Information Extraction from Fantasy Novels

Data Semantics

Kolyszko Matteo 844526
Merlo Fabrizio 847203
Serino Antonio 886757
Valente Sofia 882782

Introduction

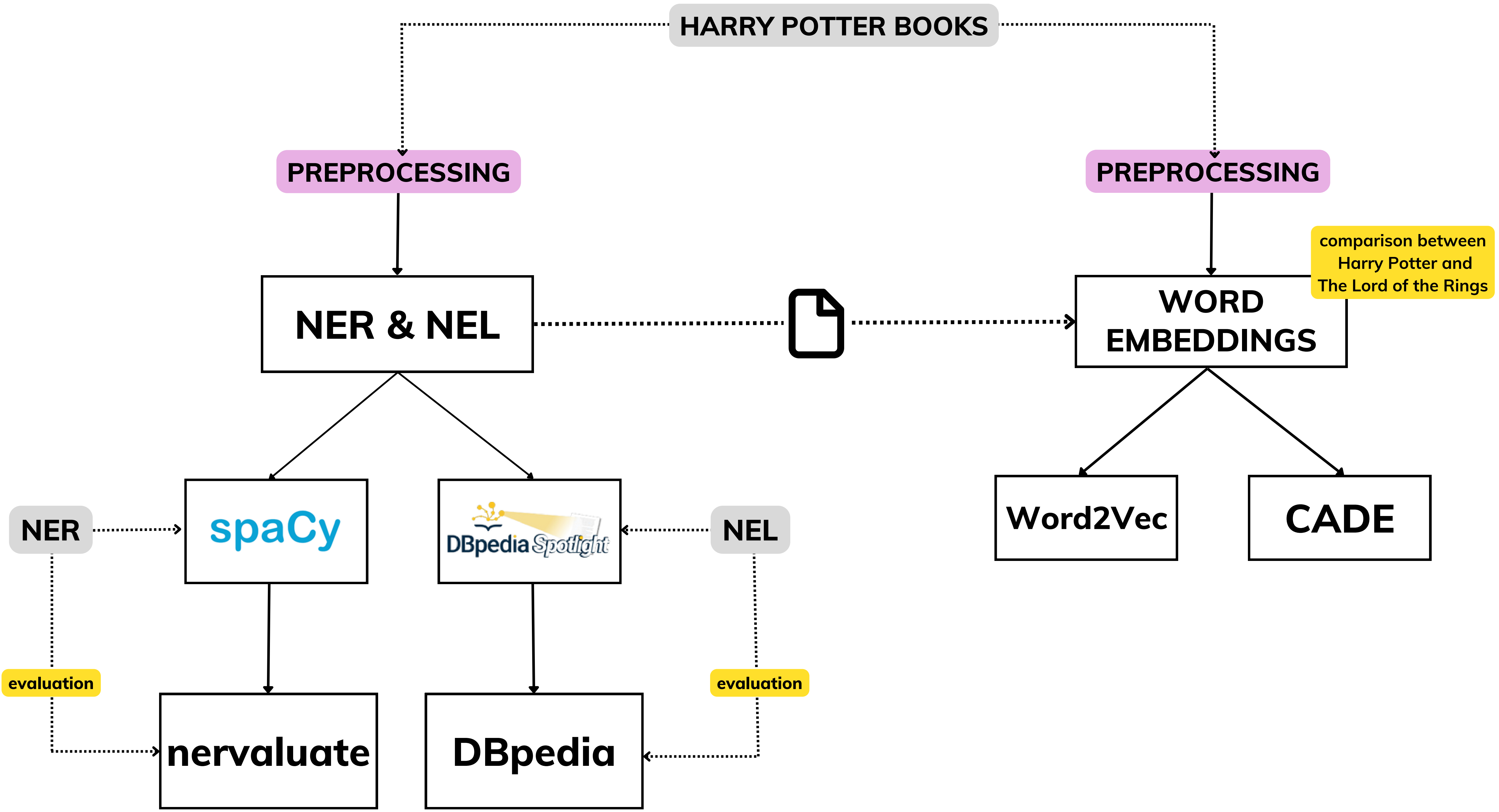
The goal of the project is to be able to extract entities and then use them for a word embeddings analysis.
In particular, we have answered these questions:

-How a pre-trained model such as en_core_web_lg performes on extracting information from fantasy novels?

-Is it possible to find similarities between known elements of the two novels although the lexical/grammatical context differs? using the CADE framework, what should be done, whether or not to use lemmatization?

-Is it possible to analyze, thanks to the use of W2V, the context in which a series of terms are allocated in the individual novels? How the context of the same term differ from HP vector space to LoR one?

Pipeline


HARRY POTTER BOOKS
PREPROCESSING
PREPROCESSING
NER & NEL
WORD EMBEDDINGS
comparison between Harry Potter and The Lord of the Rings
NER
spaCy
DBpedia Spotlight
NEL
Word2Vec
CADE
evaluation
evaluation
nervaluate
DBpedia

Preprocessing

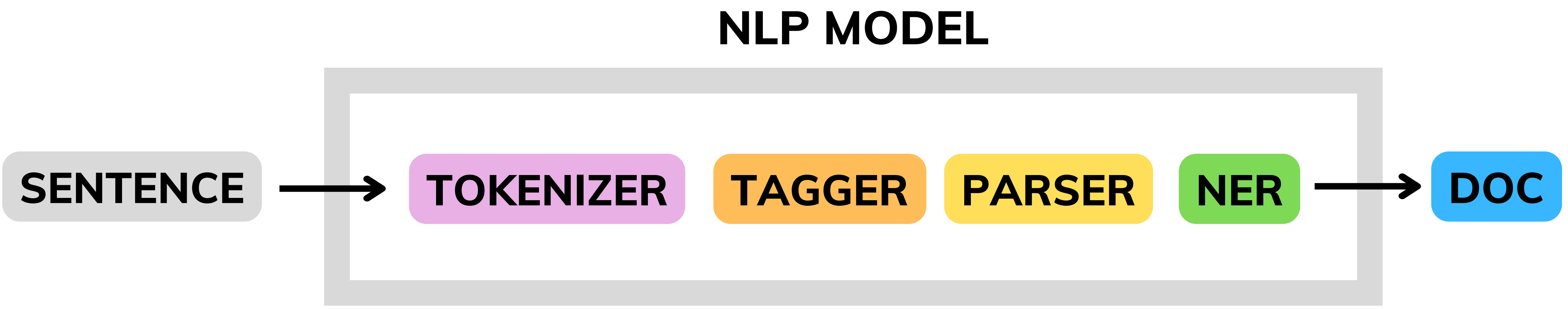
FIRST STEP	Removed chapter index
SECOND STEP	Removed dots from title names
THIRD STEP	Divided the entire corpora in sentences
FOURTH STEP	Selected 2 sentences for each character

Named Entity Recognition

with spaCy

With the help of the SpaCy library each sentence is transformed in a **Doc object**. The Doc lets you access information about the text in a structured way so that no information is lost.

Each Doc Object is structured in **token objects**. For example, a token, can be a word or a punctuation character.



