Universidade de São Paulo

TRIVIR: A Visualization System to Support Document Retrieval with High Recall

Student

Sherlon Almeida da Silva
sherlon@usp.br

Authors

Amanda G. Dias
gdias.amanda@gmail.com

Evangelos E. Milios
eem@cs.dal.ca

Maria C. F. de Oliveira
cristina@icmc.usp.br

Institute of Mathematical and Computer Sciences - ICMC/USP

Introduction

What is the Motivation?

Motivation

User

A scientist needs to discover relevant information about her research, e.g. studying the literature.

Motivation

User

Query

She already knows some interesting papers that can be useful to start the search.

Motivation

User

Query

She tries to find relevant information in a collection of papers (a corpus).

Corpus

Motivation

User

Query

She identifies relevant papers in the collection.

Corpus

Relevant Docs

Introduction

What is the Problem?

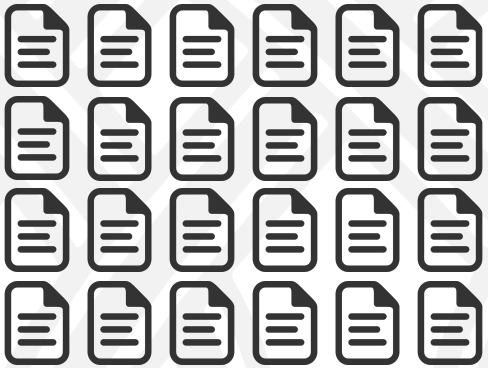
Problem

Think about this situation

Problem

Think about this situation

- There is a Corpus containing lots of documents;



Problem

Think about this situation

- There is a Corpus containing lots of documents;
- You identified one or a few 'interesting' documents;

Problem

Think about this situation

- There is a Corpus containing lots of documents;
- You identified one or a few 'interesting' documents;
- **Find other related (i.e., also interesting) documents;**

Problem

Think about this situation

- There is a Corpus containing lots of documents;
- You identified one or a few 'interesting' documents;
- Find other related (i.e., also interesting) documents;
- **How can you identify (nearly) all the relevant documents in a corpus with lots of documents?**

Problem

Think about this situation

- There is a Corpus containing lots of documents;
- You identified one or a few 'interesting' documents;
- Find other related (i.e., also interesting) documents;
- **How can you identify (nearly) all the relevant documents in a corpus with lots of documents?**

Analyze ALL documents?

Problem

Think about this situation

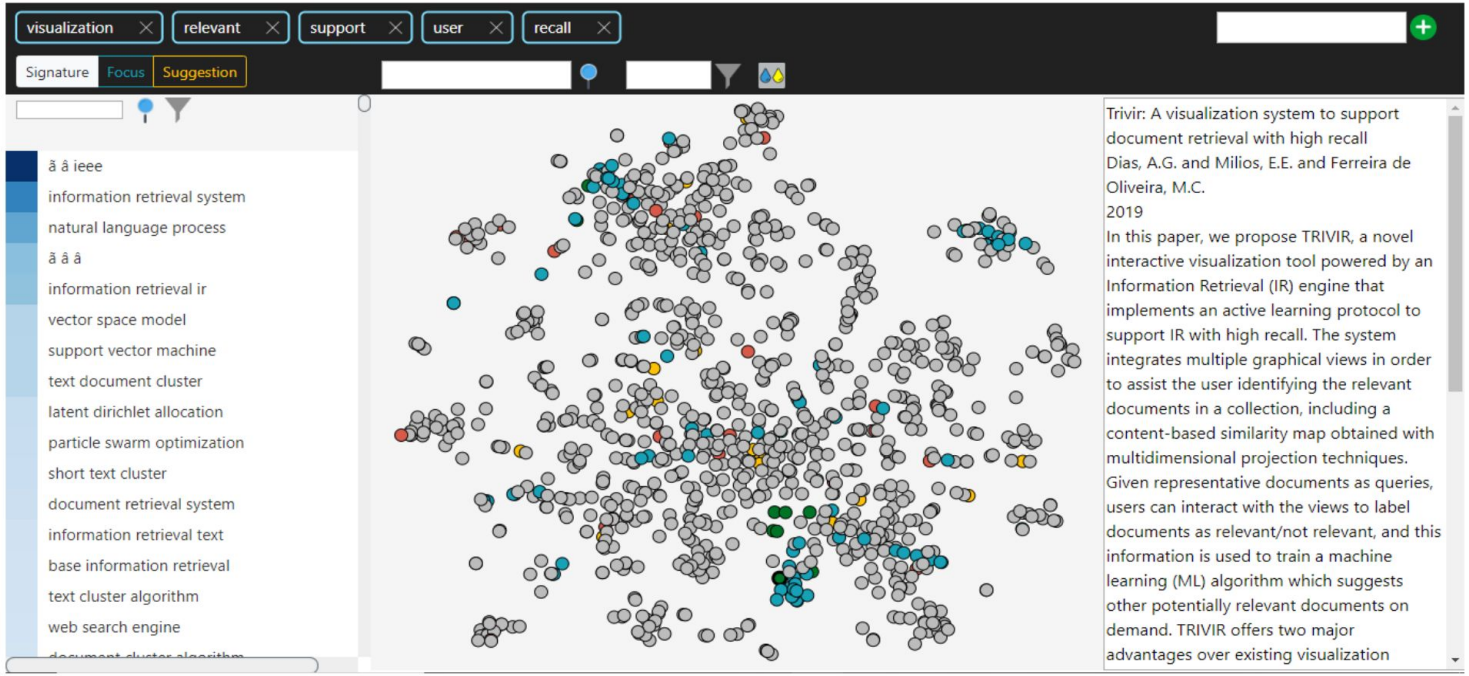
- There is a Corpus containing lots of documents;
- You identified one or a few 'interesting' documents;
- Find other related (i.e., also interesting) documents;
- **How can you identify (nearly) all the relevant documents in a corpus with lots of documents?**

Machine Learning (ML) + Visual Analytics (VA)

TRIVIR

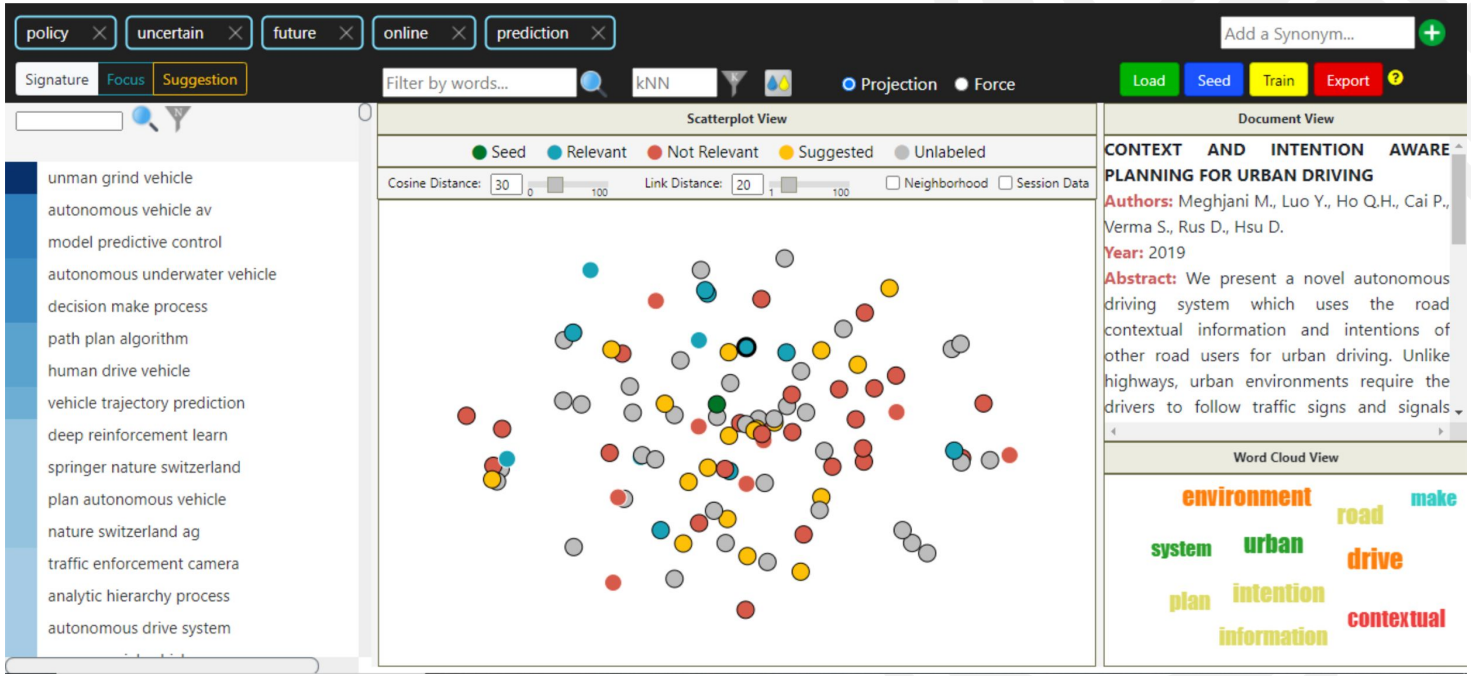
The Visualization System

TRIVIR



Visualization System

TRIVIR

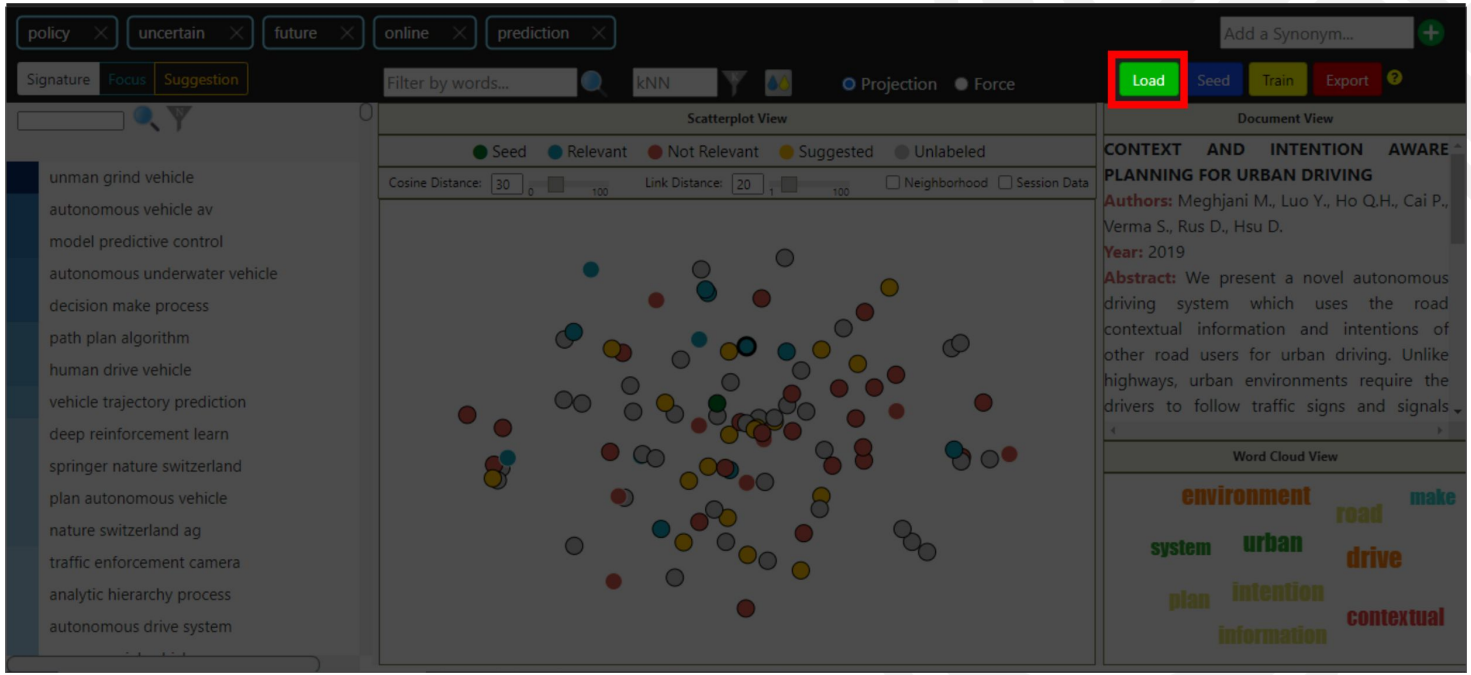


Visualization System

TRIVIR

1. Scatterplot View
2. Document View
3. Word Cloud View
4. Signature List
5. Focus List
6. Suggestion List
7. Terms View
8. Buttons

