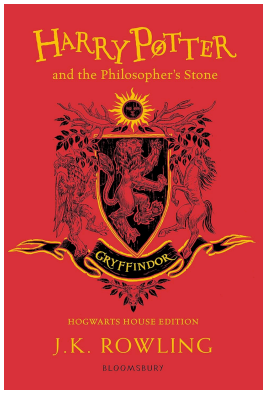
Harry Potter

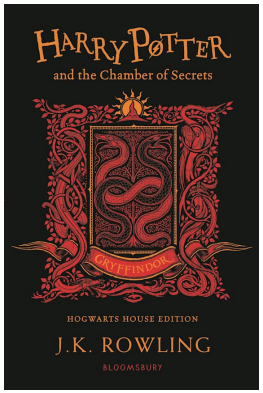
and the Text mining project

Enrico Carraro Alex Cecchetto Virginia Murru

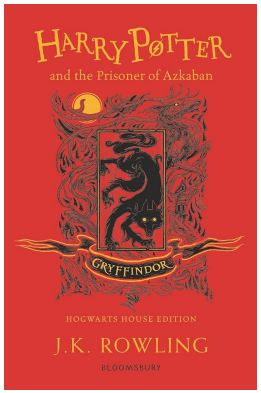
1. The Data



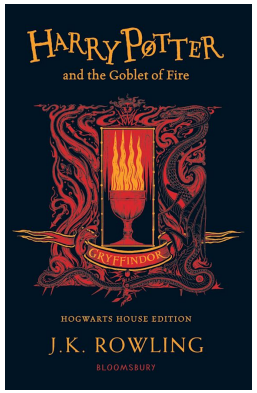
1997



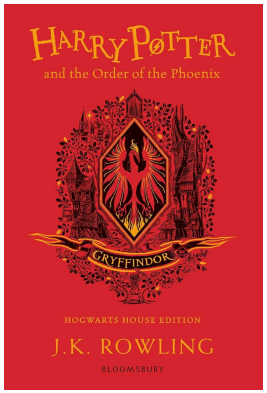
1998



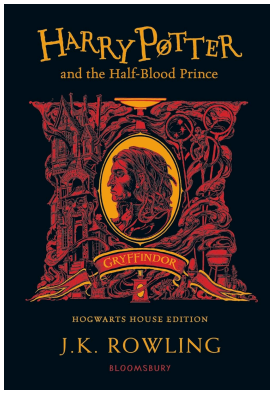
1999



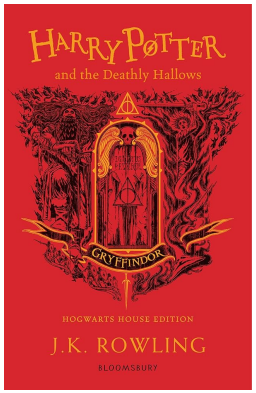
2000



2003



2005



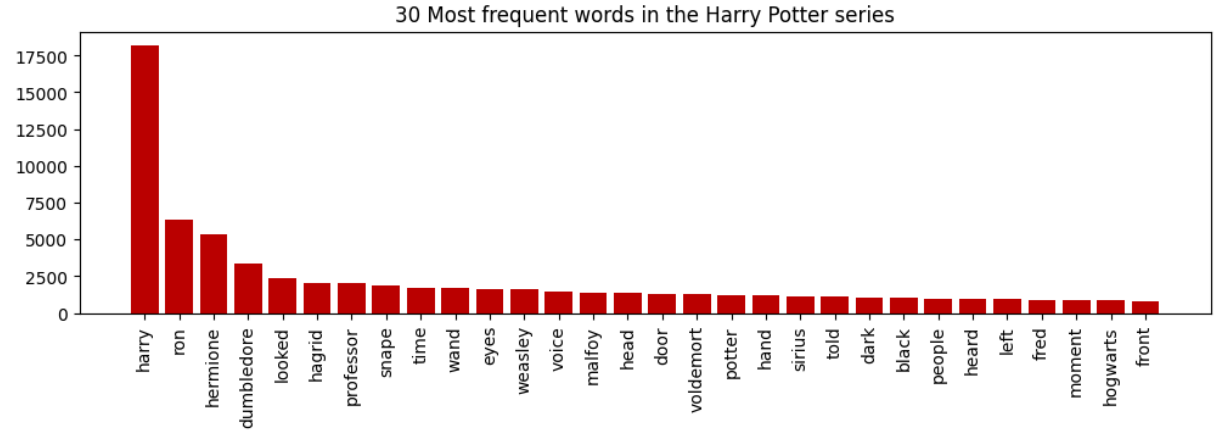
2007

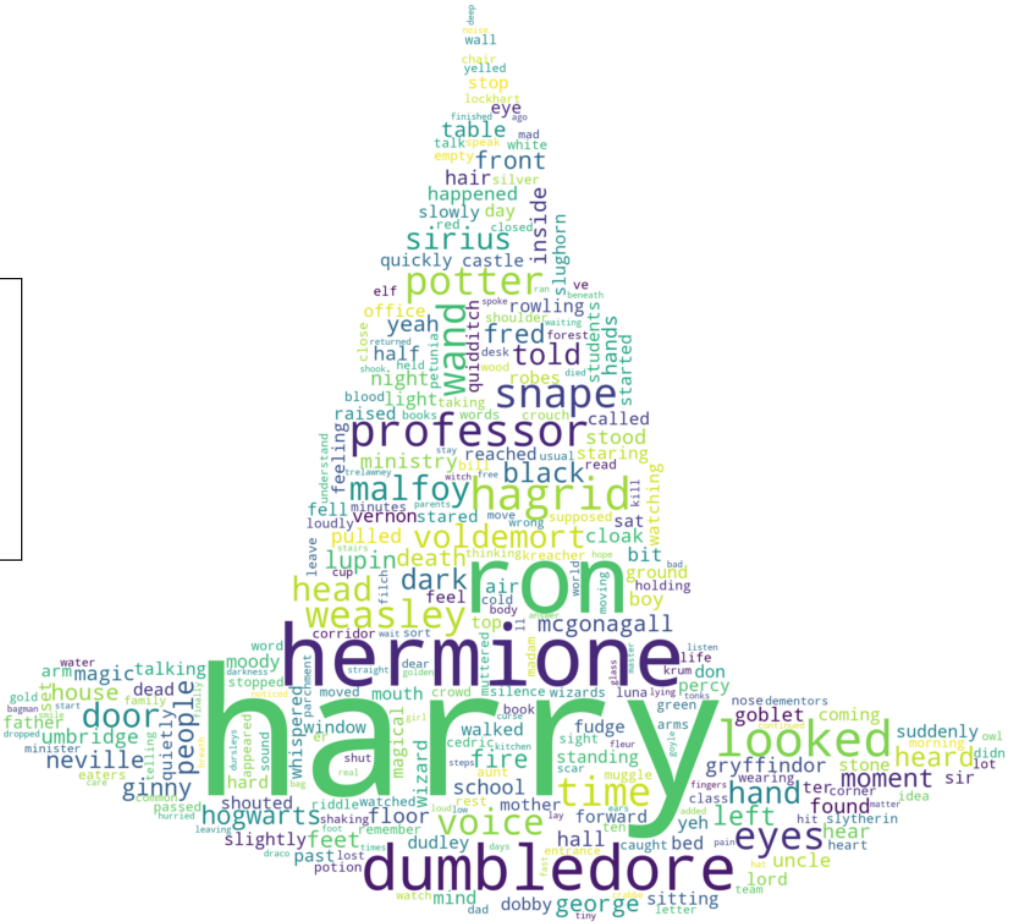
2. Preprocessing

- Obtain the tokens from the books;
- Remove stop-words and special characters;
- Obtain the stemmed words

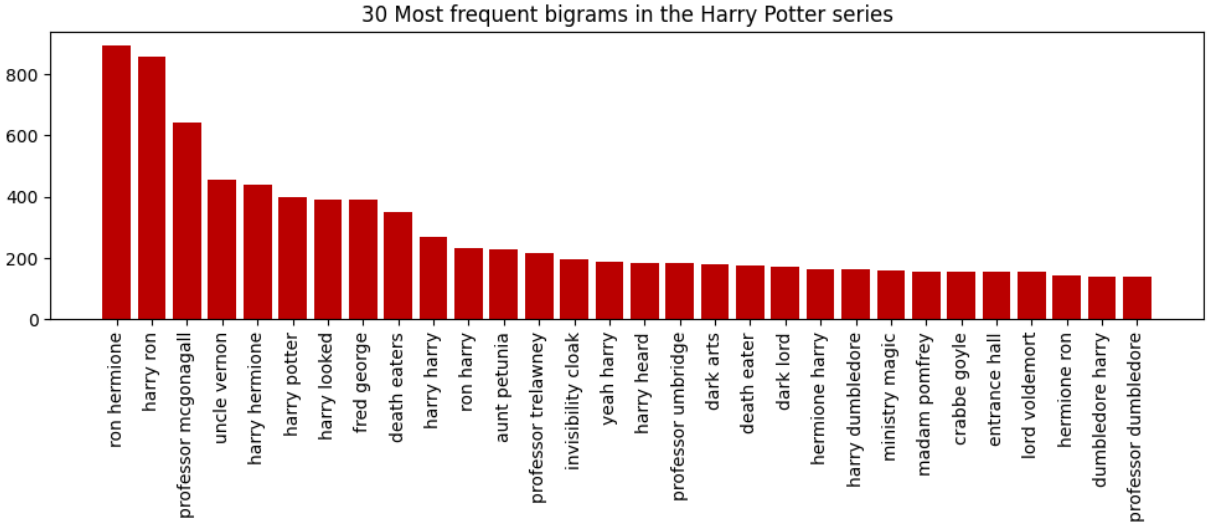
```
class CustomTokenizer:
    def __init__(self):
        self.patterns = [
            (r'[.,;!?]', 'PUNCTUATION'),  # Matches common punctuation
            (r"\b\w+'t\b|\b\w+\b|'\w+\b", "WORD") #specific pattern for words
        ]

    def tokenize(self, text):
        tokens = []
        for pattern, token_type in self.patterns:
            regex = re.compile(pattern)
            matches = regex.finditer(text)
            for match in matches:
                tokens.append((match.group(), token_type))
        return tokens
```

3. Most common words