NER application

For each sentence, the steps performed are the following:

- The entities are **identified** using "en_or_web_lg" spaCy model
- The entities are **stored in different lists**, one for each significant label

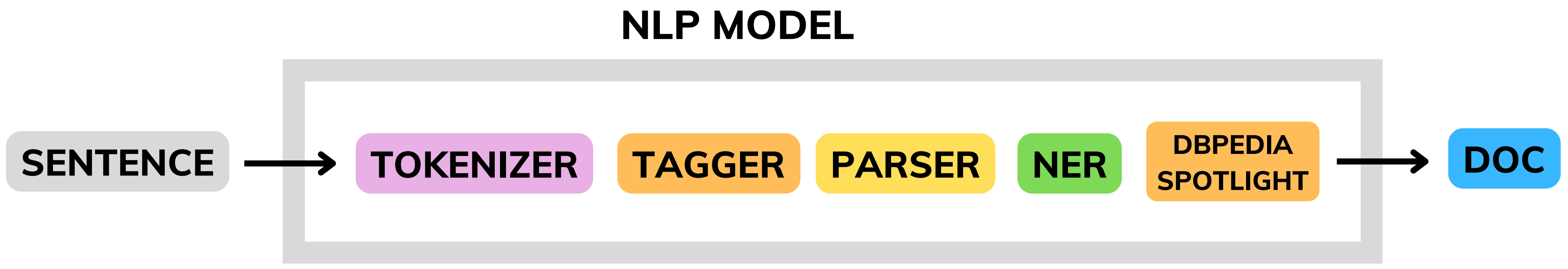
example:

“Disciplinary hearing of the twelfth of August DATE ,” said Fudge WORK_OF_ART in a ringing voice, and Percy PERSON began taking notes at once, “into offenses committed under the Decree for the Reasonable Restriction of Underage LAW Sorcery and the International Statute of Secrecy ORG by Harry James Potter PERSON , resident at number four CARDINAL , Privet Drive PERSON , Little Whinging ORG , Surrey GPE

Named Entity Linking

with DBpedia Spotlight

The library **dbpedia_spotlight** links SpaCy with DBpedia Spotlight.
You can easily get the DBpedia entities from your documents, using the public web service or by using your own instance of DBpedia Spotlight.
The doc.ents are populated with the entities and all their details (URI, type, etc.).



NEL application

For each sentence, the steps performed are the following:

- The entities are **identified** using "en_core_web_lg" spaCy model
- Each entity is then associated with a **DBpedia URL**

example:

“An’ I don’ wan’ yeh ter put yerself out too much, like I know yeh’ve got exams If yeh could jus’ nip down here in yer Invisibility Cloak maybe once a week an’ have a little chat with him I’ll wake him up, then — introduce you — ” “Wha — no!” said **http://dbpedia.org/resource/Hermione_Granger**, jumping up, “**http://dbpedia.org/resource/Rubeus_Hagrid**, no, don’t wake him, really, we don’t need — ” But **http://dbpedia.org/resource/Rubeus_Hagrid** had already stepped over the great trunk in front of them and was proceeding toward Grawp

NER Evaluation

FIRST STEP	Created dataset with SpaCy's results
SECOND STEP	Manually added how many labels have been actually found and how many were supposed to be
THIRD STEP	Manually fixed the labels
FOURTH STEP	Use library nerevaluate to compare the exact labels with SpaCy's results

NER Results

with **Nervaluate**

After the evaluation the results are the following:

71,71% entities were found (332 entities)

28,29% entities were not found (131 entities)

Tag	Correct	Incorrect	Precision
Product	7	7	50%
Person	193	33	85.9%
Event	3	2	60%
GPE	5	1	83.3%
Cardinal	22	0	100%
Ordinal	11	0	100%
Date	15	0	100%

NEL Evaluation

FIRST STEP	Created dataset with DBpedia Spotlight results
SECOND STEP	Used the character file to link them to DBpedia with a dinamic query SPARQL
THIRD STEP	Comparison between DBpedia Spotlight results and the real link of DBpedia
FOURTH STEP	Comparison between DBpedia Spotlight results and matching domain

Evaluation NEL Characters

with **DBpedia Spotlight**

Specific

Characters	Number
Found	61
Correct	11
Different URL	34
Not Recognized	16

Generic

Characters	Number
Found	61
Correct	17
Different URL	28
Not Recognized	16

WORD EMBEDDINGS

CADE & WORD2VEC

RESEARCH QUESTIONS

For this part of the project, the research questions we formulate are based on the desire to compare two texts belonging to the same literary genre, but at a distance of time, using two frameworks seen in class:

-Is it possible to find similarities between known elements of the two novels although the lexical/grammatical context differs? using the CADE framework, what should be done, whether or not to use lemmatization?

-Is it possible to analyze, thanks to the use of W2V, the context in which a series of terms are allocated in the individual novels? How the context of the same term differ from HP vector space to LoR one?

PREPROCESSING

LEMMATIZATION - NON LEMMATIZATION

in order to answer the research questions, two preprocessing were carried out:

tokenization;

- -transformation of letters to lowercase;
- -removal of alphanumeric characters;
- -Lemmatization;
- -removing stopwords.

tokenization;

- -transformation of letters to lowercase;
- -removal of alphanumeric characters;
- -removing stopwords.

SENTENCES EXPLORATION

Using Counter from Collection library to explore the sentences from Harry Potter and The Lord of the Rings.

A counter is a container that stores elements as dictionary keys, and their counts are stored as dictionary values, that allows to compute the items in an iterable list to compute:

Number of **unique** words

Most **common** words

Number of unique words:

Harry Potter	Lord of Rings
13907	8397

Most common words: **Harry**: 16111 **Ron**: 5684 **Hermione**: 4963 **Dumbledore**: 2867

Frodo: 970 **Sam**: 1277 **Gandalf**: 1097

CADE TRAINING

In addition to the preprocessing, two similar procedures were also performed for the training of the models (in the presence of lemmatization and in the absence of lemmatization) referring to what we saw in class:

- create the aligner object with the CADE(size=30) method;
- training on the concatenation of the two corpus;
- training of the first slice on the corpus of "Harry Potter";
- training of the second slice on the corpus of "the Lord of the Rings".

In order to be able to evaluate which of the two trainings, with lemmatization or without lemmatization, is better, the distance of the cosine on the term "could" was used (a very common term in both corpus).

TYPE	COSINE
Lemmatization	0,93
No Lemmatization	0,91

Having obtained a slightly higher score using the lemmatized corpus, we decided to continue the analysis using the lemmatized models.

FIRST RESEARCH QUESTION

To answer the first research question, two different types of queries were expressed:

- The former are based on the search for similarities with the main entities derived from the analysis of named entity recognition

harry	ron	hermione	dumbledo re	snape	voldemort	bellatrix	hogwarts	azkaban
frodo, 0.77	merry, 0.68	sam, 0.63	gandalf, 0.81	faramir, 0.54	enemy, 0.77	aloud, 0.69	rivendell , 0.69	sauron, 0.79
pippin, 0.69	sam, 0.65	pippin, 0.63	faramir, 0.72	aragorn, 0.54	mordor, 0.69	wormto ngue, 0.69	return, 0.69	servant, 0.78
merry, 0.64	pippin, 0,64	merry, 0.60	aragorn, 0.67	without, 0.53	saruma n, 0.64	comma nd, 0.68	enter, 0.64	death, 0.75

FIRST RESEARCH QUESTION

- the latter are based on the search for similarities with 5 terms representing particularly recurring themes in both novels

magic	war	friendship	good	evil
lord, 0.64	eagle, 0.88	endure, 0.91	good, 0.78	slave, 0.82
cell, 0.56	race, 0.88	ally, 0.90	well, 0.73	wraith, 0.81
rule, 0.50	numenor, 0.87	sorrow, 0.89	mean, 0.68	badge, 0.81
mine, 0.49	belfalas, 0.87	wholly, 0.89	likely, 0.67	torture, 0.81
war, 0.47	mundburg, 0.87	grief, 0.88	party, 0.67	deceive, 0.81

- satisfactory results for the entities;
- less satisfactory results for the relevant terms.