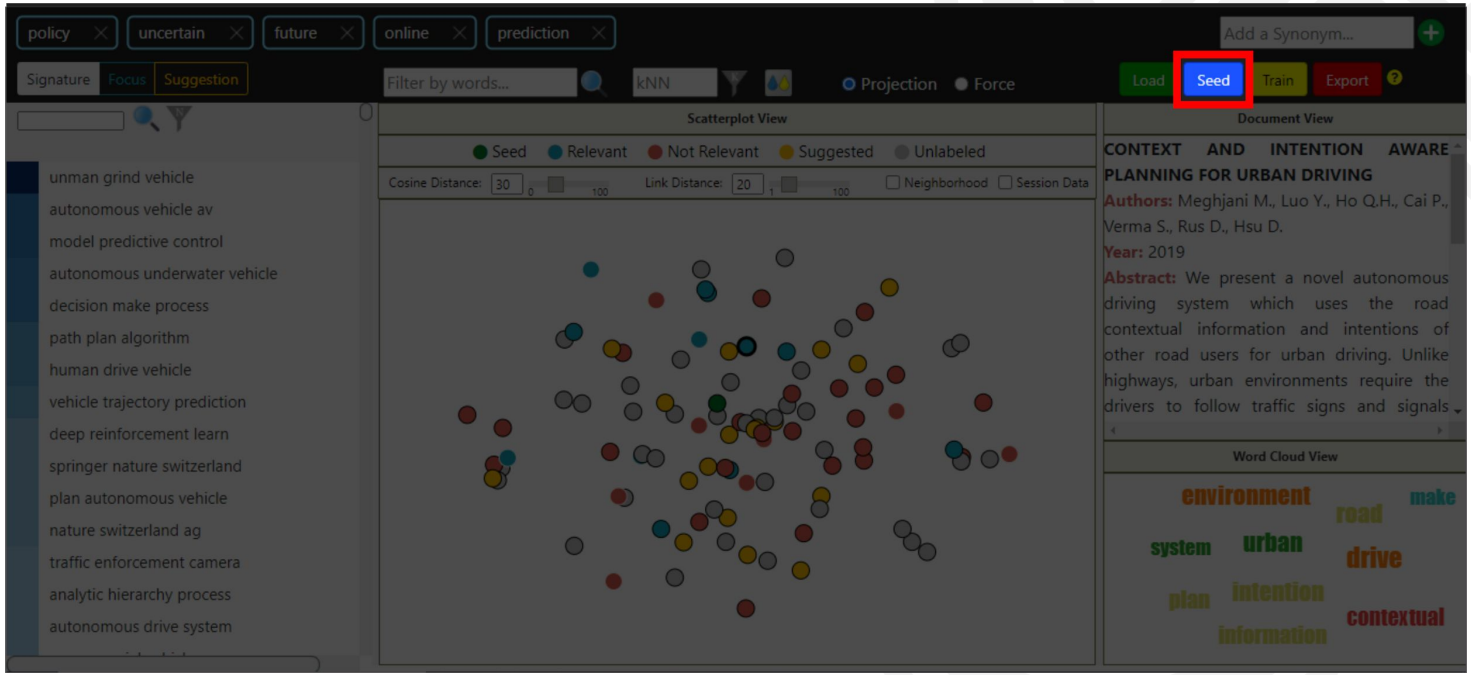
TRIVIR



Load Button:

Click Load button to insert the Corpus into the system.

TRIVIR



Seed Button:

Click Seed button to choose the initial query (SEED Document).
A SEED is a document considered highly relevant to your search.

TRIVIR

Scatterplot view:

Similarity Map of Documents

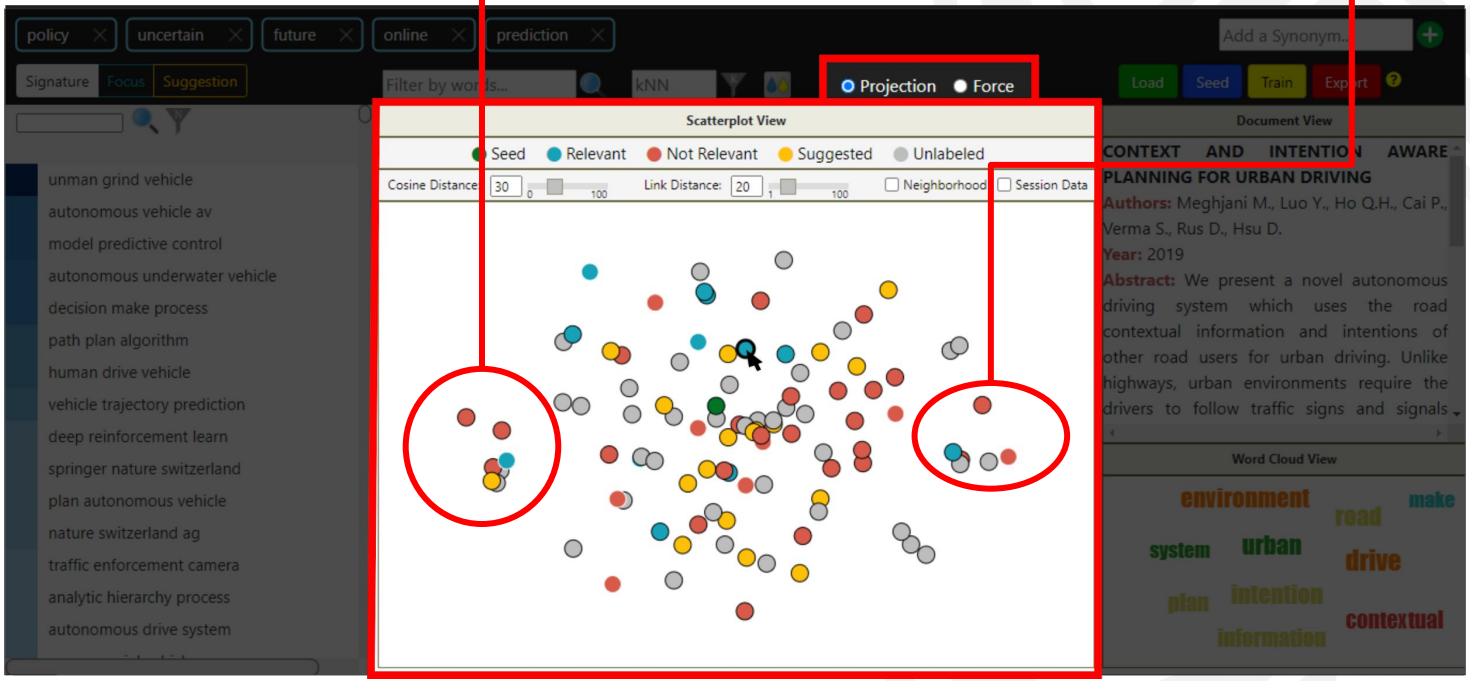
TRIVIR

Scatterplot view:

Similarity Map of Documents

TRIVIR

Clusters of possibly similar documents
Multidimensional Projection



Scatterplot view:
Similarity Map of Documents

TRIVIR

Seed

Not Relevant

Unlabeled Documents

Relevant

Suggested

TRIVIR

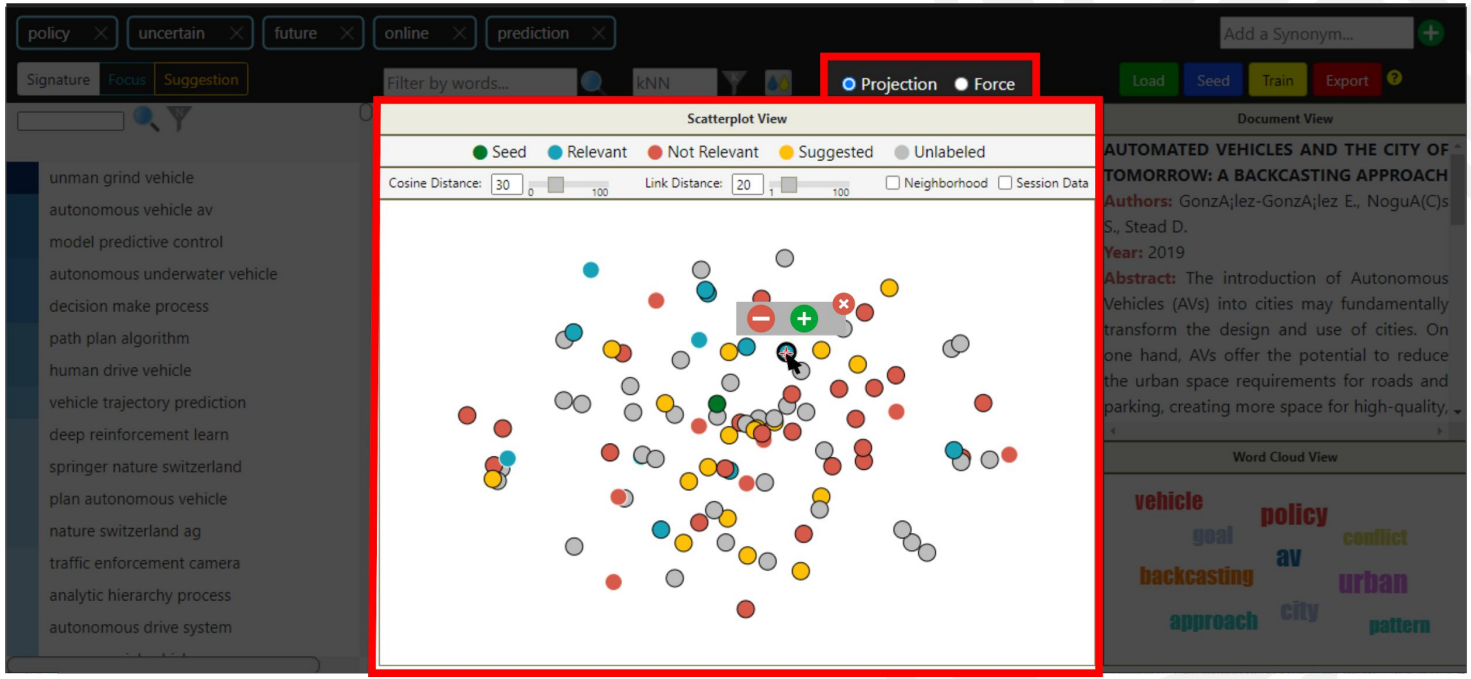
Read Documents

Unread Documents

TRIVIR

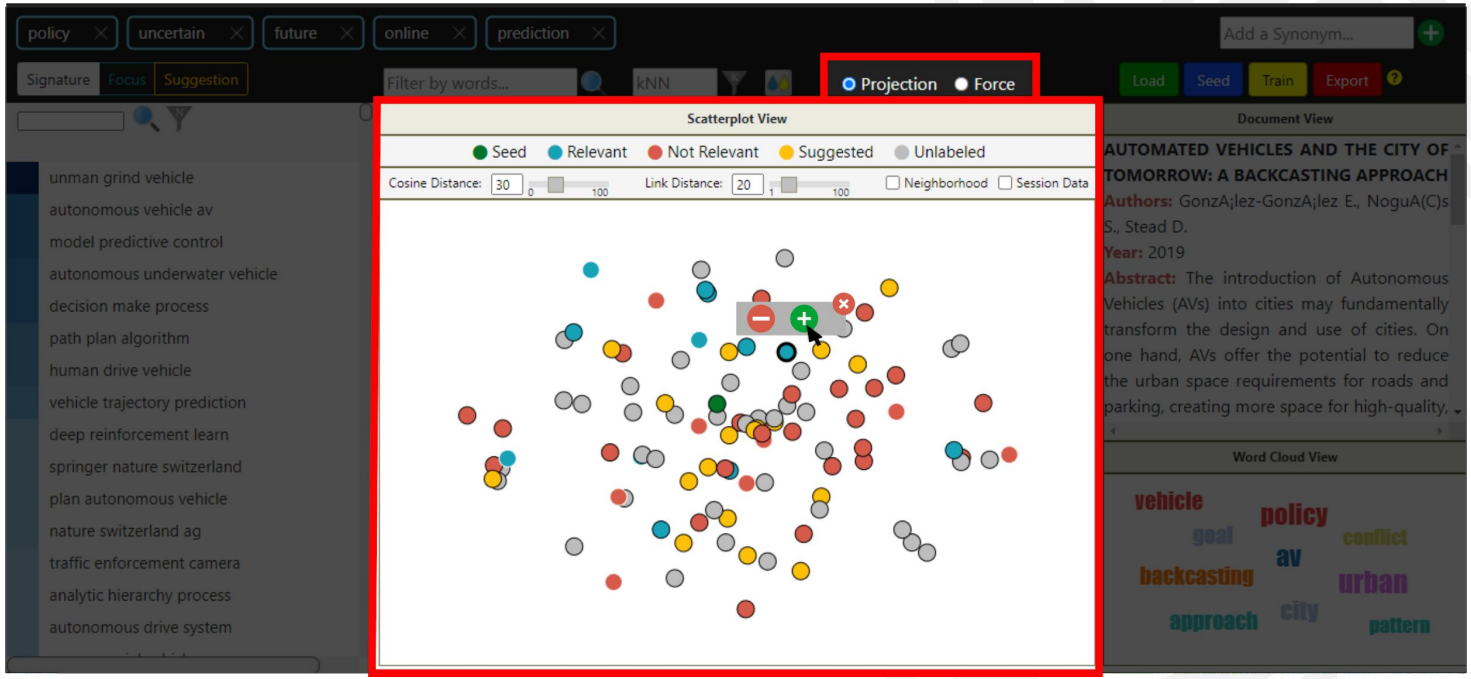
Labeling

TRIVIR



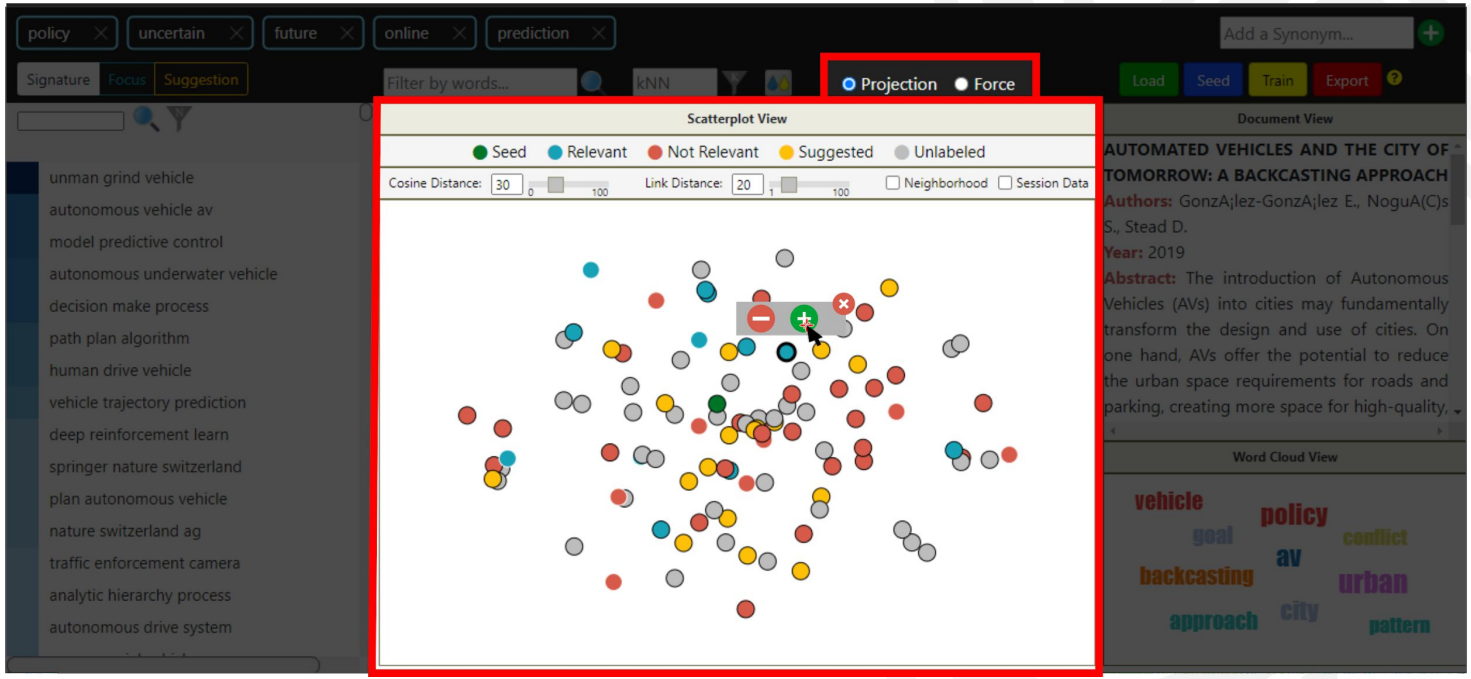
Labeling

TRIVIR



Labeling

TRIVIR



Labeling

TRIVIR

Labeling

TRIVIR

Labeling

TRIVIR

Labeling

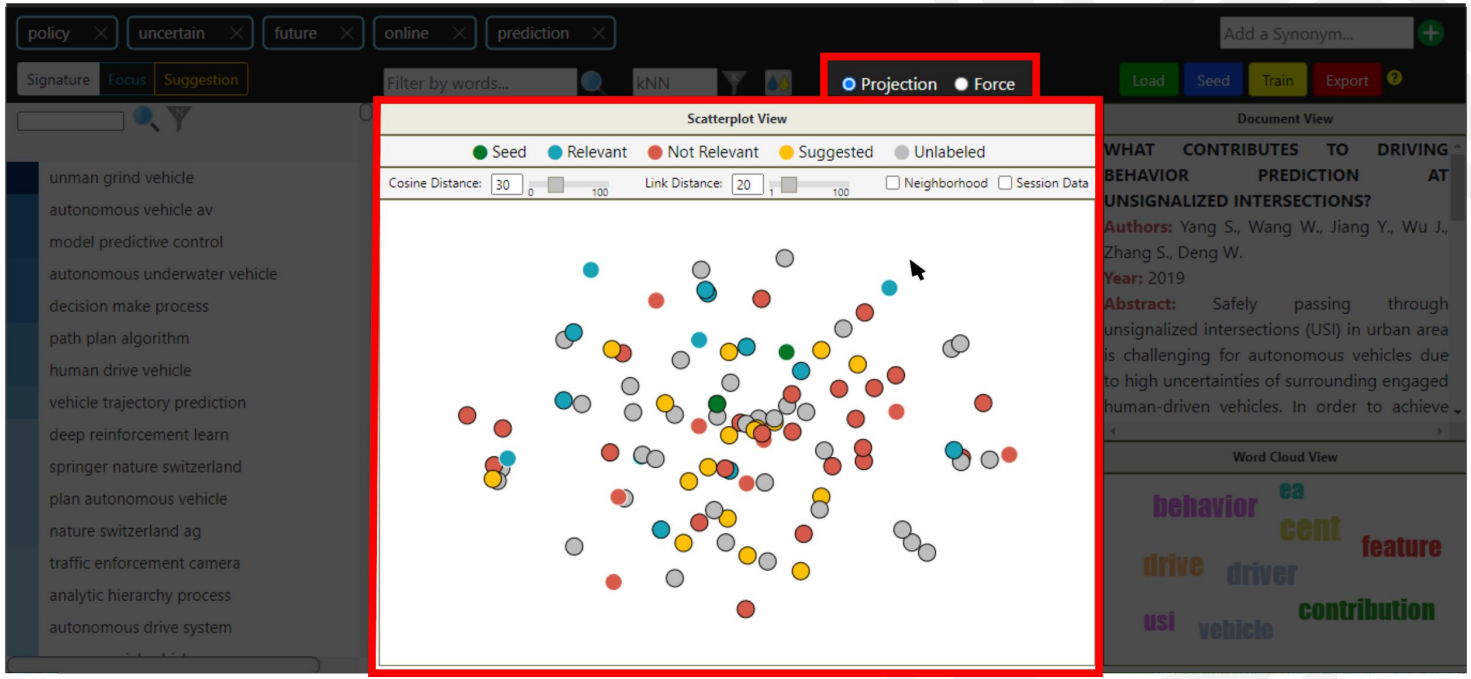
TRIVIR

Labeling

