

Cloud Computing for the Humanities

Graham Wilcock University of Helsinki



What is Cloud Computing?

- "Run your app in the cloud"
 - Using somebody else's computers
- Computing resources on-demand
 - Like electricity, or pizza delivery
- Platform-as-a-Service (PaaS)
 - Example: Google App Engine

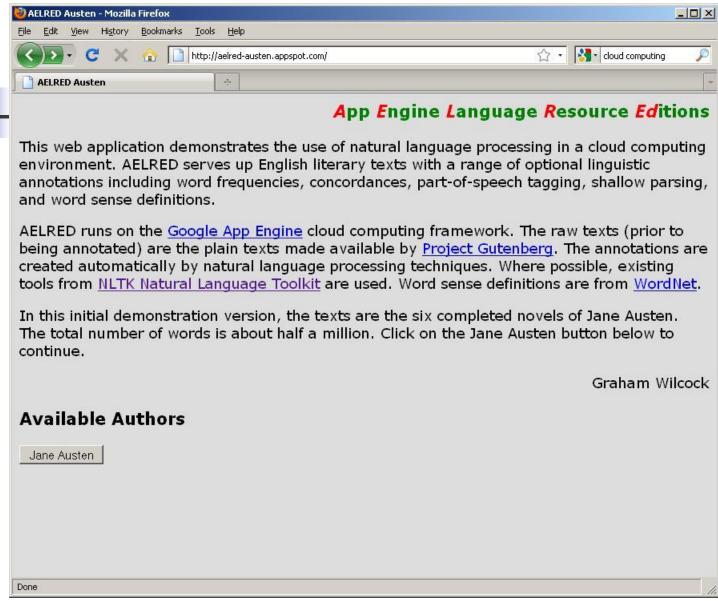




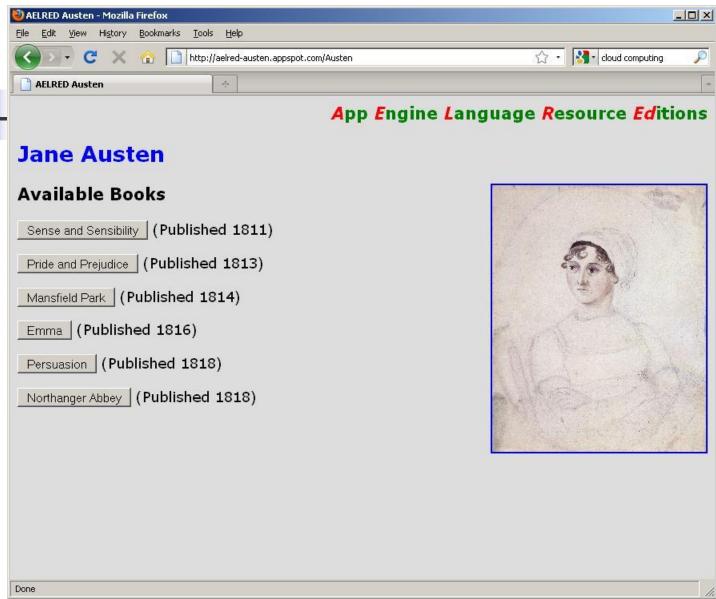
Google App Engine

- "Run your web apps on Google's infrastructure"
 - http://your-app-name.appspot.com
- My web app is AELRED:
 - App Engine Language Resource Editions
 - First version: Jane Austen novels
 - http://aelred-austen.appspot.com