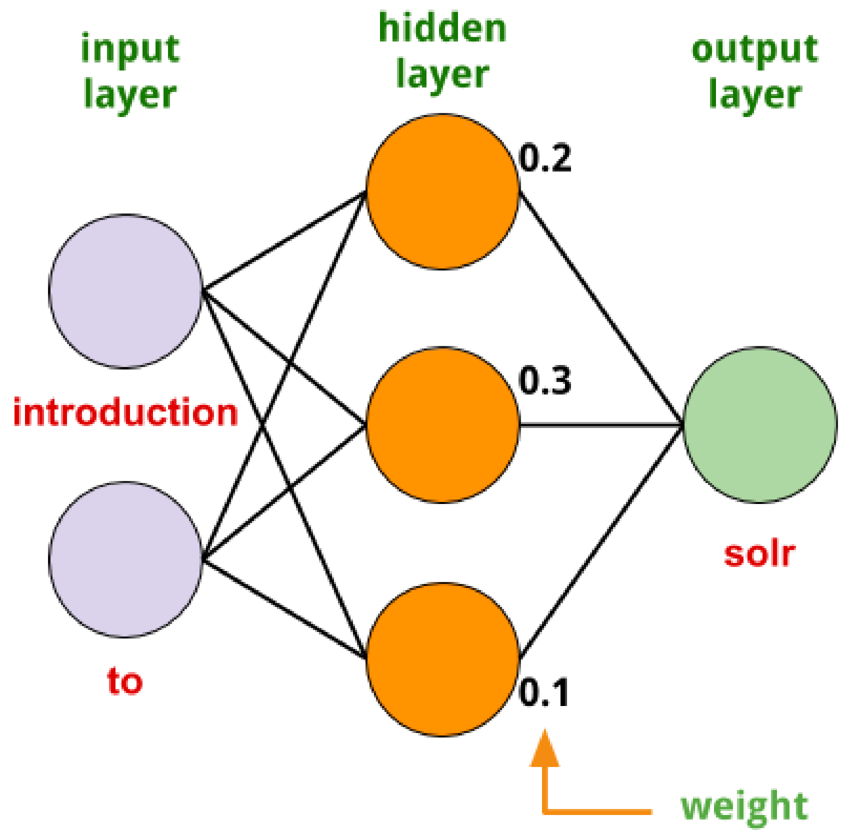
WORD2VEC TRAINING

Word2Vec: algorithm based on a neural network, that learns relationships between words automatically,embedding words in a lower-dimensional vector space.

The algorithm first creates a vocabulary from the training text data and then learns vector representations of the words.

Model parameters:

- sentences– a list of lists of tokens
- skip-gram,Training algorithm
- Min_count, pruning internal dictionary;
- workers, to speed up training:
- iter -Number of iterations (epochs) over the corpus



ANALYSIS

Using the output of the NER analysis as input for the W2V analysis→ entities

Compute the similarity between a list of token
FOCUS: main characters - antagonist characters - places

FIRST STEP
top 5 most similar words

SECOND STEP
top 5 dissimilar words for each character/place

THIRD STEP

top 5 words most similar to the sum of the vectors of each character and each character of the list.

FOURTH STEP

Cosmul similarity

COSMUL SIMILARITY & PCA AND TSNE SCATTERPLOT

most_similar_cosmul(): the `cosmul` variant uses a slightly-different comparison when using multiple positive/negative examples to compute analogies between words expressed as:

v(word) - v(word) + v(word)

output ⟶ another vector, expressing the analogy for the words considered.

PCA+t-SNE

OBJ : plot a n-dimensionsional vectors into 2 dimensional graphs for spotting interesting patterns

To make the visualizations results more relevant, we will look at the relationships between a query word, its most similar words in the model, and other words from the vocabulary.

HP ANALYSIS

SIMILARITY

Similarity for main characters		
Similar to Harry	Similar to Hermione	Similar to Ron
miserably (0.7)	excitedly (0.65)	anxiously (0.75)
hopefully (0.69)	luna (0.65)	breathlessly (0.75)
desperately (0.69)	griphook (0.65)	ginny (0.75)
awkward (0.69)	miserably (0.65)	griphook (0.72)
panicked (0.68)	awkwardly (0.65)	okay (0.72)
Dissimilar to Harry	Dissimilar to Hermione	Dissimilar to Ron
magical (-0.11)	thin (-0.06)	number (-0.08)
wizarding (-0.12)	house (-0.07)	smoke (-0.09)
international (-0.13)	number (-0.08)	grimmauld (-0.11)
smoke (-0.14)	hit (-0.09)	decree (-0.12)
hair (-0.15)	drive (-0.09)	house (-0.12)
Dissimilar to Harry + Hermione	Dissimilar to Hermione + Ron	Dissimilar to Ron + Harry
set (0.11)	beneath (0.13)	beneath (0.14)
place (0.11)	place (0.11)	hidden (0.14)
beneath (0.10)	hair (0.10)	desk (0.12)
room (0.10)	conjure (0.10)	place (0,12)
desk (0.10)	toward (0.10)	lit (0.12)

Similarity for antagonist characters		
Similar to Voldemort	Similar to Bellatrix	Similar to Draco
prophecy (0.75)	narcissa (0.89)	lucius (0.91)
wormtail (0.74)	greyback (0.84)	smirk (0.85)
bellatrix (0.71)	lucius (0.8)	sneer (0.84)
lord (0,71)	centaur (0.79)	crabbe (0.84)
connection (0.7)	woemtail (0.79)	narcissa (0.82)
Dissimilar to Voldemort	Dissimilar to Bellatrix	Dissimilar to Draco
pile (-0.07)	spent (-0.01)	owl (-0.03)
dress (-0.07)	morning (-0.01)	letter (-0.02)
chocolate (-0.05)	bed (-0.02)	parchment (-0.01)
onto (-0.04)	sat (-0.02)	bed (-0.02)
large (-0.03)	dress (-0.03)	bag (-0.02)
Dissimilar to Voldemort + Bellatrix	Dissimilar to Bellatrix + Draco	Dissimilar to Draco + Voldemort
toward (0.16)	cloak (0.19)	cloak (0.17)
jacket (0.15)	set (0.16)	set (0.13)
seize (0.15)	beneath (0.14)	desk (0.12)
cloak (0.14)	reach (0.14)	beneath (0.12)
pick(0.13)	desk (0.13)	chair (0.12)

SIMILARITY

Similarity for places / houses			
Similar to Hogwarts	Similar to Ministry	Similar to Azkaban	Similar to Gryffindor
school (0.88)	minister (0.79)	murder (0.92)	ravenclaw (0.85)
witchcraft (0.79)	improper (0.79)	faithful (0.88)	hufflepuff (0.85)
gamekeeper (0.77)	underage (0.78)	murderer (0.88)	slytherin (0.83)
may (0.77)	official (0.77)	aurors (0.88)	team (0.79)
october (0.76)	prime (0.76)	capture (0.88)	match (0.76)
Dissimilar to Hogwarts	Dissimilar to Ministry	Dissimilar to Azkaban	Dissimilar to Gryffindor
throat (-0.02)	fell (-0.33)	large (-0.097)	dark (-0.036)
fist (-0.01)	shook (-0.041)	onto (-0.093)	minister (-0.036)
ear (0.01)	sat (-0.047)	wooden (-0.088)	forehead (-0.029)
crookshanks (-0.004)	drew (-0.050)	glass (-0.085)	shut (-0.029)
tight (-0.005)	tremble (-0.056)	pink (-0.073)	fudge (-0.007)
Dissimilar to Hogwarts + Ministry	Dissimilar to Ministry + Azkaban	Dissimilar to Azkaban + Gryffindor	Dissimilar to Gryffindor + Hogwarts
tightly (0.21)	beneath (0.20)	desk (0.18)	cloak (0.15)
tip (0.17)	cloak (0.19)	cloak (0.17)	beneath (0.13)
arm (0.17)	arm (0.19)	tightly (0.17)	conjure (0.11)
sleeve (0.17)	tightly (0.18)	beneath (0.17)	thick (0.10)
grip (0.17)	seize (0.17)	toward (0.17)	arm (0.09)

