



ALICE

A PROJECT FOR DATA MODELING

DHDK 2018-19

**ANTONATOU I., APOLLONI D., GOLLINI M., MITIKU H.,
TOTARO G.**

PROJECT OVERVIEW

ALICE stands for Authors Linked: Images, works Created and vidEos

- It is a fictional application based on a database about authors of the 20th century aimed at students of literature/humanistic studies. Given a query the application will return a list of relevant authors, related works and/or materials somehow linked with the query.
- For the purposes of our course, we will present a limited sample of data, but the application may in theory accommodate a bigger amount of data as well. We decided to include 3 authors, 3 works written by each author and various visual materials.

AUTHORS-SAMUEL BECKETT

Waiting for Godot (play)

Murphy (novel)

Watt (novel)



<https://www.youtube.com/watch?v=4ffMoTfGCfY>

<https://www.youtube.com/watch?v=IohAssRQsjM>

<https://www.youtube.com/watch?v=FqpjddXaw4E>



AUTHORS-HAROLD PINTER

The Birthday Party (play)

The Room (play)

The Dumb Waiter (play)

<https://www.youtube.com/watch?v=oVchqMXobVQ>

<https://www.youtube.com/watch?v=48HbkR8z7fc>

https://www.youtube.com/watch?v=nLwC5RonO_w

<https://www.youtube.com/watch?v=-N99S8n2TiA>

<http://www.raistoria.rai.it/articoli/pinter-maestro-del-teatro-dellassurdo/11015/default.aspx>



AUTHORS-J.R.R. TOLKIEN

The Fellowship of the Ring

The Hobbit

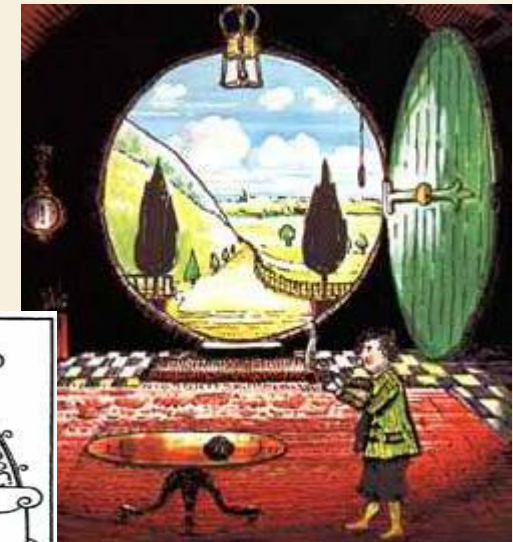
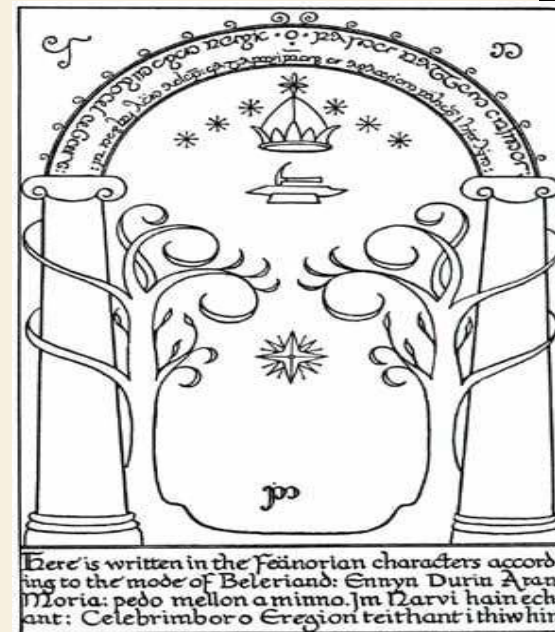
The Silmarillion

<https://www.youtube.com/watch?v=V75dMMIW2B4>

<https://www.youtube.com/watch?v=JTSOD4BBCJc>

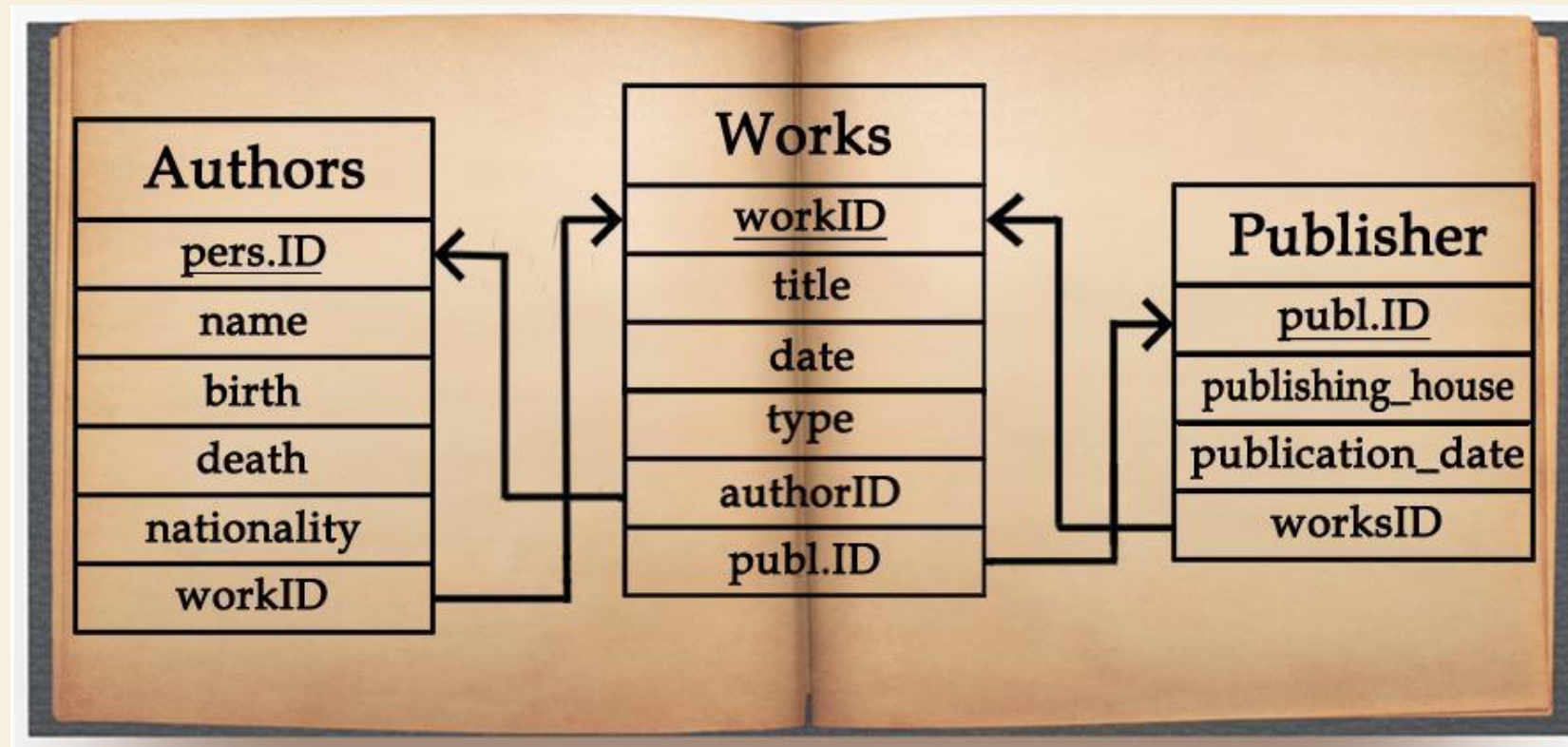
<https://www.youtube.com/watch?v=yFexwNCYenI>

