Gareth Dwyer

Supervisor: James Connan

Corpus builder and manager

- ➔ Track South African English
- ➔ Crawl pre-identified list of sites
  - ◆ Mainly online newspapers and blogs
  - ◆ mg.co.za; iol.co.za; etc.
- ➔ Match against South African word lists
- ➔ Identify new South African words
  - ◆ Manual intervention to confirm

Technology

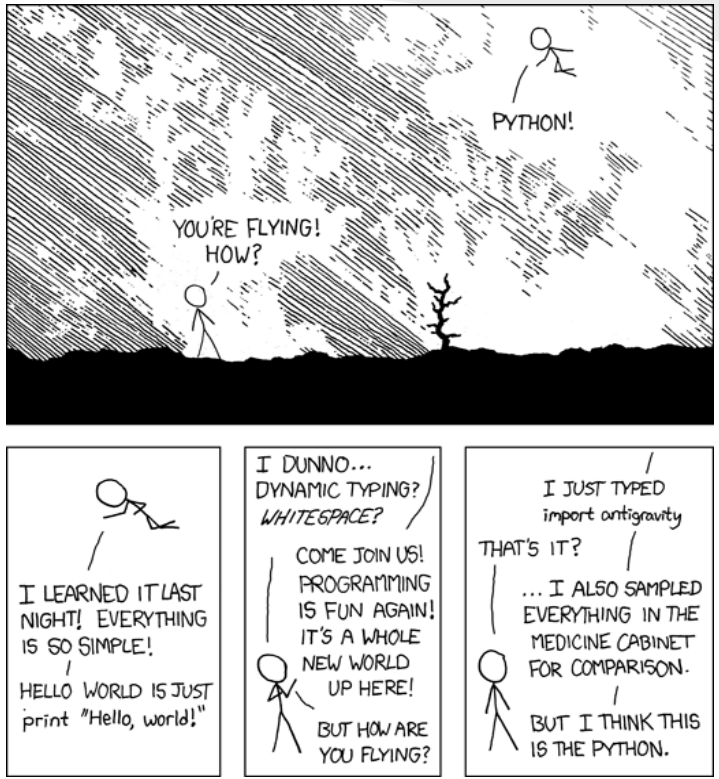
- ➔ MongoDB – NoSQL
- ➔ Python NLTK
- ➔ Scrapy
- ➔ Reporter.py
- ➔ Flask (.py) for front-end
- ➔ Custom deduplicator (.py)

Yes, I have a bit of a bias towards Python

Stages (1) Crawl for data

- Forward Crawling
  - RSS Feeds
- Backward Crawling - options
  - Manual
    - well, that is manually written Python
  - Import scrapy
  - Scrapy.scrape()

python is fun!



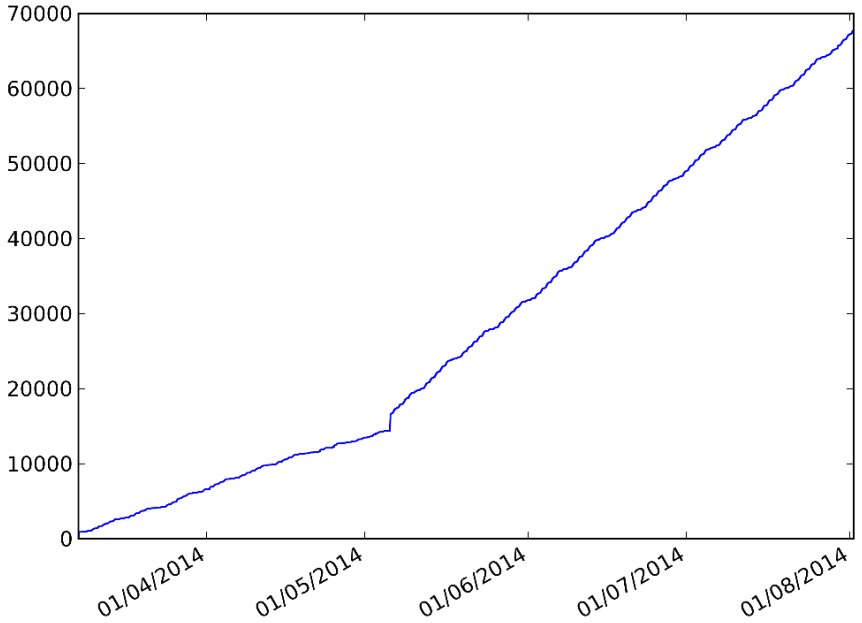
Stages (2) – Process data

- Extract text (remove boilerplate) - options
  - Identify boilerplate?
  - Identify interesting text? (reporter.py)
- Extract metadata
  - Some online content creators are nice
  - Others, not so much…
- Parse sentences
- Deduplication (Yay)

Stages (3) – Extract info

- ➔ Identify new words
- ➔ Identify new usages of words
- ➔ Allow for custom searches and analysis
  - ◆ Word frequency
  - ◆ Word in context
- ➔ AntConc.pl :/

Articles



(Sorry ‘bout [lack of] labels)

Deduplication

- Exact Deduplication - “visit cars.co.za” (<< don’t really)
  - Straight-forward
    - Hash all articles
    - Check for duplicates
- Near deduplication (SAPA and plagiarism)
  - More effort - err on keeping too much or too little
  - Done badly in the past
    - And in some cases very badly…

My(?) algorithm

For each article

    for each sentence

        hash(sentence)

        get matches(hashed_sentence)

    for each match:

        pairwise comparison (gets percentage similarity)

Only need to do pairwise comparison (expensive) on very few articles, and can be optimized further

Performance

- ➔ Still took two days to run on 70000 articles
- ➔ Until ensure_index() on the sentences hases
- ➔ one line of code
  - ◆ two days -> 20 minutes (Alan knows the feels)

More?

- Demo?
  - 404, server not found
- More articles?
  - Backwards crawling is fast
  - Server only has 50GB

Where next?

- SAE Word lists
- Concordancer (not.pl)
- Dynamic crawling
- Customizable metadata (3 or 4 pieces of info needed)
  - <meta author= “John Smith”>
  - John Smith
- Frontend.prettify()
- Documentation

That’s all folks

Questions, Suggestions, Answers, Money?

?