TRIVIR

Labeling

TRIVIR



Labeling

TRIVIR

Labeling

TRIVIR

Labeling

TRIVIR

Labeling

TRIVIR

Labeling

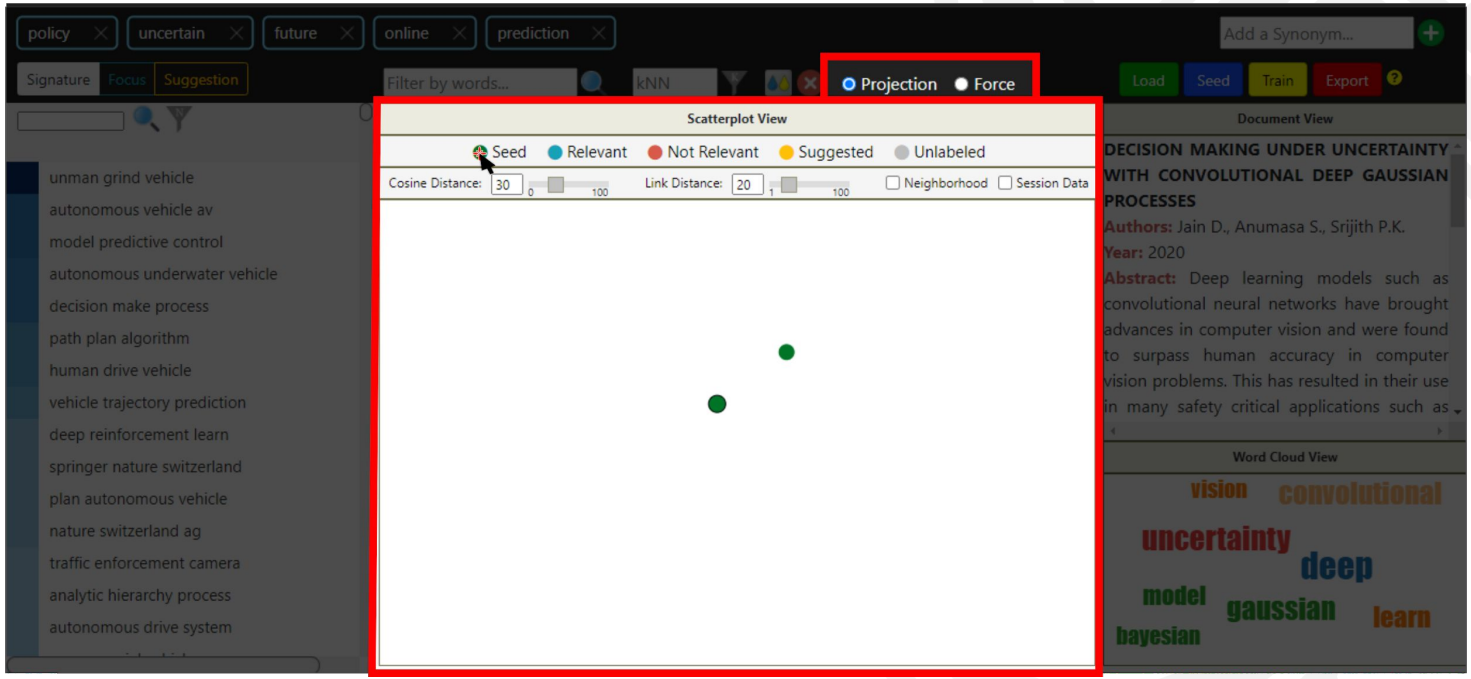
TRIVIR

Labeling

TRIVIR

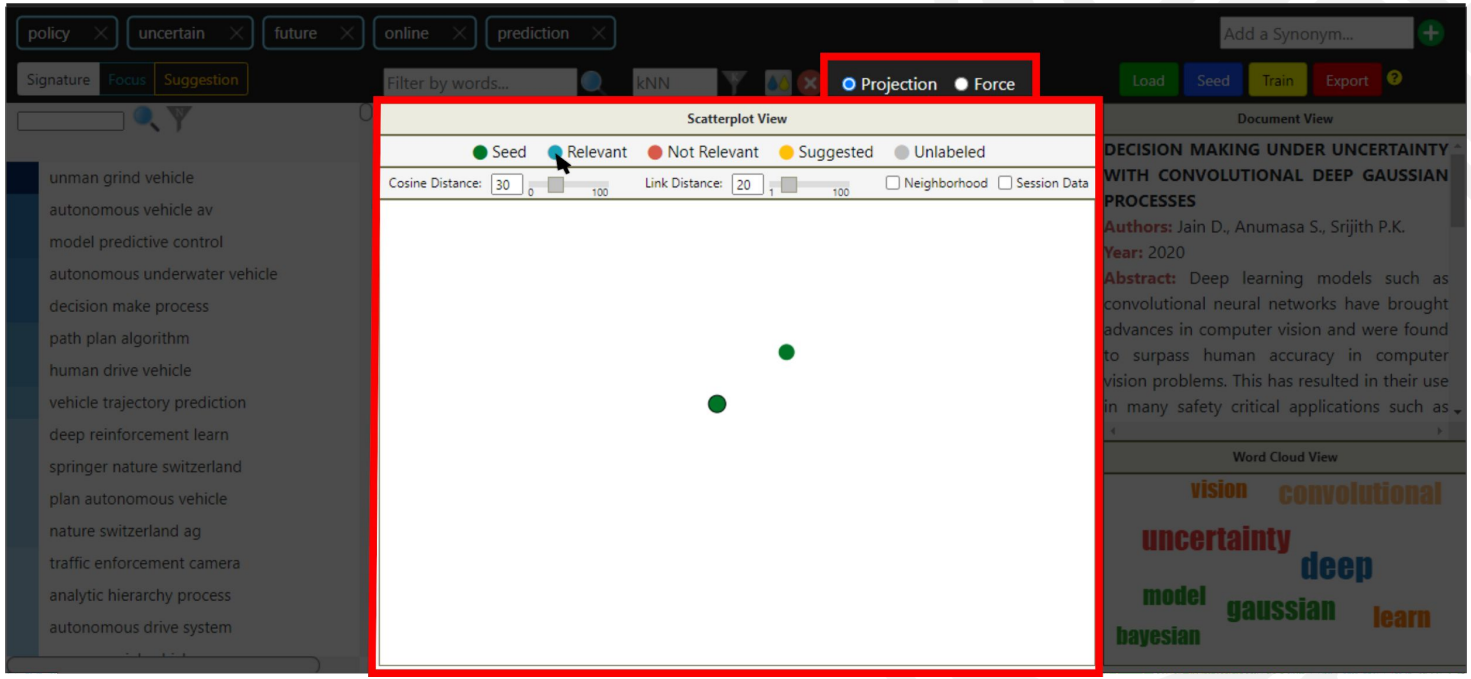
Show <u>Seed</u> documents only

TRIVIR



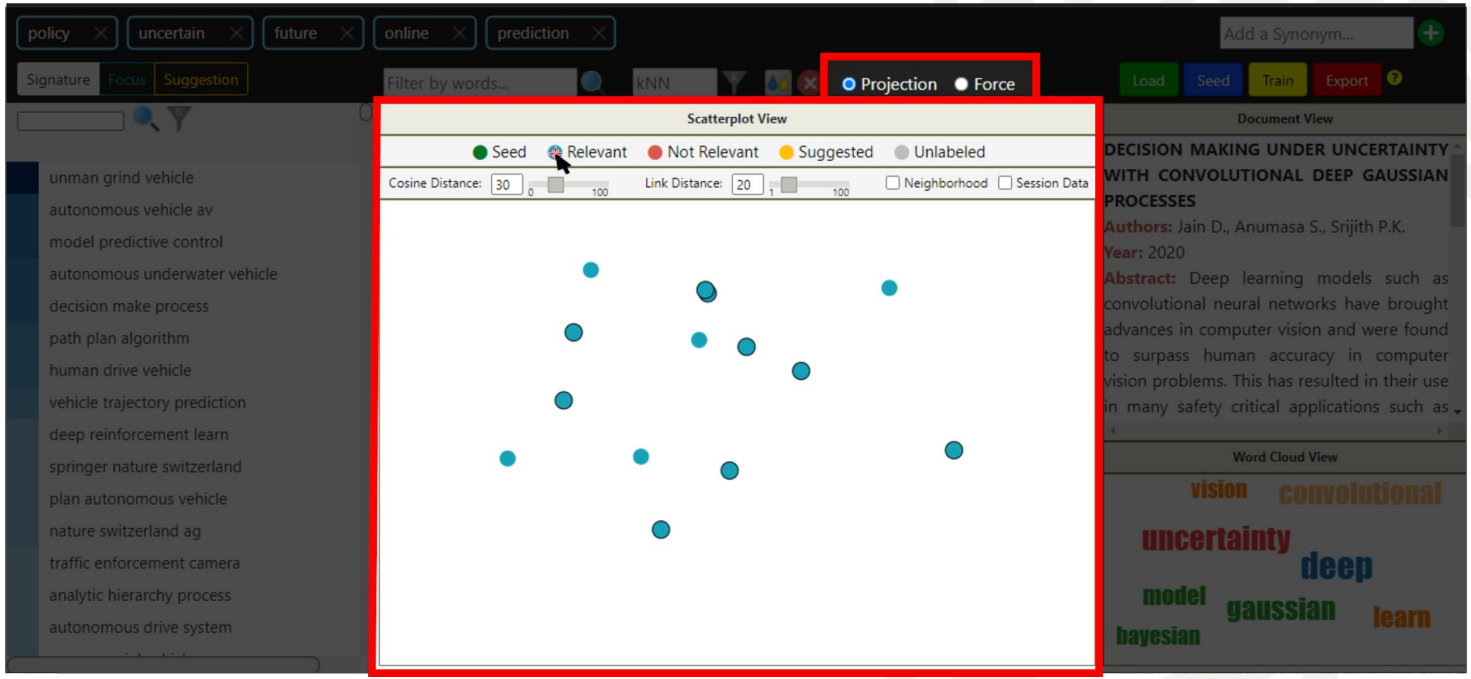
Show <u>Seed</u> documents only

TRIVIR



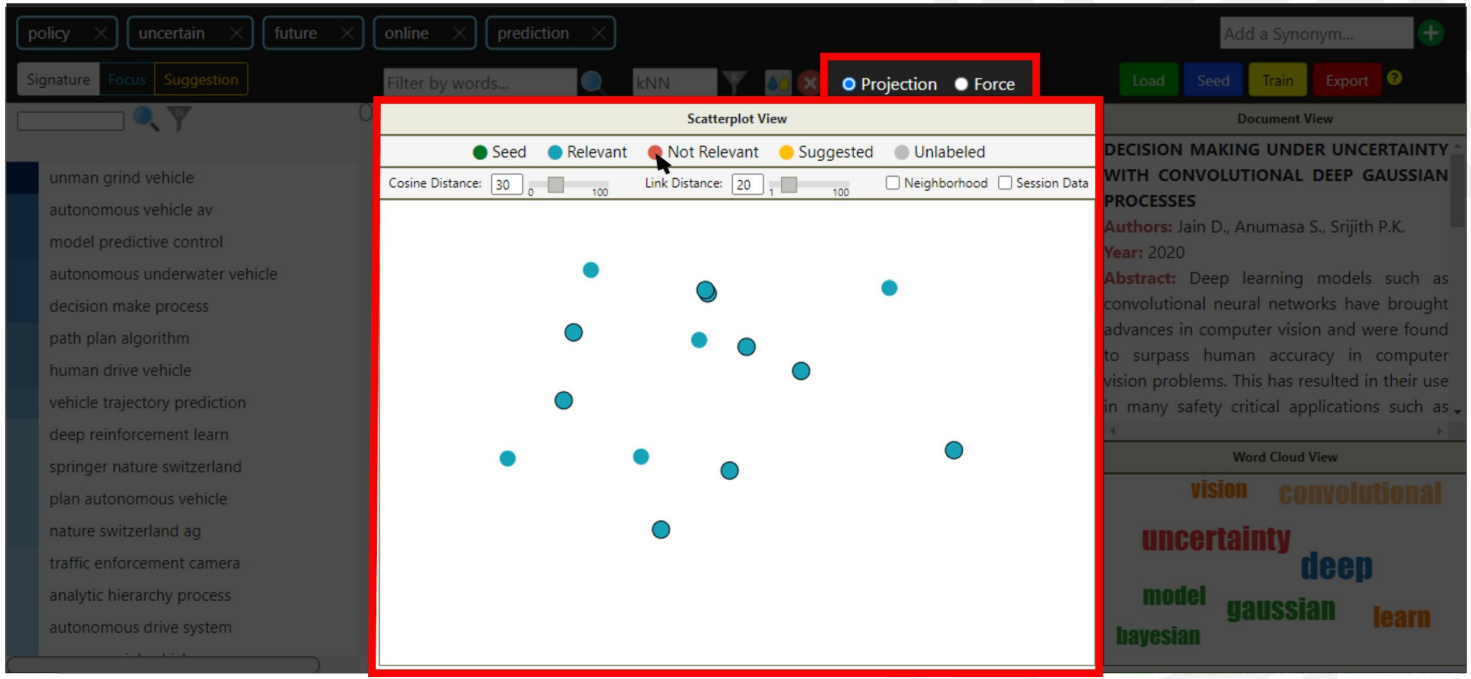
Show <u>Relevant</u> documents only

TRIVIR



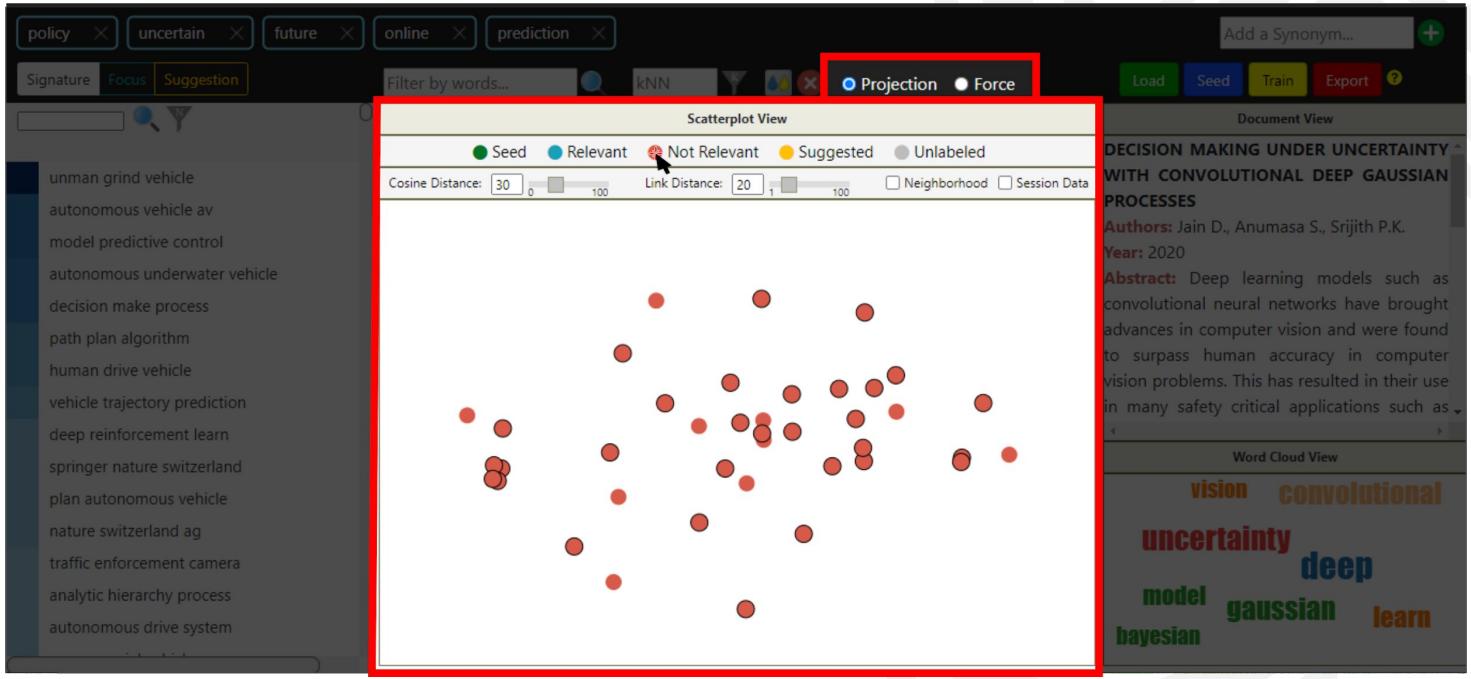
Show <u>Relevant</u> documents only

TRIVIR



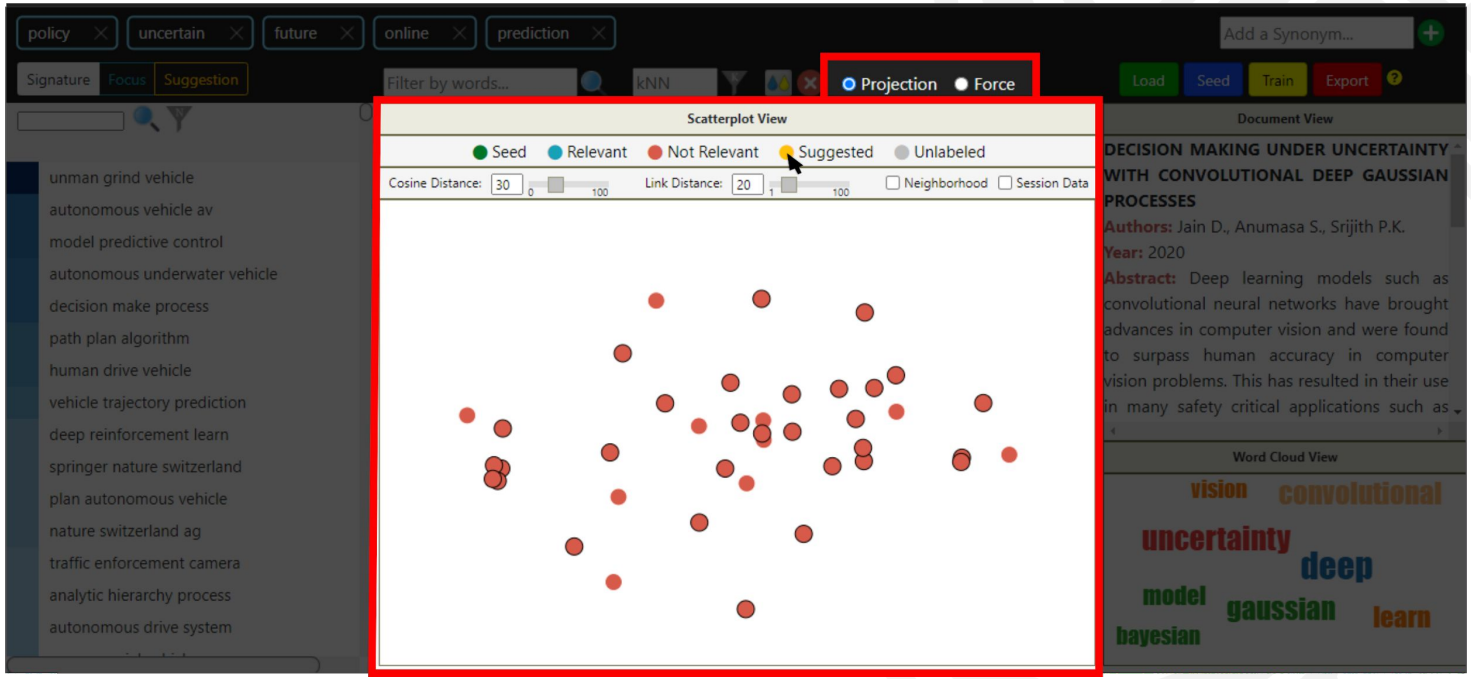
Show <u>Not Relevant</u> documents only