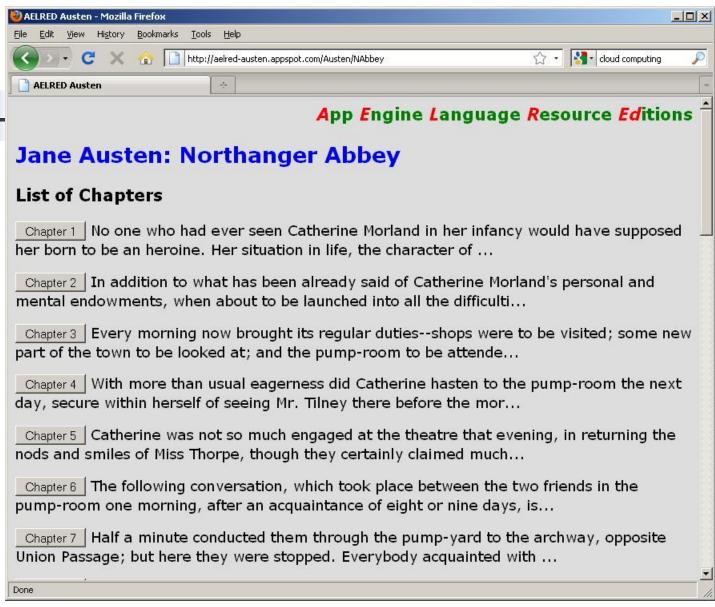




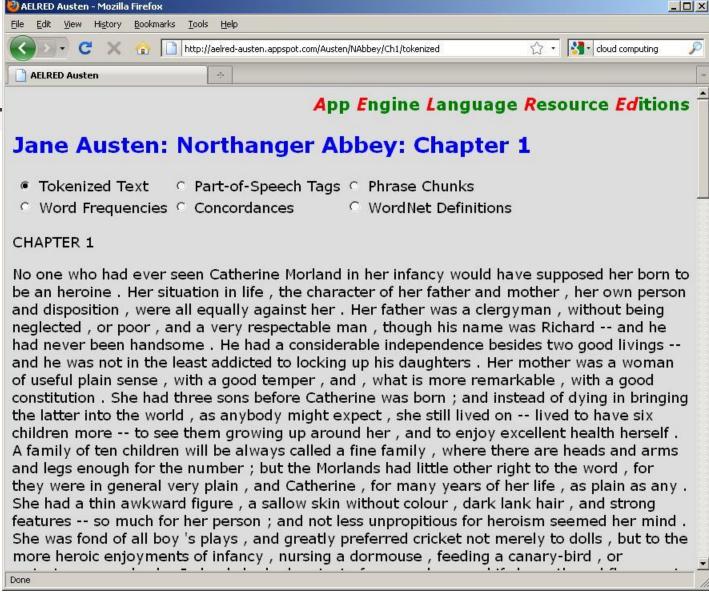




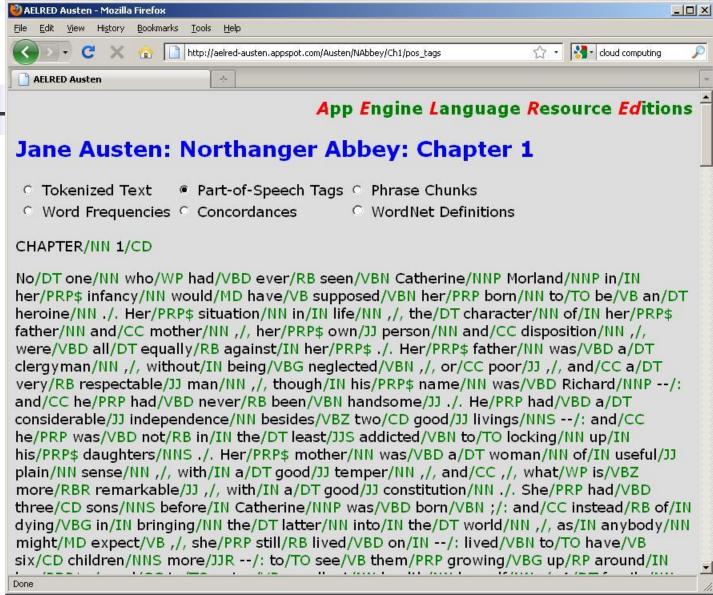




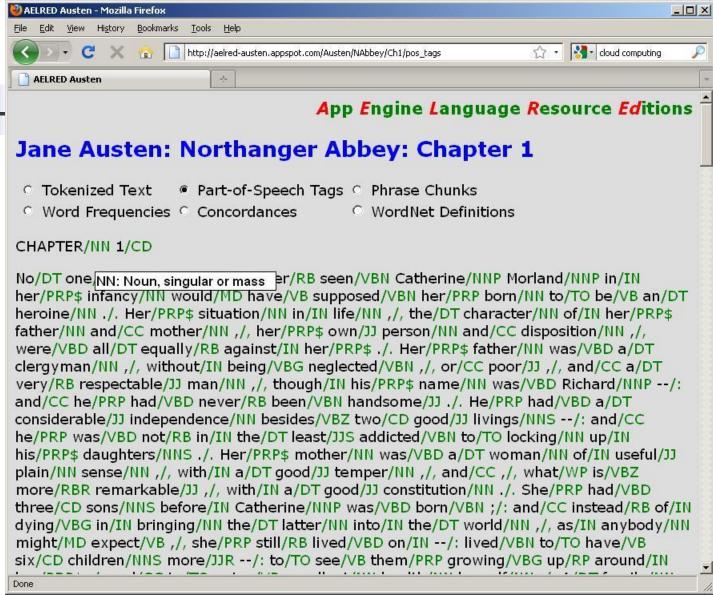


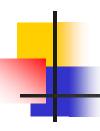


