma three years ago to investigate alleged corruption	link
17	s final report on the upgrades to President Jacob Zuma's homestead. In an application to the Pietermarit	link
18	R155 million civil claim relating to work done at Zuma's KwaZulu-Natal homestead. SIU spokesman Boy Ndal	link

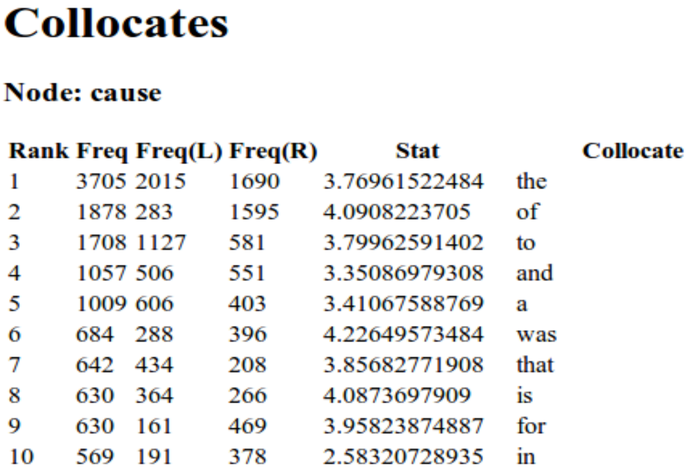
Language Analysis Tools (2)

Collocations - Word pairs which appear to be associated

Not just based on frequency though (common ones are boring)

$$log_2[frequency(n,c) \times N]/frequency(n) \times frequency(c)$$

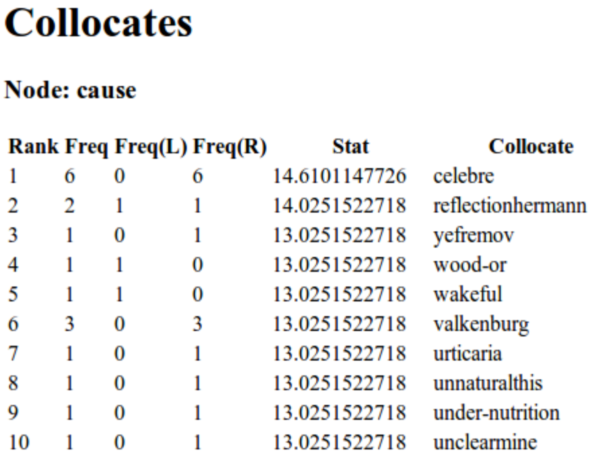
(N = size of corpus, n = ‘node word’, c = collocate)



Collocates

Node: cause

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	3705	2015	1690	3.76961522484	the
2	1878	283	1595	4.0908223705	of
3	1708	1127	581	3.79962591402	to
4	1057	506	551	3.35086979308	and
5	1009	606	403	3.41067588769	a
6	684	288	396	4.22649573484	was
7	642	434	208	3.85682771908	that
8	630	364	266	4.0873697909	is
9	630	161	469	3.95823874887	for
10	569	191	378	2.58320728935	in



Collocates

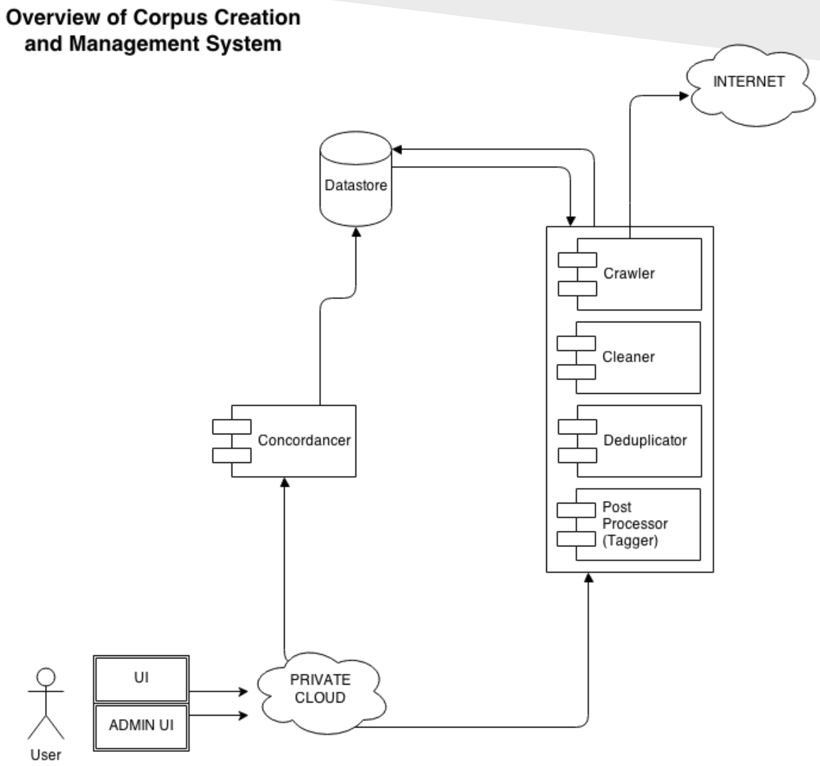
Node: cause

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	6	0	6	14.6101147726	celebre
2	2	1	1	14.0251522718	reflectionhermann
3	1	0	1	13.0251522718	yefremov
4	1	1	0	13.0251522718	wood-or
5	1	1	0	13.0251522718	wakeful
6	3	0	3	13.0251522718	valkenburg
7	1	0	1	13.0251522718	urticaria
8	1	0	1	13.0251522718	unnaturalthis
9	1	0	1	13.0251522718	under-nutrition
10	1	0	1	13.0251522718	unclearmine

System Design

- ❖ Modular design:
  - ➢ If a better way is found/built, simply swap out module
- ❖ All modules interact directly with DB
- ❖ Don’t need each other

(BA student still can’t draw stick figures)



System design (2)

- Crawl
  - RSS
  - WayBack
- Deduplicate (exact, hashes on HTML)
- Clean (automatic, saved manual settings)
- Deduplicate (exact, hashes of plain text)
- Deduplicate (near, hashes of sentences)
- Post process (POS tagging, word lists)

CronJobs and indexed database flags

(isCleaned, isDeduplicated, etc)

Conclusion

Achieved all primary goals:

- Created corpus (Linguists already using it)
- Evolves, allows for new feeds
- Crawling, Storing, Cleaning Deduplicating

Further work:

- Dynamic data not as good as hoped
  - Use Disqus API, but limited to sites that use Disqus
- Better GUIs
- Cleaning could be more automated
- More research into deduplication (clustered TF-IDF matrices?)

That’s all, folks

Questions, Suggestions, Answers, Money?

?