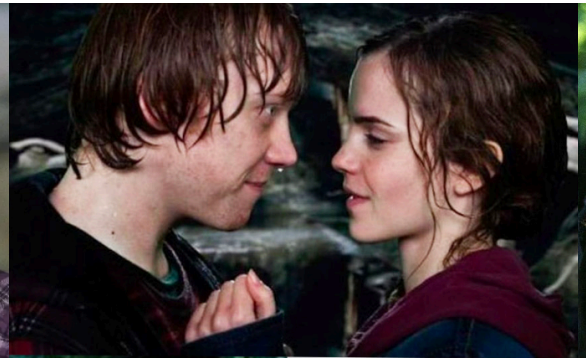


3. Bigrams


30 Most frequent bigrams in the Harry Potter series
800
600
400
200
0
ron hermione
harry ron
professor mcgonagall
uncle vernon
harry hermione
harry potter
harry looked
fred george
death eaters
harry harry
ron harry
aunt petunia
professor trelawney
invisibility cloak
yeah harry
harry heard
professor umbridge
dark arts
death eater
dark lord
hermione harry
harry dumbledore
ministry magic
madam pomfrey
crabbe goyle
entrance hall
lord voldemort
hermione ron
dumbledore harry
professor dumbledore


harry ron
ron hermione
professor mcgonagall
harry hermione
death eater
harry stared
ron looked
madam pomfrey
dark arts
professor lupin
time harry
house elf
harry told
told harry
fred george
uncle vernon
lord voldemort
rita skeeter
defense dark
ron harry
professor trelawney
invisibility cloak
hermione ron
harry quickly
hermione looked
harry dumbledore
harry potter
death eaters
dark lord
shook head
harry looked
harry harry
dumbledore harry
aunt petunia
entrance hall
yeah harry
raised wand
madame maxime
harry heard
mad eye
crabbe goyle
ministry magic
harry eyes
harry found
prime minister
professor umbridge
hermione harry
hagrid harry
albus dumbledore
voice harry
daily prophet
hospital wing
moment harry
draco malfoy
looked harry

