EX 04
Applying NLP

Introduction

Natural Language Processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

In this assignment we will apply Natural Language Processing techniques like Sentiment analysis, topic modelling and word clouds on the Harry Potter Books to learn how these techniques can give us significant insights about large text datasets.

Data preparation

In order to get significant insights from the books, we first need to prepare the data. Data preparation in natural language processing mainly involves text cleaning. In NLP we do not have any missing values, target variable or independent variables, all we have is text. Therefore, data preparation for NLP varies from our standard data preparation.

We first open the text file in python and insert it into a list. We then remove all the punctuations, numeric characters, special characters and stop words. Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words removal is critical for many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead. This helps us get meaningful insights from a given text data, rather than getting commonly used words in the English language like, ‘and’, ’the’,’ for’, ’don’t’, etc. I have not only removed the standard English stop words but I have updated the stop words list with context based stop words.

For example, as the books are about Harry Potter it is obvious that the books will mention Harry's name multiple times and we do not want our word clouds to depict something we already know. Thus, removing context-based stop words is also a crucial step.

Data cleaning was performed by defining functions to remove punctuations, stop words, special characters and numbers. After removing the unnecessary characters, we tokenize the list. Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms.
The functions used for cleaning and tokenizing are given below:

#Cleaning

```
def clean_text(text):
    text = " ".join(text)
    text = text.lower()
    text = text.strip()
    text = re.sub('\[.*?\]’', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '',
           text)
    text = re.sub('\w*\d\w*', '', text)
    text = text.replace("’","")
    text = re.sub('[^A-Za-z0-9]+', ' ', text)

    return text
```

#Tokenization

```
def final_clean(texts):
    toks = word_tokenize(texts)
    stp = [word for word in toks if word not in stop_words]
    #stpr = ' '.join(stp)
    return stp
```

Analysis and Insights

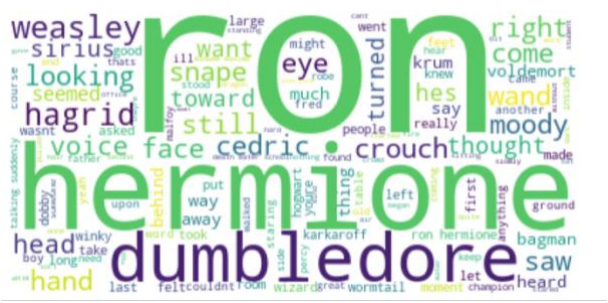
The **word clouds** for all the books are shown below:

Book 1

Book 2

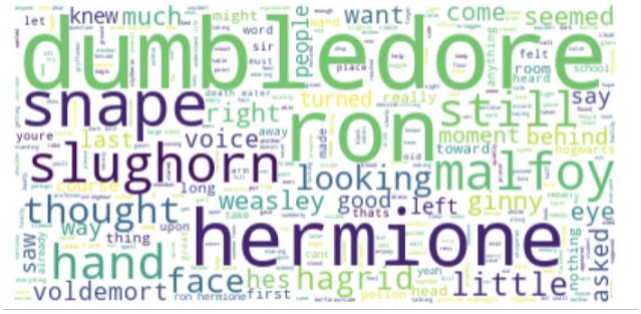
Book 3

Book 4

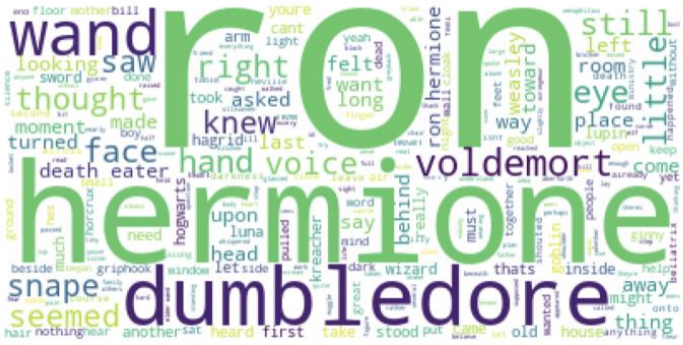


Book 5

Book 6



Book 7



The top 10 bigrams for all the books are shown below:

The top 10 bigrams for Book 2 are:

```
['uncle vernon', 'professor mcgonagall', 'aunt petunia', 'ron
hermione', 'professor dumbledore', 'common room', 'professor
quirrell', 'crabbe goyle', 'privet drive', 'fred george']
```

