Institution: The University of Edinburgh

Unit of Assessment: B11 — Computer Science and Informatics

Title of case study: The Natural Language Toolkit (NLTK)

1. Summary of the impact

The Natural Language Toolkit (NLTK) is a widely-adopted Python library for natural language processing. NLTK is run as an open source project. Three project leaders, Steven Bird (Melbourne University), Edward Loper (BBN, Boston) and Ewan Klein (University of Edinburgh) provide the strategic direction of the NLTK project.

NLTK has been widely used in academia, commercial / non-profit organisations and public bodies, including Stanford University and the Educational Testing Service (ETS), which administers widely-recognised tests across more than 180 countries. NLTK has played an important role in making core natural language processing techniques easy to grasp, easy to integrate with other software tools, and easy to deploy.

2. Underpinning research

The research described in this case study has been conducted since 2000 at the University of Edinburgh by Professor Ewan Klein (employed from 1/9/1985 to the present) and Dr Johan Bos, Research Fellow (employed from 1/7/2000 to 31/7/2005).

Natural language processing (NLP) covers any computational manipulation of natural language. Many everyday technologies use NLP. These include predictive text on mobile phones, document retrieval via search engines, and spoken dialogue systems for information services. NLP is also an increasingly important component of research in other disciplines. Examples include machine-assisted curation of biomedical text and data exploration in the digital humanities.

A key component of NLP is text understanding. While statistical methods and corpus-based methods are clearly important for this challenge, there is a growing acceptance that a firm foundation in formal theories of natural language semantics is also crucial. This even applies for relatively 'noisy' tasks such as textual entailment [1]. Research into developing semantic frameworks able to deal with discourse phenomena, anaphor resolution, and under-specification has been carried out over a long period at Edinburgh. This research has included innovative approaches to providing computational interpretations of these frameworks [2, 3, 4].

Edinburgh joined the NLTK project in 2004. Research into natural language inference and understanding at the University of Edinburgh has provided the underpinning for all the semantics modules in NLTK.

Logical inference has been made into a relatively tractable component of natural language understanding by treating first-order theorem proving as a callable service, and by pairing theorem proving with model building [5]. NLTK incorporates these techniques with the help of the Prover9 system [4, 6].

3. References to the research

1. J. Bos and K. Markert, *Recognising textual entailment with robust logical inference*, in *Machine Learning Challenges*, MLCW 2005 (J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, eds.), vol. 3944 of LNAI, pp. 404–426, Springer, 2006.

   DOI: *10.1007/11736790_23*

2. J. Bos, *Computational semantics in discourse: Underspecification, resolution, and inference*, Journal of Logic, Language and Information, vol. 13, no. 2, pp. 139-157, 2004.

   DOI: *10.1023/B:JLLI.0000024731.26883.86*

3. E. Klein, *Computational semantics in the Natural Language Toolkit*, in Proceedings of the *Australasian Language Technology Workshop* (ALTW'06) (L. Cavedon and I. Zukerman, eds.), Sydney, pp. 26-41, 2006.

   PDF: *http://www.alta.asn.au/events/altw2006/proceedings/Klein.pdf*

4. D. Garrette and E. Klein, *An extensible toolkit for computational semantics*, in IWCS-8: Proceedings of the Eighth *International Conference on Computational Semantics*, pp. 116-127, Association for Computational Linguistics, 2009.

   WWW: *http://dl.acm.org/citation.cfm?id=1693770*

5. J. Bos, *Exploring model building for natural language understanding*, in *ICoS-4, Inference in Computational Semantics*. Workshop Proceedings (P. Blackburn and J. Bos, eds.), pp. 41-55, 2003.

   WWW: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.1109*

6. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media, 2009. ISBN-10: 0596516495, ISBN-13: 978-0596516499.

   WWW: *http://shop.oreilly.com/product/9780596516499.do*
   HTML: *http://nltk.org/book/* has a freely-available Creative Commons edition of the book.

References [3], [4] and [6] are indicative of the quality of the underpinning research.

3.1. Awards

- EPSRC, Instruction-Based Learning for Mobile Robots, 2000–2003
- Edward Clarence Dyason Universitas 21 Fellowship, 2006
- Google Summer of Code (under the umbrella of the Python Software Foundation), 2008

4. Details of the impact

4.1. Quantitative overview

NLTK is widely used in higher education courses, covering topics such as natural language processing, computational linguistics, empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used in more than 150 different courses, distributed across universities in 23 different countries as reported on the NLTK web site at *https://sites.google.com/site/naturallanguagetoolkit/courses*. Two specific examples involving NLTK's semantics components are Stanford University's Natural Language Understanding course and Robin Cooper's Computational Semantics course at the University of Gothenburg.

4.2. Dissemination of NLTK

In July 2007, the NLTK project leaders (who were in California presenting a three-week course at the LSA Summer School, Stanford) were invited to give a presentation at the Bay Area Python Interest Group (BayPIGgies), held in Google's Mountainview headquarters. This attracted an audience of 70 non-academics and initiated a dialogue with non-academic users of NLTK that continues today.

