Smelly London: visualising historical smells through text-mining, geo-referencing and mapping

Deborah Leem
d.leem@wellcome.ac.uk
Wellcome Trust, United Kingdom

Overview

Wellcome Collection is one of the world's major resources for the study of health and histories. Over the past few years Wellcome have been developing a world-class digital library by digitising a substantial proportion of their holdings. As part of this effort, approximately 5,500 Medical Officer of Health (MOH) reports for London spanning from 1848-1972 were digitised in 2012. Currently Wellcome holds the most comprehensive digital collection of the London MOH reports. Since September 2016 Wellcome have been digitising 70,000 more reports covering the rest of the United Kingdom (UK).

The MOH reports were published annually by the Medical Officers of Health employed by local authorities across the UK. These reports provided vital statistics and a general statement on the health of the population. MOH reports concentrated on reporting infectious diseases and resolving the problems as well as covering other areas of social responsibilities. (Chave, 1987) They have been long regarded as an important source for the 19th and 20th century history of Public Health and stem from reaction to infectious disease in the mid-19th century. Although there were attempts at standardisation, the reports display each MOH's interest, idiosyncrasies and particular strengths. Therefore, they also provide a particular perspective on the everyday lives of Londoners over several generations. No digital techniques have yet been applied successfully to add value to this very rich resource.

As part of the Smelly London project, the OCR-ed text of the MOH London reports has been text-mined using the Python programming language. Through text mining we produced a geo-referenced dataset containing smell categories for visualisation to explore the data. At the end of the Smelly London project the MOH smell data will also be available through other platforms such as Good City Life and Layers of London. This will allow the public and other researchers to compare smells in London from the 19th century to present day. This has the further potential benefit of engaging with the public. This is a collaborative, interdisciplinary project which will allow us to enhance and demonstrate the capabilities of innovative text mining tools we design to allow the automatic extraction of information from OCR-ed text. This paper presents the intended aims of the project; how this was achieved; an analysis of the findings; an interactive map of the results and a browser game of smells and disease.

Data and visualisation

As Roy Porter famously remarked that "todays history comes deodorised", sensory history is a relatively new historical approach. Historians rarely provide us an opportunity to hear, taste or smell the past. Medical historians have incorporated some aspects of sensory history into their research and explored the past belief that bad smells were causes of disease. However, there is very little research carried out covering this period.

Furthermore, smell has a great influence over how we perceive places and contributes to the construction of a place's identity (Quercia et al., 2015). During the 19th century the paranoia surrounding smells associated with poor hygiene heightened in many European cities (Reinarz, 2014). The Great Stink of 1858 resulted in the discussion of moving Parliament outside London for example. Despite the rise of germ theory (Pasteur and Koch) in the 1880s, concerns with disease-causing miasma (smells) did not disappear entirely. The MOH reports are one of the richest available sources on local public health administration and patterns of disease.

We enriched the text-mining pipeline with Natural Language Processing (NLP), including lemmatisation and part-of-speech tagging. The first iteration of the project has a feature to identify the category of the smells found by using a mapping table to work out the most common smell types. This step complements the close reading analysis and enables us to scale up the amount of information extracted from the texts. Our next research plan is to work on automatic identification of smell terms based on their contextual features to discover new categories that escaped previous classifications. This will allow us to identify smell categories in a data-driven fashion.

As the data becomes more structured, they can be more readily overlaid with other maps and images such as Charles Booth's London Poverty Map and 19th

century disease maps. Having multiple layers will enable us to run various comparisons and assess if there are any correlations between smells and diseases as well as links to the socio-economic identity of areas in London.

During the first phase of the project we created a smelly map based on the number of smell hits to visualise the first set of results.

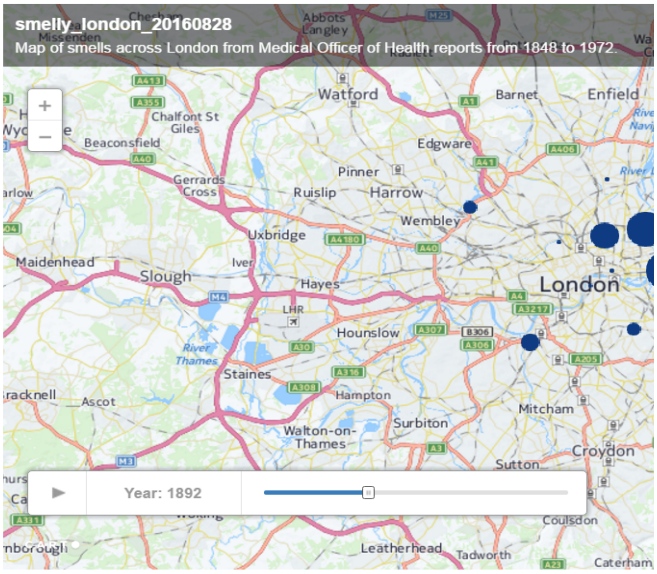

Figure 1. Smelly Map of London showing all smells

From the list of the existing London local authorities for the MOH reports we compiled, the geographic coordinates of present-day equivalents were extracted using an API. For the places that did not exist in the API, we manually added the geographic coordinates from Wikipedia. On the map each of the points marks the number of smells occurring at the centroid of each of the locations. We grouped the number of smells into sets of ten (e.g. 1- 10, 11 - 20) to avoid having giant points on the map for the places where there are almost 100 smells recorded. Finally, the map scrolls through the years. The data displayed in the mapping visualisation was obtained using text-mining via Python scripting. Python was the language of choice due to its high productivity rate and the fact that there are a large amount of third party libraries that offer highly useful functionality with just a few lines of code. For example, NLTK is a popular Python set of libraries that can achieve advanced NLP.

