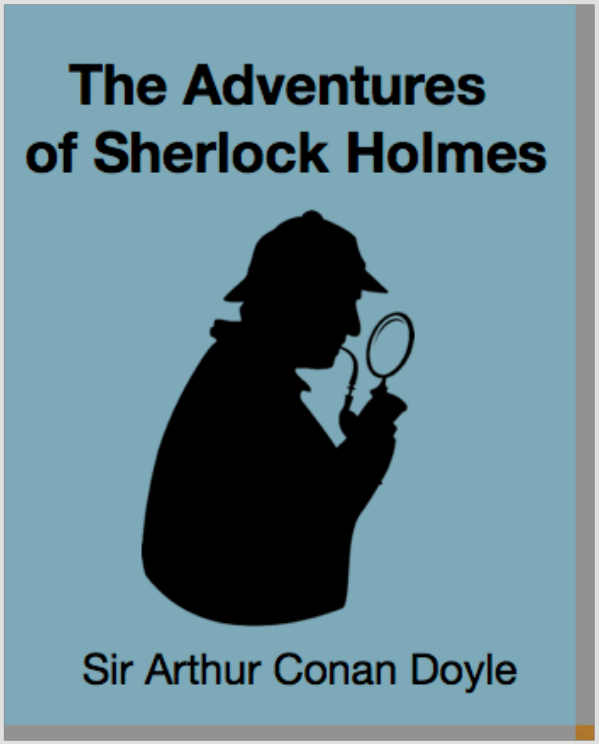
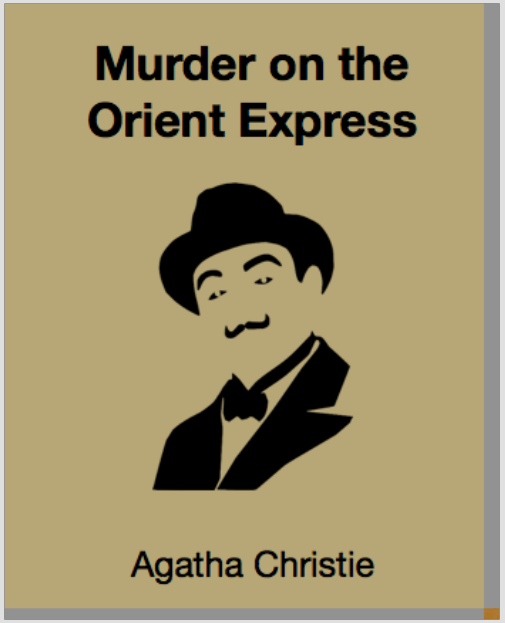
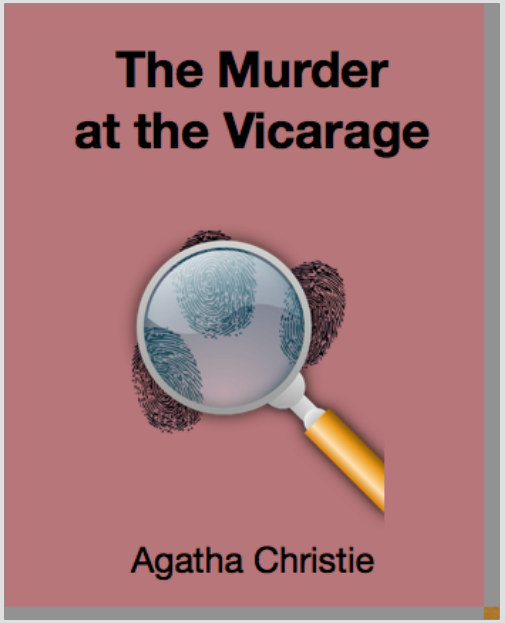
Topic modeling

Recommender system

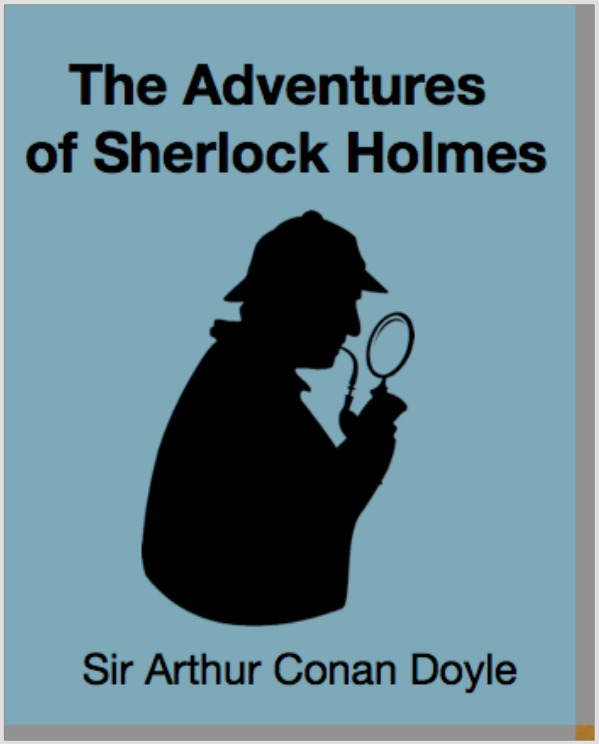






Topic modeling

We can say that sherlock holmes has the following topics in order of relevancy.