TRIVIR



Show <u>Not Relevant</u> documents only

TRIVIR



Show <u>Suggested</u> documents only

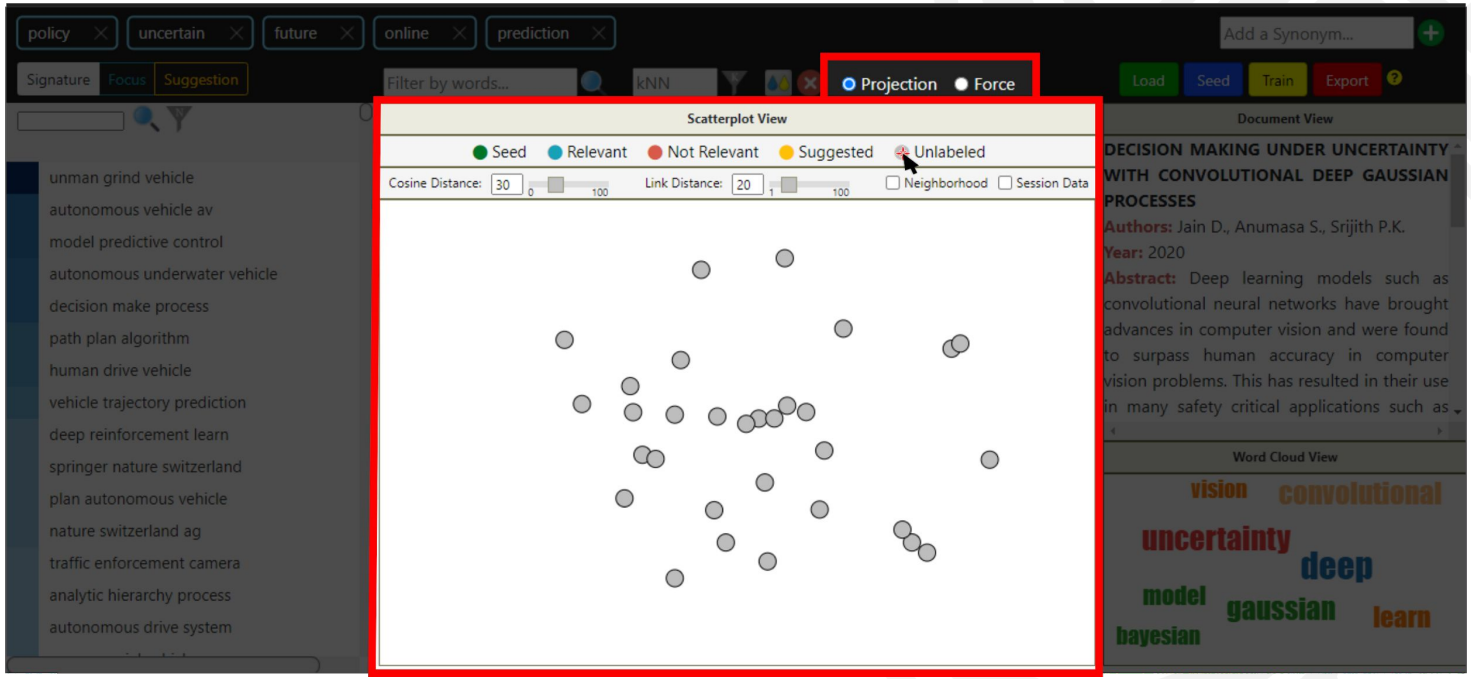
TRIVIR

Show <u>Suggested</u> documents only

TRIVIR

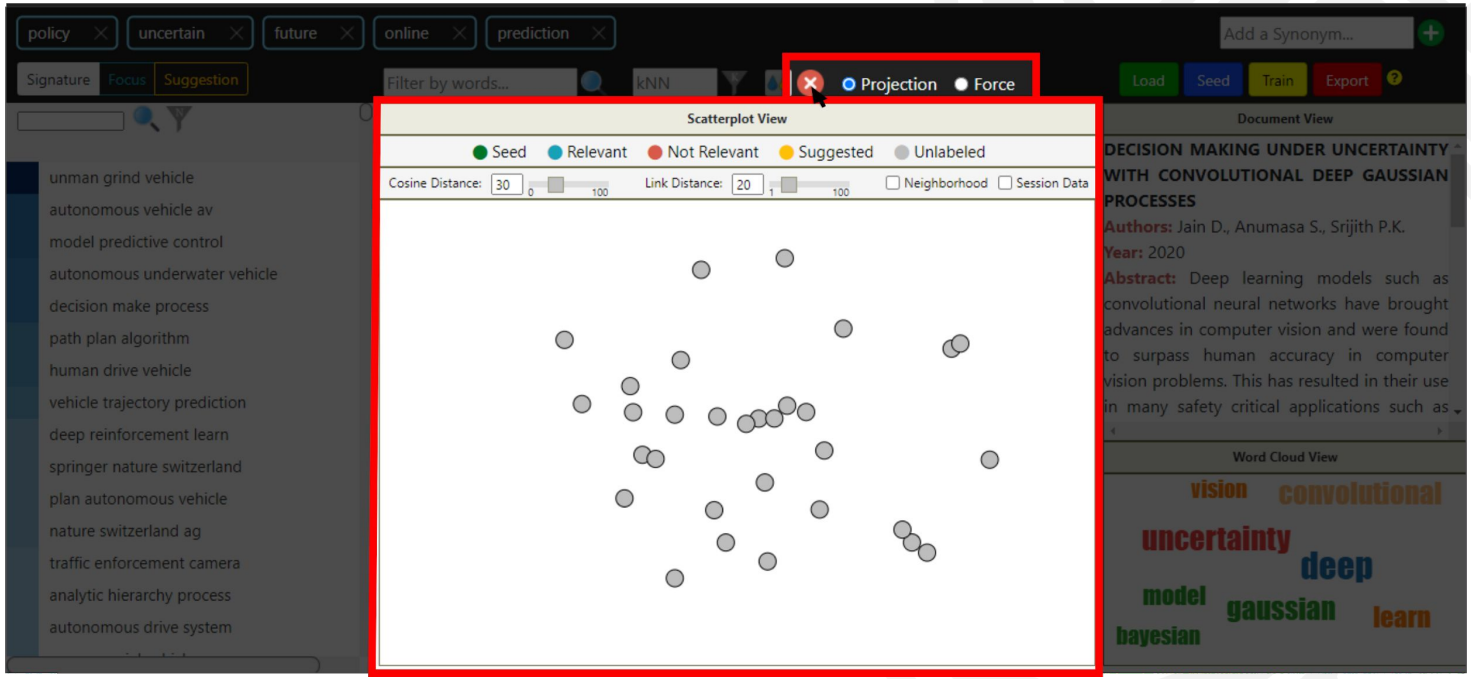
Show <u>Unlabeled</u> documents only

TRIVIR



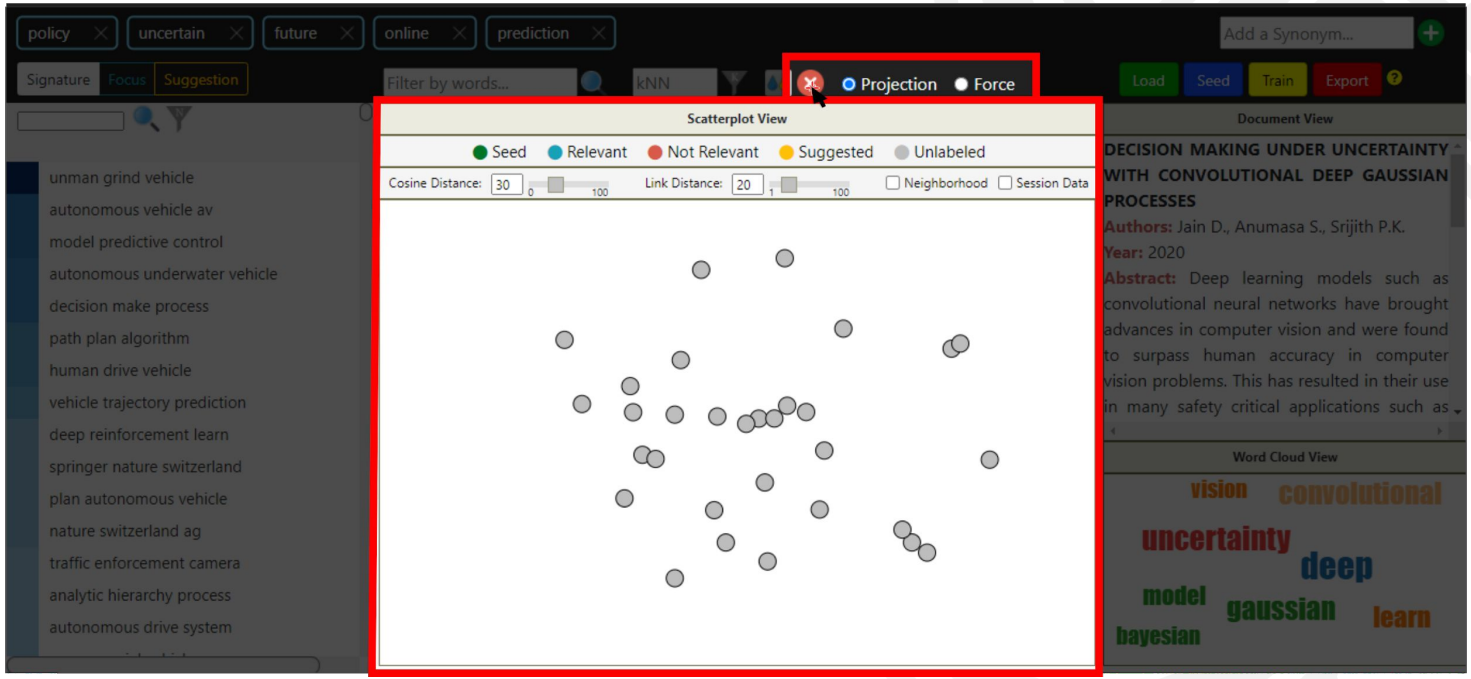
Show <u>Unlabeled</u> documents only

TRIVIR



Clear selected filters

TRIVIR



Clear selected filters

TRIVIR

Clear selected filters

TRIVIR

Parameters

Cosine Distance: as the value increases, the documents connected are less similar.

