***May 16, 2012***

*Similarity score* is the cosine similarity between the scores of terms extracted from the article and the scores of terms representing the matching category.

- In probability ratios score, 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.

- Should look at the most frequent terms for O.J. and figure out to which category they belong.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **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)

**May 17, 2012**

Using the new String/Regex cleaner, we get almost the same results at a fraction of the time it takes the POS cleaner. String/Regex actually did better on Agatha Christie! This is even before removing proper nouns (i.e., capitalized words not at the beginning of a sentence) using String/Regex and Unicode support.

Two features that probably help the Regex cleaner get better results are:

1. The former discards tokens containing non-English characters, unlike the latter, which occasionally stores fragments of such characters that only add noise.
2. The former converts all characters to lowercase.

It’s main drawback, however, is the lack of proper noun removal.

I’m pretty happy with the results, but we need to evaluate more articles.

**Using String/Regex cleaner**

Time to compute TF-IDF for entire corpus: 2.24 sec

|  |  |  |  |
| --- | --- | --- | --- |
| **Article** | **Suggested Category** | **Sim Score** | **Elapsed time** |
| Cillian\_Murphy | Media biographies | (1.900e-05) | 0.056 |
| AC\_DC | Music biographies | (2.173e-05) | 0.132 |
| Anton\_Chekhov | Literature and theatre biographies | (3.438e-05) | 0.188 |
| Tim\_Duncan | Sport and recreation biographies | (4.681e-05) | 0.254 |
| Tiger\_Woods | Sport and recreation biographies | (1.274e-05) | 0.321 |
| Karl\_Marx | Literature and theatre biographies | (4.317e-06) | 0.398 |
| Agatha\_Christie | Literature and theatre biographies | (5.552e-06) | 0.463 |
| Toni\_Blair | Politics and government biographies | (5.485e-06) | 0.537 |
| Napoleon | Warfare biographies | (5.296e-06) | 0.625 |
| O.\_J.\_Simpson | Sport and recreation biographies | (6.115e-06) | 0.683 |

**Using POS cleaner**

Time to compute TF-IDF for entire corpus: 3.45548892021

|  |  |  |  |
| --- | --- | --- | --- |
| **Article** | **Suggested Category** | **Sim Score** | **Elapsed time** |
| Cillian\_Murphy | Media biographies | (2.029e-05) | 8.339 |
| AC\_DC | Music biographies | (4.024e-05) | 16.478 |
| Anton\_Chekhov | Literature and theatre biographies | (9.143e-06) | 21.421 |
| Tim\_Duncan | Sport and recreation biographies | (5.567e-05) | 31.222 |
| Tiger\_Woods | Sport and recreation biographies | (1.729e-05) | 41.500 |
| Karl\_Marx | Literature and theatre biographies | (5.819e-06) | 54.521 |
| Agatha\_Christie | Media biographies | (8.413e-06) | 63.898 |
| Toni\_Blair | Politics and government biographies | (7.505e-06) | 75.626 |
| Napoleon | Warfare biographies | (6.540e-06) | 93.386 |
| O.\_J.\_Simpson | Sport and recreation biographies | (9.550e-06) | 98.410 |