**GRAY — Textal**

<2 images>

**Textal: Unstructured Text Analysis Workflows through Interactive Smartphone Visualisations**

Gray, S., Terras, M., Ammann, R. and Hudson-Smith, A.

Working from user location and source document metadata, this paper seeks to understand how users analyse text through mobile devices. It will present our initial findings from the first 18 months of the project and catalogue the next steps for Textal as an open service. We will discuss the technical challenges of creating bespoke analysis workflows as well as providing insight into promoting digital humanities to the general public through mobile platforms. This paper presents a collaborative project between UCL Centre for Digital Humanities (UCL DH) and the Bartlett Centre for Advanced Spatial Analysis (UCL CASA).

Textual analysis aims to return relevant logical data requests to users’ queries from digital documents using natural language and a common vocabulary (Hobbs et al., 1982). Unstructured text analysis still remains a challenge for today’s automated algorithms (Borodkin et al., 2014) and high-performance computing systems as these systems attempt to classify relevant data from human-created datasets. With the exponential increase of user-submitted content on social media sites and the vast amount of information contained on the Internet in various file formats, providing an overview of large text corpuses through automated processes still remains a challenge.

Resources available to researchers, such as TAPoR (Text Analysis Portal for Research) (TAPoR, 2014), programming libraries such as Natural Language Toolkit (NLTK) (Bird, 2006), and a Python library for processing natural language from human language data (NLTK Project, 2014), to name a few, allow researchers to analyse complex or large datasets but often neglect the visualisation aspect of analysis. The creation of data visualisations is often left to the researcher after analysis is completed. These tools provide interfaces via web-based portals, purpose-built desktop applications, or APIs that may be slow, complex, or hard to use, which can often become a barrier for people who may need extensive knowledge before a tool, or API, becomes useful.

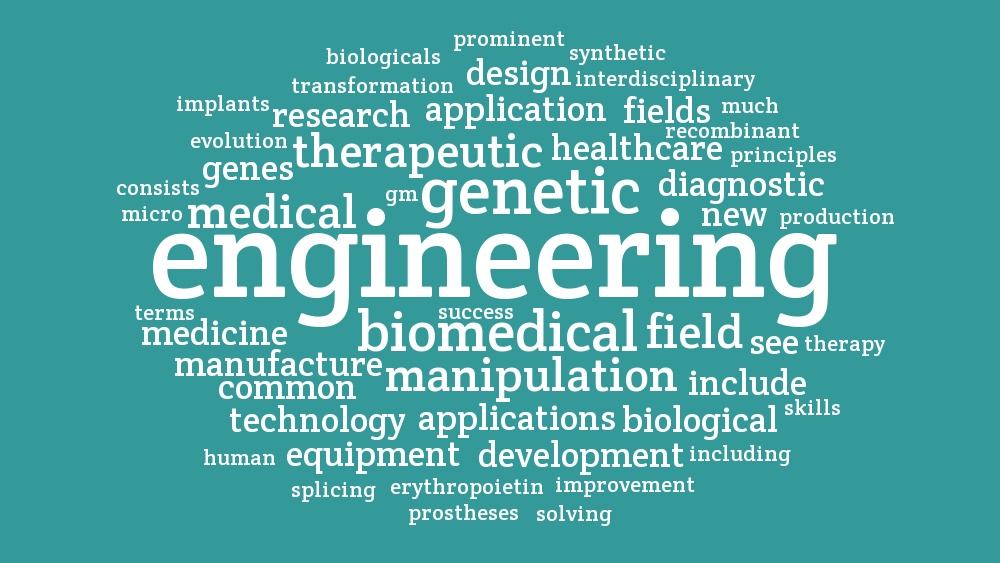


Figure 1. Textal word cloud visualisation of user’s biomedical paper (Cooper, 2014).

Word clouds are graphical representations of textual data depicting the frequency of a given word in relation to other words within a source document (Seifert, 2008). This visualisation method derives from tag clouds, which are weighted lists of keyword metadata, first introduced as a search technique through the photo-sharing web service Flickr in 2004. Word clouds became a widely used technique among researchers analysing large texts after Wordle popularized the automated creation of such visualisations through its online tool (Feinberg, 2009). However, word clouds hide important information about the structure of the original document, context (McNaught and Lam, 2010), and statistics about individual words, which are hidden behind the visualisation.

