Me

Gareth Dwyer

BA – Philosophy, German, Computer Science

gareth@dwyer.co.za

Hamilton Honours Lab

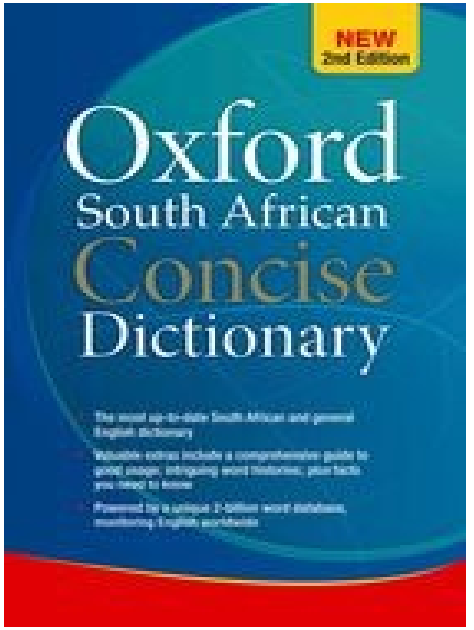
Supervisor: James Connan

Web-crawler and Corpus 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

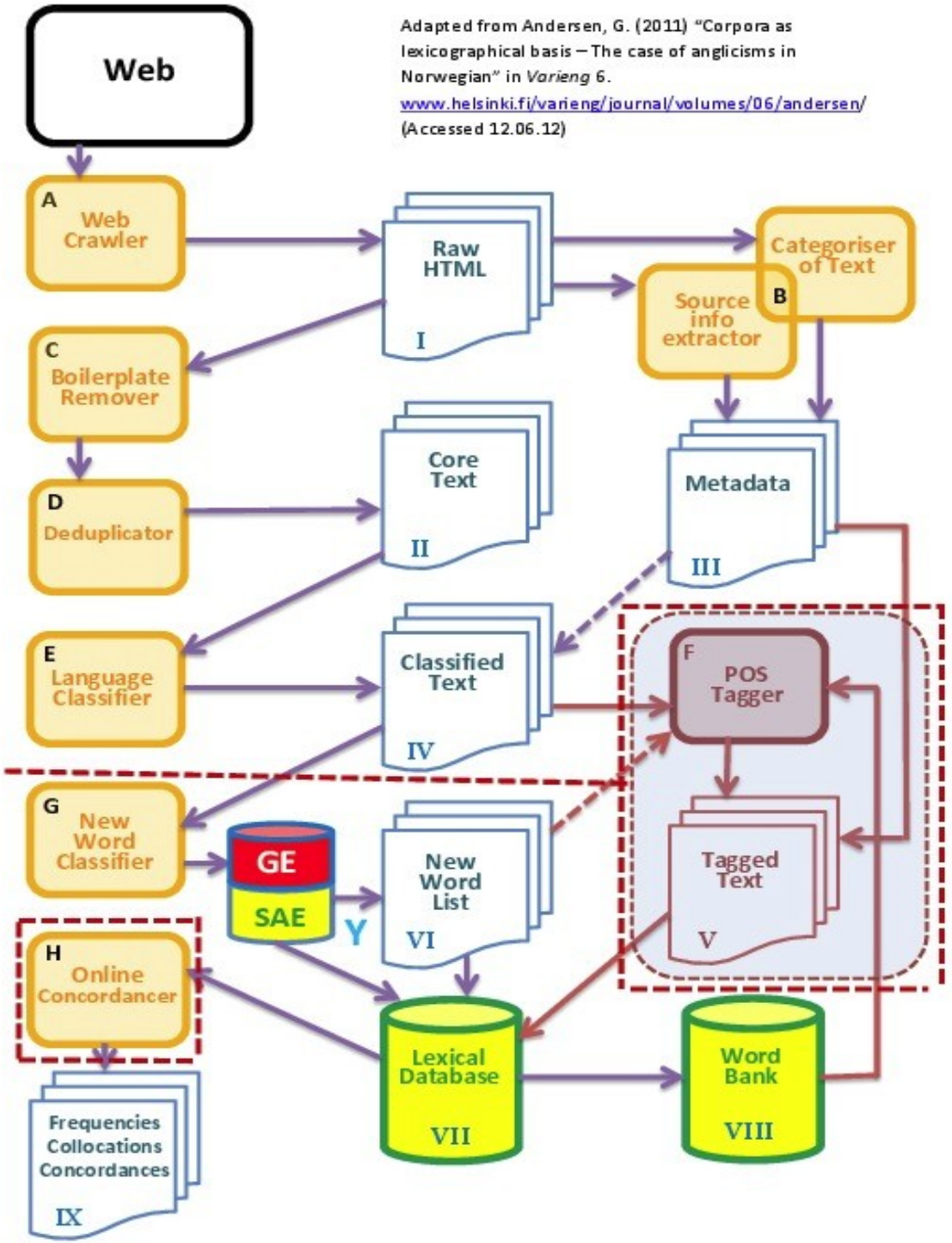
Dictionaries

- Descriptive vs Prescriptive
- Dictionaries need to be descriptive
- Oxford South African Concise Dictionary
  - Compiled at Rhodes by the Dictionary Unit for South African English (DSAE)



Corpora

- Body of text
- Taken from a variety of sources
  - Social media to scholarly articles
- Used for analysis, tracking language use
- Oxford English Corpus
  - Word popularity, word usage
  - Selfie
  - Minuscule (miniscule)



Adapted from Andersen, G. (2011) "Corpora as lexicographical basis – The case of anglicisms in Norwegian" in *Varieng* 6. www.helsinki.fi/varieng/journal/volumes/06/andersen/ (Accessed 12.06.12)

Web-crawler

Challenges

- Flexibility
- Deduplication
  - Near Duplication
- Ajax
  - Comments on blogs and newspapers

Crawling Ajax

- Disqus
  - Used widely now, but that could cha
  - API
- CrawlJax
  - More flexible, open source
  - Creates DOM graph
  - Written in Java
- Selenium
  - Can simulate any browser movements

Databases

- MongoDB 

  - Document Database
  - Scalable
  - Flexible (Fast development)
  - Unstructured data