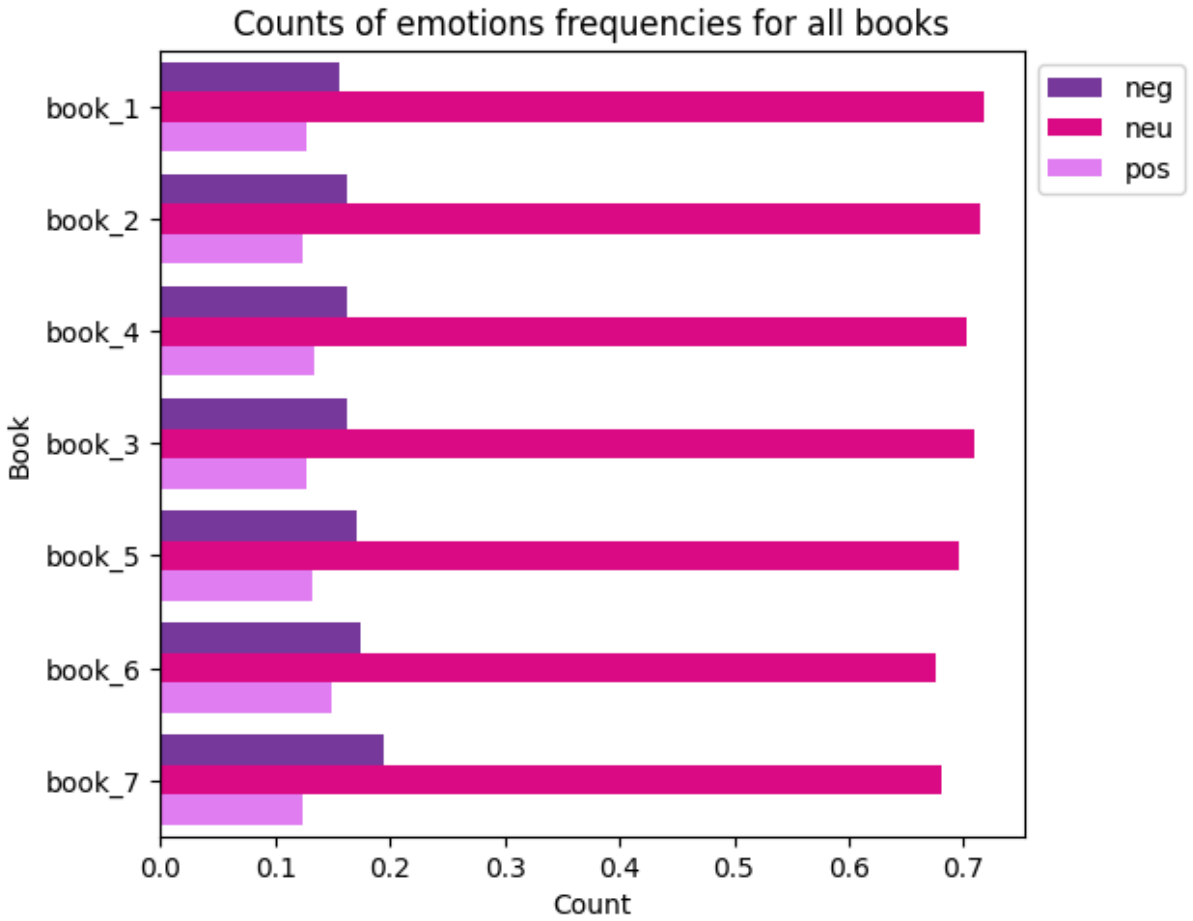




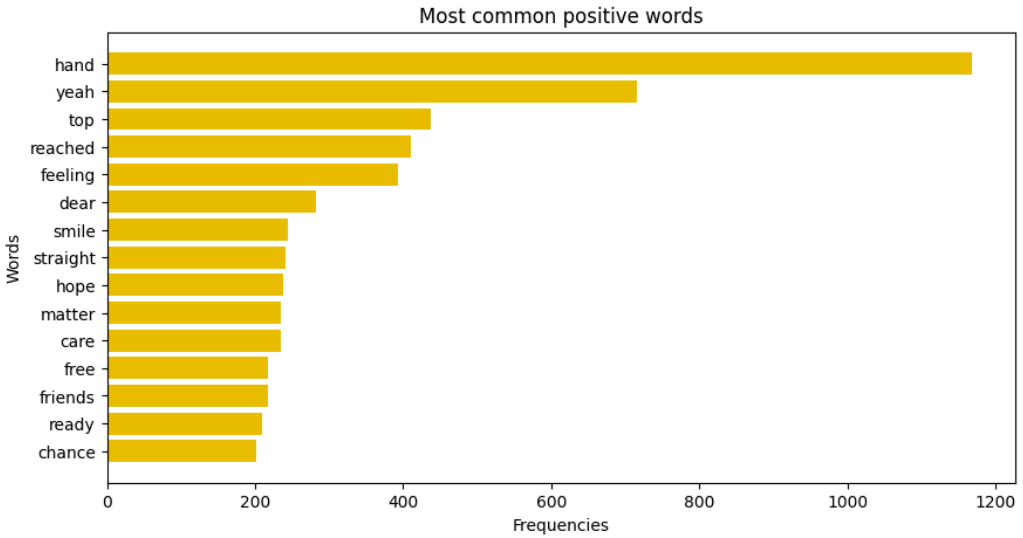

SENTIMENT ANALYSIS

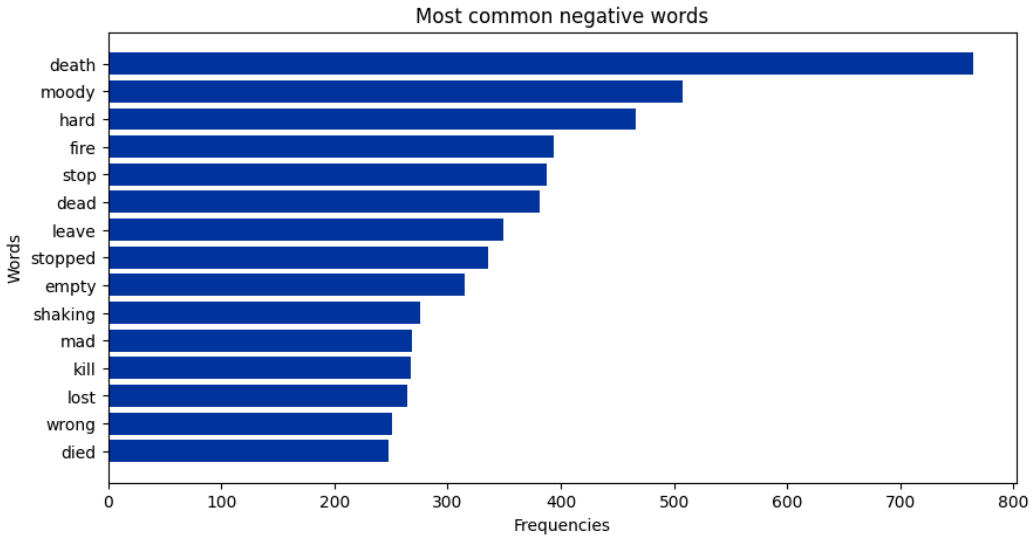



3. Sentiment Analysis