60% detective

30% adventure

10% horror

Document is a distribution over topics

yes. good observation.

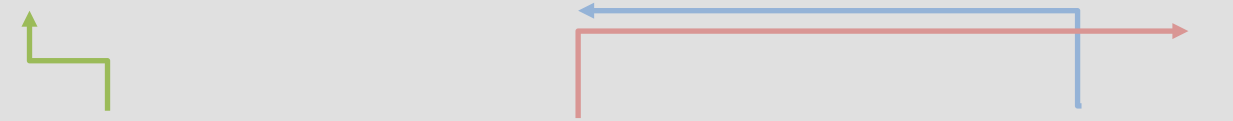
Topic modeling

Suppose these percentages are the 'most popular words'.

Sports	Economy	Politics
20% Football	24% Money	10% President
10% Hockey	9% Dollar	4% USA
5% Goal	7% Euro	3% Union
1% Score	3% Bank	1% Law
…	…	…

Topic modeling

Sports	Economy	Politics
20% Football	24% Money	10% President
10% Hockey	9% Dollar	4% USA
5% Goal	7% Euro	3% Union
1% Score	3% Bank	1% Law
…	…	…



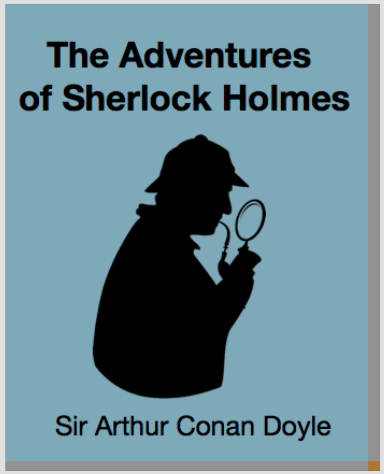
Football player from USA has salary in dollars

This shows the belongings of a sentence's words to a topic.

Topic is a distribution over words

Yes! Each topic has probabilities of the words appearing.

Similarity



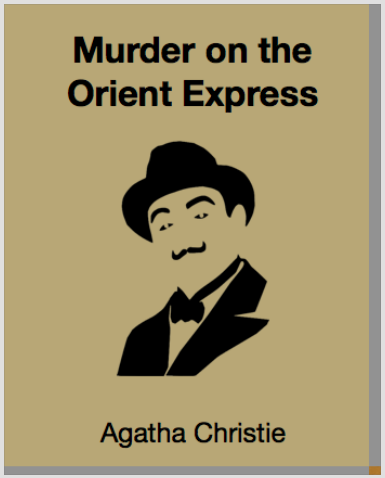
We can define some topic vector for each book.

60% detective

30% adventure

10% horror

$$= \begin{pmatrix} 0.6 \\ 0.3 \\ 0.1 \end{pmatrix} = a$$



62% detective

33% adventure

5% horror

$$= \begin{pmatrix} 0.62 \\ 0.33 \\ 0.05 \end{pmatrix} = b$$

Similarity/distance

These are some distance metrics for the topics.

$$a = \begin{pmatrix} 0.6 \\ 0.3 \\ 0.1 \end{pmatrix} \qquad b = \begin{pmatrix} 0.62 \\ 0.33 \\ 0.05 \end{pmatrix}$$

Euclidean distance

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \approx 0.004$$

smaller euclidean distance means they are very similar.

Cosine similarity

$$\cos(a, b) = \frac{a^T b}{\|a\| \cdot \|b\|} \approx 0.997$$

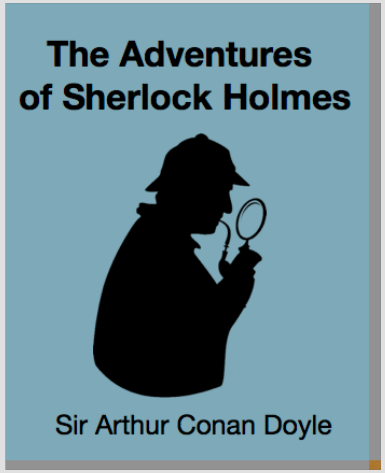
Why cosine similarity more relevant?
Because they are roughly in the same 'direction'.

Goals

1. Construct topics

Sports	Economy	Politics
20% Football	24% Money	10% President
10% Hockey	9% Dollar	4% USA
…	…	…

2. Assign topics to texts



- 60% detective
- 30% adventure
- 10% horror