The University of Manchester

AHA! Workshop on Information Discovery in Text - Dublin, 23rd August 2014

Mining temporal footprints from Wikipedia

Michele Filannino*, Goran Nenadic

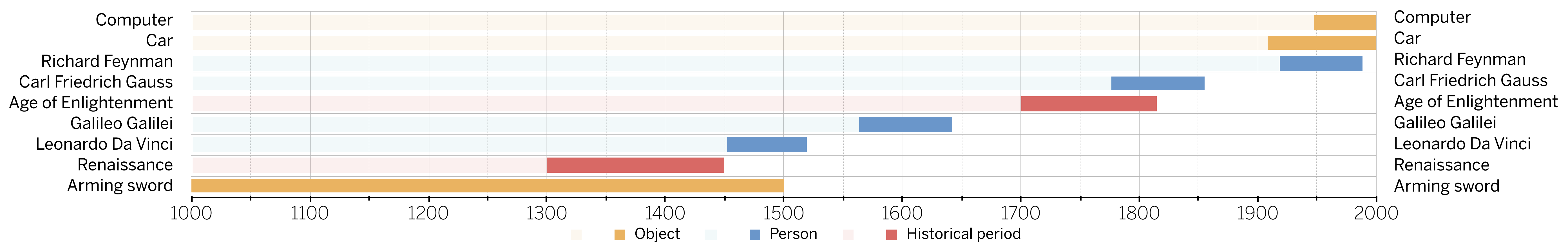
School of Computer Science
University of Manchester, UK

*Presenting author, e-mail: filannim@cs.man.ac.uk

http://www.cs.man.ac.uk/~filannim

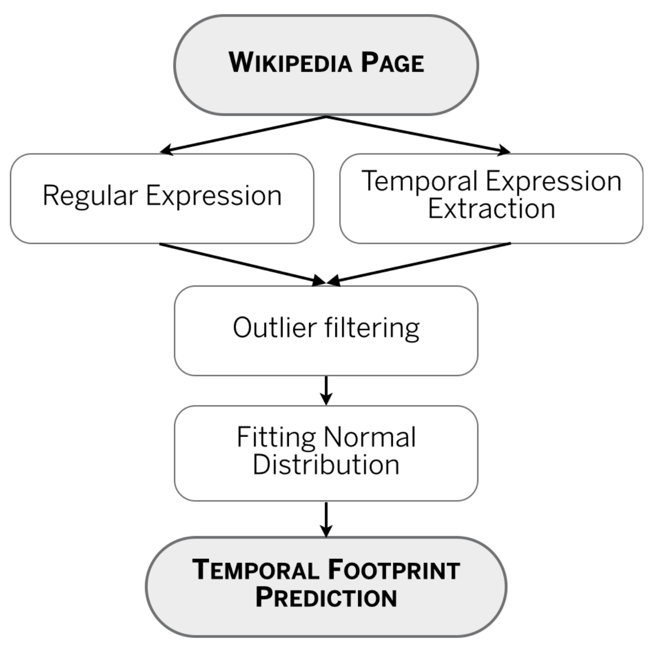
Temporal Footprint

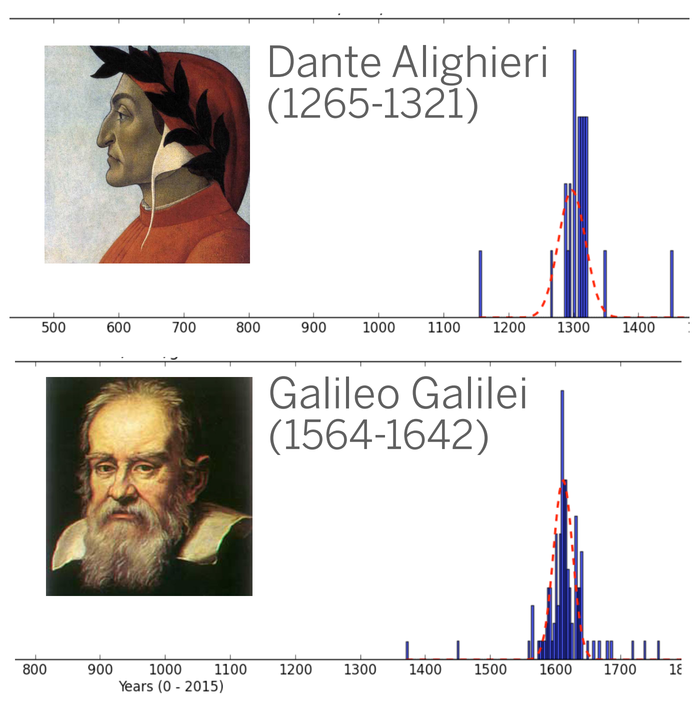
Temporal footprints are **time-line periods** that are associated to the existence of specific concepts. For example, the temporal footprint of people lies between their birth and death, whereas the temporal footprint of a business company lies between its constitution and extinction.



Methodology

We propose to predict footprint's **lower and upper bound** using temporal expressions appearing in the text. The approach has three steps: (I) extracting mentions of **temporal expressions**, (II) **filtering outliers** from the previously obtained probability mass function, and (III) fitting a **normal distribution** to this function.





We used DBpedia to obtain **228,824** Wikipedia pages about people born since 1000AD along with their birth and death dates.