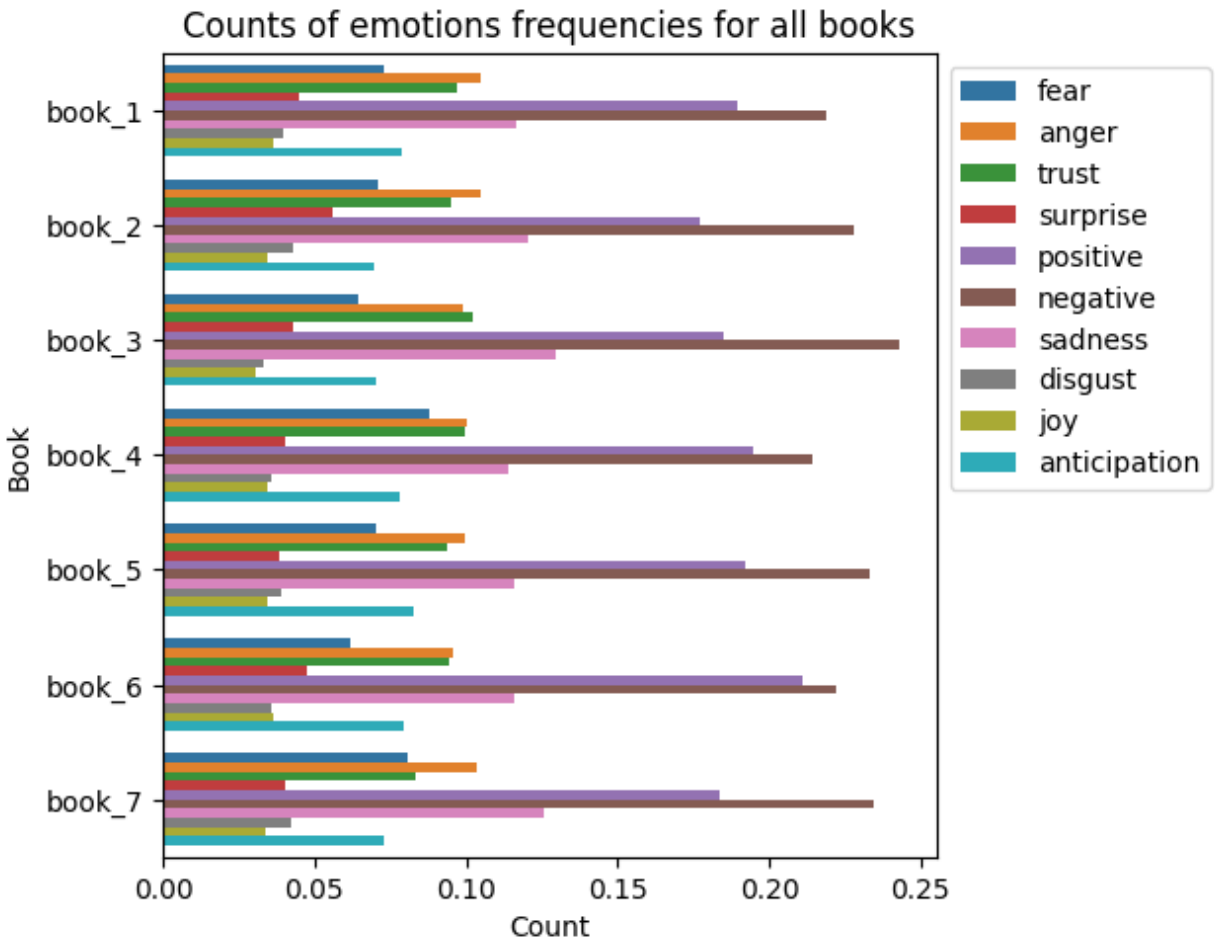


3. Sentiment Analysis

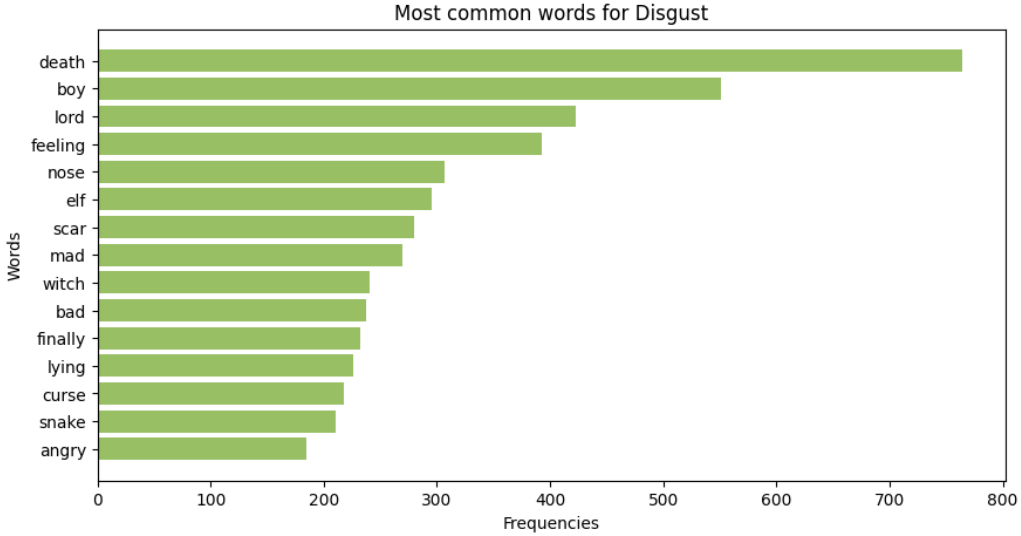


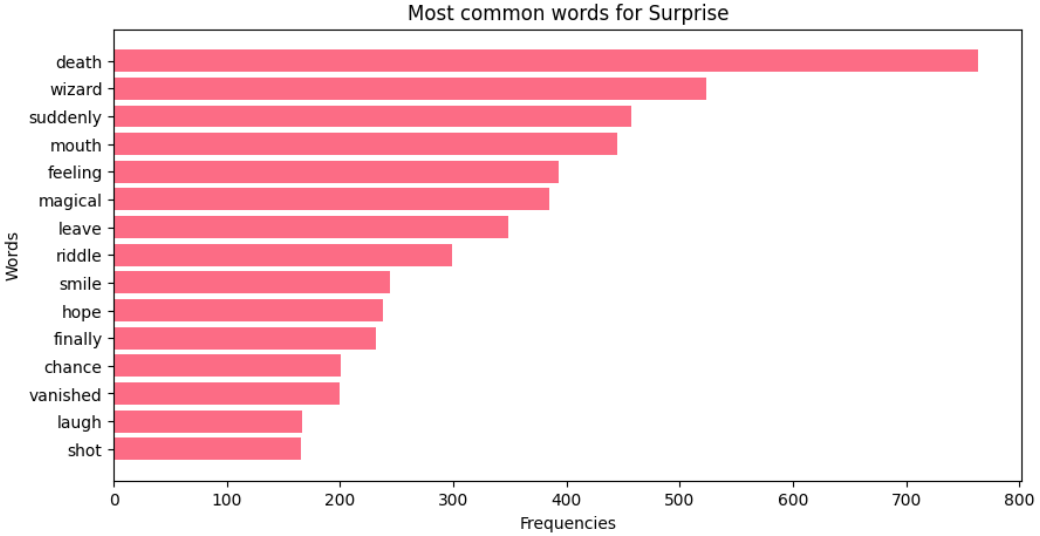


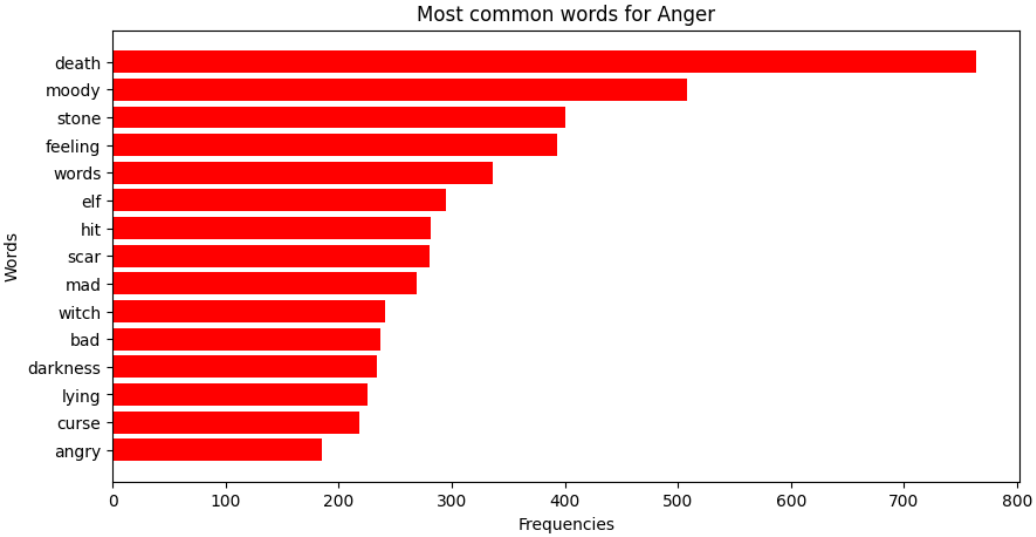