COSMUL

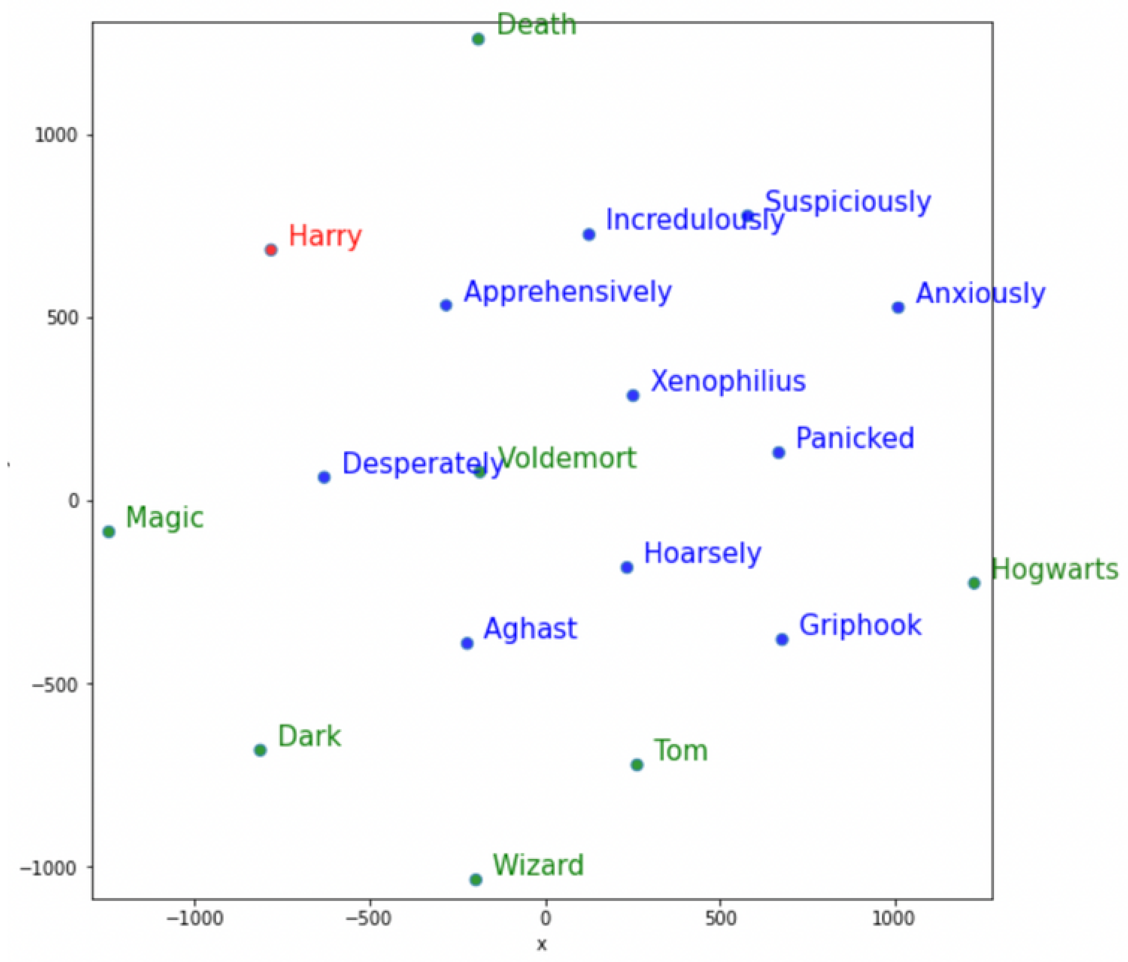
v(Voldemort) - v(friendship) + v(Harry)
Harry is related to friendship, as **Pettigrew** is related to Voldemort

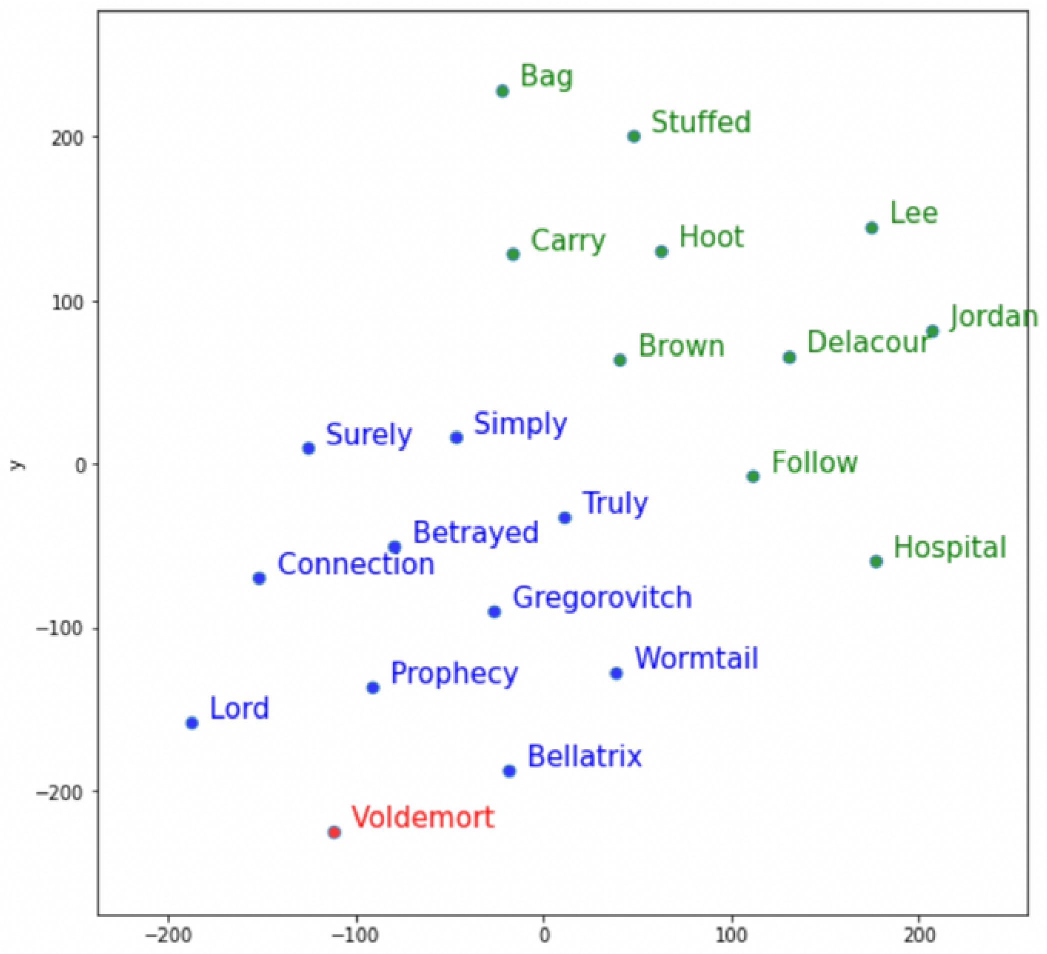
v(Hogwarts) - v(wand) + v(Dumbledore)
Dumbledore is related to wand, as **headmaster** is related to Hogwarts

HP: TSNE SCATTERPLOT

The vector representation of **Harry**, its 10 most similar terms from the model and other words in a 2D chart:

The vector representation of **Voldemort** and its 10 most similar terms of the model are compared with the vector representation of the 20 most dissimilar words in Voldemort





LEGEND:

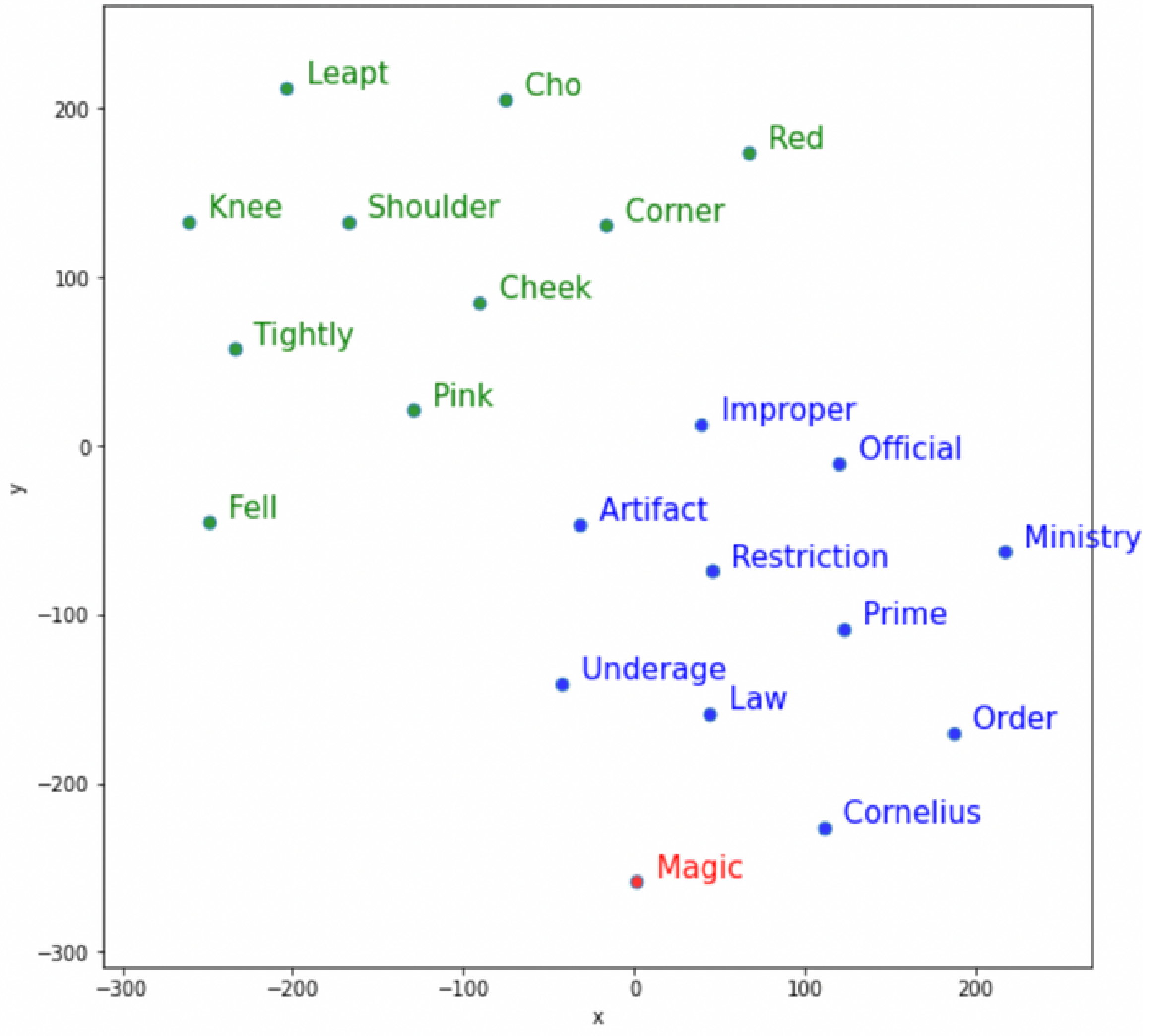
most similar terms

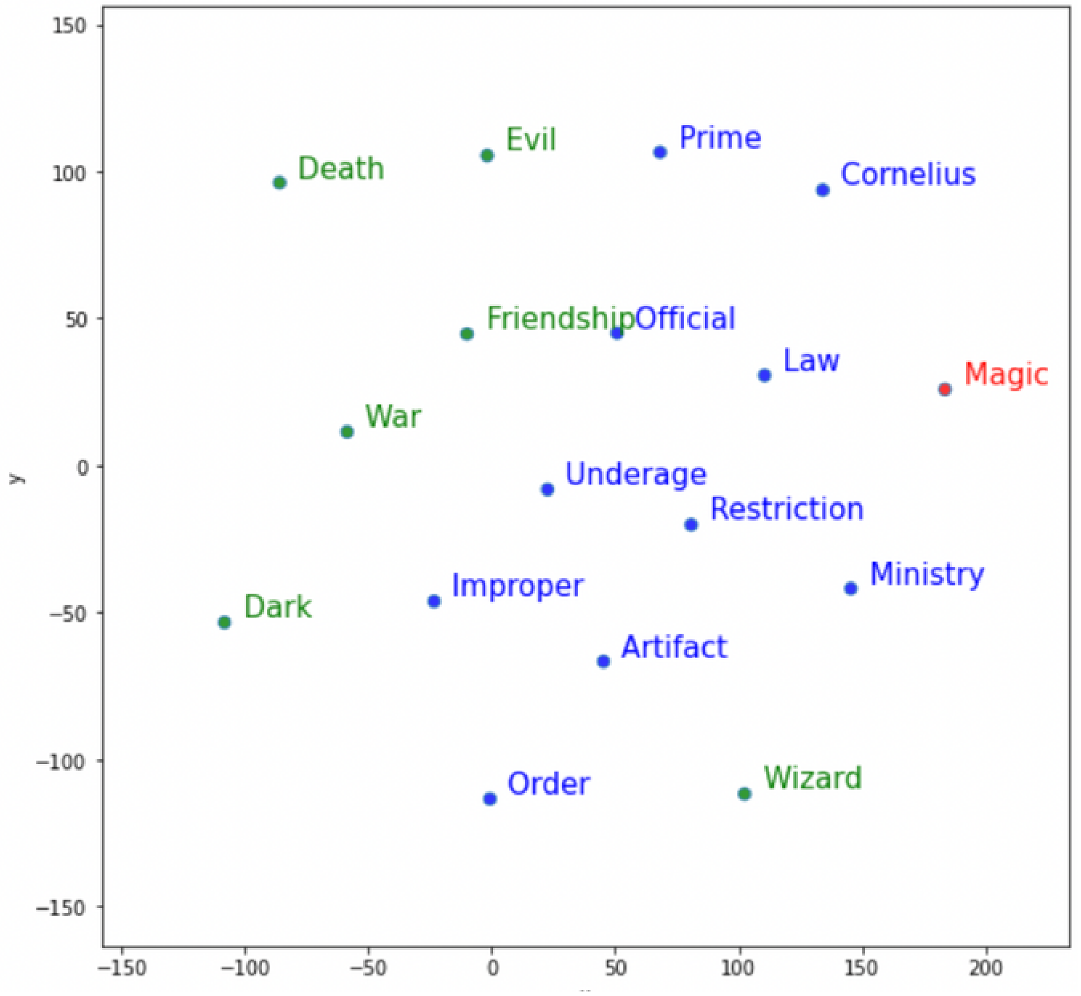
other terms from the vocab

HP: TSNE SCATTERPLOT

The vector representation of **Magic**, its 10 most similar terms from the model and other words in a 2D chart:

The vector representation of **Magic** and its 10 most similar terms of the model are compared with the vector representation of the 5 terms common in the two fantasy novels





LEGEND:

- most similar terms
- other terms from the vocab

LORD OF THE RINGS: MODEL

SIMILARITY

Similarity for main characters		
Similar to Frodo	Similar to Gandalf	Similar to Legolas
sam (0.96)	aragon (0.93)	gimli (0.96)
pippin (0.92)	legolas (0.93)	aragon (0.95)
gollum (0.92)	faramir (0.93)	éomer (0.94)
strider (0.91)	strider (0.93)	gandalf (0.93)
treebeard (0.91)	beregond (0.91)	faramir (0.92)
Dissimilar to Frodo	Dissimilar to Gandalf	Dissimilar to Legolas
sea (-0.08)	flow (-0.15)	leaf (-0.15)
mina (-0.11)	leaf (-0.15)	flow (-0.18)
field (-0.11)	stream (-0.2)	power (-0.19)
gondor (-0.11)	green (-0.22)	beyond (-0.2)
mountain (-0.14)	smoke (-0.02)	year (-0.21)
Dissimilar to Frodo + Gandalf	Dissimilar to Gandalf + Legolas	Dissimilar to Legolas + Frodo
white (0.04)	flow (0.01)	flow (-0.0026)
mina (0.03)	hung (-0.01)	dark (-0.02)
tirith (0.02)	misty (-0.01)	misty (-0.02)
flow (0.003)	bare (-0.01)	red (-0.03)
silver (-3.64)	silver (-0.02)	hung (-0.03)

Similarity for antagonist characters		
Similar to Sauron	Similar to Saruman	Similar to Gollum
evil (0.95)	council (0.09)	sam (0.94)
destroyed 0.95)	wisdom (0.96)	frodo (0.92)
deed (0.95)	service (0.96)	try (0.92)
peril (0.94)	halfling (0.96)	sleep (0.92)
work (0.94)	receive (0.95)	something (0.91)
Dissimilar to Sauron	Dissimilar to Saruman	Dissimilar to Gollum
forward (-0.07)	stream (-0.24)	gondor (-0.07)
sprang (-0.08)	steep (-0.25)	mina (-0.11)
peer (-0.1)	slope (-0.25)	sea (-0.12)
suddenly (-0.12)	climb (-0.27)	white (-0.12)
step (-0.12)	along (-0.27)	king (-0.13)
ssimilar to Sauron + Sarum	similar to Saruman + Goll	Dissimilar to Gollum + Sauron
sprang (0.07)	ran (0.005)	white (0.05)
drew (0.05)	climb(0.004)	mina (0.03)
forward (0.04)	across (0.004)	upon (0.02)
peer (0.04)	steep (0.004)	tirith (0.02)
bent (0.03)	towards (-0.003)	silver (0.007)

COSMUL

v(Frodo) - v(ring) + v(Gollum)

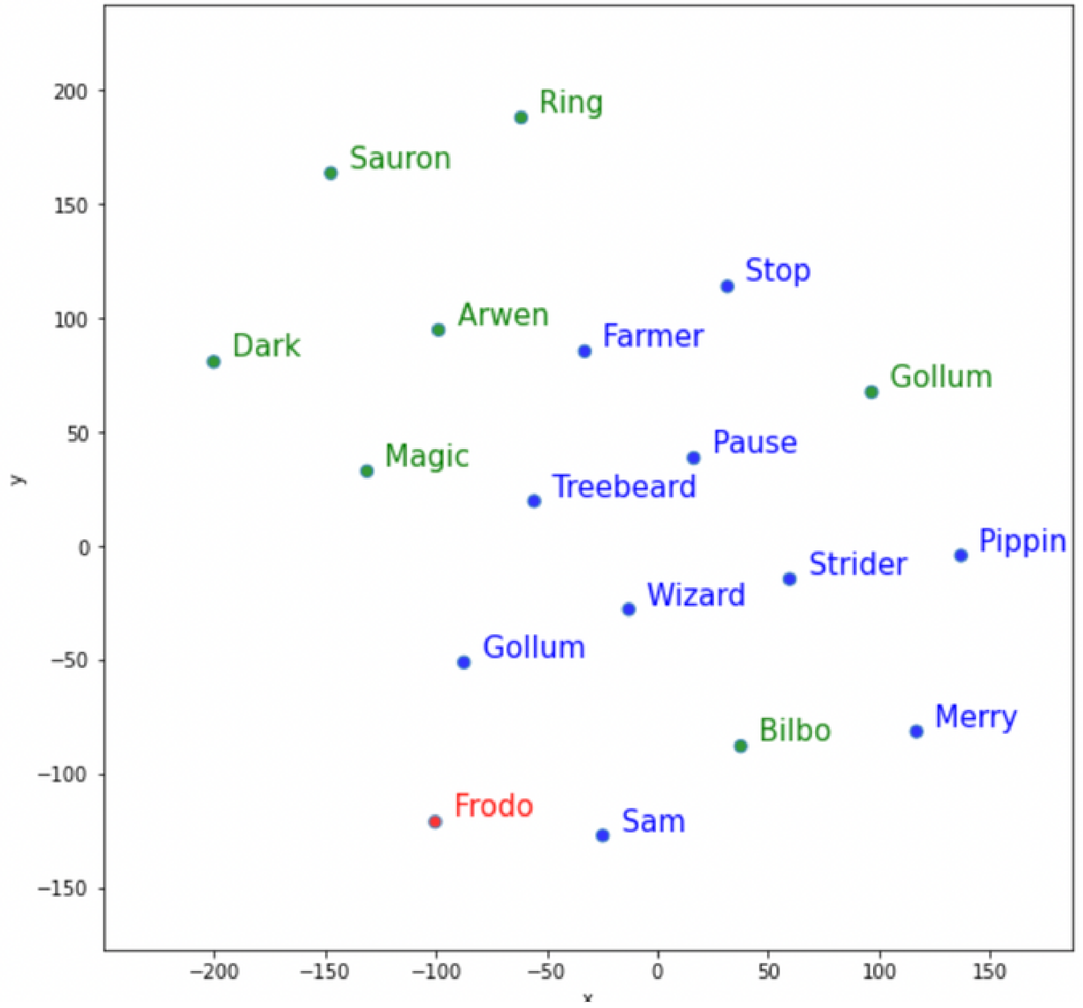
gollum is related to ring, as **Sam** is related to Frodo
v(Arwen) - v(Gimli) + v(Sauron)

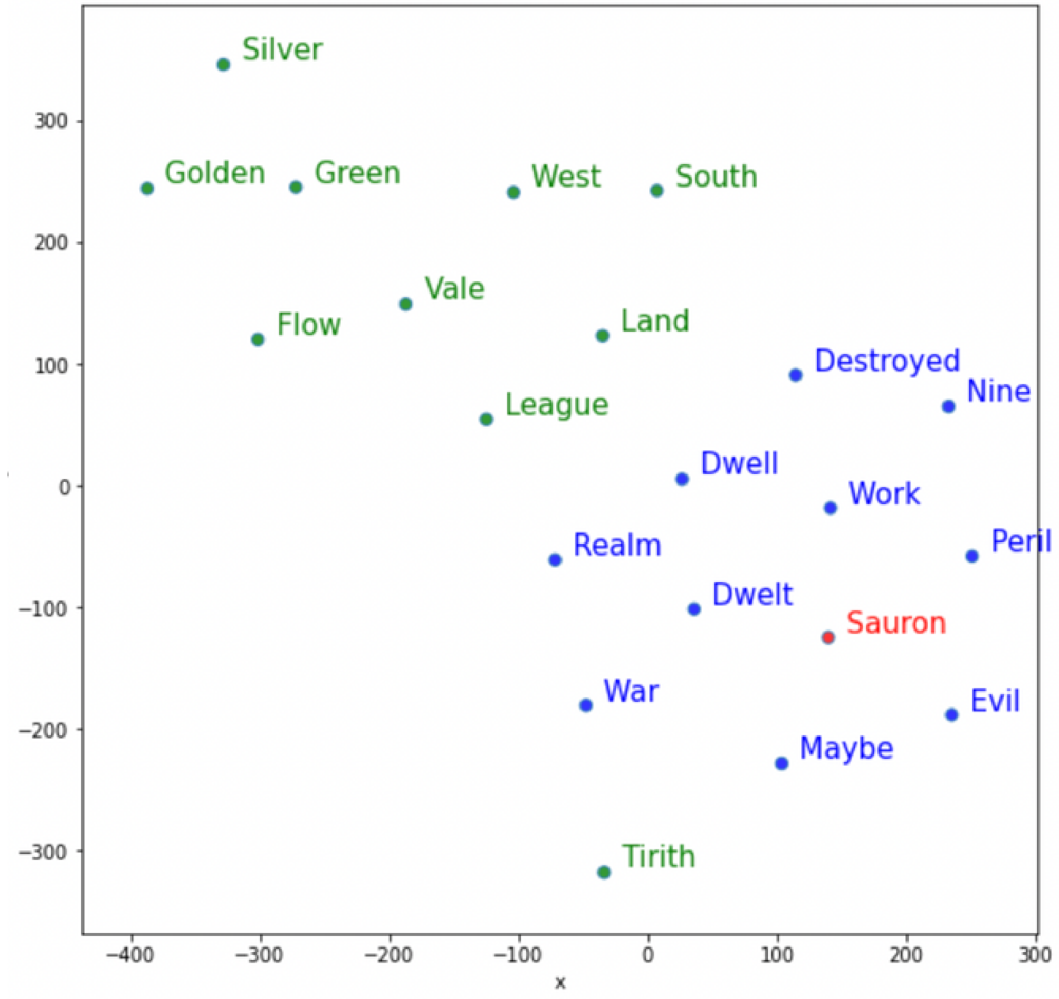
Sauron is related to Gimli, as **power** is related to Arwen