The next generation of map we produced during the second phase displays different smell categories that are colour-coded. The smell categories used for this map are Sewer; Waste-rubbish; Waste-excrement; Thames; Water; Food; Trade; Animal; Factory-fuel; Disinfectant; School; Air; Decomposition; Habitation; and Absence of smell. These categories were obtained through manual inspection of the data produced from searching for sentences containing smell-related words. In our codebase, we first analysed 5500 MOH London reports to find sentences that contained smell related words. Once the sentences were further analysed and categorised manually, the results were stored down into a local database by year, borough and a unique ID programmatically.

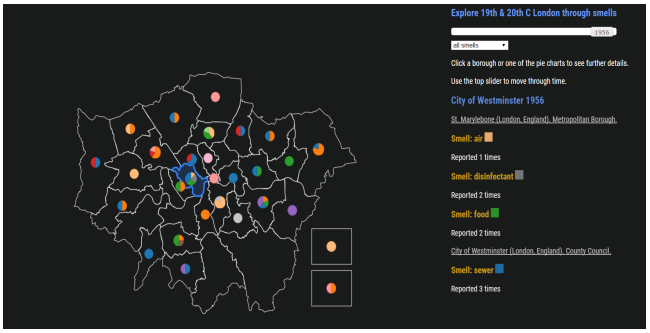

Figure 2. Smelly map of London showing smell categories

Computer programming can be used to perform tasks thousands of times faster than humans. In the Python code written to extract the data from the MOH reports, parallel processing was employed to speed up the running time of the program. Inside a computer there is a CPU which runs the tasks given by the program. Modern CPUs have multiple cores which allows the calculations to be run concurrently. In our project the CPU had four cores which allowed the running time of the program to be shortened by as much as three times. The next objective for the project is to scale up the size of the text-mining from 5,500 reports to over 70,000 reports covering the entire UK. In order to process such large datasets we are investigating the use of distributed computing resources such as Amazon Web Service (AWS). The code written for this project has been made open source under the MIT license along with documentation so that other programmers or researchers can use the codebase in their own text mining projects. The code has already been used in another project at Wellcome to investigate the idea of women's right to work during the 19th and 20th century London

Vision

The Smelly London project aims to bring together historical data with modern digitisation and visualisation to give us a unique, revealing and visceral glimpse into a London of the past and what it tells us about

London today. Analysing the MOH reports tells the intimate narratives of the everyday experiences of 19th and 20th century Londoners through the 'smellscape'. The Smelly London project provides a great opportunity to demonstrate how new knowledge and insights have risen from the use of powerful digital applications. This project will produce models that facilitate new kinds of humanities research. All outputs generated from the project will be open access and open source. Our data is available in a public repository on GitHub and other platforms.

Bibliography

Bynum, W. F. (1993) *Medicine and the Five Senses*, Cambridge; New York: Cambridge University Press.

Chave, S.. *Recalling the Medical Officer of Health: Writings by Sydney Chave,* London: King's Fund Publishing Office.

Classen, C, et al. (1994) *Aroma: The Cultural History of Smell*, London; New York: Routledge.

Cockayne, E. (2007). *Hubbub: Filth, Noise and Stench in England 1600 – 1700*, New Haven [Conn.]; London: Yale University Press.

Corbin, A. (1986) *The Foul and the Fragrant: odor and the French Social Imagination*, Leamington Spa: Berg.

Dobson, M. (1994). Malaria in England: A Geographical and Historical Perspective, *Parassitologia* 36 (1994): 35-60

Dobson, M. (1980)"Marsh fever"-The geography of malaria in England, *Journal of Historical Geography* 6(4) : 357-89.

Jenner, M. (2011) 'Follow your nose? Smell, smelling, and their histories', *The American Historical Review*, 116, 350

Quercia, D., Schifanella, R., Aiello, L. M., McLean, K. (2015). Smelly Maps: The Digital Life of Urban Smellscapes, *Proceeding of the 9th International AAAI Conference on Web and Social Media (ICWSM)*.

Reinarz, J. (2014) *Past Scents: Historical Perspectives on Smell,* Chicago: University of Illnois Press, 2014.

Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, et al. (2016) Text Mining the History of Medicine, *PLoS ONE* 11(1): e0144717. doi:10.1371/journal.pone.0144717