- SQL
  - 1NF, 2NF, 3NF, EKNF, BCNF, 4NF, 5NF, DKNF, 6NF

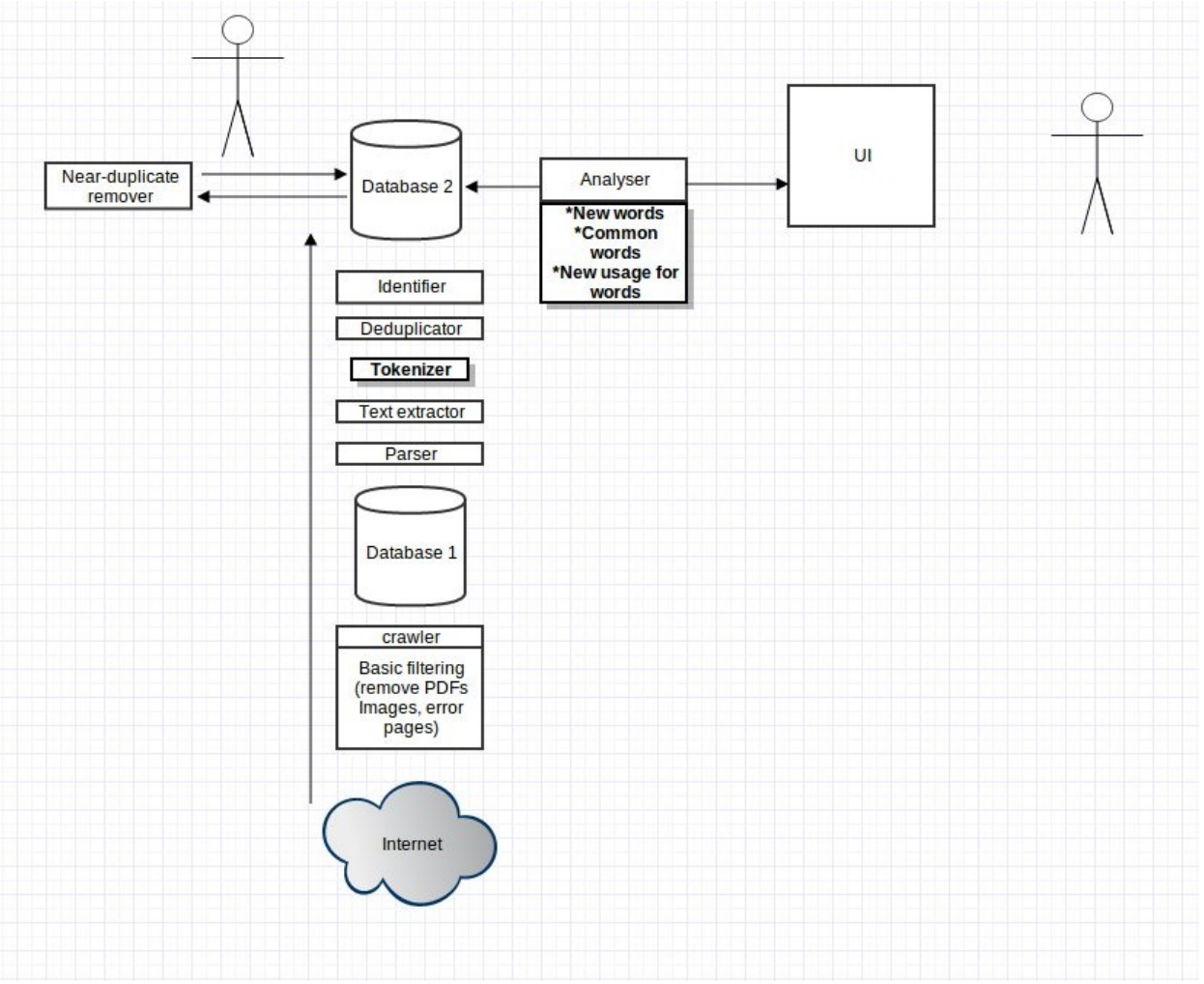
Natural Language Processing

- NLTK
  - Python, open source
  - Tokenization (break into words intelligently)
    - Deal with punctuation properly
  - Word matching (different tenses etc, same word)
  - Parse trees

