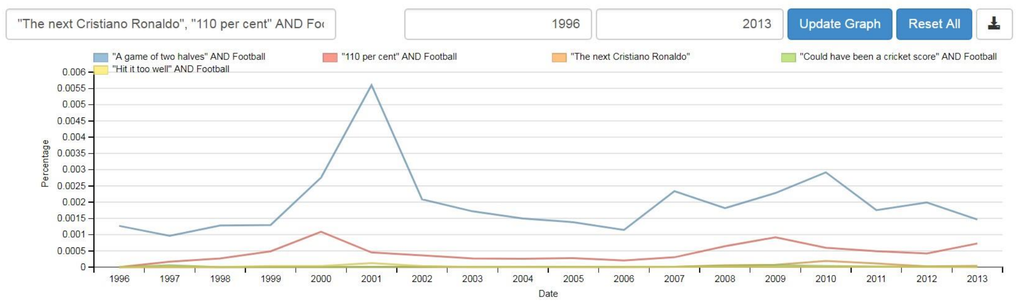
From the sidelines to the archived web: What are the most annoying football phrases in the UK?

Introduction

The UK Web Archive in partnership with JISC and the Internet Archive acquired the JISC UK Web Domain Dataset (1996–2013). In 2015, the UK Web Archive as part of the Big UK Domain Data for the Arts and Humanities project (BUDDAH), launched a historical search engine service called Shine which sits over the JISC dataset and links out to the Internet Archive. The Shine interface searches across 3,520,628,647 distinct records from .uk domain that were captured by the Internet Archive from January 1996 to the 6th April 2013. Shine can be interrogated in two ways, either through the faceted Search function or through the Trends function which produces a graph based on the percentage of the resources archived over a given year between 1996 and 2013. This could be used to study wider patterns in the collection based on a variety of subjects in almost any language. Through the lens of football (soccer), this research assesses how useful the Trends function on the Shine interface is to determine the popularity of a sample of selected phrases on the archived .uk web. Although the methods could be applied to any subject area, it is hoped that the findings from this study will encourage further research in sports and linguistics using the UK Web Archive.

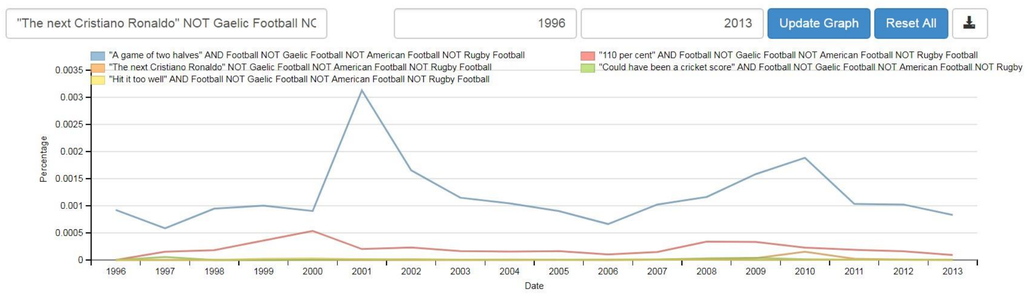
Shine Trends

Men's Football Clichés Search One



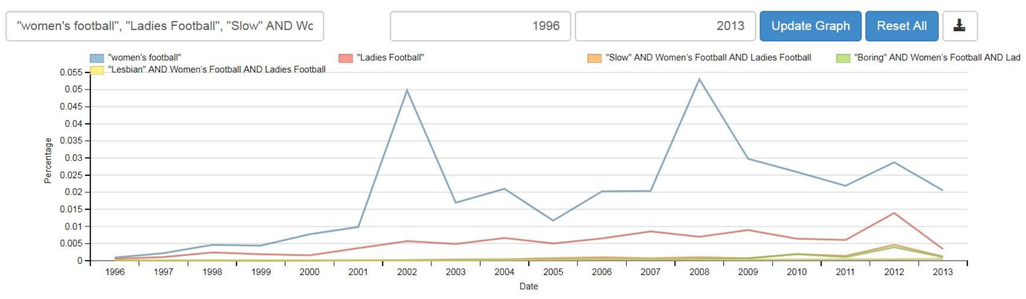
Search query - "The next Cristiano Ronaldo", "110 per cent" AND Football, "Hit it too well" AND Football, "A game of two halves" AND Football, "Could have been a cricket score" AND Football.