Lord of Rings: TSNE SCATTERPLOT

The vector representation of **Frodo**, its 10 most similar terms from the model and other words in a 2D chart:

The vector representation of **Sauron** and its 10 most similar terms of the model are compared with the vector representation of the 20 most dissimilar words in Sauron





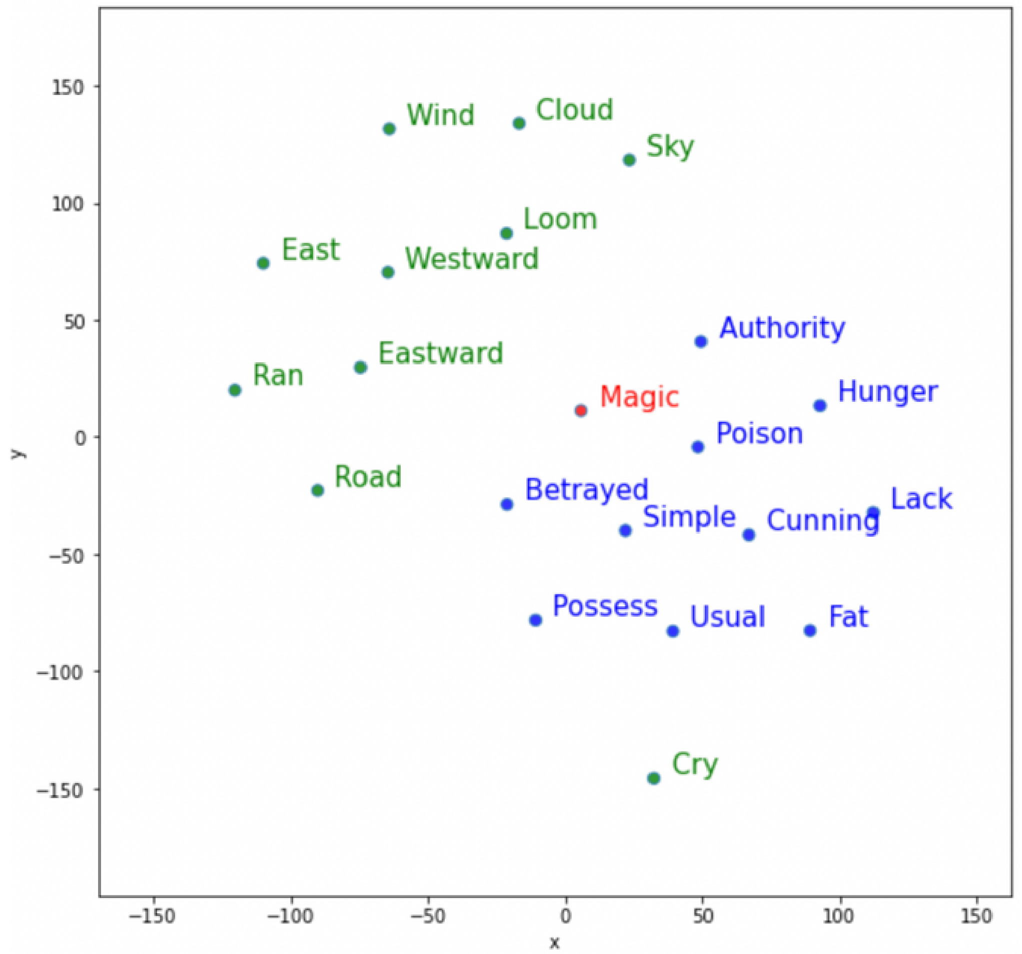
LEGEND:

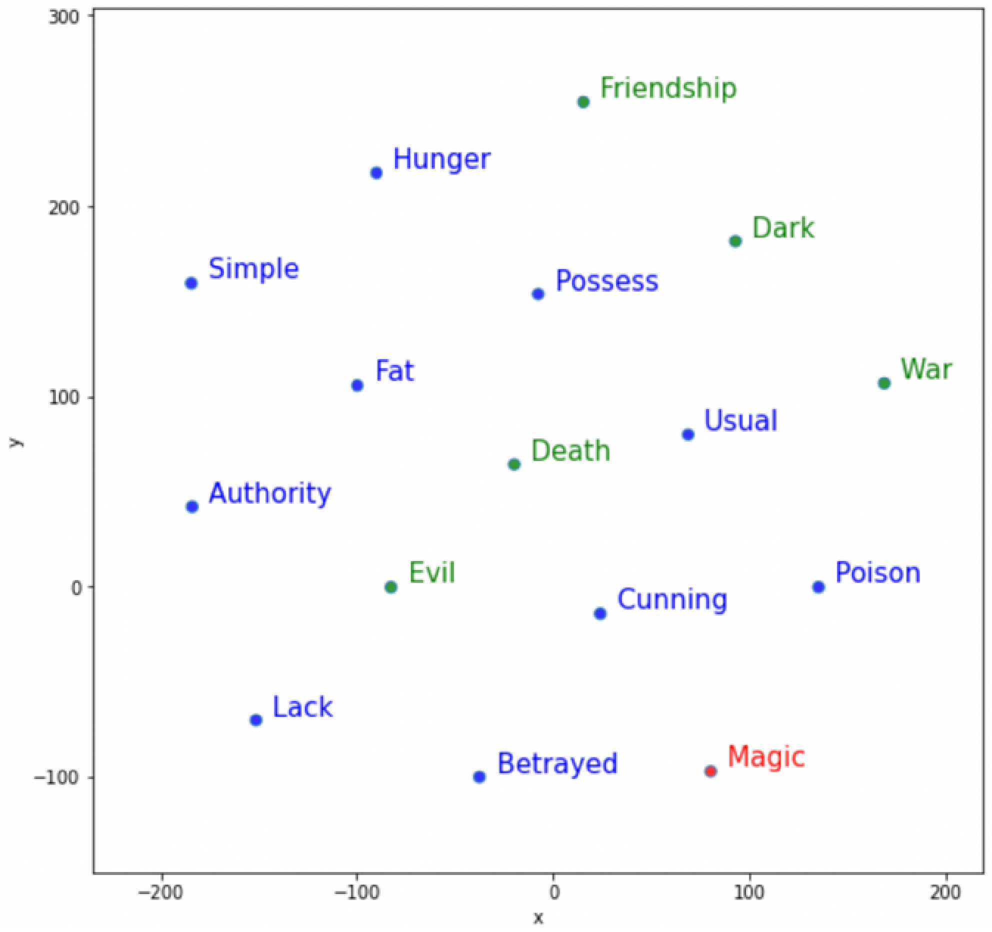
- most similar terms
- other terms from the vocab

