The value after the category name is the cosine similarity score between the probability ratio vectors of the article and the leading category. Notice the huge difference in similarity between documents in the training set and documents not in that set. This is probably because I created the latter set by manually copying text from Wikipedia articles, removing only the table of contents and all footnotes, and thus introducing a lot of noise.

The cases of Napoleon and O.J. Simpson are particularly interesting. We’d expect Napoleon’s bio to be classified as either "Politics",  "Warfare", or "Royalty", and apparently it's writing style resembles the latter the most. O.J. Simpson’s bio is even more intriguing, as it was classified as “Politics” (perhaps because of the trial references; we don’t have a Law category), with “Sport” coming in second (0.00485) and “Media” third (0.00462). Both examples demonstrate how our tool can be used to identify the tone (better word?) of an article.

- Look at the most frequent terms for O.J. and check their affiliation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **TF\*IDF** |  |  | **Probability** |  |  |
| **Article** | **Category** | **Similarity Score** | **Time** | **Category** | **Similarity Score** | **Time** |
| AC/DC |  |  |  | Music | 0.0937 | 8.596 |
| Cillian Murphy | Media | 0.028  (2.029e-05) | 6.75  (5.94) | Media | 0.1787 | 8.527 |
| Tim Duncan | Sport and recreation | (5.567e-05) | (10.12) | Sport and recreation | 0.1471 | 10.215 |
| Anton Chekhov |  |  |  | Literature and theatre | 0.1088 | 5.356 |
| Tiger Woods | Sport and recreation | 0.0158  (1.729e-05) | 10.515  (10.69) | Sport and recreation | 0.0078 | 10.644 |
| Karl Marx |  |  |  | Politics and government | 0.0079 | 13.281 |
| Agatha Christie |  |  |  | Literature and theatre | 0.0085 | 9.374 |
| Toni Blair |  |  |  | Politics and government | 0.0027 | 11.423 |
| Napoleon | Warfare | 0.068  (6.54e-06) | 18.02  (20.72) | Royalty, nobility and heraldry | 0.0126 | 18.791 |
| O. J. Simpson | Sport and recreation | -0.0437  (9.55e-06) | 5.213  (5.27) | Politics and government | 0.0076 | 5.188 |

Highlighted articles are in the training set.

Values in () were achieved after replacing Term Count with Term Frequency (i.e., dividing number of occurrences in a document by total number of words in a document)