The top 10 bigrams for Book 2 are:

```
 ['professor mcgonagall', 'ron hermione', 'chamber secrets',
'fred george', 'uncle vernon', 'nearly headless', 'headless
nick', 'gilderoy lockhart', 'madam pomfrey', 'common room']
```

The top 10 bigrams for Book 3 are:

```
 ['ron hermione', 'professor lupin', 'professor mcgonagall',
'professor trelawney', 'uncle vernon', 'sirius black', 'aunt
marge', 'common room', 'madam pomfrey', 'crabbe goyle']
```

The top 10 bigrams for Book 4 are:

```
 ['ron hermione', 'madame maxime', 'professor mcgonagall', 'fred
george', 'rita skeeter', 'uncle vernon', 'death eaters', 'great
hall', 'world cup', 'ludo bagman']
```

The top 10 bigrams for Book 5 are:

```
 ['uncle vernon', 'aunt petunia', 'privet drive', 'ministry
magic', 'uncle vernons', 'little whinging', 'ron hermione',
'magnolia crescent', 'kingsley shacklebolt', 'telephone box']
```

The top 10 bigrams for Book 6 are:

```
 ['ron hermione', 'prime minister', 'death eaters', 'professor
mcgonagall', 'common room', 'dark lord', 'death eater', 'dark
arts', 'invisibility cloak', 'lord voldemort']
```

The top 10 bigrams for Book 7 are:

```
 ['ron hermione', 'death eaters', 'death eater', 'elder wand',
'invisibility cloak', 'godrics hollow', 'professor mcgonagall',
'dark lord', 'asked hermione', 'deathly hallows']
```

The sentiment analysis for each book is given below:

(Book I) Harry Potter and the Sorcerer's Stone

It is positive for 14.1%, It is negative for 13.5%, It is neutral for 72.4%

(Book II) Harry Potter and the Chamber of Secrets

It is positive for 13.8%, It is negative for 14.3%, It is neutral for 71.9%

(Book III) Harry Potter and the Prisoner of Azkaban

It is positive for 14.1%, It is negative for 14.4%, It is neutral for 71.5%

(Book IV) Harry Potter and the Goblet of Fire

It is positive for 14.7%, It is negative for 14.3%, It is neutral for 71.0%

(Book V) Harry Potter and the Order of the Phoenix

It is positive for 12.3%, It is negative for 15.7%, It is neutral for 72.1%

(Book VI) Harry Potter and the Half-Blood Prince

It is positive for 16.2%, It is negative for 15.3%, It is neutral for 68.5%

(Book VII) Harry Potter and the Deathly Hallows

It is positive for 13.7%, It is negative for 16.5%, It is neutral for 69.8%

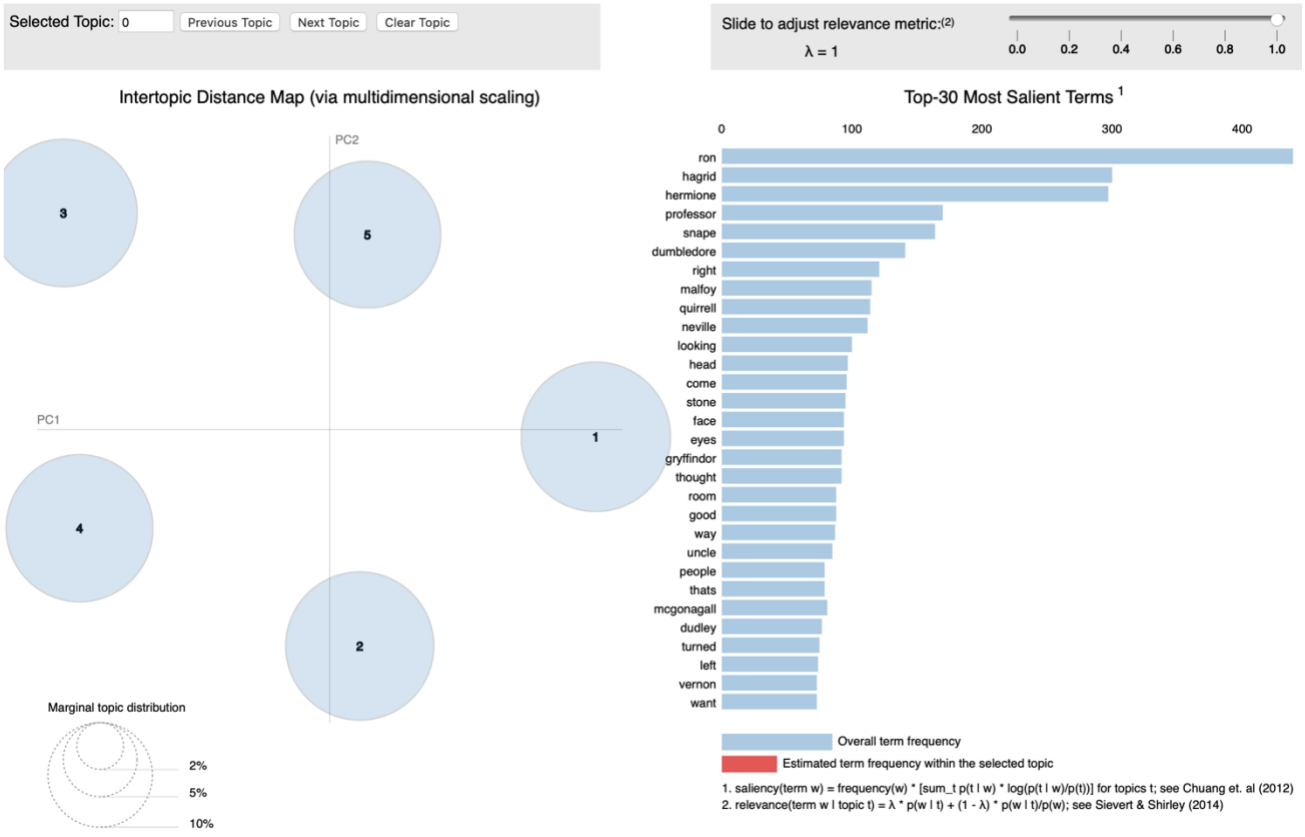
It is evident that all the books generally carry a neutral sentiment which tells that the book is written in a way where the author is narrating the events taking place.

The **Topic detection** was done using LDA for all the books. The Topics detected using LDA for Book 1 are shown below as an example.

```
##############THE 5 TOPICS FOR BOOK 1 ARE AS FOLLOWS###############

LDA Model:
Topic 0:
[('right', 127.7950013880537), ('come', 101.53753614306345), ('head', 100.86405490230533), ('turned', 77.8547889949
2105), ('wasnt', 74.38776942313481), ('boy', 69.42809390244054), ('long', 66.81520558033382), ('voice', 62.29926560
646637), ('wood', 55.31502628685677), ('slytherin', 52.78722173236256)]


Topic 1:
[('professor', 179.456747502407), ('dumbledore', 149.966715957311), ('stone', 98.8035567581346), ('good', 91.274390
9099658), ('mcgonagall', 88.13175544022147), ('want', 77.72675959211084), ('hogwarts', 70.662142519804), ('bit', 65
.40367663151267), ('cloak', 65.08321708149312), ('saw', 61.75379222029426)]


Topic 2:
[('hermione', 326.20376963077524), ('malfoy', 119.77067976976701), ('quirrell', 119.20119167029944), ('neville', 11
8.76807149364345), ('eyes', 99.30474362392765), ('gryffindor', 96.24973781150317), ('uncle', 88.80914313389928), ('
vernon', 77.09115089814652), ('heard', 75.49468426873447), ('let', 71.89279782606309)]


Topic 3:
[('snape', 173.6814635002374), ('face', 99.70529354068421), ('thought', 95.76654434368034), ('way', 91.438605381482
28), ('people', 82.36900374310346), ('dudley', 80.8784500595305), ('left', 78.62748037874454), ('hes', 75.274068682
34219), ('really', 73.2953785074758), ('came', 71.72596899605641)]


Topic 4:
[('ron', 466.7082003708645), ('hagrid', 329.5882191528413), ('looking', 104.41203090355783), ('room', 91.6372056799
7073), ('thats', 83.6634910155122), ('ter', 73.08645980894674), ('away', 72.8949525684498), ('suddenly', 70.7792291
5023336), ('great', 70.23712279849842), ('house', 69.99658994869704)]
```