TRIVIR

Parameters

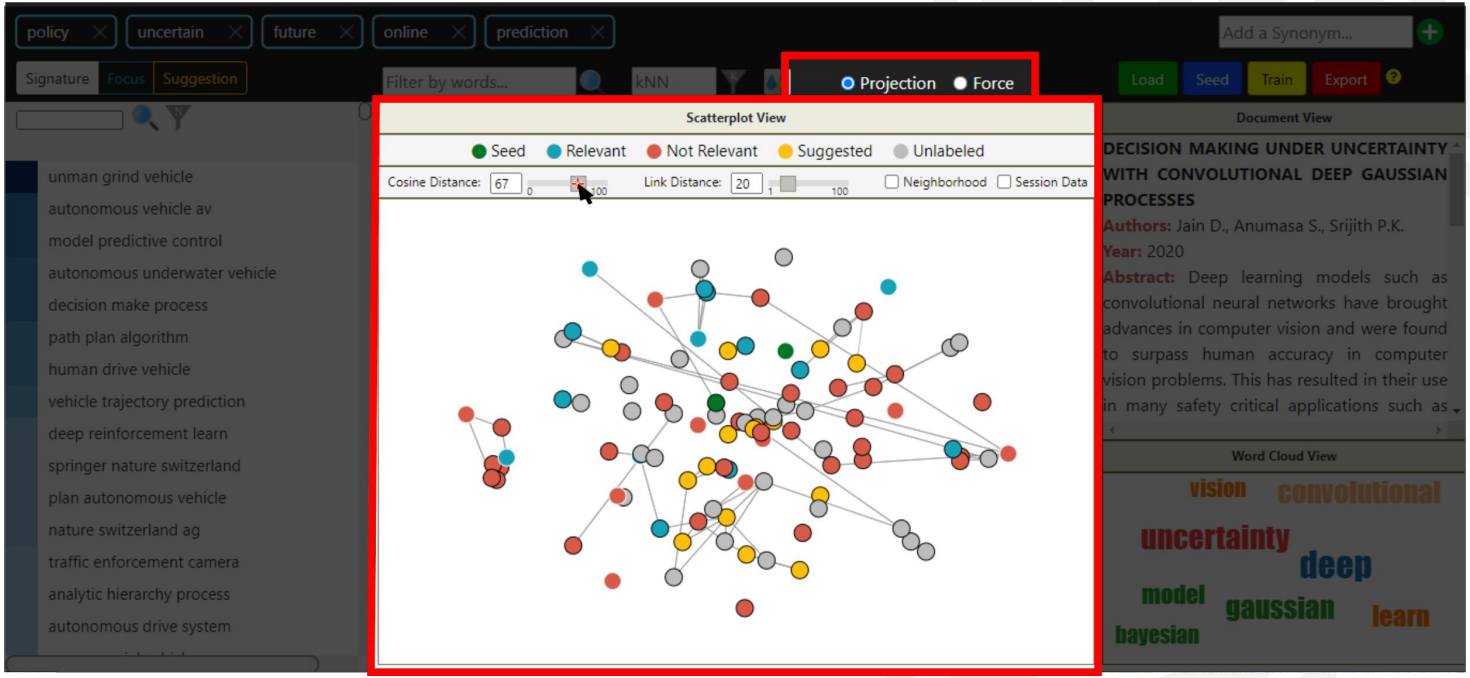
Cosine Distance: as the value increases, the documents connected are less similar.

TRIVIR

Parameters

Cosine Distance: as the value increases, the documents connected are less similar.

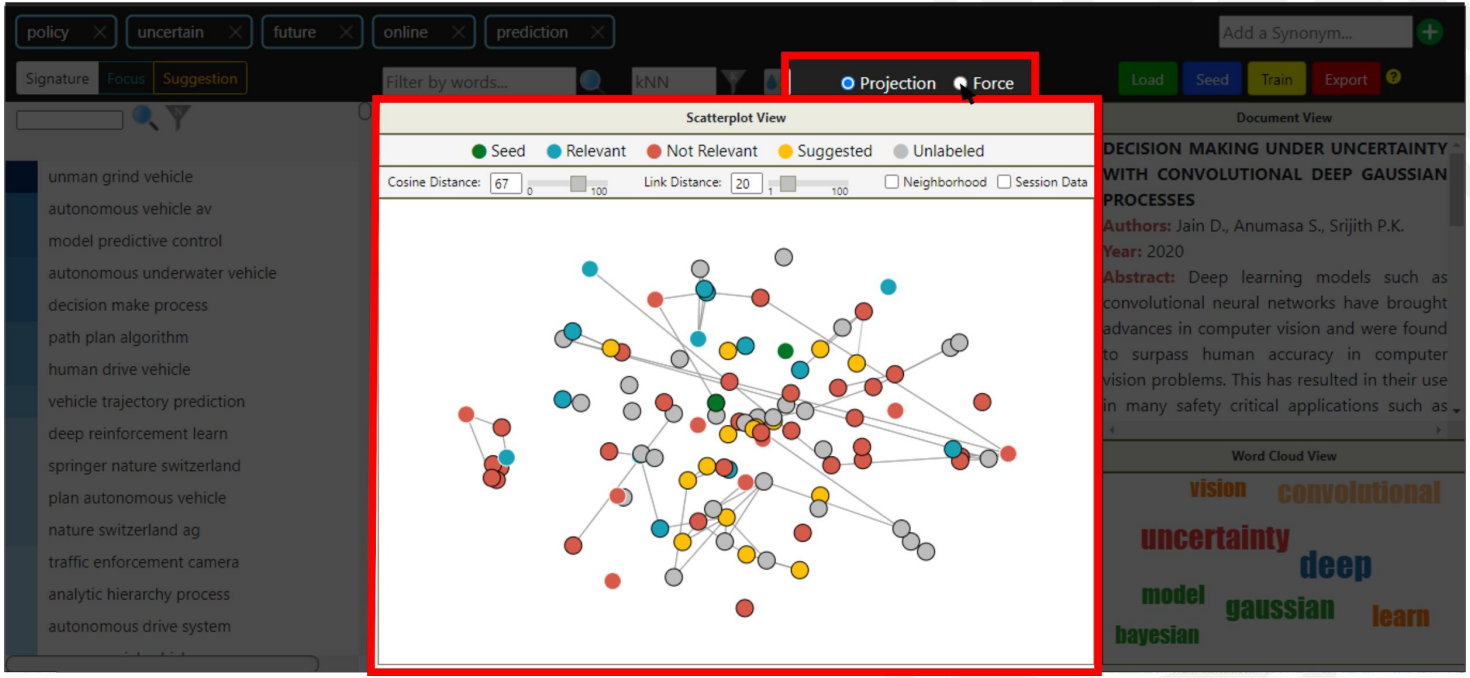
TRIVIR



Parameters

Cosine Distance: as the value increases, the documents connected are less similar.

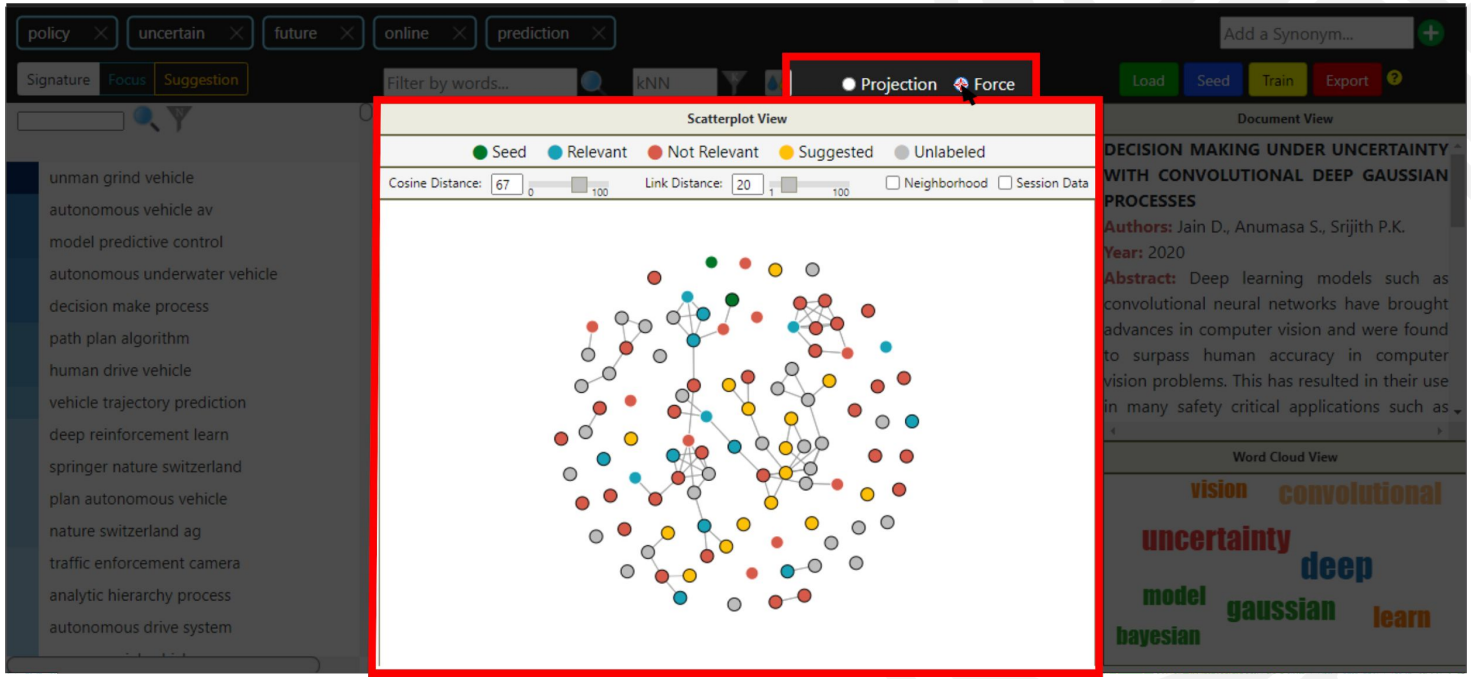
TRIVIR



Visualization Techniques

Projection: the placement of the circles is computed with a Multidimensional Projection technique.

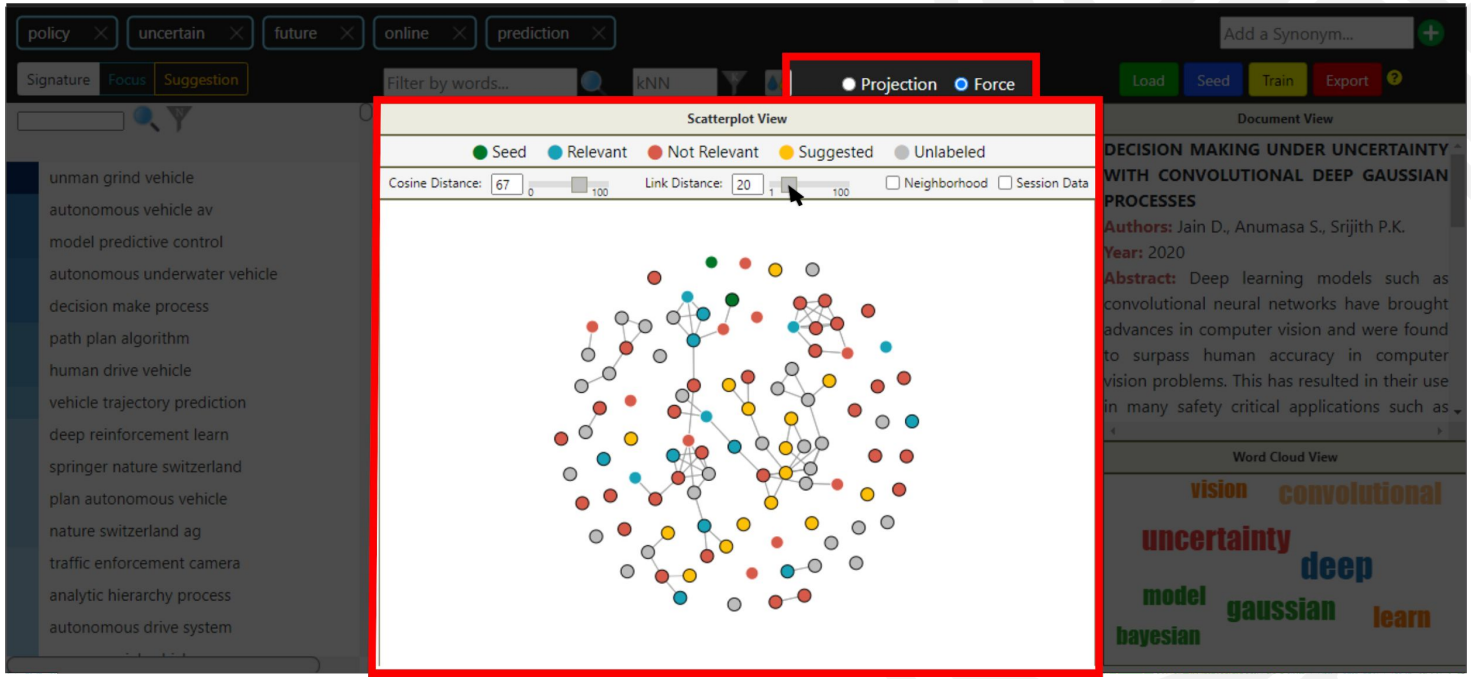
TRIVIR



Visualization Techniques

Force: the placement of the circles is computed with a Force-Layout technique.