<https://www.youtube.com/watch?v=0JlJSLzja7E>



STRUCTURED DATA

In our case, structured data concerns information about the author (name, dates of birth and death etc.), their works (titles, date of publication etc.) and their publishers. Below follows the relational modeling of the structured data:



SEMI-STRUCTURED DATA

In our case semi-structured data refers to information about the content of the works in the form of semantic keywords taken from descriptions of the plots (from sources like Goodreads, Wikipedia etc.) The XML modeling of the semi-structured data is presented here:

<period>20th century

<movement>modernism

<submovement>theatre of absurd

<technique>stream of consciousness

<theme>absurd</theme>

<theme>violence</theme>

<theme>nightmare</theme>

<theme>modernism</theme>

<theme>humor</theme>

<theme>waiting</theme>

<theme>tragic</theme>

<theme>satire</theme>

<theme>existentialism</theme>

</technique>

1/2

</submovement>

<submovement>fantasy

<theme>magic</theme>

<theme>journey</theme>

<theme>battle</theme>

<theme>dragons</theme>

<theme>prequel</theme>

<theme>deities</theme>

<theme>jewels</theme>

</submovement>

</movement>

<movement>postmodernism

</movement>

<movement>structuralism

</movement>

</period>

2/2

UNSTRUCTURED DATA

We define as unstructured data the collection of texts and low-level features, mainly shapes and color distribution, extracted from images and videos related to our content. We used one of these images to perform the analysis of the low-level features of our interest.

UNSTRUCTURED DATA-IMAGE ANALYSIS

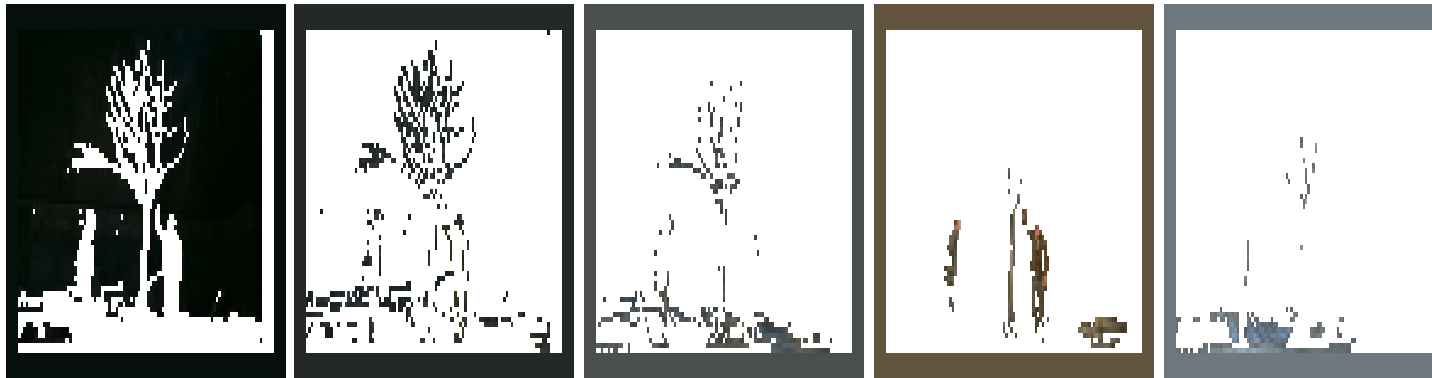


- **Media type:** discrete/still, captured from real world, 2D (2 dimensions on space)
- **3 levels of MM representation:** 1st level: low level features: a) shapes (trees, people, ground) → from image segmentation
 - b) Colors (marshland, charleston green, cathedral, wombat, storm grey)
 - c) HSV
 - d) Other features

Using the site <http://mkweb.bcgsc.ca/color-summarizer/> we obtained:

a) Shapes











Pixels of the image assigned to each cluster. The border is the color of the cluster as calculated by the average value of its pixels.



