# Interactive System for Reasoning about Document Age

Adam Jatowt and Ricardo Campos
Kyoto University and Polytechnic Institute of Tomar, LIAAD – INESC TEC
adam@dl.kuis.kyoto-u.ac.jp
ricardo.campos@ipt.pt

http://tinyurl.com/timestamping

## Overview

Recently, many historical texts have become digitized and made accessible for search and browsing. Professionals who work with collections of such texts often need to verify the correctness of documents' key metadata - their creation dates. In this paper, we demonstrate an interactive system for estimating the age of documents. It may be useful not only for tagging a large number of undated documents, but also for verifying already known timestamps. In order to infer probable dates, we rely on a large scale lexical corpora, *Google Books Ngrams*. Besides estimating the document creation year, the system also outputs evidences to support age detection and reasoning process and allows testing different hypotheses about document's age.

## System

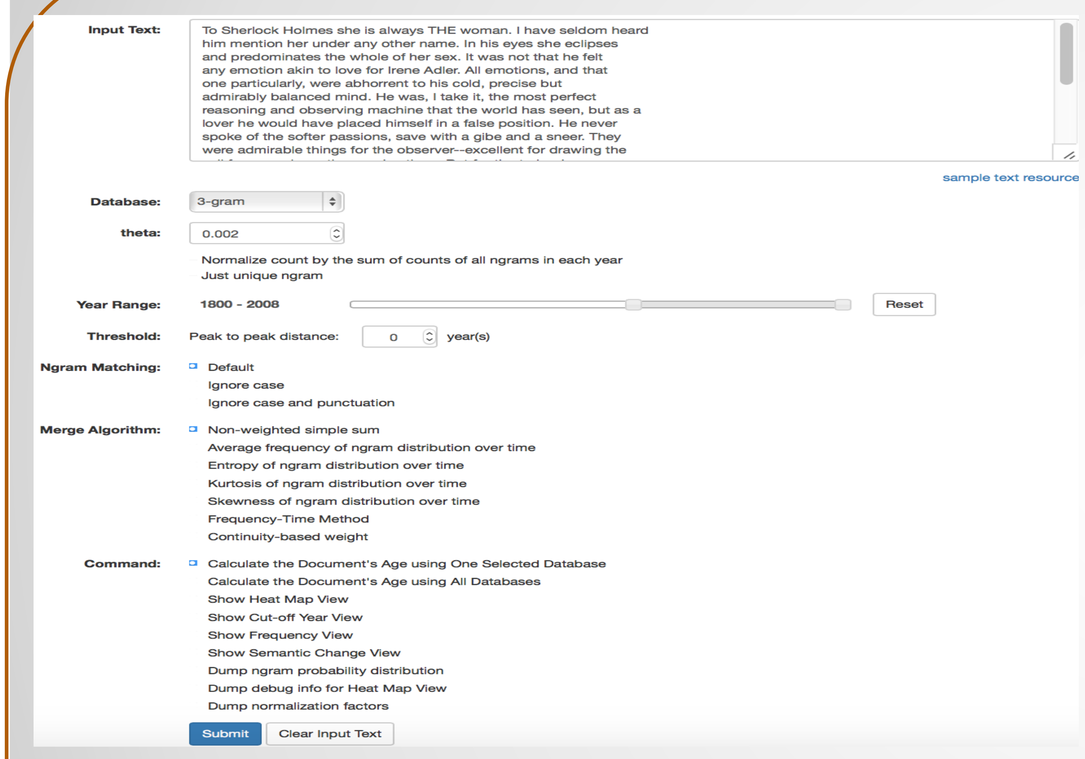



**Figure 1. Snapshot of the main interface.**

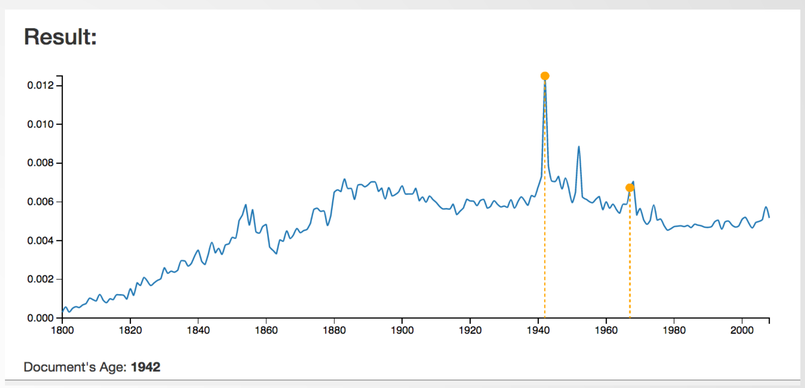



**Figure 2. Creation date probability distribution plot and the detected year (1942) of the first part (809 words) of W. Churchill's speech "Address to Joint Session of US Congress, 1941" based on 3-grams (non-weighted sum).**