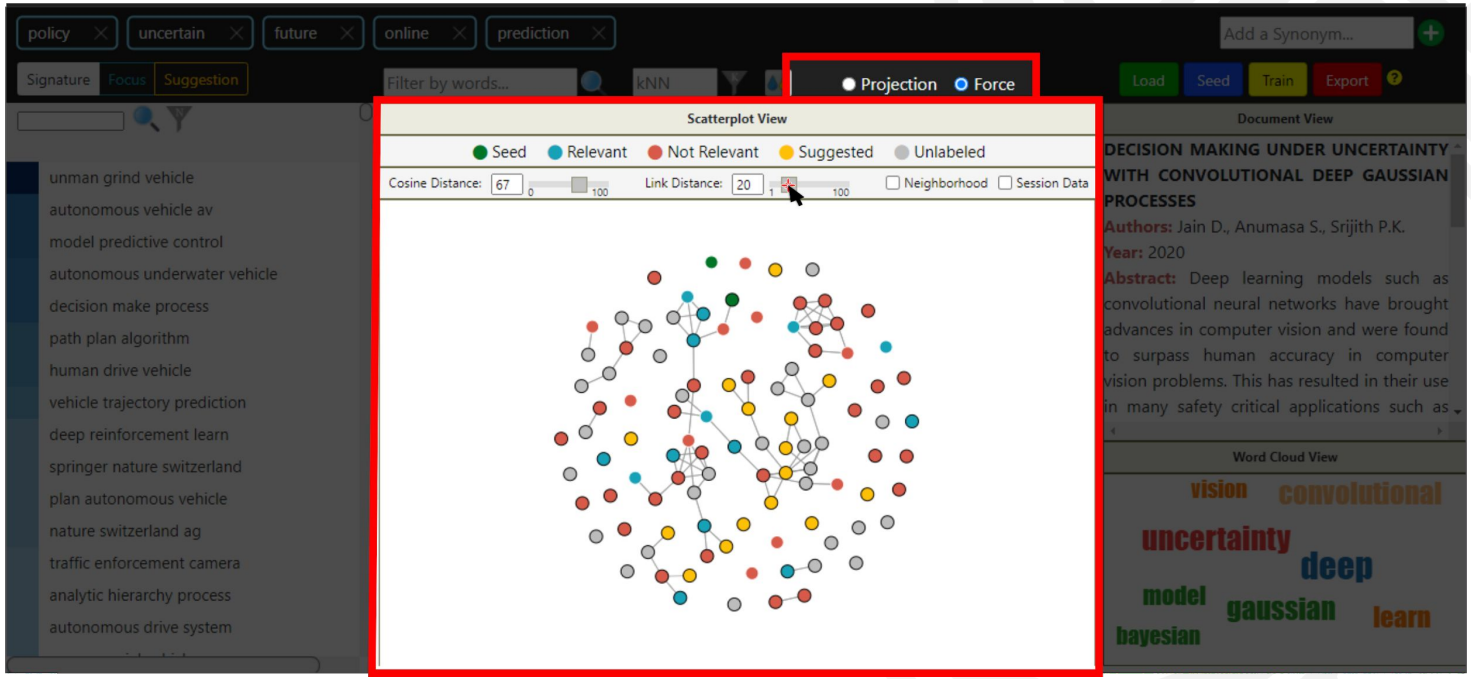
TRIVIR



Parameters

Link Distance: the length of the links.

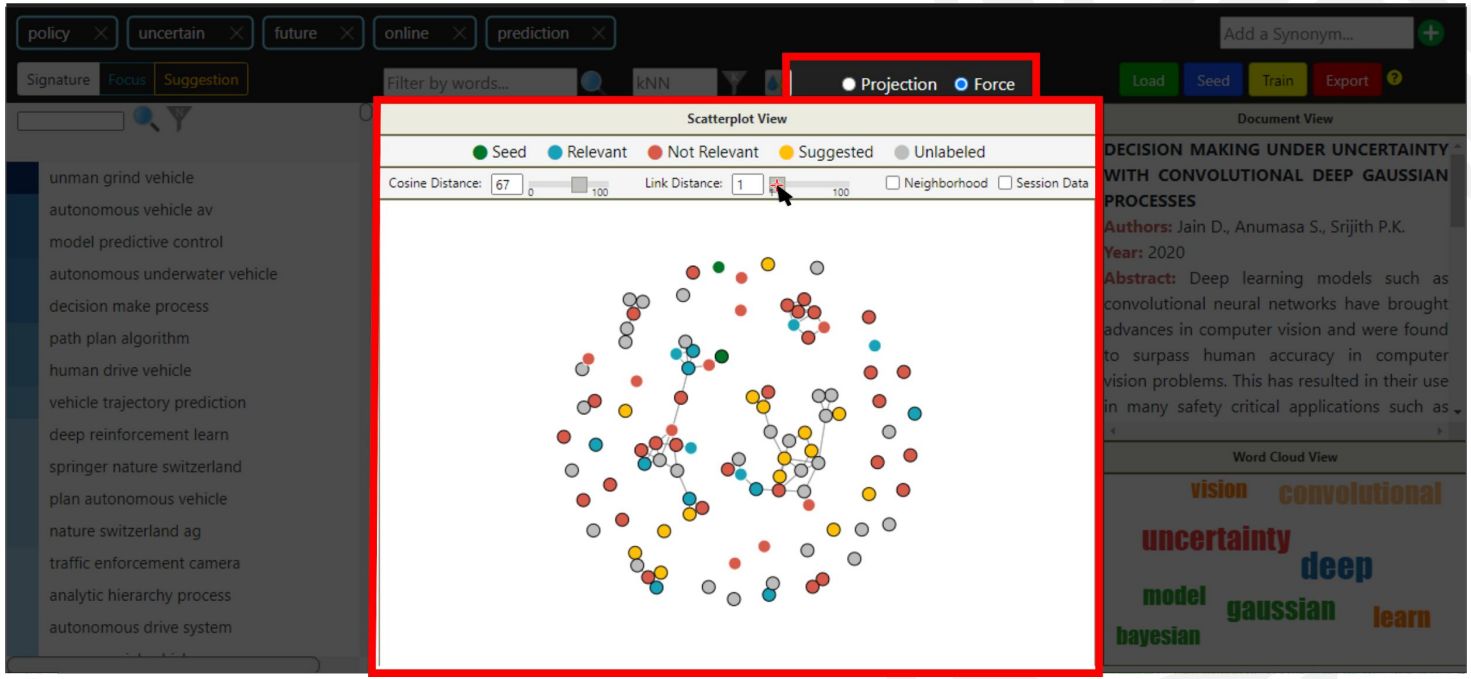
TRIVIR



Parameters

Link Distance: bring similar closer/move similar farther.

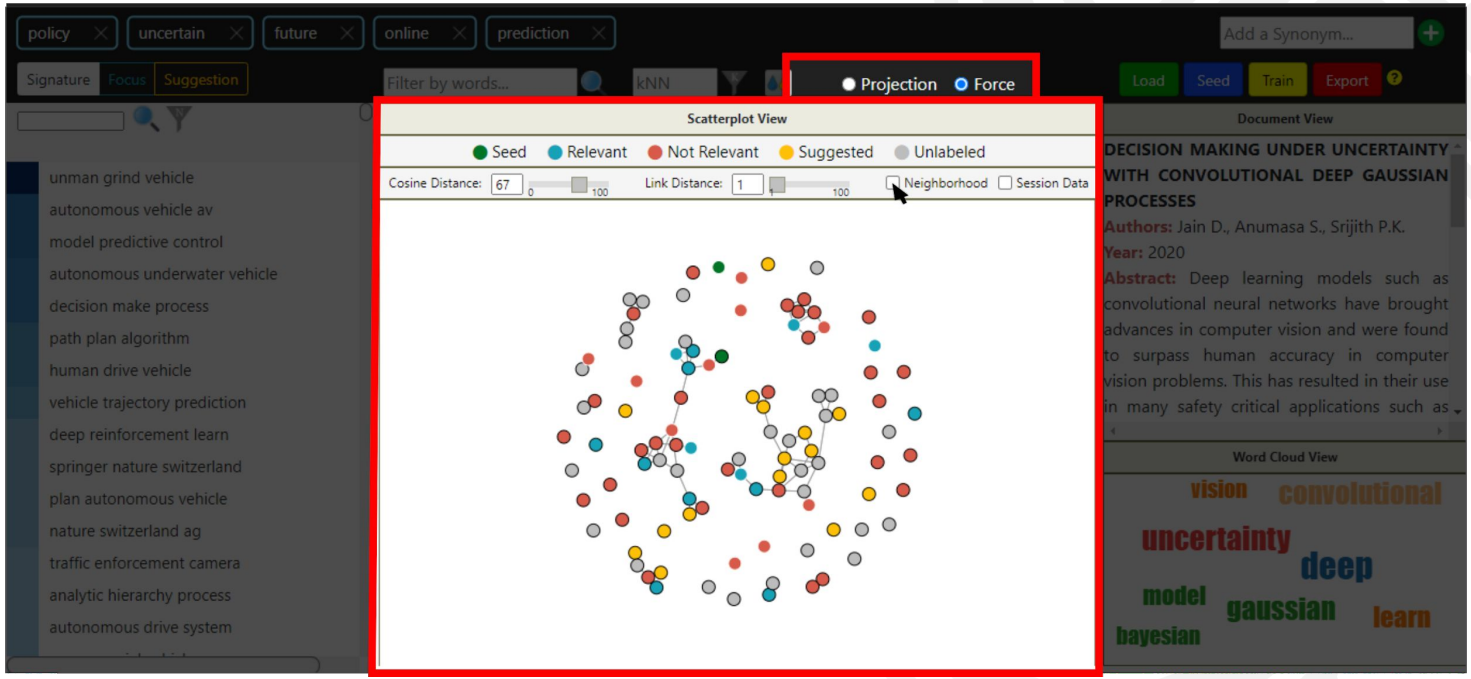
TRIVIR



Parameters

Link Distance: bring similar closer/move similar farther.