Men's Football Clichés Search Two



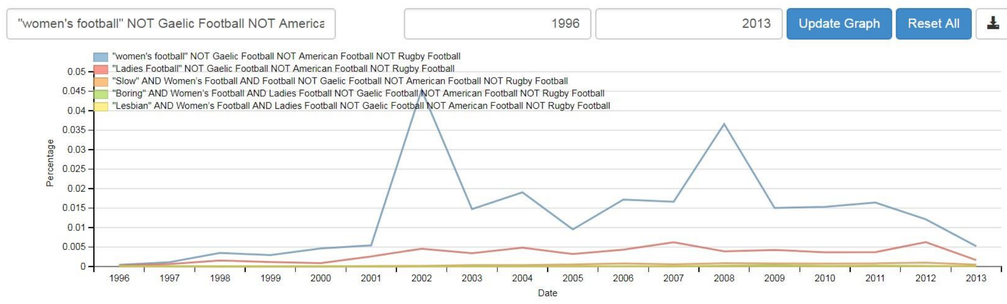
Search query - "The next Cristiano Ronaldo" NOT Gaelic Football NOT American Football NOT Rugby Football, "110 per cent" AND Football NOT Gaelic Football NOT American Football NOT Rugby Football, "Hit it too well" AND Football NOT Gaelic Football NOT American Football NOT Rugby Football, "A game of two halves" AND Football NOT Gaelic Football NOT American Football NOT Rugby Football, "Could have been a cricket score" AND Football NOT Gaelic Football NOT American Football NOT Rugby Football

Women's Football Clichés Search One



Search query - "women's football", "Ladies Football", "Slow" AND Women's Football AND Ladies Football, "Lesbian" AND Women's Football AND Ladies Football, "Boring" AND Women's Football AND Ladies Football

Women's Football Clichés Search Two



Search query - "women's football" NOT Gaelic Football NOT American Football NOT Rugby Football, "Ladies Football" NOT Gaelic Football NOT American Football NOT Rugby Football, "Slow" AND Women's Football AND Football NOT Gaelic Football NOT American Football NOT Rugby Football, "Lesbian" AND Women's Football AND Ladies Football NOT Gaelic Football NOT American Football NOT Rugby Football, "Boring" AND Women's Football AND Ladies Football NOT Gaelic Football NOT American Football NOT Rugby Football

References

Helena Byrne, What do you think are the most annoying phrases to describe women's football? https://footballcollective.org.uk/2018/05/18/what-do-you-think-are-the-most-annoying-phrases-to-describe-womens-football/

Niels Brügger, *The Archived Web, Doing History in the Digital Age*. 2018. 124

Andrew Jackson, Building a 'Historical Search Engine' is no easy thing. https://britishlibrary.typepad.co.uk/webarchive/2015/02/building-a-historical-search-engine-is-no-easy-thing.html

Andrew Jackson, Introducing SHINE 2.0 - A Historical Search Engine. http://blogs.bl.uk/webarchive/2016/02/updating-our-historical-search-service.html

Jane Winters, 'Web Archives for Humanities Research: Some Reflections', in *The Web as History*, eds. Niels Brügger and Ralph Schroeder (London: UCL Press, 2017), 242. https://www.uclpress.co.uk/products/84010

Conclusion

When working with big data sets the biggest issue is retrieving precise results to research questions. It has been noted that there needs to be a shift in how we perceive an acceptable answer to research questions, 'with big data, we'll often be satisfied with a sense of general direction rather than knowing a phenomenon down to the inch, the penny, the atom'. Consequently, it is impossible to give a definitive answer to 'what are the most annoying football phrases on the archived UK web?' Based on the simple search strategy above the most annoying phrase in the men's category is 'a game of two halves' and the women's is 'women's football'. If we change the search parameters to include the string 'NOT Gaelic Football NOT American Football NOT Rugby Football', then the results are the same in both categories but generate fewer results. However, this does not take into account any duplications that might exist in the collection or the fact that most news websites will have a tag for the football section of their website. As the phrase 'a game of two halves' is generic and used in different ways, many results would be discussing something else but as football is also on that page it would come back as a valid result. Therefore, as long as the search parameters are clearly outlined then the results to a search query should be seen as valid.

UKWA
UK WEB ARCHIVE

@UKWebArchive UKWA UK WEB ARCHIVE www.webarchive.org.uk/blog www.webarchive.org.uk/shine data.webarchive.org.uk/opendata/