Textal is a smartphone application that incorporates the visual style of word clouds, the interactivity provided by mobile devices, and the power of natural language processing workflows into a single easy-to-use tool (Figure 2). The application utilises the ubiquitous ‘pinch and zoom’ feature of touchscreen mobile interfaces, allowing users to explore the data behind the word cloud by touching individual words. Textal hides the complexity of unstructured text analysis through various cloud endpoints and distributed workflows. These workflows allow users to create and interact with word clouds in real time while the statistics of the document are processed on cloud servers in the background. As demand increases on the system, the processing engine can scale dynamically using on-demand computing available from cloud platforms as well as local virtualisation hardware to reduce latency.

The system was created in response to an increasing trend towards emergent mobile and web-based technologies and seeks to understand how mobile technologies can be harnessed within digital humanities research. Textal is the first attempt to build a standalone application that brings together tools and workflows for use by researchers to analyse unstructured text as well as giving the general public a tool to easily create word clouds. The app aspires to tap the potential for public engagement by capitalising on the popularity of the word clouds as a ‘gateway’ to text analysis (Meloni, 2009).



Figure 2. Textal iOS Application main screen, word cloud visualisation showing underlying statistics.

Textal was launched on the iOS App Store on 14 June 2013, along with the associated website http://www.textal.org (Textal, 2015), and analysis of the data has been compiled up until 24 February 2015, the first one year, eight months, and 10 days (640 days) of the service. In this time, Textal has been downloaded onto 5,749 devices in five separate territories and has been translated into six languages. Users have created over 3,000 word clouds with over 100 million words processed, showing a need for such a tool from both the general public and as an analysis tool for researchers. This paper discusses not only the bespoke creation of a distributed workflow engine for text analysis but also the insights gained from the usage data of the application’s global user base and the textual data crowdsourced from users of the application. We will explore the use of the Textal API within the wider context of research into real-time social media analysis using distributed systems and provisioned cloud computing being carried out at the Bartlett Centre for Advanced Spatial Analysis. We will also address the ramifications for the academic field in embracing mobile platforms, such as pace of technology upgrades, uptake, and dissemination in a crowded mobile application marketplace.

**References**

**Bird. S.** (2006). NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Association for Computational Linguistics, Sydney, July, 69–72.

**Borodkin, A., Lisin, E. and Strielkowski, W.** (2014). Data Algorithms for Processing and Analysis of Unstructured Text Documents. *Applied Mathematical Sciences,* **8**(25): 1213–22.

**Cooper, L.** (2014). Textal—Word Cloud—Biomedical Paper. http://textal.org/gallery/200e0cf6d177.

**Feinberg, J.** (2009). Wordle. http://www.wordle.net/.

**Hobbs, J. R., Walker, D. E. and Amsler, R. A.** (1982, July). Natural Language Access to Structured Text. In *Proceedings of the Ninth Conference on Computational Linguistics*, Prague, 5–10 July 1982. Amsterdam: North-Holland, pp. 1:127–32.

**McNaught, C. and Lam, P.** (2010). Using Wordle as a Supplementary Research Tool. *Qualitative Report,* **15**(3): 630–43, http://www.nova.edu/ssss/QR/QR15-3/mcnaught.pdf.

**Meloni, J.** (2009). Wordles, or the Gateway Drug to Textual Analysis. ProfHacker. *Chronicle of Higher Education*, 21 October, http://chronicle.com/blogs/profhacker/word les-or-the-gateway-drug-to-textual-analysis/22781.

**NLTK Project.** (2014). Natural Language Toolkit Library. http://www.nltk.org.

**Seifert, C., Kump, B., Kienreich, W., Granitzer, G. and Granitzer, M.** (2008). On the Beauty and Usability of Tag Clouds. In *Proceedings of the 12th International Conference on Information Visualisation*, IEEE, London, 8–11 July 2008, pp. 17–25.

**TAPoR—Text Analysis Portal for Research.** (2014). http://www.tapor.ca.

**Textal Website.** (2014). http://www.textal.org.

**UCL CASA.** (2014). Bartlett Centre for Advanced Spatial Analysis. http://casa.ucl.ac.uk.

**UCL DH.** (2014). Centre for Digital Humanities. http://www.ucl.ac.uk/dh.