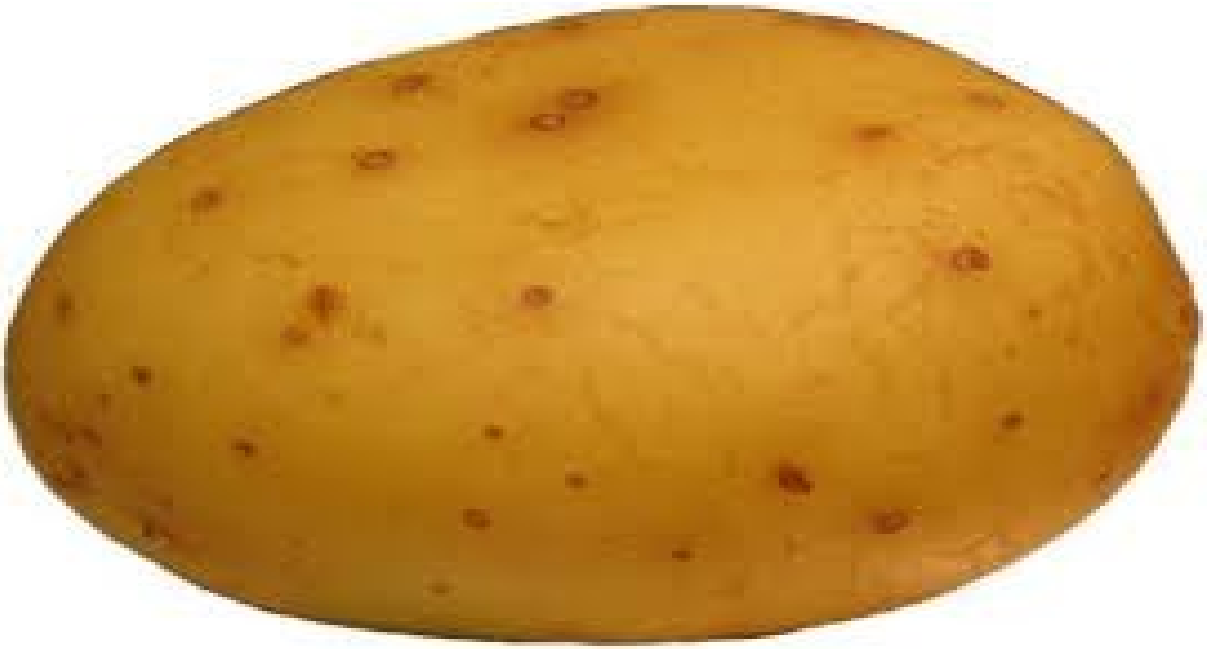
- Reporter
  - Python, open source
  - Identifies main text of article based on div scoring

Design

- Two databases – raw data, extracted text
- Modularizatio
  - Crawler
  - Parser
  - Text extracto
  - Tokenizer
  - Deduplicator
  - Metadata ide
  - Analyser
  - UI



Potato



Questions?