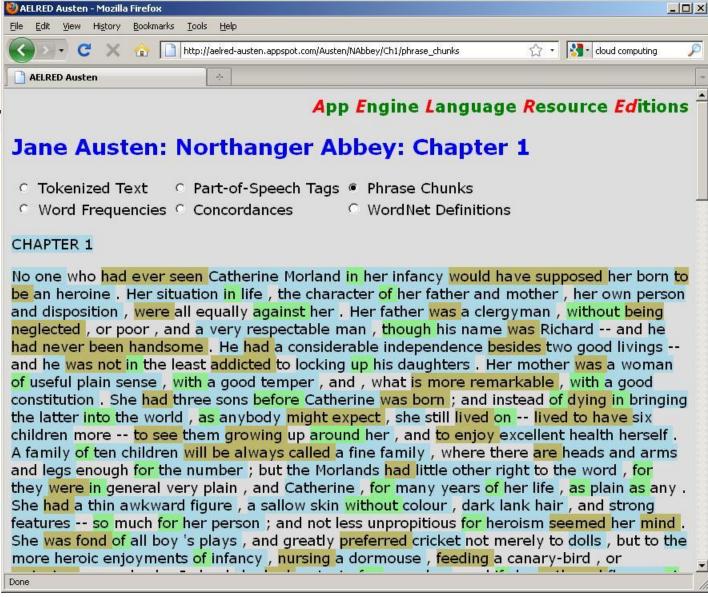




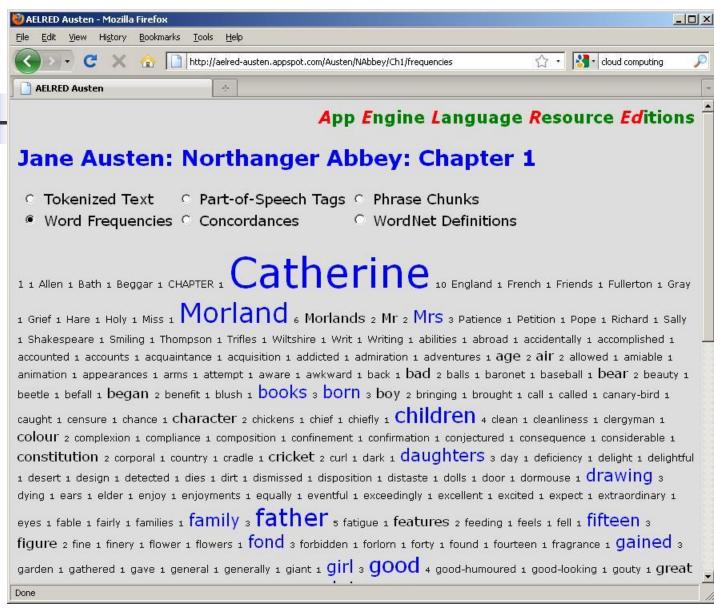


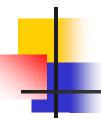


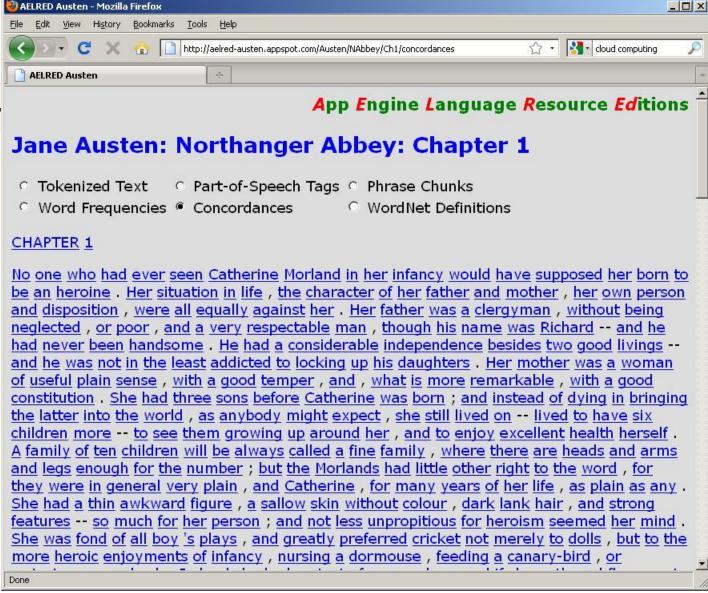




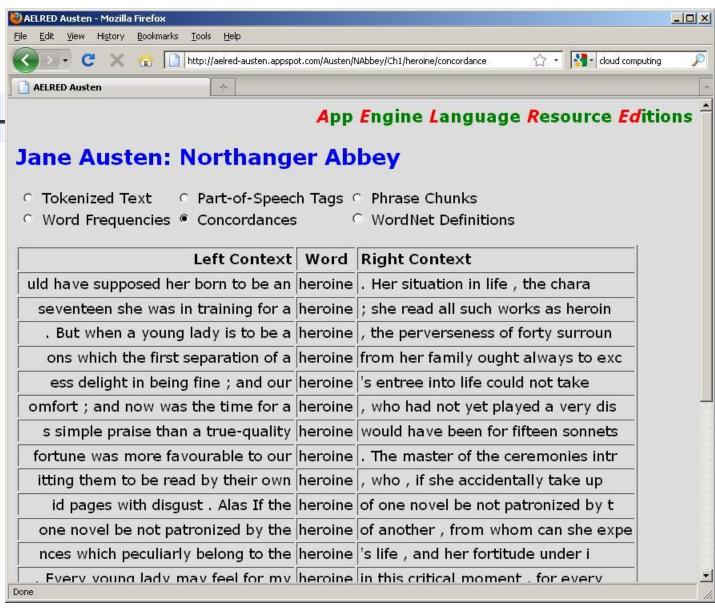




