**Step1:**

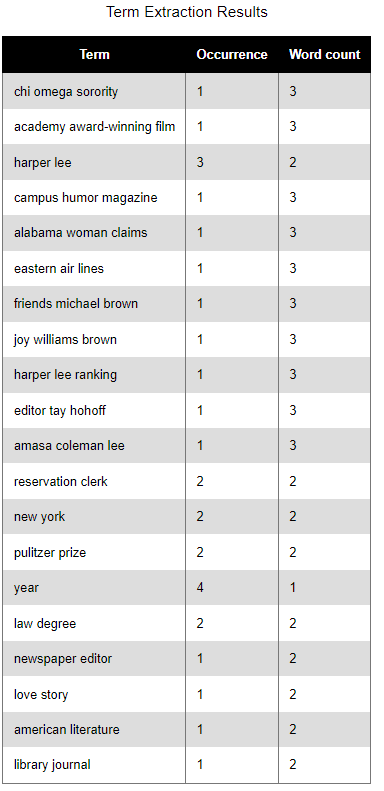
Ground Truth: According to me, the important terms in the ten documents, and their occurrences are:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S. No. | Term | Doc 1 Count | Doc 2 Count | Doc 3 Count | Doc 4 Count | Doc 5 Count | Doc 6 Count | Doc 7 Count | Doc 8 Count | Doc 9 Count | Doc 10 Count |
| 1 | Source | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | List | 2 | 4 | 2 | 3 | 4 | 2 | 2 | 2 | 2 | 2 |
| 3 | Title | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 |
| 4 | Author | 2 | 5 | 2 | 2 | 4 | 2 | 2 | 4 | 4 | 2 |
| 5 | Genre | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 6 | 1 |
| 6 | Classics | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 7 | Fiction | 3 | 3 | 2 | 2 | 3 | 3 | 4 | 4 | 1 | 3 |
| 8 | Literature | 3 | 2 | 0 | 1 | 2 | 2 | 0 | 0 | 2 | 2 |
| 9 | Adult | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| 10 | Novel | 4 | 5 | 1 | 7 | 3 | 6 | 6 | 5 | 6 | 3 |
| 11 | Total Word Count | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12 | Key Words | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | Brief Summary | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 14 | love | 3 | 2 | 0 | 1 | 0 | 1 | 1 | 7 | 5 | 6 |
| 15 | place | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 16 | bestseller | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 1 |
| 17 | About the Author | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 18 | children | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | best | 6 | 3 | 1 | 2 | 3 | 2 | 4 | 4 | 3 | 5 |
| 20 | critical | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

**Step 2:**

Term Extractor 1: Five Filters <http://termextract.fivefilters.org/>

Document 1:



Document 2:

A table of text with black text

Description automatically generated with medium confidence

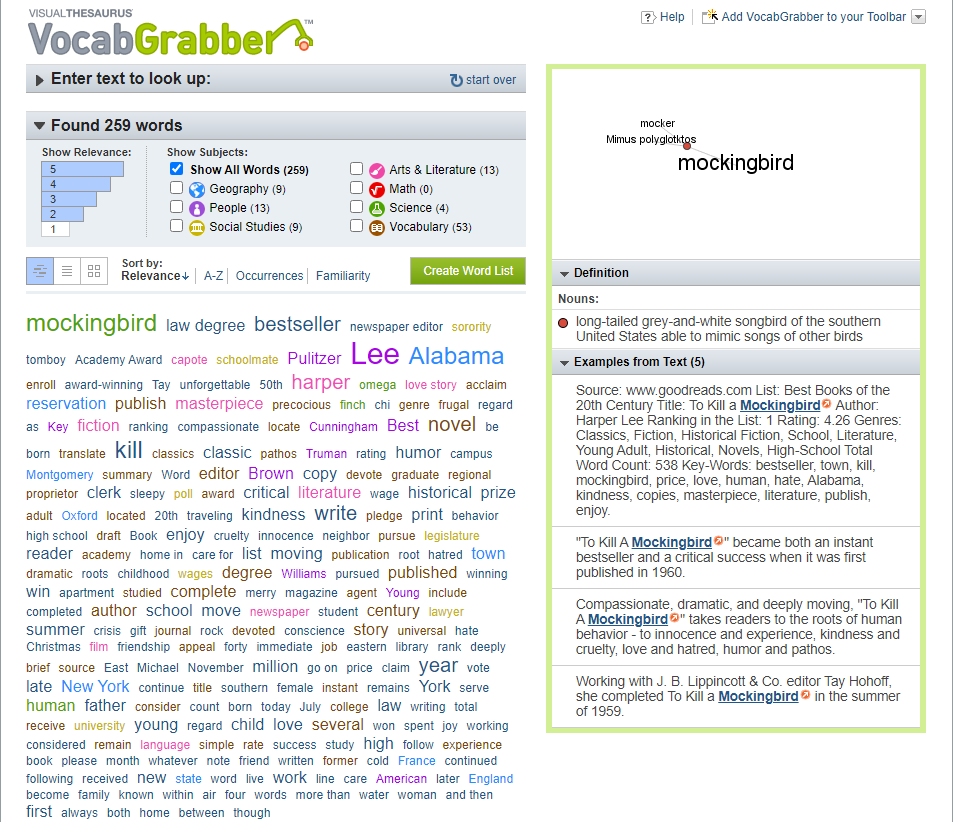
Document 3:

A table with text and numbers

Description automatically generated

Term Extractor 2: Vocab Grabber (<https://www.visualthesaurus.com/vocabgrabber/#>)

Document 1:



Document 2:

A screenshot of a computer

Description automatically generated

Document 3:

A screenshot of a computer

Description automatically generated

**Step 3:**

Applying NLTK tokenizer method (after lower-casing), to all ten documents, returns:

File: Doc1.docx

Most frequent 1-word terms:

the: 25

and: 19

in: 18

of: 15

to: 14

lee: 9

she: 9

was: 7

her: 7

it: 6

Most frequent 2-word terms:

kill mockingbird: 5

to kill: 4

harper lee: 4

of the: 3

in the: 3

in it: 3

at the: 3

the century: 2

author harper: 2

novel of: 2

Most frequent 3-word terms:

to kill mockingbird: 4

of the century: 2

author harper lee: 2

the pulitzer prize: 2

million copies in: 2

copies in print: 2

as reservation clerk: 2

source list best: 1

list best books: 1

best books of: 1

--------------------------------------------------

File: Doc2.docx

Most frequent 1-word terms:

the: 23

and: 20

of: 14

in: 11

orwell: 10

his: 9

is: 6

he: 6

by: 6

was: 6

Most frequent 2-word terms:

of the: 3

george orwell: 3

with the: 3

in the: 2

big brother: 2

have entered: 2

dystopian novel: 2

for any: 2

orwell was: 2

language and: 2

Most frequent 3-word terms:

the spanish civil: 2

spanish civil war: 2

source list best: 1

list best books: 1

best books of: 1

books of the: 1

of the century: 1

the century title: 1

century title author: 1

title author george: 1

--------------------------------------------------

File: Doc3.docx

Most frequent 1-word terms:

the: 18

her: 17

she: 13

of: 12

rowling: 12

and: 10

was: 9

harry: 8

in: 8

name: 8

Most frequent 2-word terms:

harry potter: 5

her mother: 5

she had: 3

of her: 3

in the: 2

infamous school: 2

mystical world: 2

he is: 2

under the: 2

pen name: 2

Most frequent 3-word terms:

me when was: 2

said of her: 2

she had difficult: 2

source list best: 1

list best books: 1

best books of: 1

books of the: 1

of the century: 1

the century title: 1

century title harry: 1

--------------------------------------------------

File: Doc4.docx

Most frequent 1-word terms:

the: 25

of: 15

he: 9

in: 8

his: 8

and: 8

fitzgerald: 6

story: 5

to: 5

jazz: 4

Most frequent 2-word terms:

of the: 9

in the: 3

jazz age: 3

new york: 3

the jazz: 3

the century: 2

the great: 2

great gatsby: 2

scott fitzgerald: 2

lavish parties: 2

Most frequent 3-word terms:

of the jazz: 3

of the century: 2

the great gatsby: 2

the jazz age: 2

as one of: 2

one of the: 2

writers of the: 2

source list best: 1

list best books: 1

best books of: 1

--------------------------------------------------

File: Doc5.docx

Most frequent 1-word terms:

the: 24

and: 19

of: 13

in: 11

orwell: 9

his: 9

was: 8

to: 7

is: 6

by: 6

Most frequent 2-word terms:

of the: 4

animal farm: 3

george orwell: 3

with the: 3

in the: 2

progress justice: 2

to totalitarianism: 2

orwell was: 2

language and: 2

the spanish: 2

Most frequent 3-word terms:

the spanish civil: 2

spanish civil war: 2

source list best: 1

list best books: 1

best books of: 1

books of the: 1

of the century: 1

the century title: 1

century title animal: 1

title animal farm: 1

--------------------------------------------------

File: Doc7.docx

Most frequent 1-word terms:

the: 37

in: 20

and: 20

of: 12

mitchell: 11

her: 10

with: 9

to: 8

american: 7

was: 6

Most frequent 2-word terms:

with the: 5

of the: 4

gone with: 4

the wind: 4

in the: 4

civil war: 3

in and: 3

her stories: 3

the best: 2

romance novels: 2

Most frequent 3-word terms:

gone with the: 4

with the wind: 4

american civil war: 2

her stories were: 2

source list the: 1

list the best: 1

the best of: 1

best of the: 1

of the best: 1

the best romance: 1

--------------------------------------------------

File: Doc8.docx

Most frequent 1-word terms:

the: 22

her: 21

of: 16

and: 16

she: 14

to: 12

in: 10

rivers: 9

romance: 7

love: 6

Most frequent 2-word terms:

of the: 5

romance novels: 3

redeeming love: 3

francine rivers: 3

in the: 3

michael hosea: 3

the best: 2

author francine: 2

christian fiction: 2

fiction historical: 2

Most frequent 3-word terms:

author francine rivers: 2

romance writers of: 2

hall of fame: 2

source list the: 1

list the best: 1

the best of: 1

best of the: 1

of the best: 1

the best romance: 1

best romance novels: 1

--------------------------------------------------

File: Doc10.docx

Most frequent 1-word terms:

the: 41

of: 15

with: 13

to: 12

is: 11

in: 10

and: 8

notebook: 7

that: 7

his: 6

Most frequent 2-word terms:

the notebook: 5

of the: 4

in the: 4

the best: 3

nicholas sparks: 3

story of: 3

with the: 3

best of: 2

author nicholas: 2

fiction contemporary: 2

Most frequent 3-word terms:

the best of: 2

author nicholas sparks: 2

the story of: 2

story of noah: 2

over million copies: 2

source list the: 1

list the best: 1

best of the: 1

of the best: 1

the best romance: 1

--------------------------------------------------

File: Doc9.docx

Most frequent 1-word terms:

the: 29

her: 19

of: 17

to: 15

in: 10

and: 10

mcnaught: 9

historical: 8

with: 8

romance: 7

Most frequent 2-word terms:

in the: 4

of the: 3

whitney my: 3

my love: 3

historical romance: 3

the first: 3

the regency: 3

the best: 2

romance novels: 2

author judith: 2

Most frequent 3-word terms:

whitney my love: 3

author judith mcnaught: 2

the first female: 2

first female executive: 2

female executive producer: 2

executive producer at: 2

producer at cbs: 2

at cbs radio: 2

cbs radio station: 2

the regency period: 2

--------------------------------------------------

File: Doc6.docx

Most frequent 1-word terms:

the: 24

and: 17

in: 14

to: 12

of: 11

gabaldon: 8

her: 8

she: 8

for: 8

war: 5

Most frequent 2-word terms:

of the: 5

the best: 2

author diana: 2

in the: 2

science fiction: 2

outlander scotland: 2

scotland torn: 2

the year: 2

from the: 2

the war: 2

Most frequent 3-word terms:

outlander scotland torn: 2

one of the: 2

gabaldon decided to: 2

of the doctor: 2

source list the: 1

list the best: 1

the best of: 1

best of the: 1

of the best: 1

the best romance: 1

--------------------------------------------------

**Step 4:**

Applying NLTK tokenizer method (after lower-casing and removing stop words), to all ten documents, returns:

File: Doc1.docx

Most frequent 1-word terms:

lee: 9

kill: 5

mockingbird: 5

alabama: 5

harper: 4

bestseller: 4

year: 4

fiction: 3

literature: 3

young: 3

Most frequent 2-word terms:

kill mockingbird: 5

harper lee: 4

author harper: 2

pulitzer prize: 2

million copies: 2

copies print: 2

law degree: 2

new york: 2

reservation clerk: 2

source list: 1

Most frequent 3-word terms:

author harper lee: 2

million copies print: 2

source list best: 1

list best books: 1

best books century: 1

books century title: 1

century title kill: 1

title kill mockingbird: 1

kill mockingbird author: 1

mockingbird author harper: 1

--------------------------------------------------

File: Doc2.docx

Most frequent 1-word terms:

orwell: 10

author: 5

war: 4

best: 3

george: 3

fiction: 3

novel: 3

freedom: 3

police: 3

literary: 3

Most frequent 2-word terms:

george orwell: 3

tale relevant: 2

big brother: 2

dystopian novel: 2

spanish civil: 2

civil war: 2

source list: 1

list best: 1

best books: 1

books century: 1

Most frequent 3-word terms:

spanish civil war: 2

source list best: 1

list best books: 1

best books century: 1

books century title: 1

century title author: 1

title author george: 1

author george orwell: 1

george orwell ranking: 1

orwell ranking list: 1

--------------------------------------------------

File: Doc3.docx

Most frequent 1-word terms:

rowling: 12

harry: 8

name: 8

mother: 6

potter: 5

school: 5

young: 3

world: 3

book: 3

born: 3

Most frequent 2-word terms:

harry potter: 5

infamous school: 2

mystical world: 2

pen name: 2

name joanne: 2

rowling said: 2

source list: 1

list best: 1

best books: 1

books century: 1

Most frequent 3-word terms:

source list best: 1

list best books: 1

best books century: 1

books century title: 1

century title harry: 1

title harry potter: 1

harry potter sorcerer: 1

potter sorcerer stone: 1

sorcerer stone author: 1

stone author rowling: 1

--------------------------------------------------

File: Doc4.docx

Most frequent 1-word terms:

fitzgerald: 6

story: 5

jazz: 4

gatsby: 3

scott: 3

novels: 3

age: 3

new: 3

york: 3

novel: 3

Most frequent 2-word terms:

jazz age: 3

new york: 3

great gatsby: 2

scott fitzgerald: 2

achievement career: 2

lavish parties: 2

drink sex: 2

tale america: 2

story writer: 2

short story: 2

Most frequent 3-word terms:

source list best: 1

list best books: 1

best books century: 1

books century title: 1

century title great: 1

title great gatsby: 1

great gatsby author: 1

gatsby author scott: 1

author scott fitzgerald: 1

scott fitzgerald ranking: 1

--------------------------------------------------

File: Doc5.docx

Most frequent 1-word terms:

orwell: 9

farm: 5

author: 4

best: 3

animal: 3

george: 3

fiction: 3

literary: 3

culture: 3

list: 2

Most frequent 2-word terms:

animal farm: 3

george orwell: 3

progress justice: 2

justice equality: 2

spanish civil: 2

civil war: 2

source list: 1

list best: 1

best books: 1

books century: 1

Most frequent 3-word terms:

progress justice equality: 2

spanish civil war: 2

source list best: 1

list best books: 1

best books century: 1

books century title: 1

century title animal: 1

title animal farm: 1

animal farm author: 1

farm author george: 1

--------------------------------------------------

File: Doc7.docx

Most frequent 1-word terms:

mitchell: 11

american: 7

stories: 6

gone: 4

wind: 4

historical: 4

fiction: 4

war: 4

romance: 3

civil: 3

Most frequent 2-word terms:

gone wind: 4

civil war: 3

romance novels: 2

clayton county: 2

county atlanta: 2

atlanta georgia: 2

georgia american: 2

american civil: 2

daughter plantation: 2

plantation owner: 2

Most frequent 3-word terms:

clayton county atlanta: 2

county atlanta georgia: 2

atlanta georgia american: 2

georgia american civil: 2

american civil war: 2

daughter plantation owner: 2

means disposal claw: 2

disposal claw way: 2

claw way poverty: 2

source list best: 1

--------------------------------------------------

File: Doc8.docx

Most frequent 1-word terms:

rivers: 9

romance: 7

love: 6

christian: 5

time: 5

angel: 5

sold: 5

novels: 4

author: 4

fiction: 4

Most frequent 2-word terms:

romance novels: 3

redeeming love: 3

francine rivers: 3

michael hosea: 3

author francine: 2

christian fiction: 2

fiction historical: 2

historical romance: 2

time men: 2

bag gold: 2

Most frequent 3-word terms:

author francine rivers: 2

bag gold women: 2

sold prostitution child: 2

member romance writers: 2

romance novels rivers: 2

source list best: 1

list best best: 1

best best romance: 1

best romance novels: 1

romance novels twentieth: 1

--------------------------------------------------

File: Doc10.docx

Most frequent 1-word terms:

notebook: 7

story: 5

noah: 5

sparks: 4

love: 4

best: 3

romance: 3

nicholas: 3

man: 3

tender: 3

Most frequent 2-word terms:

nicholas sparks: 3

author nicholas: 2

fiction contemporary: 2

notebook open: 2

open lap: 2

lap woman: 2

morning ritual: 2

ritual understand: 2

tender story: 2

stay forever: 2

Most frequent 3-word terms:

author nicholas sparks: 2

notebook open lap: 2

open lap woman: 2

morning ritual understand: 2

source list best: 1

list best best: 1

best best romance: 1

best romance novels: 1

romance novels twentieth: 1

novels twentieth century: 1

--------------------------------------------------

File: Doc9.docx

Most frequent 1-word terms:

mcnaught: 9

historical: 8

romance: 7

whitney: 7

regency: 7

novels: 6

love: 5

first: 4

romances: 4

genre: 4

Most frequent 2-word terms:

whitney love: 3

historical romance: 3

romance novels: 2

author judith: 2

judith mcnaught: 2

stone mother: 2

france relatives: 2

childhood love: 2

maintain engagement: 2

also first: 2

Most frequent 3-word terms:

author judith mcnaught: 2

also first female: 2

first female executive: 2

female executive producer: 2

executive producer cbs: 2

producer cbs radio: 2

cbs radio station: 2

set regency period: 2

source list best: 1

list best best: 1

--------------------------------------------------

File: Doc6.docx

Most frequent 1-word terms:

gabaldon: 8

war: 5

time: 5

claire: 4

novel: 4

outlander: 3

fiction: 3

science: 3

scotland: 3

torn: 3

Most frequent 2-word terms:

author diana: 2

science fiction: 2

outlander scotland: 2

scotland torn: 2

torn war: 2

time forces: 2

life shatter: 2

shatter heart: 2

james fraser: 2

gabaldon decided: 2

Most frequent 3-word terms:

outlander scotland torn: 2

scotland torn war: 2

life shatter heart: 2

source list best: 1

list best best: 1

best best romance: 1

best romance novels: 1

romance novels twentieth: 1

novels twentieth century: 1

twentieth century title: 1

--------------------------------------------------

**Step5:**

Handling Synonyms and Equivalent Terms in Term Extraction

When analyzing text data, particularly from literary documents, it becomes evident that natural language is inherently rich and diverse. Humans often express the same concept using different terms, phrases, or even variations in wording. This linguistic variability poses a challenge for term extraction engines. Below, we discuss how to handle this issue and improve the accuracy of term extraction by identifying synonymous terms and equivalent phrases.

Observations

1. Synonymous Terms:

* Different terms can refer to the same concept. For instance, "novel" and "book," "author" and "writer," or specific variations like "Harper Lee" and "Lee."
* In the provided data, terms like "George Orwell" and "Orwell," or "harry potter" and "potter" are treated as distinct terms but refer to the same entities.

2. Phrase Variations:

* Phrases like "to kill a mockingbird" might be written as "kill mockingbird" or "harper lee's novel." Such variations make it difficult for simple frequency-based methods to recognize them as the same entity.

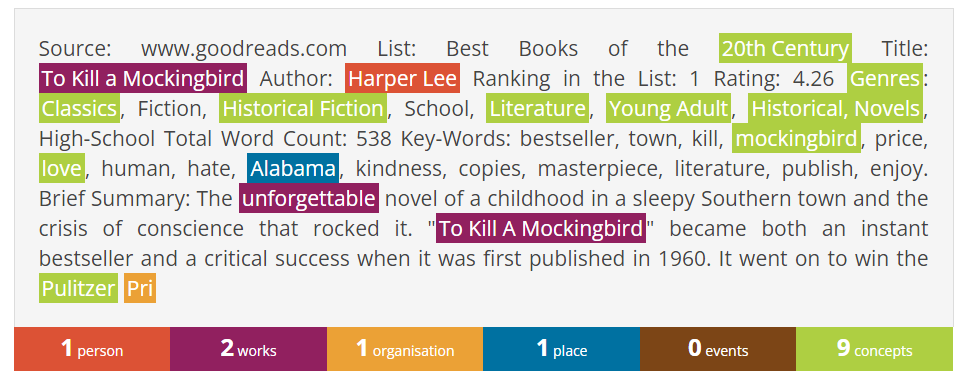
3. Contextual Equivalence:

* The context in which terms are used can imply their equivalence. For example, "Pulitzer Prize" and "award" might not be identical in a general sense but are equivalent in the context of literary achievements.

**Step 6:**

Document 1:

Entity Extractor: [www.dandelion.eu](http://www.dandelion.eu)



Entity Extractor: <https://www.visualthesaurus.com/vocabgrabber/#>

A screenshot of a computer

Description automatically generated

Document 2:

Entity Extractor: [www.dandelion.eu](http://www.dandelion.eu)

A screenshot of a computer

Description automatically generated

Entity Extractor: <https://www.visualthesaurus.com/vocabgrabber/#>

A screenshot of a computer

Description automatically generated

Document 3:

Entity Extractor: [www.dandelion.eu](http://www.dandelion.eu)

A screenshot of a computer

Description automatically generated

Entity Extractor: <https://www.visualthesaurus.com/vocabgrabber/#>

A screenshot of a computer

Description automatically generated