b) Colors

Cluster colors, sized by number of pixels:



cluster	pixels	name	HEX	RGB	HSV	LCH	Lab	tags
	74.51%	 11,15,8 marshland $\Delta E=1.6$	#060F09	6 15 9	137 59 6	4 4 153	4 -3 2	almost cod gordons marshland midnight moss onyx black green grey
	10.08%	 35,43,43 charleston green $\Delta E=2.4$	#242926	36 41 38	137 12 16	16 3 153	16 -3 2	light dark charleston filmpro jungle nero pasifika sentry style uhi black green grey
	6.52%	 75,77,74 cathedral $\Delta E=1.6$	#4B4F4D	75 79 77	153 6 31	33 2 166	33 -2 1	armadillo blast cape cathedral cod ironsand quarter ship silver streak thunder grey
	4.49%	 103,87,68 wombat $\Delta E=2.2$	#62543F	98 84 63	37 36 38	37 15 82	37 2 15	dark amber arrowtown grayish groundbreaker iroko pravda spark stonewall tobacco triple wombat brown
	4.40%	 116,120,128 storm grey $\Delta E=2.2$	#6E787F	110 120 127	205 14 50	50 6 247	50 -2 -5	pale battleship bluff eighth infinity nevada raven rolling sky stone storm tuna weathered blue grey

c) HSV

	avg	med	min	max
HSV:H	132 1.00	132	16	360
HSV:S	56	56	1	100
HSV:V	12	6	1	82

d) Other features

Dimension: 600x380 (from image properties)

Size: 31, 5 KB (from image properties)

Color space: RGB for display, HSV for representation

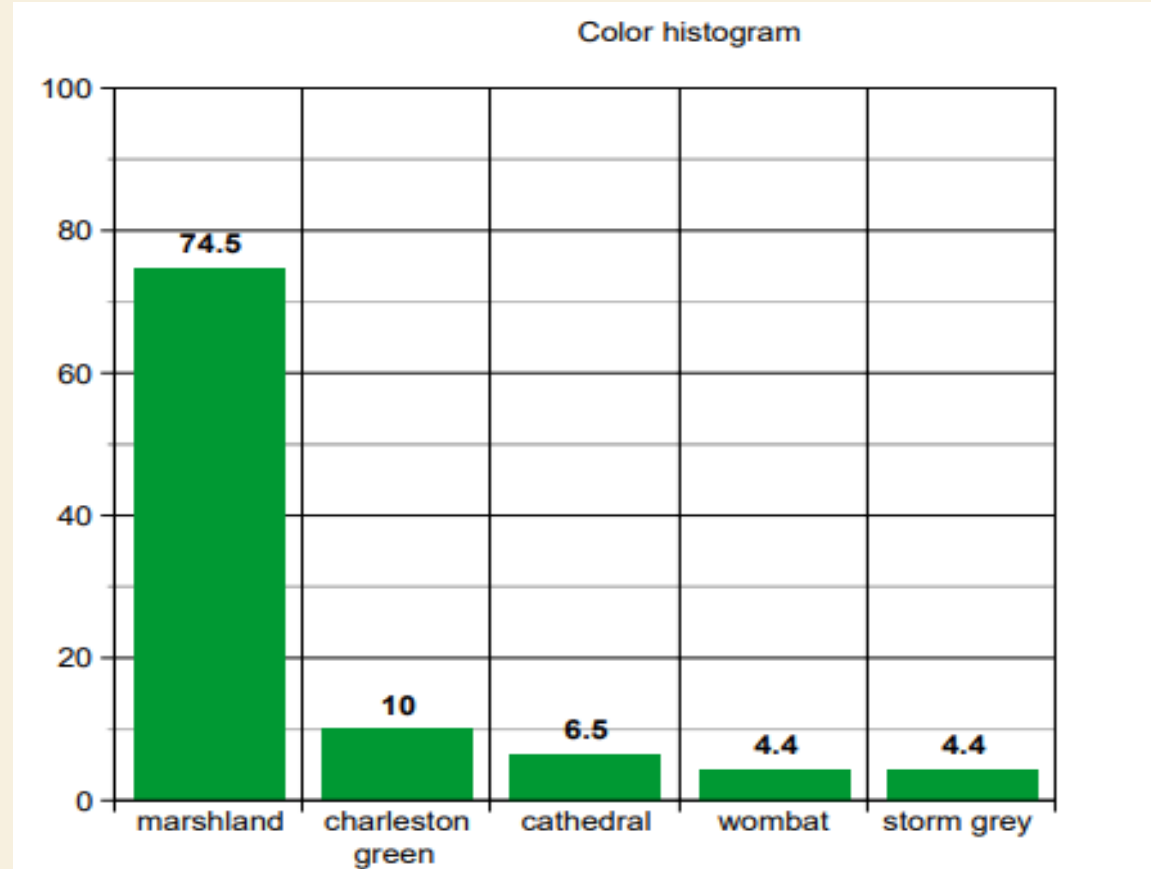
Spatial relations (proposed by us): i) local properties: position, area, perimeter

ii) global properties (relations): person on the left= Object A; tree= Object B; person on the right= Object C; ground=Object D → Object A is left to Object B, Object C is right to Object B, Object D is below Objects A,B,C