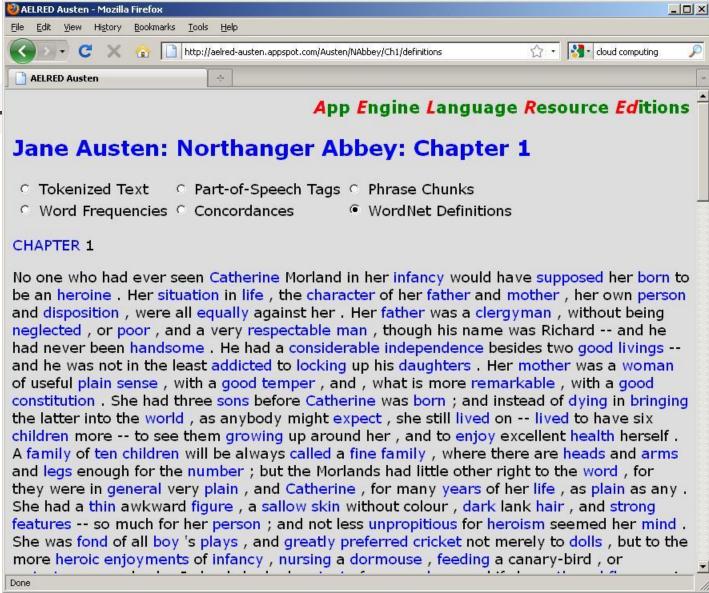




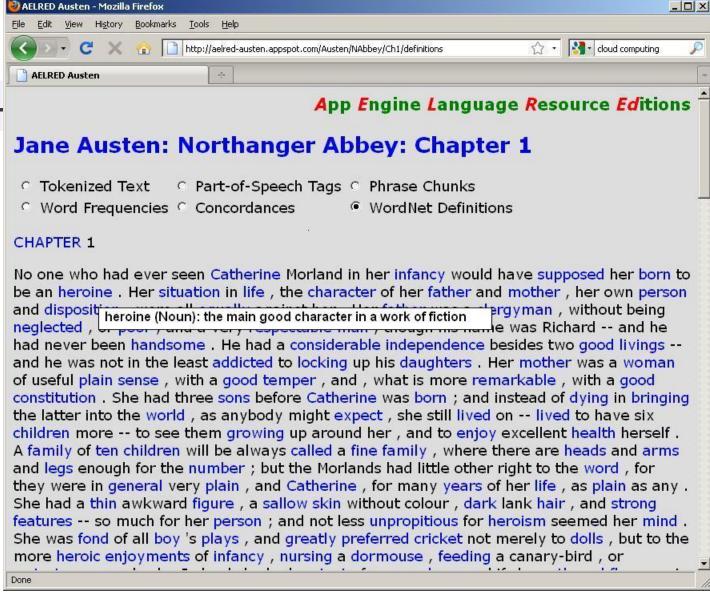








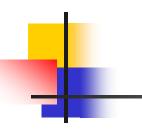


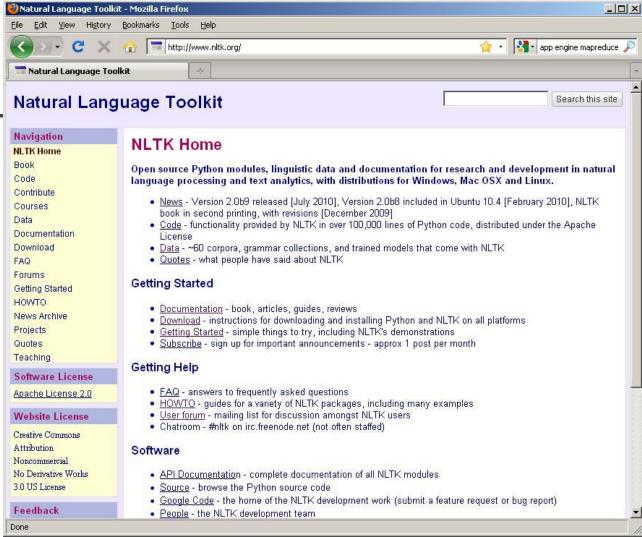




Key Ideas: Easy, Big, Free

- Easy: use Python
 - NLTK Natural Language Toolkit
 - Django HTML Template Engine
- Big: Google's scalable infrastructure
 - BigTable non-relational datastore
 - MapReduce data-intensive processing
- Free: App Engine has free quotas
 - Only pay if high demand for app







