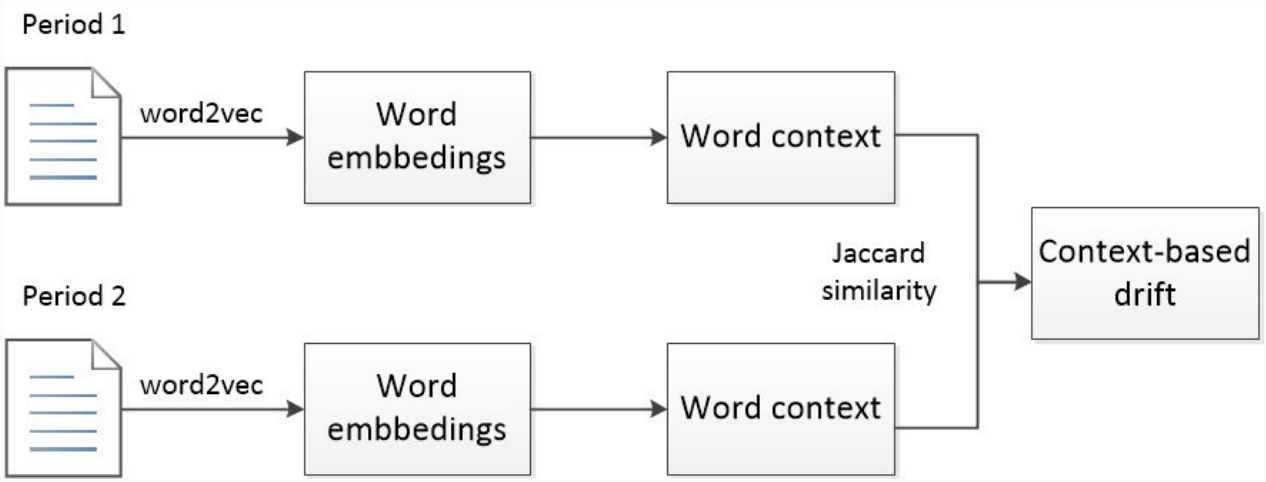
Study concept drift in 150-year English literature

Ruiyuan Li, Pin Tian, and Shenghui Wang

University of Twente

1st Workshop on AI + Informetrics (AII2021) at the iConference2021

Detecting context-based drift



Dataset

33679 English books from **Gutenberg Project**.

There is no specific publication time, so we use

Publication time

=(the author's birth year + death year)/2

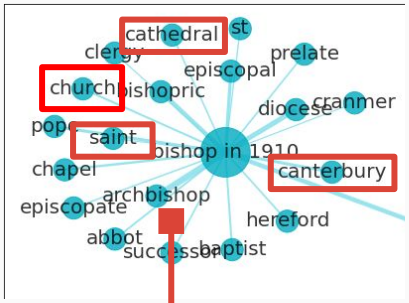
Group	Time span	Book
1800	1790–1810	1230
1830	1820–1840	2214
1850	1840–1860	3712
1870	1860–1880	6595
1890	1880–1900	8661
1910	1900–1920	7721
1930	1920–1940	1734
1950	1940–1960	1926

Table 1. Number of Books in each period

Example: Bishop

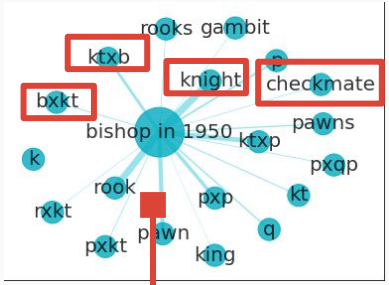
The first generally recognized world championship

“Cathedral”,”church”, ”saint”,”Canterbury”

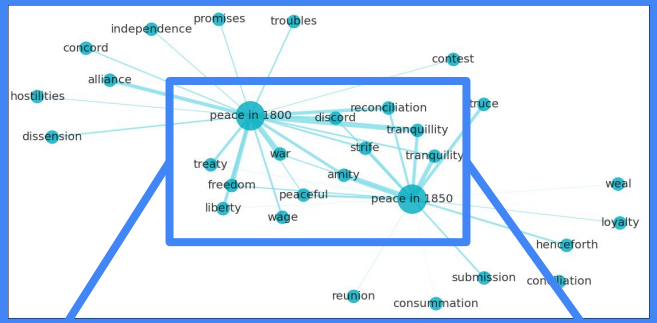


Birth of FIDE's World Championship cycle

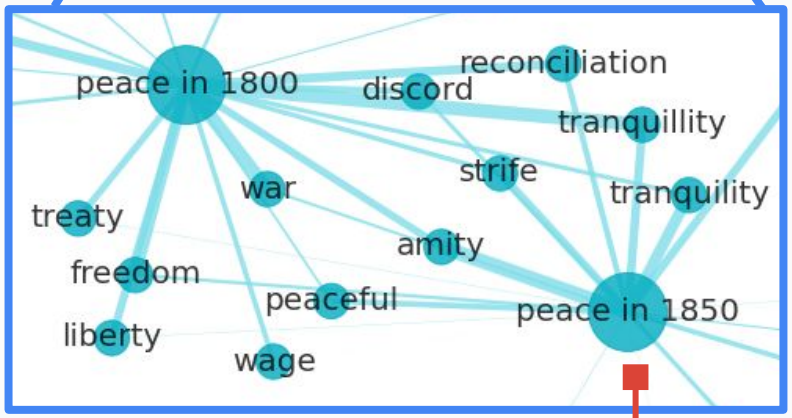
Chess notation:“ktxb”,”bxkt” “knight”,”checkmate”

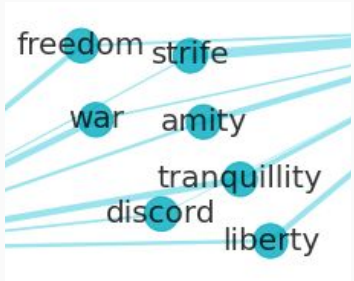


19th 1886 20th 1900-1920 1948 1940-1960 21th



Visualisting drift





Number of Same words: 12 7 4

1850 1910 1950

Future work

- Apply dynamic word embeddings to detect semantic drift

- Work with more recent corpus(e.g., newspaper, twitter) to identify more modern drift

- Work with scientific publications and study the drift in scientific concepts

Study concept drift in 150-year English literature

Q&A

Ruiyuan Li*, Pin Tian, and Shenghui Wang

University of Twente

* r.li-3@student.utwente.nl