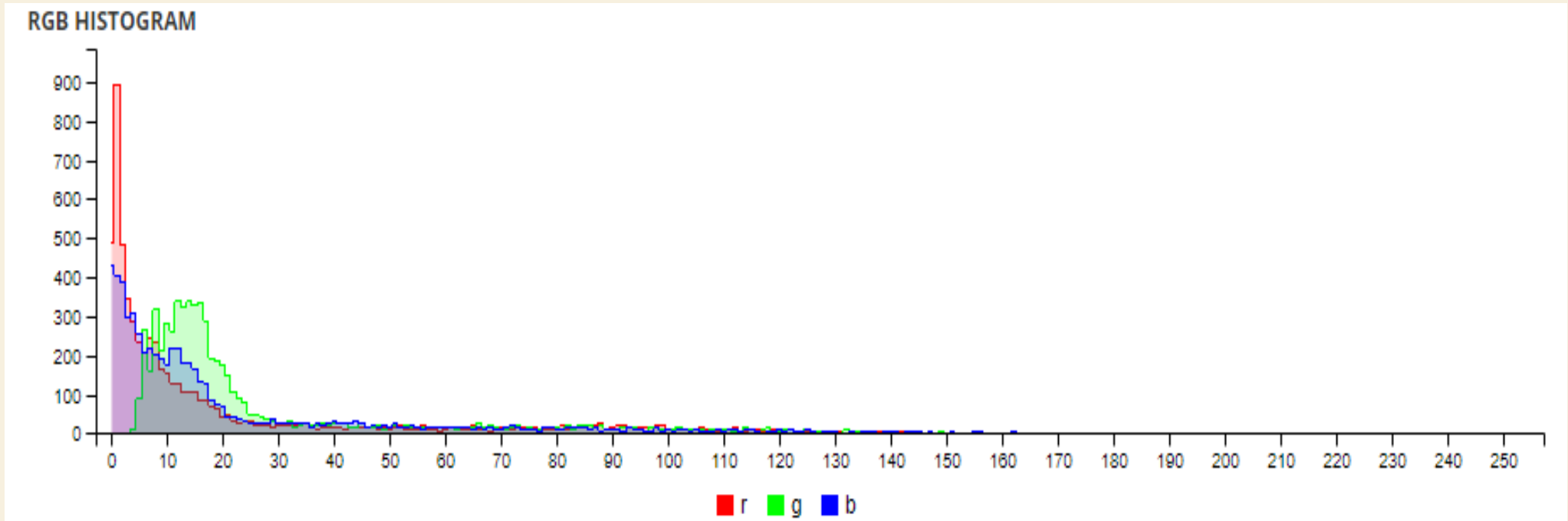
Content-based search: what features are relevant in this case? ☐ Colors, shapes

Histograms

- Based on percentage of colors per pixel provided by the [site](#) (see slide 12)

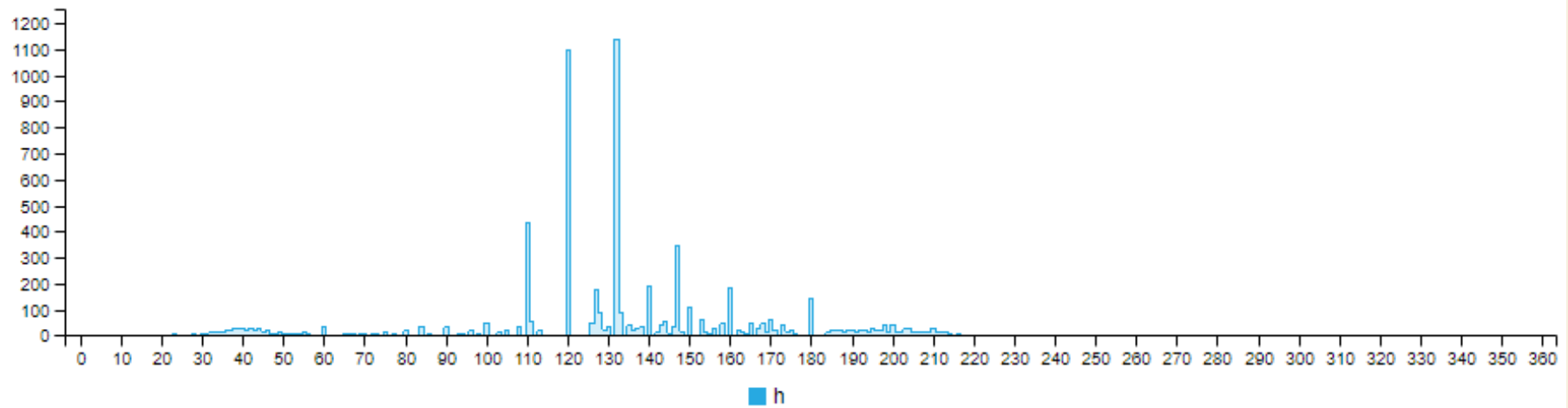
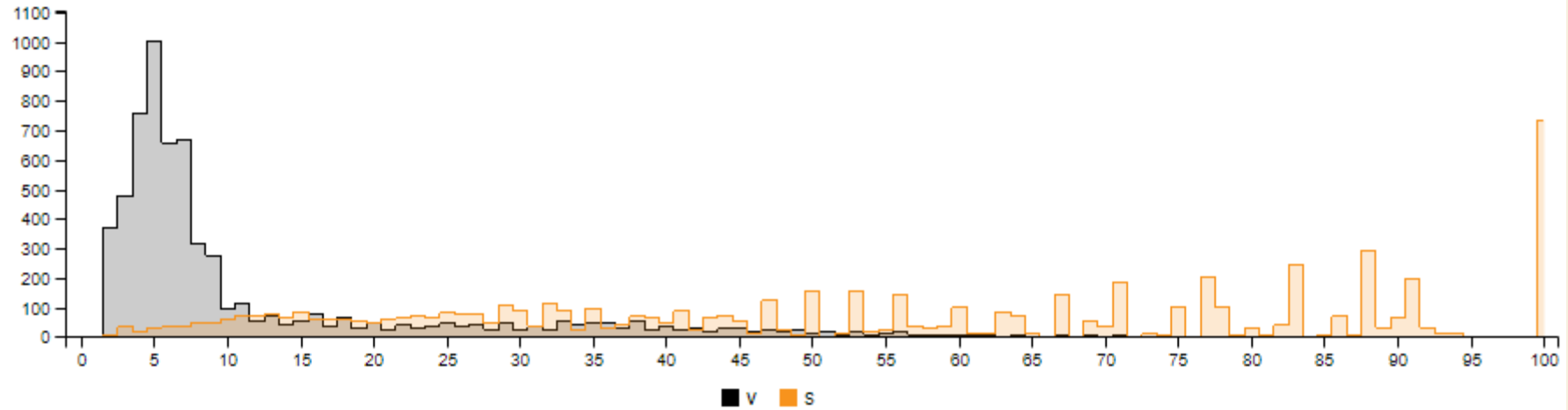