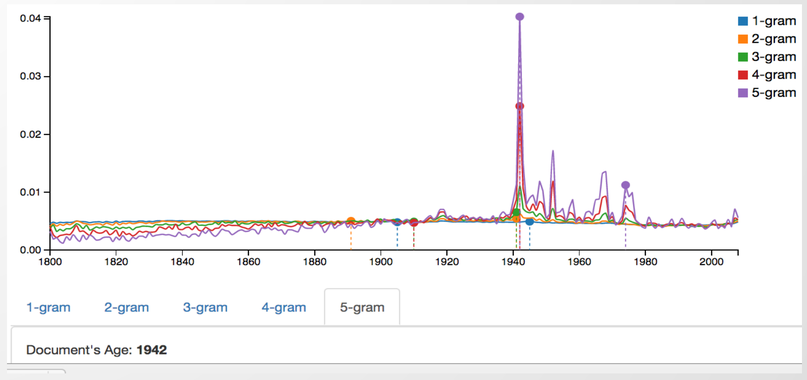
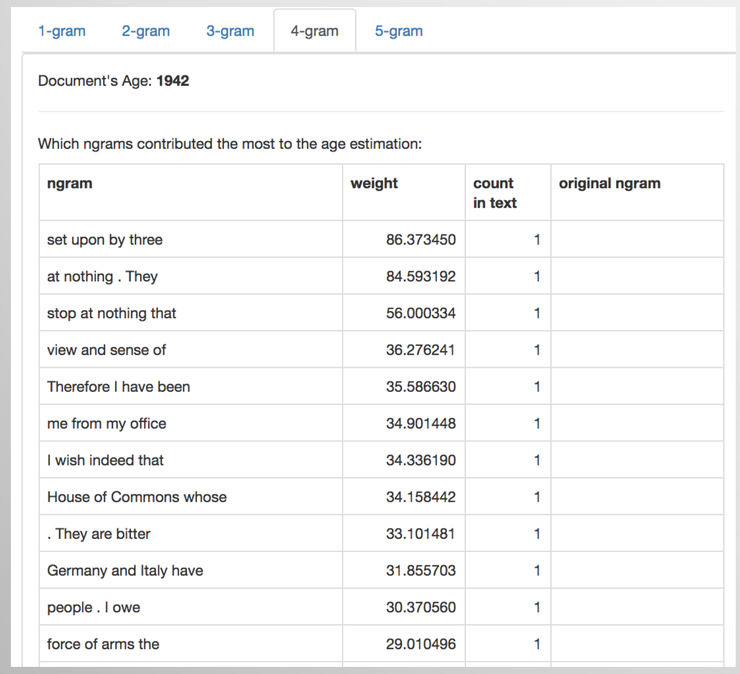



**Figure 3. Creation date probability distribution plots for the sample text used in Fig. 2 based on ngrams for all *n* values (non-weighted sum).**



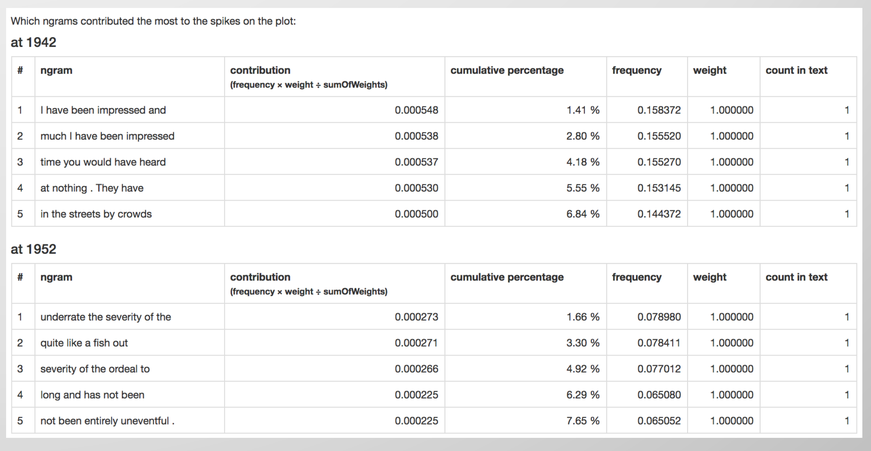
1-gram 2-gram 3-gram **4-gram** 5-gram

Document's Age: **1942**

Which ngrams contributed the most to the age estimation:

| ngram | weight | count in text | original ngram |
|---|---|---|---|
| set upon by three | 86.373450 | 1 | |
| at nothing . They | 84.593192 | 1 | |
| stop at nothing that | 56.000334 | 1 | |
| view and sense of | 36.276241 | 1 | |
| Therefore I have been | 35.586630 | 1 | |
| me from my office | 34.901448 | 1 | |
| I wish indeed that | 34.336190 | 1 | |
| House of Commons whose | 34.158442 | 1 | |
| . They are bitter | 33.101481 | 1 | |
| Germany and Italy have | 31.855703 | 1 | |
| people . I owe | 30.370560 | 1 | |
| force of arms the | 29.010496 | 1 | |

**Figure 4. Contributing top-scored 4-grams of the sample text used in Fig. 2.**



Which ngrams contributed the most to the spikes on the plot:

at 1942

| # | ngram | contribution (frequency × weight ÷ sumOfWeights) | cumulative percentage | frequency | weight | count in text |
|---|---|---|---|---|---|---|
| 1 | I have been impressed and | 0.000548 | 1.41 % | 0.158372 | 1.000000 | 1 |
| 2 | much I have been impressed | 0.000538 | 2.80 % | 0.155520 | 1.000000 | 1 |
| 3 | time you would have heard | 0.000537 | 4.18 % | 0.155270 | 1.000000 | 1 |
| 4 | at nothing . They have | 0.000530 | 5.55 % | 0.153145 | 1.000000 | 1 |
| 5 | in the streets by crowds | 0.000500 | 6.84 % | 0.144372 | 1.000000 | 1 |

at 1952

| # | ngram | contribution (frequency × weight ÷ sumOfWeights) | cumulative percentage | frequency | weight | count in text |
|---|---|---|---|---|---|---|
| 1 | underrate the severity of the | 0.000273 | 1.66 % | 0.078980 | 1.000000 | 1 |
| 2 | quite like a fish out | 0.000271 | 3.30 % | 0.078411 | 1.000000 | 1 |
| 3 | severity of the ordeal to | 0.000266 | 4.92 % | 0.077012 | 1.000000 | 1 |
| 4 | long and has not been | 0.000225 | 6.29 % | 0.065080 | 1.000000 | 1 |
| 5 | not been entirely uneventful . | 0.000225 | 7.65 % | 0.065052 | 1.000000 | 1 |

**Figure 5. Explanation of the two highest peaks of the text sample used in Fig. 2.**

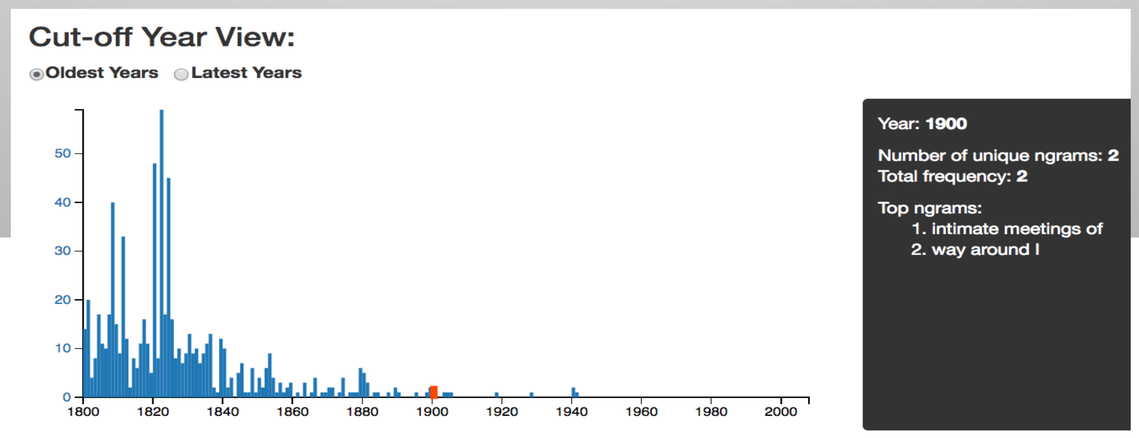



**Figure 6. Cut-off view with oldest years of the text used in Fig. 2 with the year 1900 being highlighted for detailed data.**