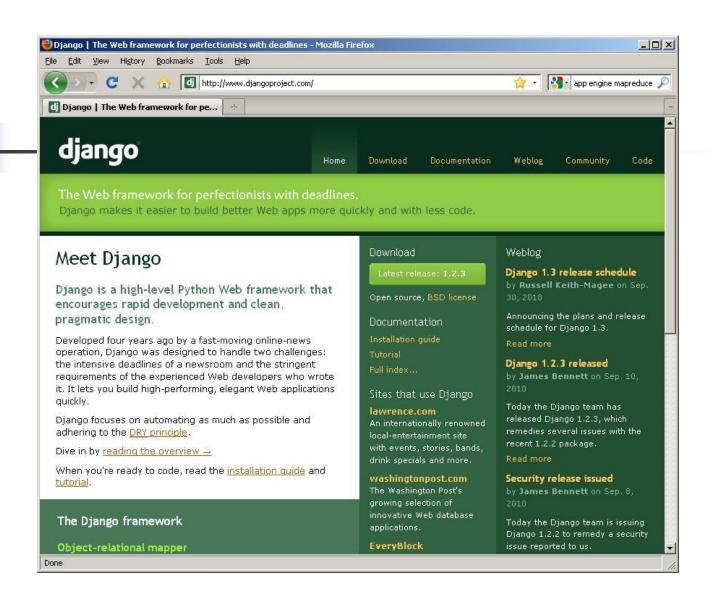
NLTK Natural Language Toolkit

- Open source Python tools
 - Taggers, chunkers, parsers, classifiers ...
- Many major corpora and resources
 - Brown Corpus, Penn Treebank, WordNet ...
- Excellent free online textbook
 - Natural Language Processing with Python
 - Stephen Bird, Ewan Klein, Edward Loper



NLTK and App Engine

- App Engine code must be pure Python
- Normal "import nltk" does not work
 - Some NLTK code is not pure Python
 - E.g. uses Numpy with C for speed
- Use "import aelred" instead
 - Aelred code is pure Python
 - Other customization, e.g. tokenization





Django Web App Framework

- Open source Python
 - Model-View-Controller design pattern
 - Models defined easily by Python classes
- HTML Template Engine
 - Web pages generated using contexts
 - Excellent "template inheritance" facility
- Free online textbook
 - Django: The Book



Google BigTable Datastore

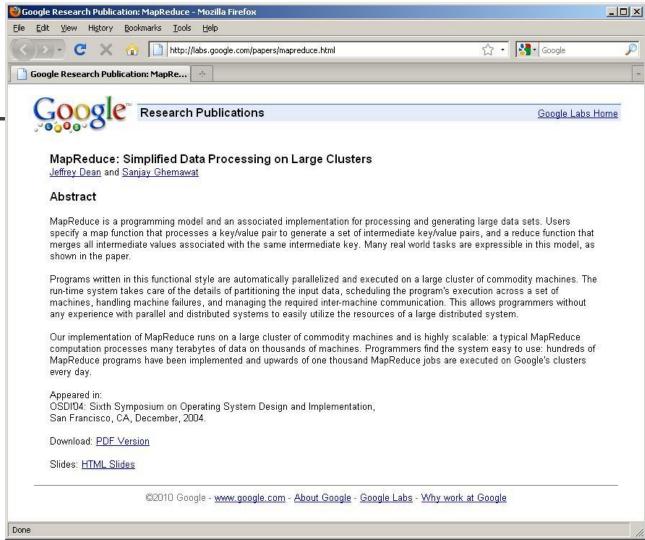
- Non-relational database
 - Different thinking from SQL databases
 - Designed for massive scalability
- My current way of using the datastore:
 - Serialize complex objects to YAML
 - Store/retrieve YAML as big text strings



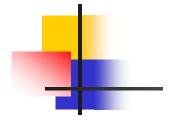
MapReduce Algorithms

- Data-intensive distributed processing
 - Different thinking from usual algorithms
 - Designed for massive scalability
- My current way of using MapReduce:
 - Iterate over all entities in datastore
 - Delete entity, or update and save











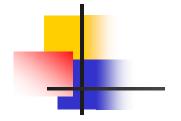
Data-Intensive Text Processing with MapReduce

Jimmy Lin Chris Dyer

Synthesis Lectures on Human Language Technologies

Graeme Hirst, Series Editor

Converginted Material





Introduction to Linguistic Annotation and Text Analytics

Graham Wilcock

Synthesis Lectures on Human Language Technologies

Graeme Hirst, Series Editor