- RGB



- HSV

HSV HISTOGRAMS



After having extracted the color histograms and the shapes, we can define the methods for comparison in case of content-based search, which would in theory return similar objects to the query object. To address the notion of similarity, we explore the possibilities presented during the course:

Comparing histograms: Over the 1-to-1 matching offered by lp-norms and the precise, cross-talk effect provided by the expensive quadratic distance (essential in cases of local properties comparison), we believe that the weighted Euclidean distance is a good option. The weighted version of the Euclidean distance allows us to choose the most important colors in the image and is cheaper than quadratic. It does not offer the possibility for cross-talk, but it is enough for global properties comparison, which is of our interest.

Comparing shapes: extract contour → choose most interesting point with max curvature points → switch with Fourier to frequency domain → Euclidean (1-to-1 matching, if there is shift in shape, the two shapes will be returned as different even if they are the same) or Dynamic Time Warping (1-to-many, useful when there is a shift in the shapes)

UNSTRUCTURED DATA-IMAGE ANALYSIS

2nd level of MM representation:

Annotations:

- **Suggested from the [site](#):** black, grayish, midnight, stonewall
- **Proposed by us:** people, pointing, tree, deserted

3rd level of MM representation:

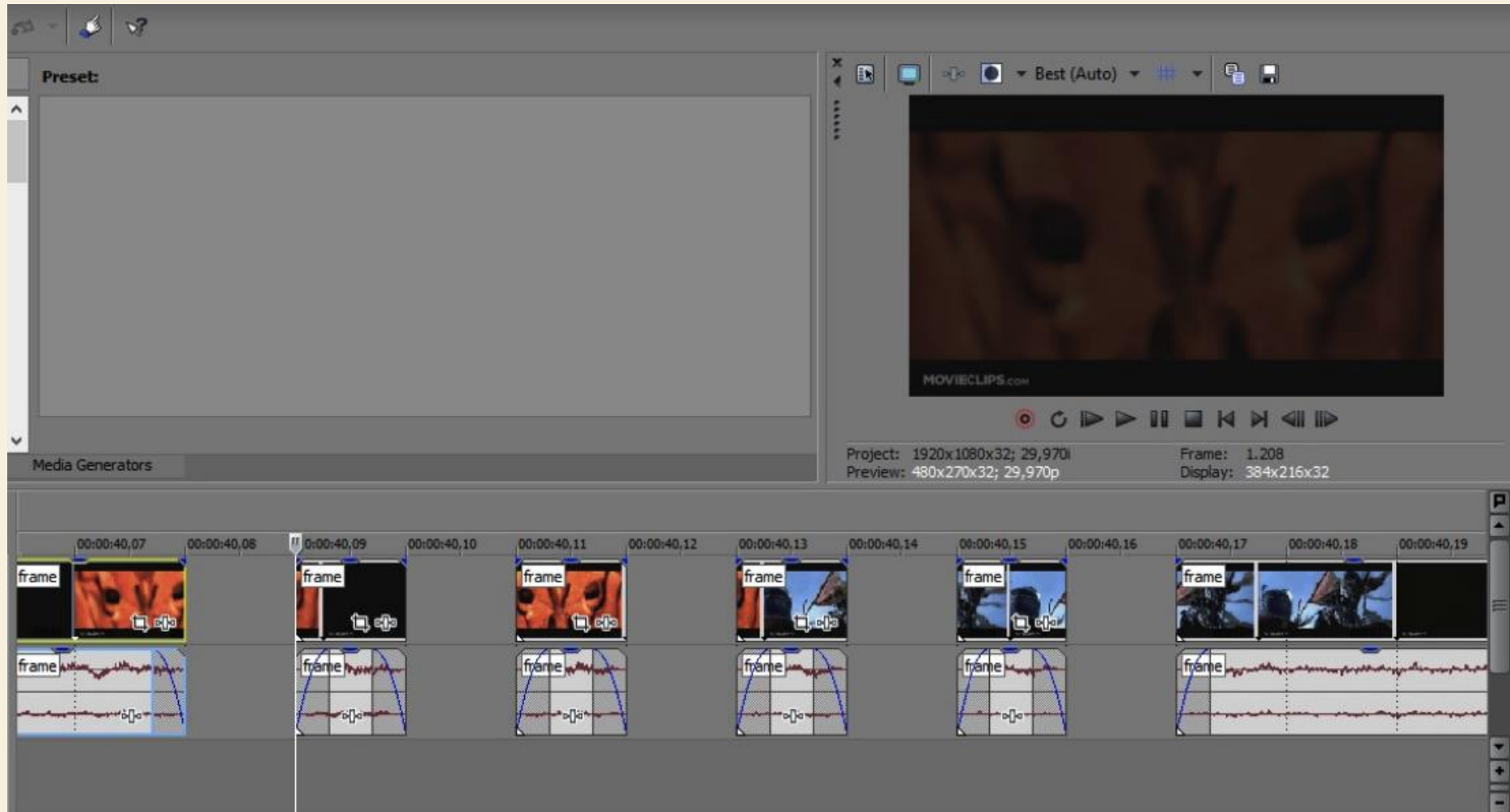
Possible semantic interpretation: people talking about the tree

VIDEO FRAMES

Videos are 3-D arrays of color pixels and they are built according to two the horizontal and the vertical dimensions.

A frame is the set of all pixels that generate the still image of the complete moving video. Frames are put sequentially together to create the video. Each frame is displayed on the screen for a very short time. Therefore, the sequence of frames, that are nothing but still images, produces the illusion of the moving image.