4. Sentiment analysis

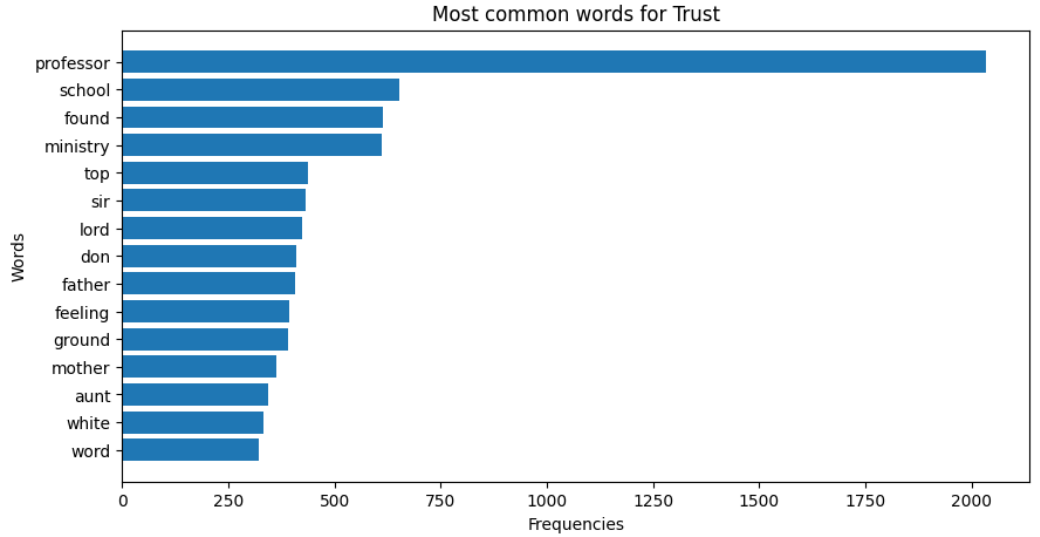


3. Sentiment Analysis

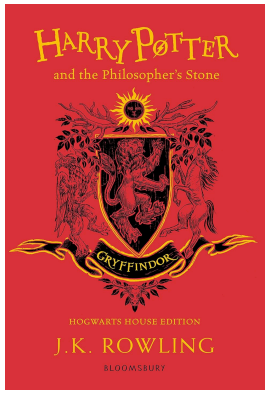




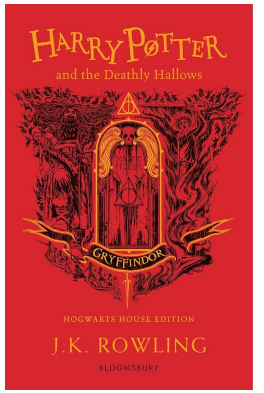




\



? ? ? ? ?



