

Topic modeling



Recommender system

The Adventures of Sherlock Holmes



Sir Arthur Conan Doyle

Murder on the Orient Express



Agatha Christie

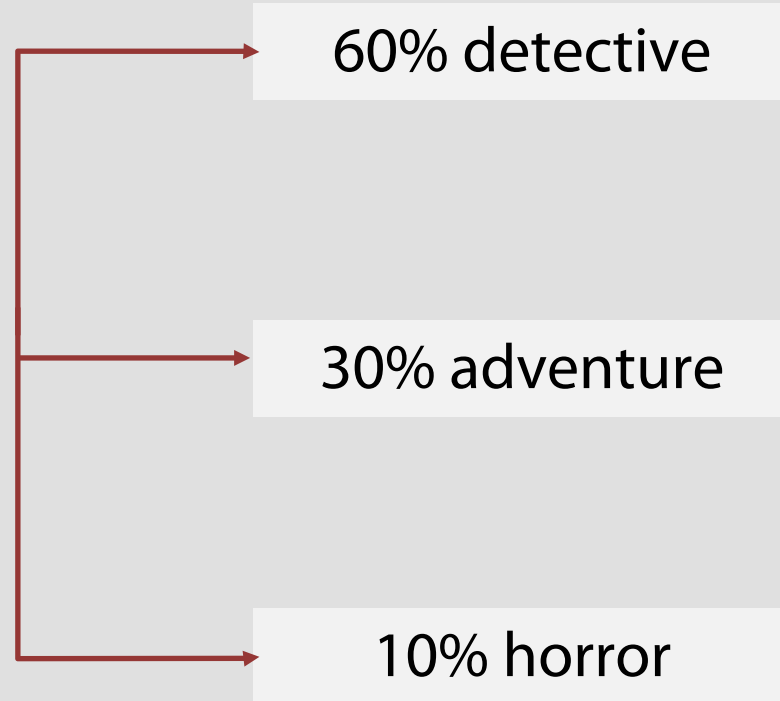
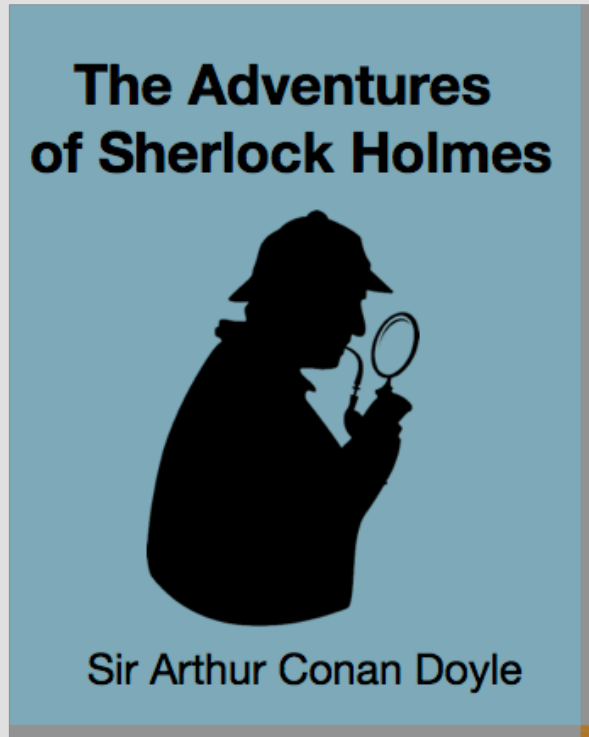
The Murder at the Vicarage



Agatha Christie



Topic modeling



Document is a distribution over topics



Topic modeling

Sports

20% Football

10% Hockey

5% Goal

1% Score

...



Topic modeling

Sports

20% Football
10% Hockey
5% Goal
1% Score
...

Economy

24% Money
9% Dollar
7% Euro
3% Bank
...



Topic modeling

Sports

20% Football
10% Hockey
5% Goal
1% Score

...

Economy

24% Money
9% Dollar
7% Euro
3% Bank

...

Politics

10% President
4% USA
3% Union
1% Law

...



Topic modeling

Sports

20% Football
10% Hockey
5% Goal
1% Score
...

Economy

24% Money
9% Dollar
7% Euro
3% Bank
...

Politics

10% President
4% USA
3% Union
1% Law
...

Football player from USA has salary in dollars



Topic modeling

Sports

20% Football
10% Hockey
5% Goal
1% Score
...

Economy

24% Money
9% Dollar
7% Euro
3% Bank
...

Politics

10% President
4% USA
3% Union
1% Law
...

...
↑
Football player from USA has salary in dollars
← ... →



Topic modeling

Sports

20% Football
10% Hockey
5% Goal
1% Score
...

Economy

24% Money
9% Dollar
7% Euro
3% Bank
...

Politics

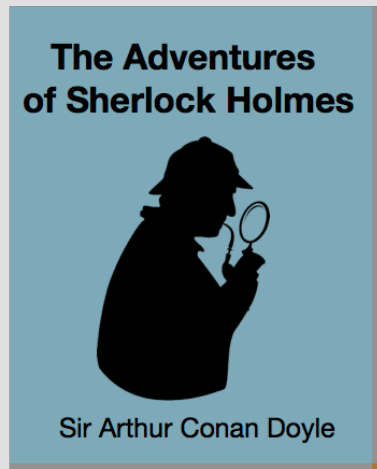
10% President
4% USA
3% Union
1% Law
...

...
↑
Football player from USA has salary in dollars
... ← ... → ...

Topic is a distribution over words



Similarity

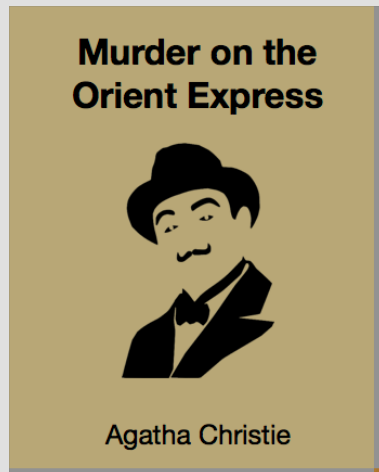


60% detective

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

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

Euclidean distance

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

Cosine similarity

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



Goals

1. Construct topics

Sports

20% Football
10% Hockey
...

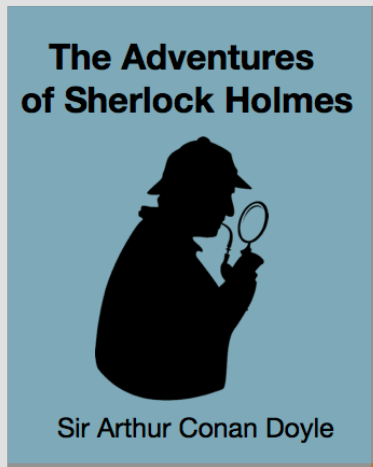
Economy

24% Money
9% Dollar
...

Politics

10% President
4% USA
...

2. Assign topics to texts



60% detective

30% adventure

10% horror