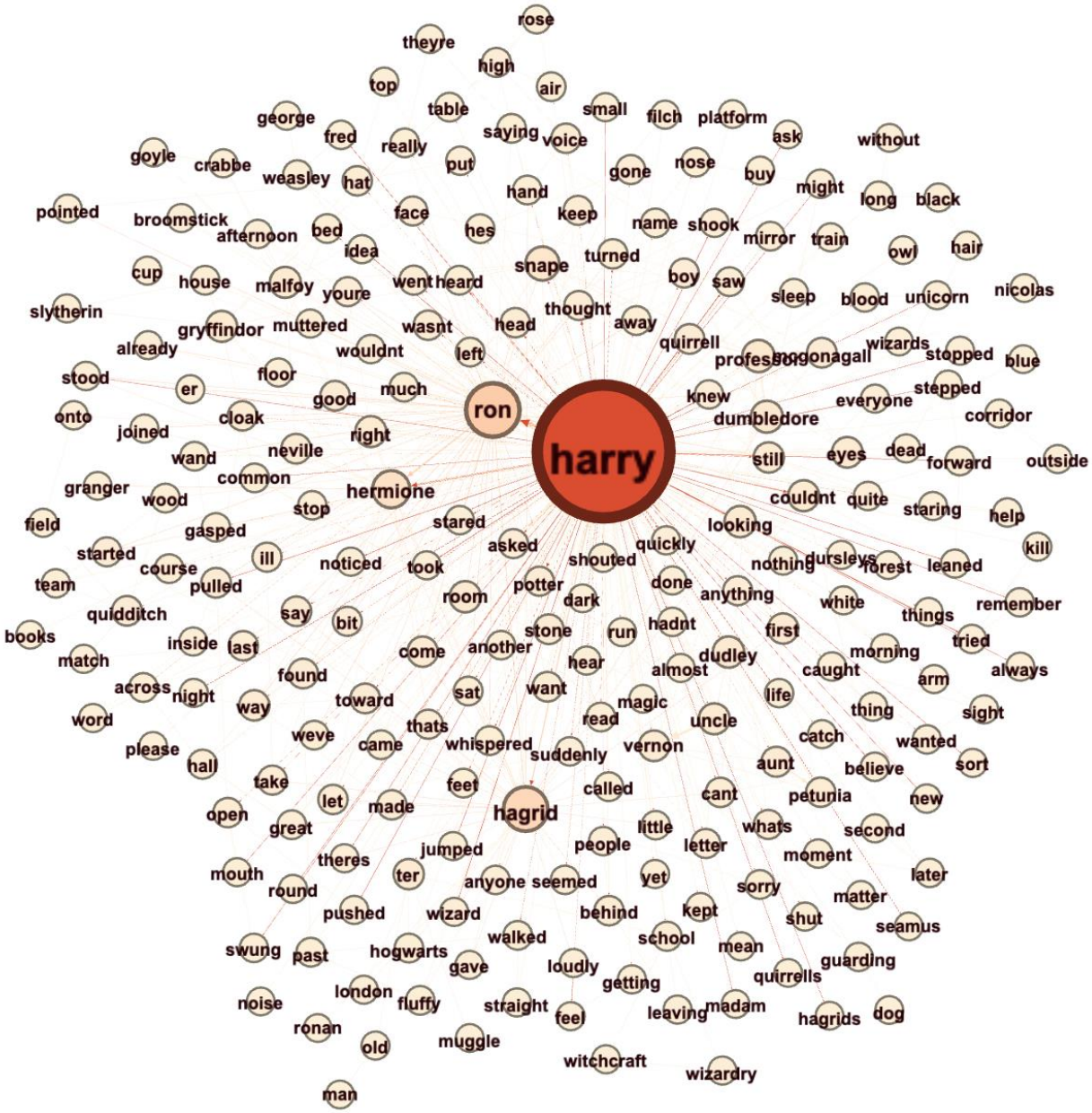
I have also generated a html file for visualizing these topics and they keywords in those topics with respect to their weightage. This gives us a clean understanding of the topics and their overlap, along with the weightage for each word associated with those topics. This is crucial for finding meaningful insights from unstructured text data.

The html file for book 1 is shown below as an example:



All the other html files will be generated on your system when you run the python file attached. You can then interact within those html files to deep dive into the detected topics to find interesting insights.

The network diagram of Book 1 is shown below:



Gephi was used to create the network diagram above. It is an interactive diagram so you can hover over any of the nodes and see the other nodes which are connected to it. This allows us to gain significant insights about how a word(node) is connected to others. Harry is the center of the entire network which is quite obvious.

As an example, when I hover over the word ‘stone’ it can be seen from the image below that Harry, Ron, Snape, Quirrell have some connection with the stone. As we know from having read the Harry Potter Book 1, the **sorcerer’s stone** had a lot to do with **Professor Quirrell** and the network diagram below confirms this connection.