By using Sony Vegas Pro, a software devoted to video editing, we could take the trailer of *The Fellowship of the Ring* (i.e. unstructured data) and divide it into frames so as to show the exact point where the image changes and is substituted by another one.



TEXT ANALYSIS

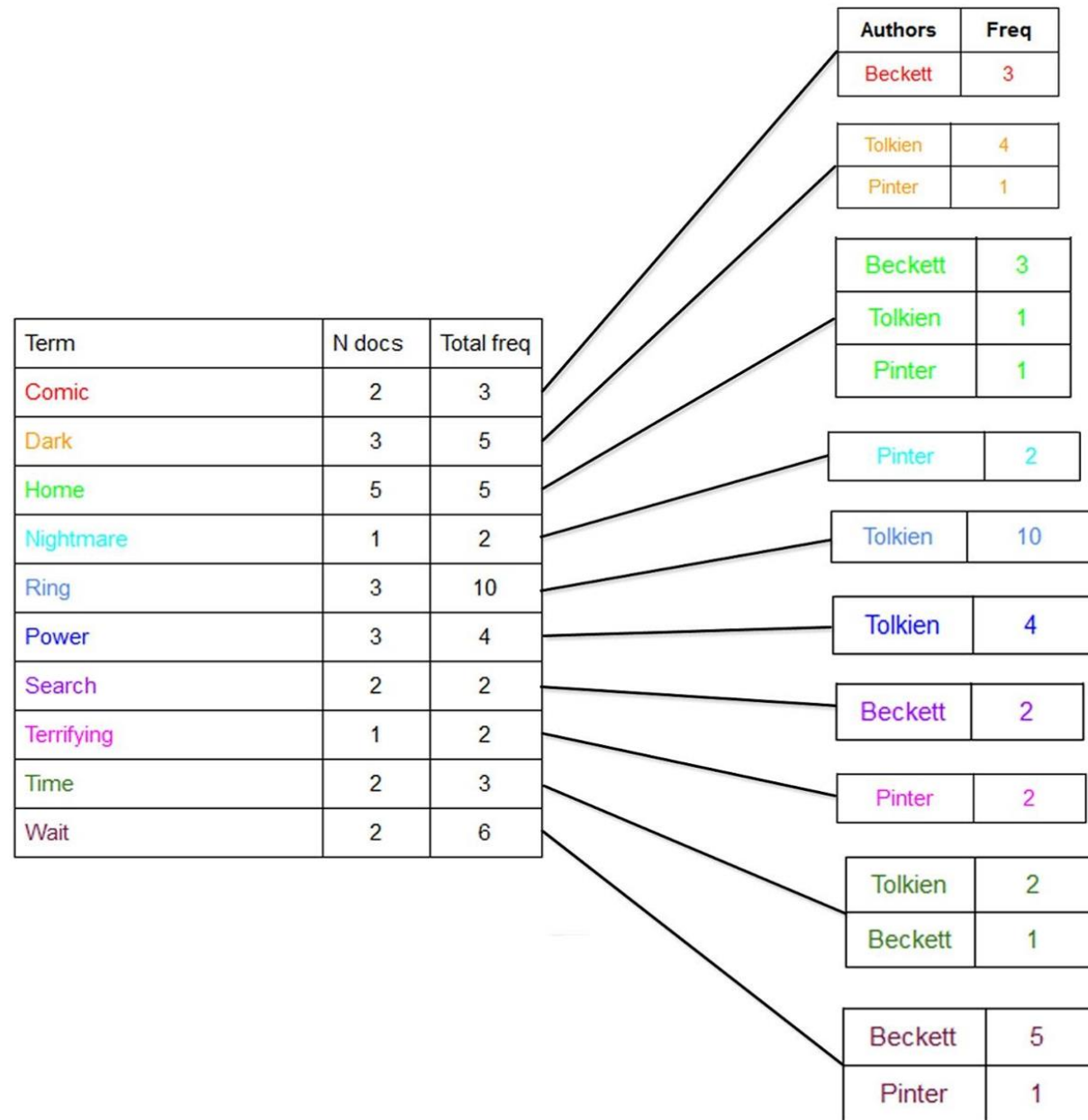
Term-document matrix

For the matrix and the inverted index we chose 10 words from the descriptions of the plots. The matrix represents the absence (0) and presence (1) of each of the terms in each of the three authors of our application.

	Pinter	Beckett	Tolkien
comic	0	1	0
dark	1	0	1
home	1	1	1
nightmare	1	0	0
ring	0	0	1
power	0	0	1
search	0	1	0
terrifying	1	0	0
time	1	1	1
wait	1	1	0

Inverted index

The inverted index includes information about the frequency of the terms in the documents. The posting list (instead of the posting file which takes up a lot of space) is integrated to offer information about the specific document containing the terms and the corresponding frequencies.



Vector Space Model

We computed the weights ($w = tf * idf$, $idf = \log(N/N_i)$) of each of the terms for each document, so that in case of a keyword query (the keyword being among the 10 words presented here), the weights would be utilized to compute cosine similarity and present to the user the results in a decreasing order. An example of this can be found in the next slide.