4.3. Route to impact

NLTK enthusiasts in academia have taken the toolkit on to their subsequent work in industry. In 2007, Nitin Madnani, then a PhD student at the University of Maryland, published an article in ACM *Crossroads* magazine [A] reporting on his positive experiences with using the Natural Language Toolkit. In 2010, he took up a position with the Text, Language and Computation group of the Educational Testing Service (ETS, *http://www.ets.org/*) at Princeton, New Jersey and continued to work with NLTK there. The Educational Testing Service is a non-profit organization, which operates across more than 180 countries. NLTK performs core NLP tasks at ETS and forms an integral part of their TextEvaluator project. Nitin Madnani revised his ACM Crossroads article when at the ETS, making an updated version of his ACM Crossroads paper available on his web page in August 2012 [B].

4.4. Importance of NLTK to the Python community

In November 2012, the Python Software Foundation awarded funding to help port NLTK to Python 3, and made the following comments to explain its decision [C]:

> *For many, NLTK is one of the major remaining roadblocks to Python 3 adoption. As many projects have been ported and many more are working on it, getting NLTK on Python 3 will be huge for the community.*
>
> *…Not only will the NTLK port be a boon to wider Python 3 adoption, but it should provide a good story for others to lean on when porting large codebases, especially when it comes to working with Python 3's Unicode implementation.*

This support led to the release of the NLTK 3.0 alpha version in January 2013.

4.5. Availability and use of NLTK

NLTK code is available under the Apache License, Version 2.0, and the documentation is available under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States license. NLTK is distributed via numerous routes, including

- PyPI — *http://pypi.python.org/pypi/nltk/2.0.4*
- Ubuntu — *http://packages.ubuntu.com/lucid/python-nltk* and
- Github — *http://github.com/nltk/nltk*.

Between 1st January 2008 and end of July 2013, we have evidence of 546,936 downloads of the code (in both source and binary formats), comprising:

- 50,854 downloads from Sourceforge as reported on the web page *http://sourceforge.net/projects/nltk/files/stats/timeline?dates=2008-01-01+to+2013-07-31*

- 217,471 downloads from Google Code as reported on the web page *http://code.google.com/p/nltk/downloads/list* and

- 278,611 downloads from PyPI. This figure is obtained by appending each of the version numbers 2.04, 2.0.3, 2.0.2, 2.0.1, 2.0.1rc4, 2.0.1rc3, 2.0.1rc2-git, 2.0.1rc1, 2.0b9, 2.0b8, 2.0b7, 2.0b6, 2.0b5 and 2.0b4 to the URL *http://pypi.python.org/pypi/nltk/*

These download figures do not include project downloads from Github, since the relevant information is not made available.

4.6. Reach of the NLTK project

The publication of the book *Natural Language Processing in Python* [6] greatly aided the dissemination of NLTK. The book has undergone its third printing. A Japanese translation is available [D]. Independently, a developer in the NLTK community has written a companion NLTK 'cookbook' [E].

The website *stackoverflow.com* is a popular forum where users can post technical questions about software and receive answers from relevant experts. It lists more than 2,400 questions concerning uses of NLTK, the earliest of which is from August 2008 [F].

The mailing lists that are specifically for NLTK users and developers are active and the list maintainers see a steady stream of requests to join these groups. Between 14 December 2011 and 12 December 2012, 964 people joined the NLTK-users mailing list [G], and since 11 November 2011, 240 have joined the mailing list for NLTK developers [H].

5. Sources to corroborate the impact

A. *Getting started on natural language processing with Python*, Nitin Madnani, ACM *Crossroads* magazine, 13(4), June 2007. *http://dx.doi.org/10.1145/1315325.1315330*

B. Revised version of the "Getting started on natural language processing with Python" ACM *Crossroads* paper made available on Nitin Madnani's web page after his move to ETS. Updated in August 2012 to refer to v2.0.2 of NLTK. *http://desilinguist.org/pdf/crossroads.pdf*

C. Details of the Python Foundation award to NLTK *http://pyfound.blogspot.com.es/2012/11/grants-to-assist-kivy-nltk-in-porting.html*

D. NLTK book in Japanese — *http://www.oreilly.co.jp/books/9784873114705/*

E. *Python Text Processing with NLTK 2.0 Cookbook*, Jacob Perkins, PACKT Press — *http://www.packtpub.com/python-text-processing-nltk-20-cookbook/book*

F. Questions about the NLTK software — *http://stackoverflow.com/search?q=nltk*, retrieved 2nd May 2013.

G. The NLTK users mailing list — *http://groups.google.com/group/nltk-users*

H. The NLTK developers mailing list — *http://groups.google.com/group/nltk-dev*

Archive copies of these webpages are available from *http://ref2014.inf.ed.ac.uk/impact/*