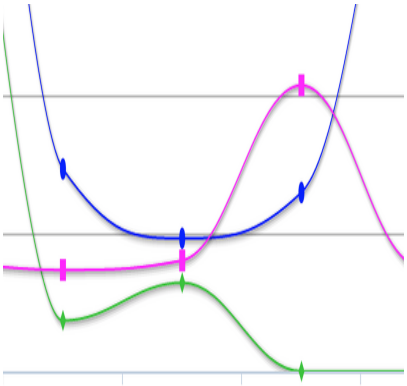


Big corpus

Description

default watermark

UK publishers produce over 180,000 books each year. (About one third are in digital formats.) So thatâ??s a lot of words, even before the outputs of other countries are taken into account, and all the other words generated online â?? self published, or unpublished â?? and journal, magazine and newspaper articles.

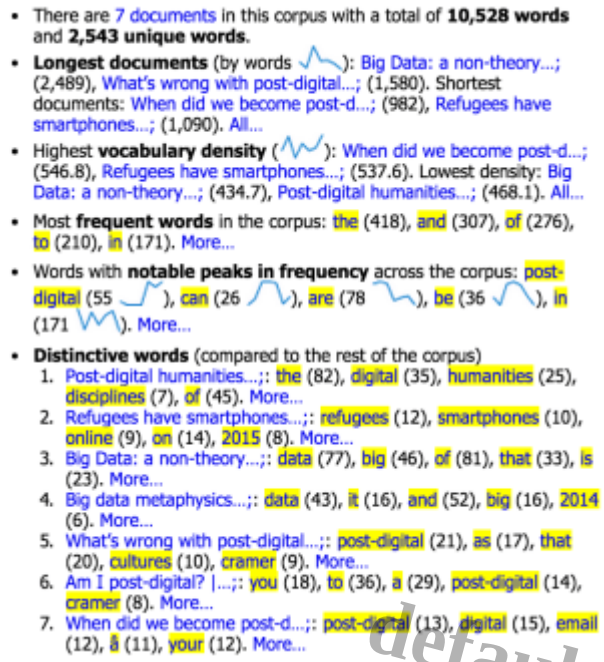
These large text corpuses are more than big data, but can be treated as such â?? counted, mined, probed, analysed, compared, correlated and turned into tables, graphs and network diagrams, without the need for anyone to understand any of it.

More precisely, scholars can use computer programs to transform literary content into different formats in order to understand it better â?? or at least differently. Thatâ??s ***distant* reading**, as opposed to *close* reading. The scholar stands back as if from afar and reviews a whole corpus (collection) of works, and combinations of corpuses. Itâ??s less about singular texts, and more about whole collections (e.g. the complete works of William Shakespeare, all nineteenth century English novels, or the Hansard Reports).

Literary theory

Kathryn Schulz in the *New York Times* (2011) is suitably skeptical about this kind of study. Franco Moretti of the Stanford Literary Lab hopes to find the â??unified theory of plot and styleâ?• (229), as if gathering data from the natural world. Schulz makes the obvious point that literary data is created â??by design,â?• and not subject to the independent, distant readings science claims to make of natural phenomena. So dispassionate analysis of texts can only get us so far before we have to commit to the meaning of what it is we are reading, or donâ??t have time to read.

As a trial I ran my last 7 blog posts through the free-to-use voyant-tools.org for analysing corpuses of texts. Hereâ??s some of what it comes up with.



- There are 7 documents in this corpus with a total of **10,528 words** and **2,543 unique words**.
- **Longest documents** (by words): Big Data: a non-theory...; (2,489), What's wrong with post-digital...; (1,580). Shortest documents: When did we become post-d...; (982), Refugees have smartphones...; (1,090). All...
- Highest **vocabulary density** ( ): When did we become post-d...; (546.8), Refugees have smartphones...; (537.6). Lowest density: Big Data: a non-theory...; (434.7), Post-digital humanities...; (468.1). All...
- Most **frequent words** in the corpus: the (418), and (307), of (276), to (210), in (171). More...
- Words with **notable peaks in frequency** across the corpus: post-digital (55), can (26), are (78), be (36), in (171). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. Post-digital humanities...;: the (82), digital (35), humanities (25), disciplines (7), of (45). More...
  2. Refugees have smartphones...;: refugees (12), smartphones (10), online (9), on (14), 2015 (8). More...
  3. Big Data: a non-theory...;: data (77), big (46), of (81), that (33), is (23). More...
  4. Big data metaphysics...;: data (43), it (16), and (52), big (16), 2014 (6). More...
  5. What's wrong with post-digital...;: post-digital (21), as (17), that (20), cultures (10), cramer (9). More...
  6. Am I post-digital? |...;: you (18), to (36), a (29), post-digital (14), cramer (8). More...
  7. When did we become post-d...;: post-digital (13), digital (15), email (12), â (11), your (12). More...

default watermark

Screen Shot 2015-10-03 at 12.32.06

The postings are ordered, so I guess thereâ??s some sense here to the idea of a trend. I look forward to discovering more, but Iâ??m reluctant to commit whole manuscripts to an online text analysis tool. At present I donâ??t think automated text analysis provides a substitute for reading, or vicarious reading through other readersâ?? interpretations.

For the interpretive scholar any text operates at a distance anyway. See posts tagged hermeneutics.

Bibliography

- Moretti, Franco. 2013. *Distant Reading*. London: Verso
- Schulz, Kathryn. 2011. What is distant reading? *The New York Times*, (June 24) online.

Notes

- Voyant Tools is a web-based reading and analysis environment for digital texts http://voyant-tools.org
- UK Book Industry Statistics (2014) are from the Publishers Association.
- Iâ??m grateful to Fabio Ciotti for alerting us to the idea of distant reading at the recent Digital Humanities Autumn School at Trier.

Category

1. Research related

Tags

1. Big Data
2. digital humanities

3. reading
4. text
5. writing

Date Created
October 3, 2015
Author
rcoyne99

default watermark