We experimented with the following settings:

(A) **TEE RegEx:** extracts all possible dates by using a simple regular expression (DDDD) and by assigning to the lower and upper bound the earliest and the latest extracted year respectively.

(B) **TEE RegEx + Filtering:** outliers are discarded from the extracted dates and then the earliest and latest dates are used for lower and upper bounds.

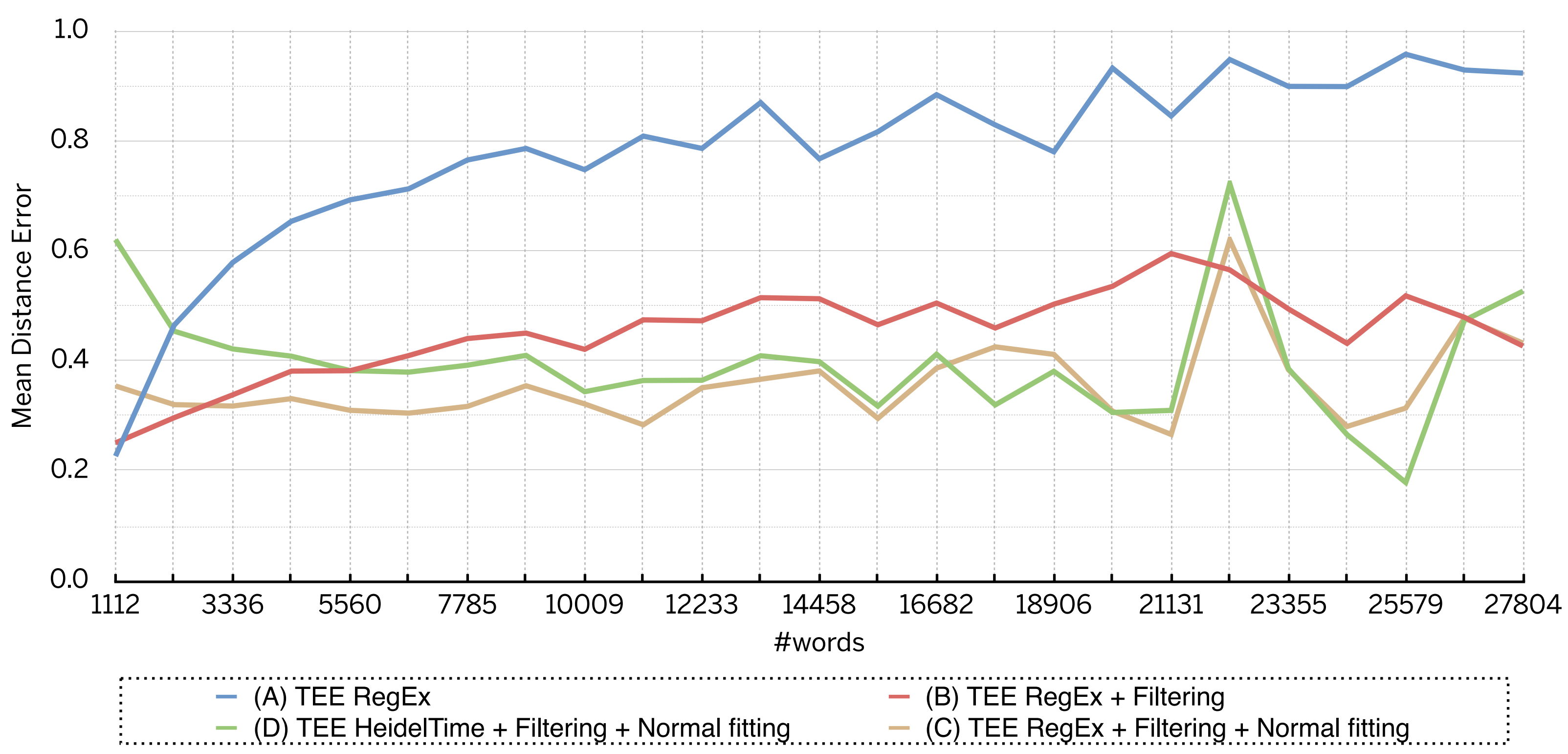
(C) **TEE RegEx + Filtering + Fitting Normal Distribution:** we use the regular expression-based extraction method and then apply filtering and normal fitting.

(D) **TEE HeidelTime + Filtering + Fitting Normal Distribution:** we use HeidelTime [1] to extract dates from the associated articles. We than apply filtering and normal fitting.

Centre for Doctoral Training in Computer Science (CDT CS)

Results

Filtering outliers positively affects the performances, where **fitting a normal distribution** helps only with long textual descriptions. Setting C slightly outperforms setting D. Regardless to the length distribution of Wikipedia articles, the aggregate results indicate the setting B outperforming the others. The use of Gaussian fitting seems to provide more stable results.



Setting	MDE	Std. dev.
(A) TEE RegEx	0.2636	0.3409
(B) TEE RegEx + Filtering	**0.2596**	0.3090
(C) TEE RegEx + Filtering + Normal fitting	0.3503	**0.2430**
(D) TEE HeidelTime + Filtering + Normal fitting	0.5980	0.2470

Aggregate results

[1] Jannik Strötgen, Julian Zell, and Michael Gertz. 2013. HeidelTime: Tuning english and developing spanish resources for TempEval-3. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 15–19, Atlanta, Georgia, USA, June. Association for Computational Linguistics.