Lord of Rings: TSNE SCATTERPLOT

The vector representation of **Magic**, its 10 most similar terms from the model and other words in a 2D chart:

The vector representation of **Magic** and its 10 most similar terms of the model are compared with the vector representation of the 5 terms common in the two fantasy novels





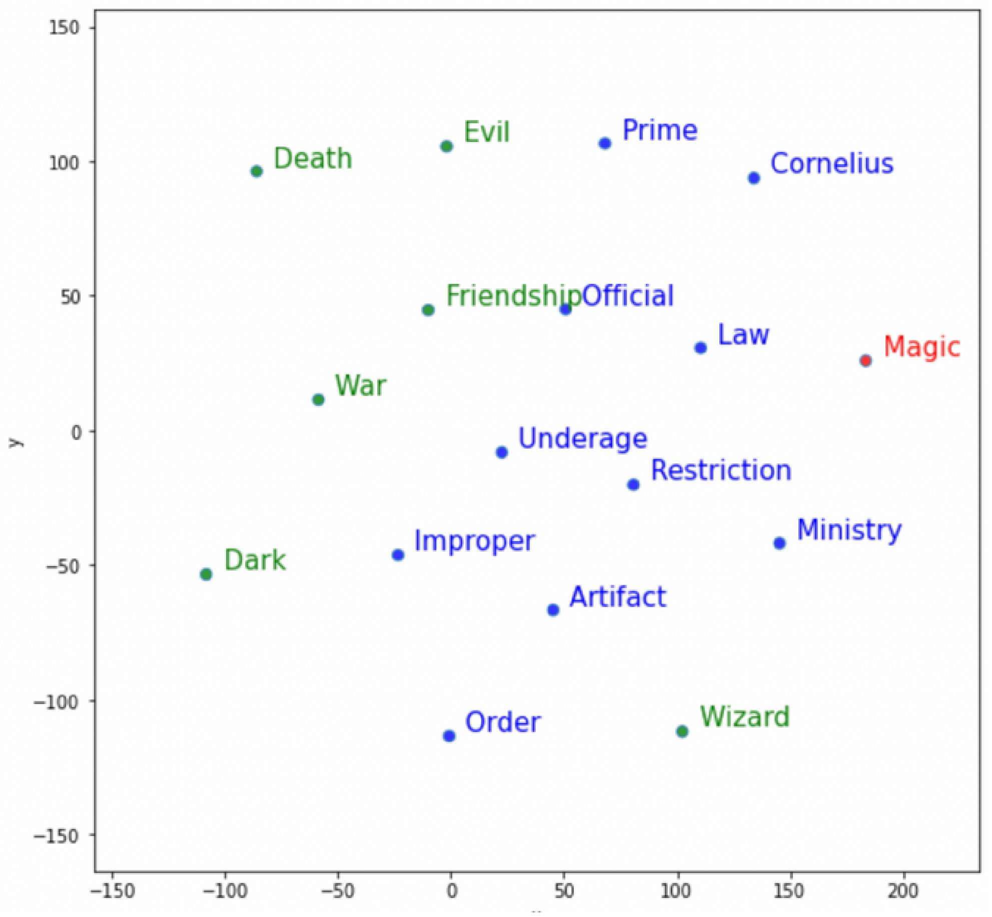
LEGEND:

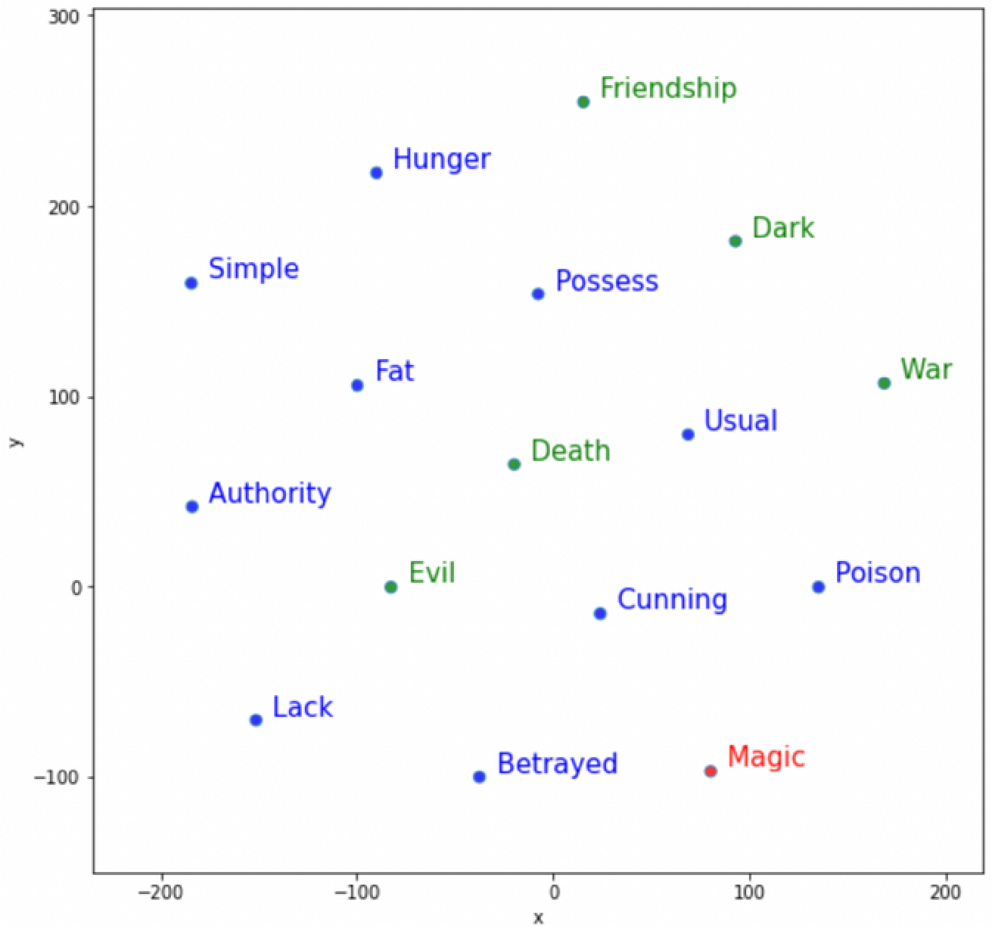
- most similar terms
- other terms from the vocab

Harry Potter vs Lord of Rings: TSNE SCATTERPLOT

The vector representation of **Magic** in **HP** and its 10 most similar terms of the model are compared with the vector representation of the 5 terms common in the two fantasy novels

The vector representation of **Magic** in **LoR**. and its 10 most similar terms of the model are compared with the vector representation of the 5 terms common in the two fantasy novels





LEGEND:

- most similar terms
- other terms from the vocab

RESOURCES

Notebook with preprocessing analysis, NER & NEL:
https://colab.research.google.com/drive/1ZjkKnVTY5NFOEGM15DzGLk2aD8n_975z?usp=sharing

Notebook with preprocessing analysis, Word2Vec & Cade:
https://colab.research.google.com/drive/1RIdBXcNH5OUl_-Bd0GGAyQ5asHoXmoBB?usp=sharing

REFERENCES

[1] Neural Entity Linking: A Survey of Models Based on Deep Learning - Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, Chris Biemann ;

[2] A Survey on Deep Learning for Named Entity Recognition - Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li;

[3] Compass-Aligned Distributional Embeddings For Studying Semantic Differences Across Corpora - Bianchi F., Di Carlo V., Nicoli P. and Palmonari M.;

[4] WEIGHTED WORD2VEC BASED ON THE DISTANCE OF WORDS - CHIA-YANG CHANG1, SHIE-JUE LEE1, CHIH-CHIN LAI.

THANKS FOR YOUR ATTENTION