Harry Potter

and the philosopher stone

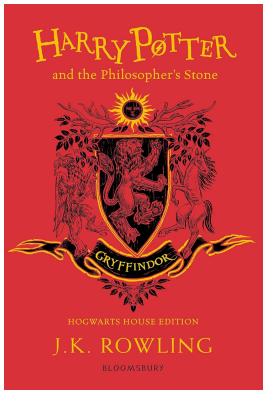
Harry Potter

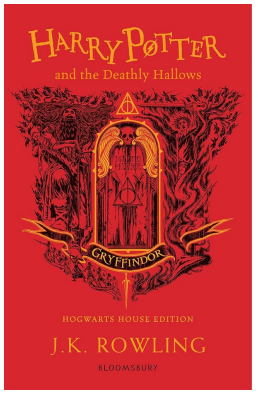
and the deathly Hallows

4. Books order - The algorithm

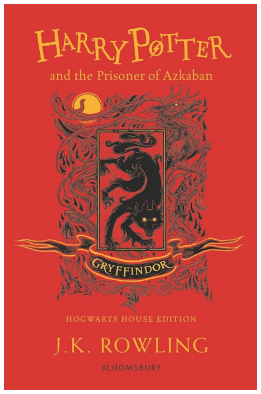
- Populations: W_2, W_3, W_4, W_5, W_6;
- Scores: $\overline{s_k} = \frac{1}{N_k} \sum_{i=1}^{N} n_i \log \frac{\theta_{7i}}{\theta_{1i}} = \sum_{i=1}^{N} n_i s_i$
- Variances: $\hat{V}(\overline{s}_k) = \frac{1}{N_k(N_k-1)} \left(\sum_{i=1}^{N} n_i s_i^2 - \frac{1}{N_k} \left(\sum_{i=1}^{n} n_i s_i \right)^2 \right)$

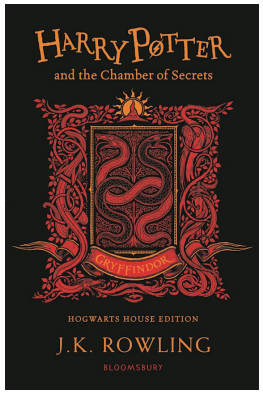
4. Books order - Results

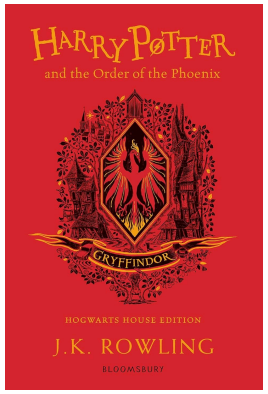


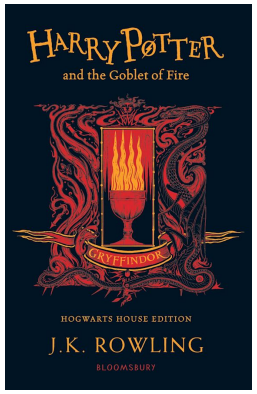


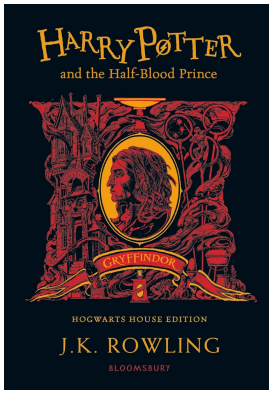
1997 — 1999 — 1998 — 2003 — 2000 — 2005 — 2007



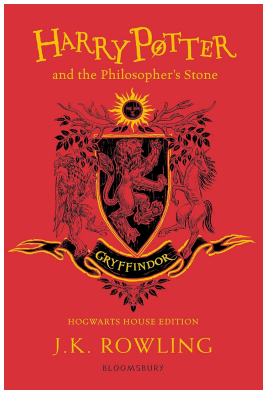


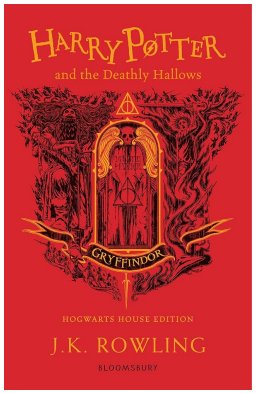




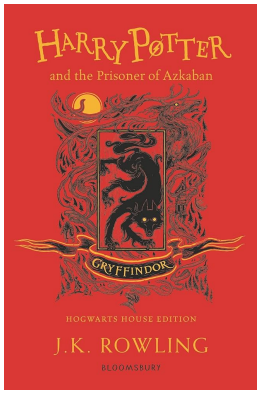


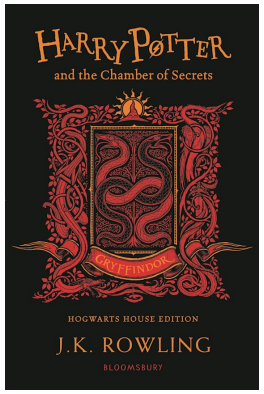
4. Harry Potter and the order of the books

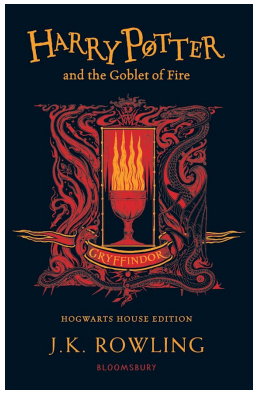


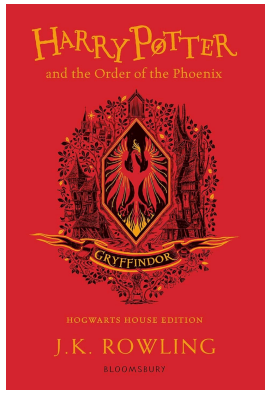


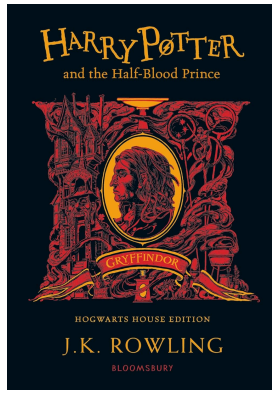
1997 — 1999 — 1998 — 2000 — 2003 — 2005 — 2007