Term	N docs	Author	w _{i,j}
Comic	1	Beckett	0.52
Dark	2	Tolkien	0.70
Home	3	Pinter	0.17
Nightmare	1	Beckett	0
Ring	1	Tolkien	0
Power	1	Pinter	0
Search	1	Pinter	0.35
Terrifying	1	Tolkien	4.77
Time	2	Tolkien	1.90
Wait	2	Beckett	0.95
		Pinter	0.95
		Tolkien	0.35
		Beckett	0.17
		Beckett	0.88
		Pinter	0.17

QUERIES

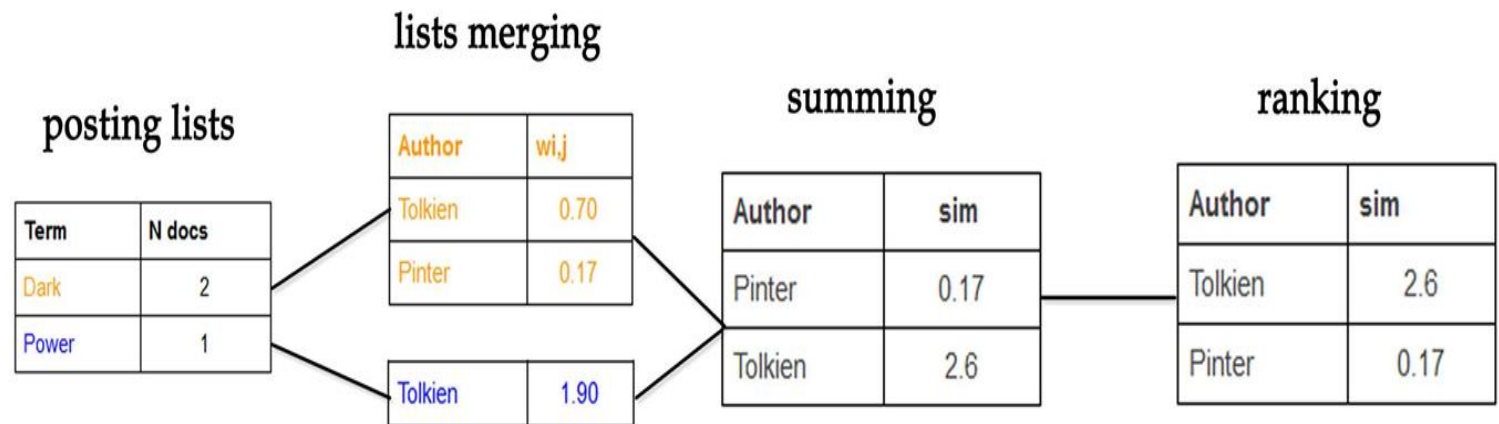
A) Query by (existing) keyword

Given a Boolean query like:

dark AND power

The procedure presented here will be followed (none of which will be shown to the final user, who is only interested in the ranked list of the results):

Query: dark AND power



B) Metadata query:

Given a query that makes use of structured data like this:

“Titles of works written by Samuel Beckett”

we can use SQL to access the corresponding records from our database (presented in slide 6):

```
SELECT W.title
```

```
FROM Works W, Authors A
```

```
WHERE A.name= “Samuel Beckett”
```

```
AND A.persID=W.AuthorID
```

LAST QUERY + EVALUATION OF PERFORMANCE

C) Query by keyword (that is not part of the application)

Given a query like:

“forest”

we expect the system to return the three works by Tolkien, which contain references to forests, and the picture from slide 10, which contains a tree. Using thesauri information, the system allows for taxonomy reasoning and therefore will return (incorrectly in this case) the image of the tree, since it is a hyponym of forest.

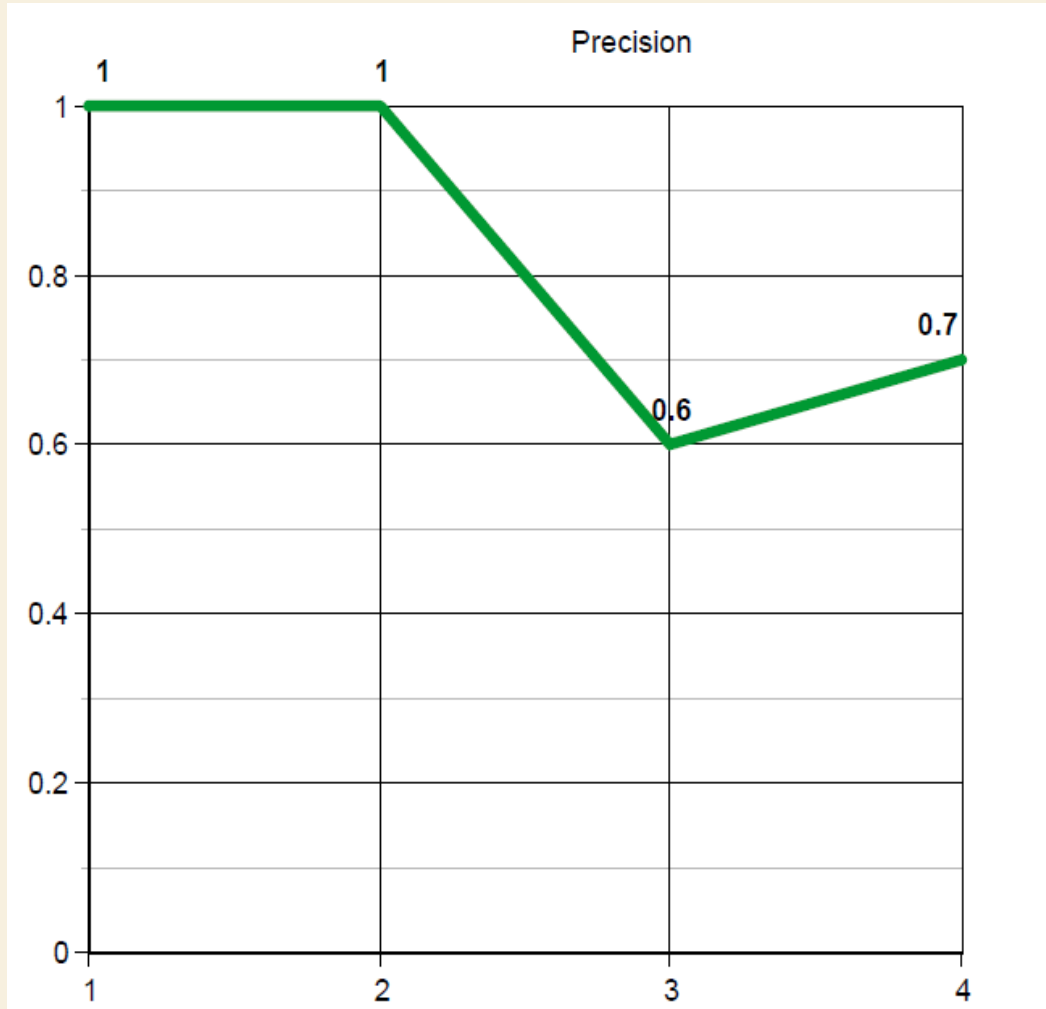
The Fellowship of the Ring
The Hobbit
The Silmarillion