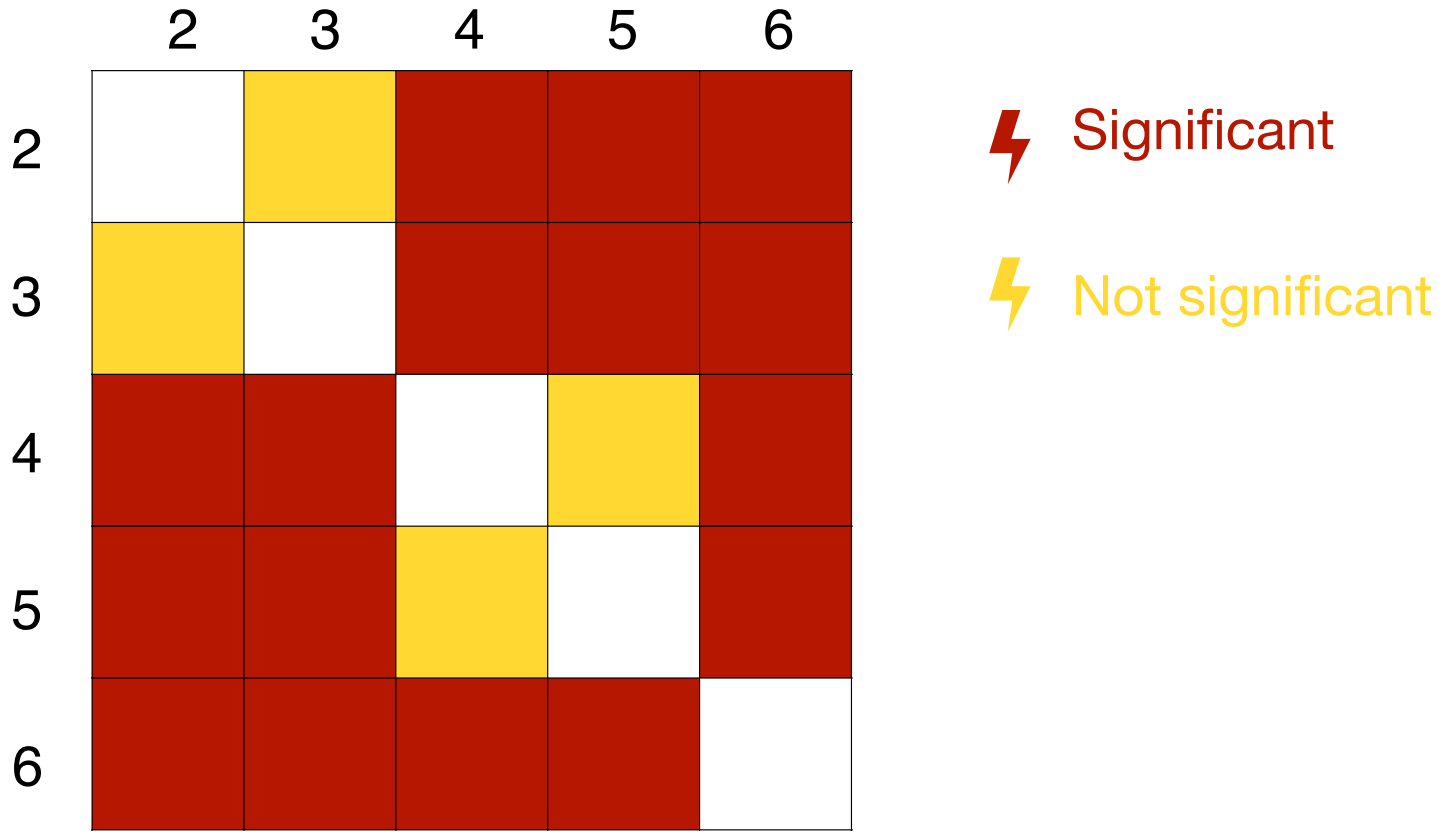








4. Harry Potter and the order of the books



5. Clustering

The similarities between books has been evaluated using the cosine similarity:

$$d_{cos}(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2} \sqrt{\sum_{i=1}^{p} y_i^2}}$$

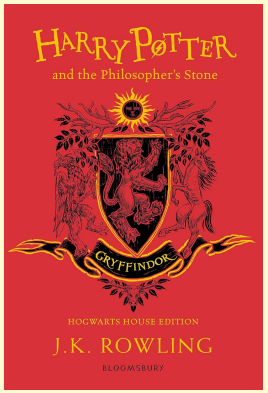
a convenient measure to evaluate the homogeneity level in a group is:

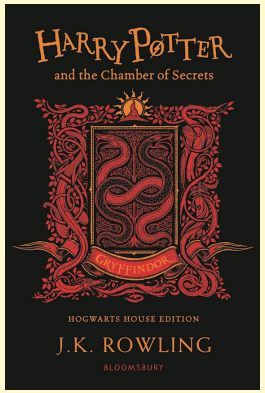
$$Cohesion = \sum_{x \in C} d_{cos}(x, c)$$

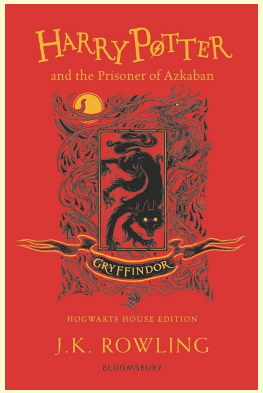
and finally, the total cohesion among groups can be calculated as:

$$Total\ cohesion = \sum_{r=1}^{k} \sum_{x \in C_r} d_{cos}(x, c_r)$$

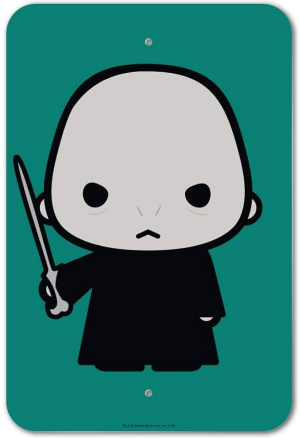
6. Clustering k-means

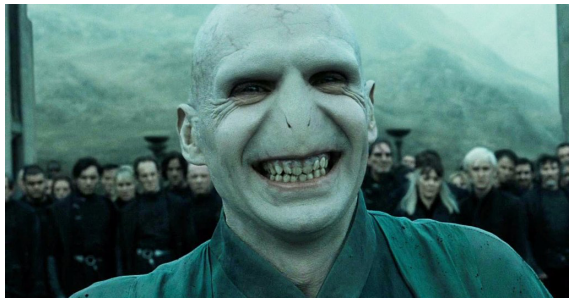




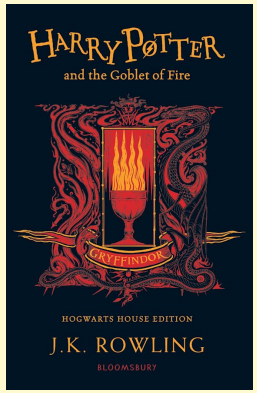


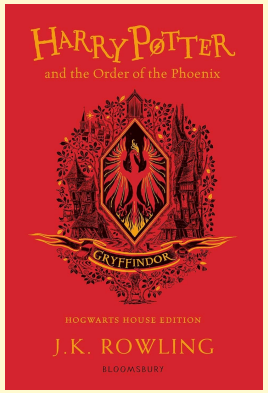
B.V

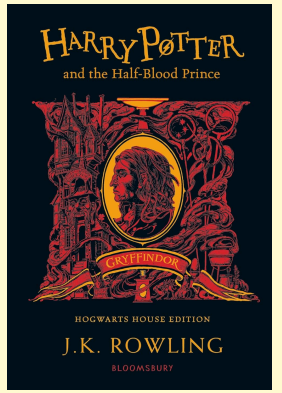


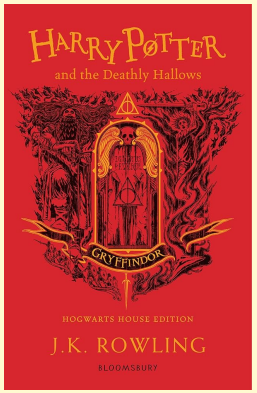


A.V









7. Spectral Clustering

First we need to represent the data as a graph, so we calculate distances between the books. We define the matrices

$$W = (d_{cos}(x_i, x_j))_{ij} \qquad \text{and} \qquad D = diag(d_i)$$

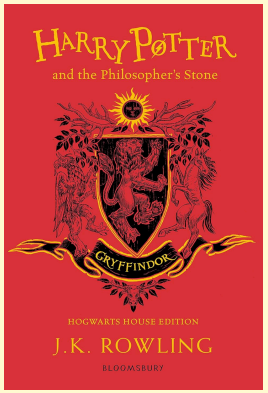
where D is a diagonal matrix, whose diagonal elements correspond to the degree associated to each book.

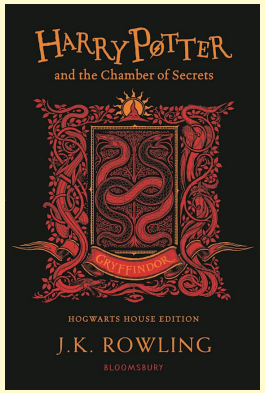
Then we compute:

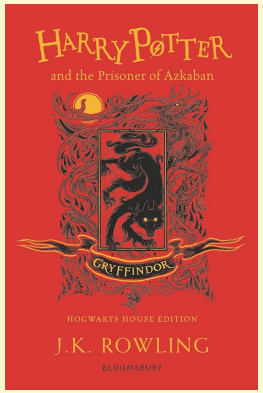
$$L = D - W \qquad \text{and} \qquad L_{sym} = D^{-1/2} L D^{-1/2}$$

We run a clustering algorithm (k-means) on the first k eigenvectors associated to the k smallest eigenvalues, where k is equal to the number of clusters desired. In this case we choose to divide the books in only two clusters.

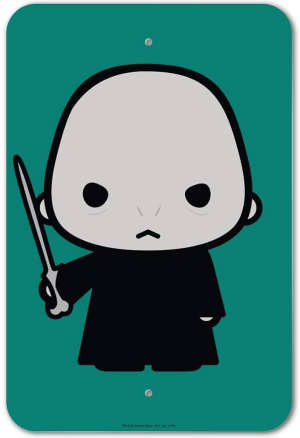
7. Spectral Clustering

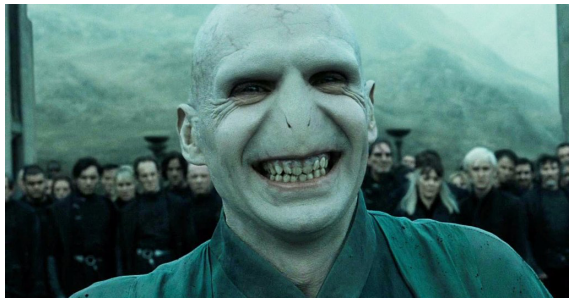




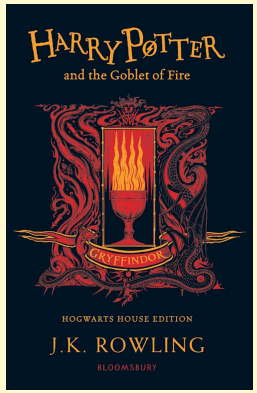


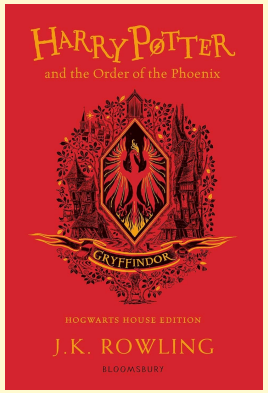
B.V

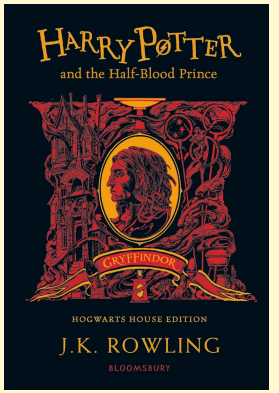


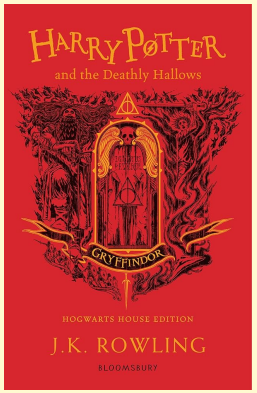


A.V



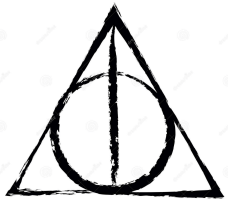






19 minutes later… (we hope)

THANK YOU FOR THE ATTENTION



Enrico Carraro Alex Cecchetto Virginia Murru