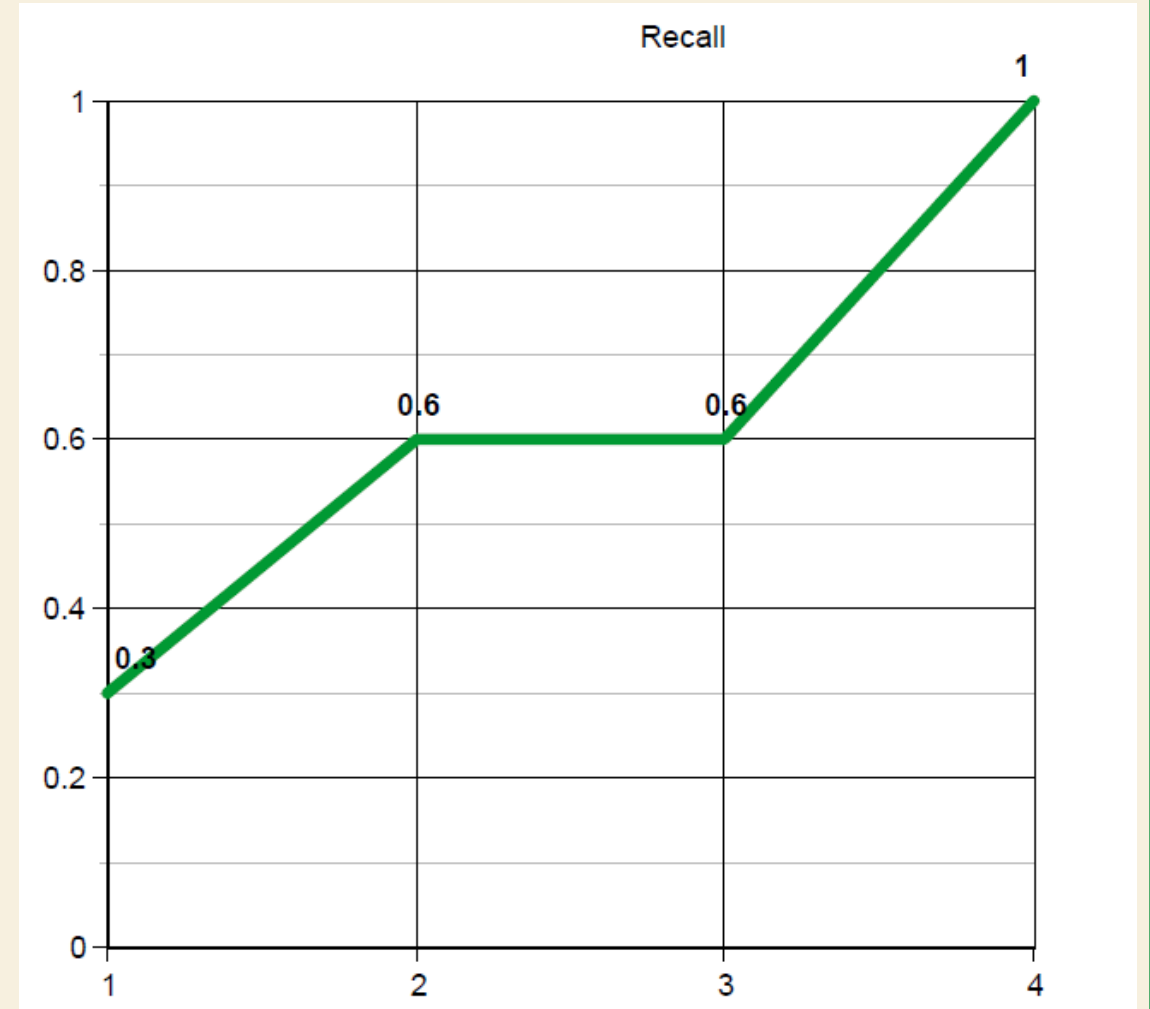
Results	Fellowship of the Ring (#1)	The Hobbit (#2)	Beckett picture (#3)	Silmarillion (#4)
relevant	✓	✓		✓

EVALUATION GRAPHS

Precision



Recall



Our sources of information and tools

In order to develop this project, we used:

Websites:

- BuzzFeed News (<https://www.buzzfeednews.com/>)
- CultureVulture (<https://culturevulture.net/>)
- Create a Graph (<https://nces.ed.gov/nceskids/createagraph/>)
- Goodreads (<https://www.goodreads.com/>)
- Image Color Summarizer (<http://mkweb.bcgsc.ca/color-summarizer/>)
- OpenCulture (<http://www.openculture.com/>)
- Rai Storia (<http://www.raistoria.rai.it/>)
- The Guardian (<https://www.theguardian.com>)
- Tolkien Gateway (http://www.tolkiengateway.net/wiki/Main_Page)
- Voyant Tools (<https://voyant-tools.org/>)
- Wikipedia (https://en.wikipedia.org/wiki/Main_Page)
- Worldcat (<https://www.worldcat.org/>)
- YouTube (<https://www.youtube.com/>)

Softwares:

- Adobe Photoshop
- Sony Vegas Pro