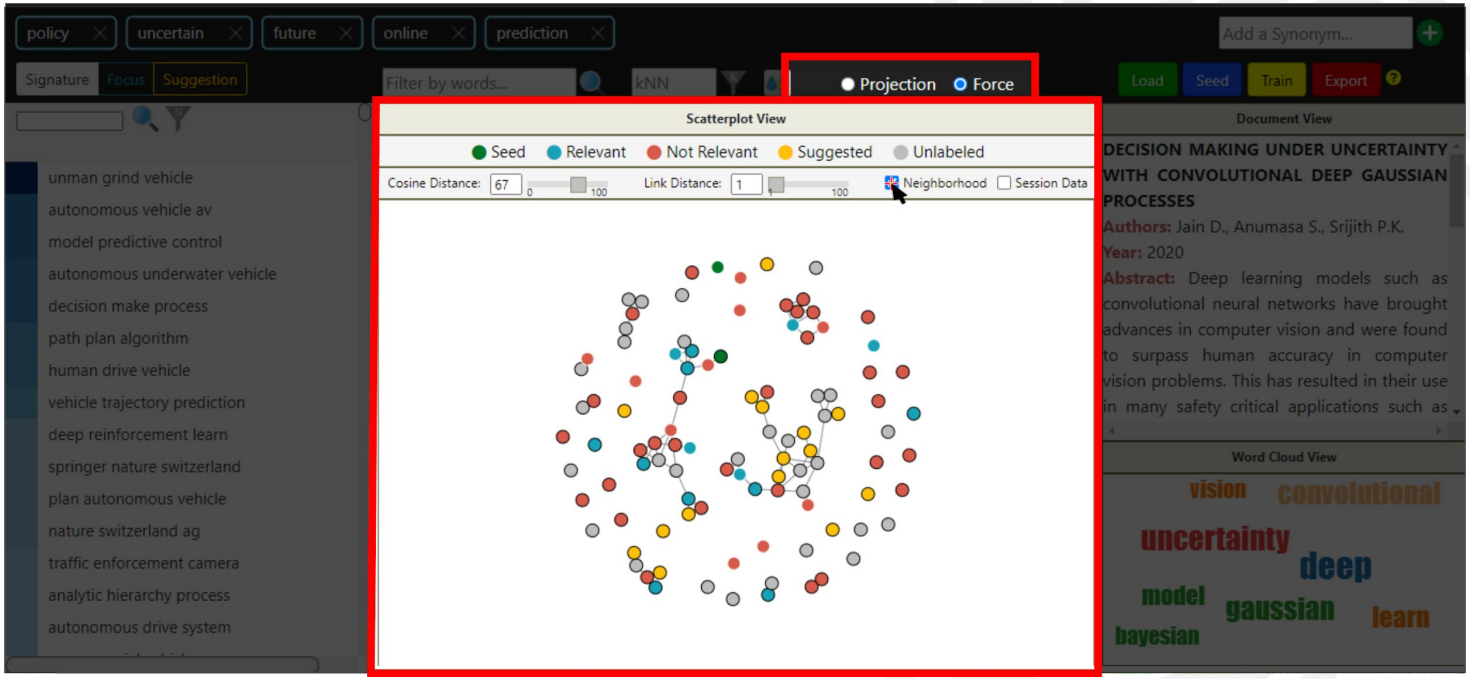
TRIVIR



Parameters

Neighborhood: Represents the 'closeness' between documents.

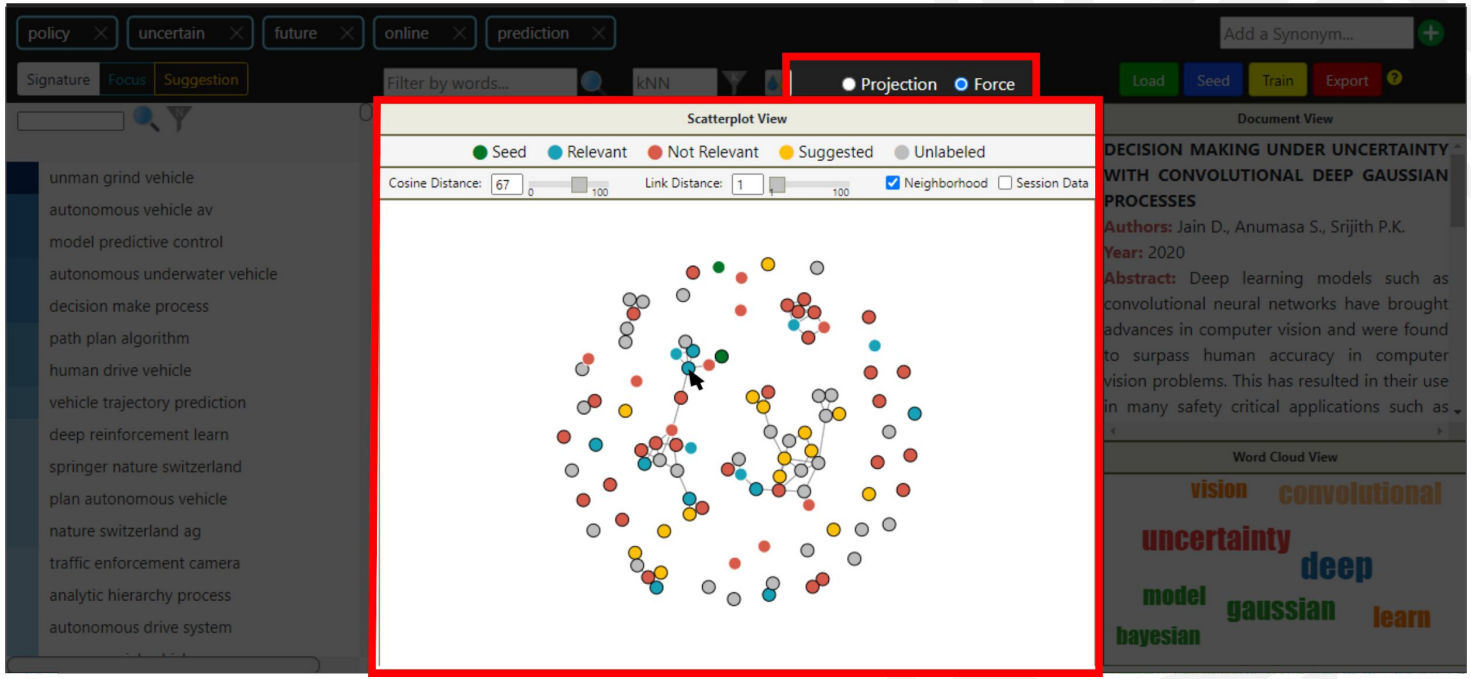
TRIVIR



Parameters

Neighborhood: Show 'closest neighbors' to a document, i.e., its most similar documents.

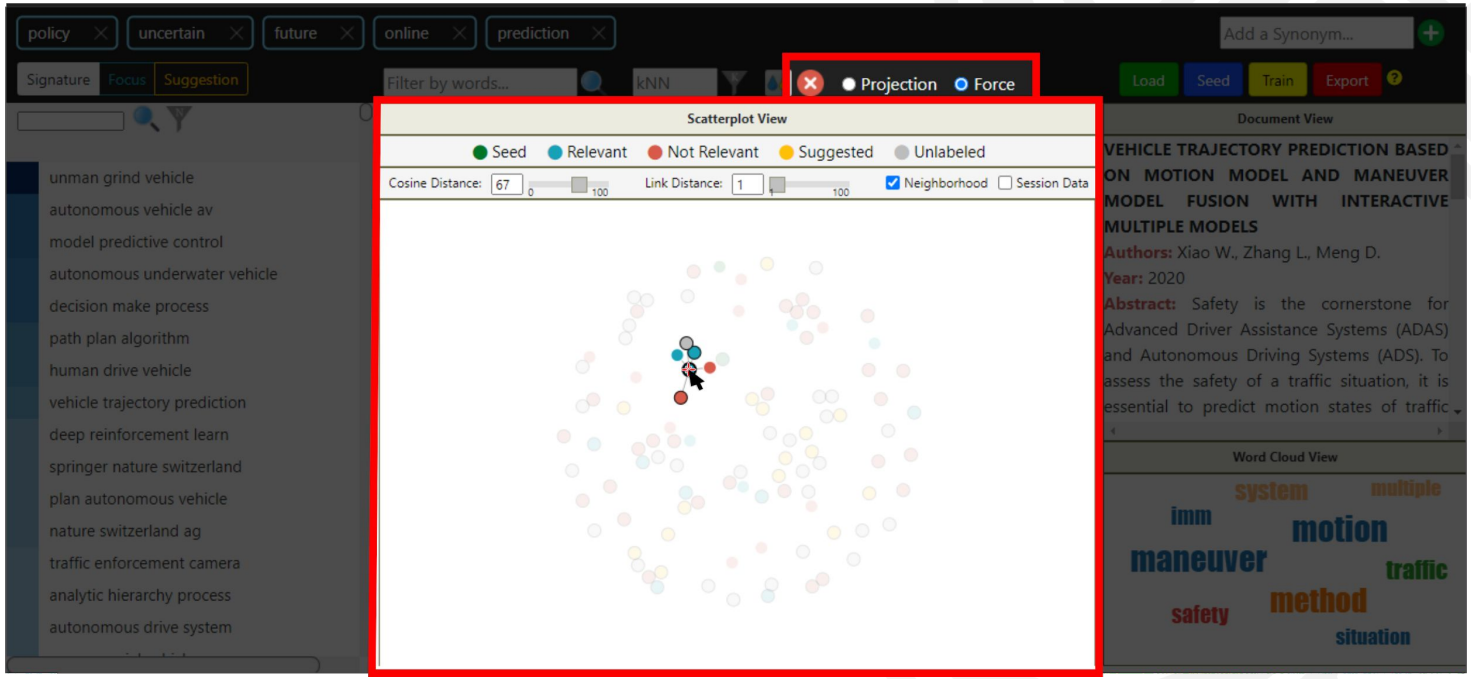
TRIVIR



Parameters

Neighborhood: Show 'closest neighbors' to a document, i.e., its most similar documents.

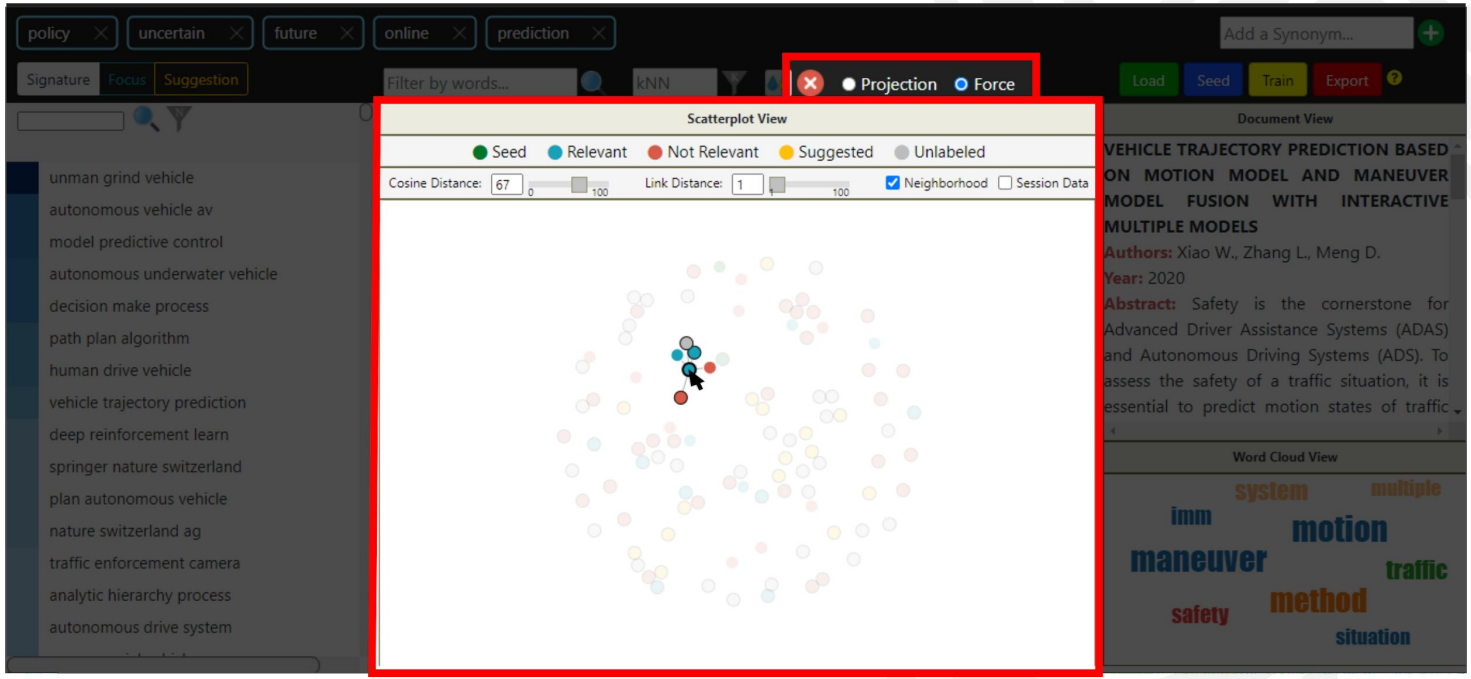
TRIVIR



Parameters

Neighborhood: Show 'closest neighbors' to a document, i.e., its most similar documents.

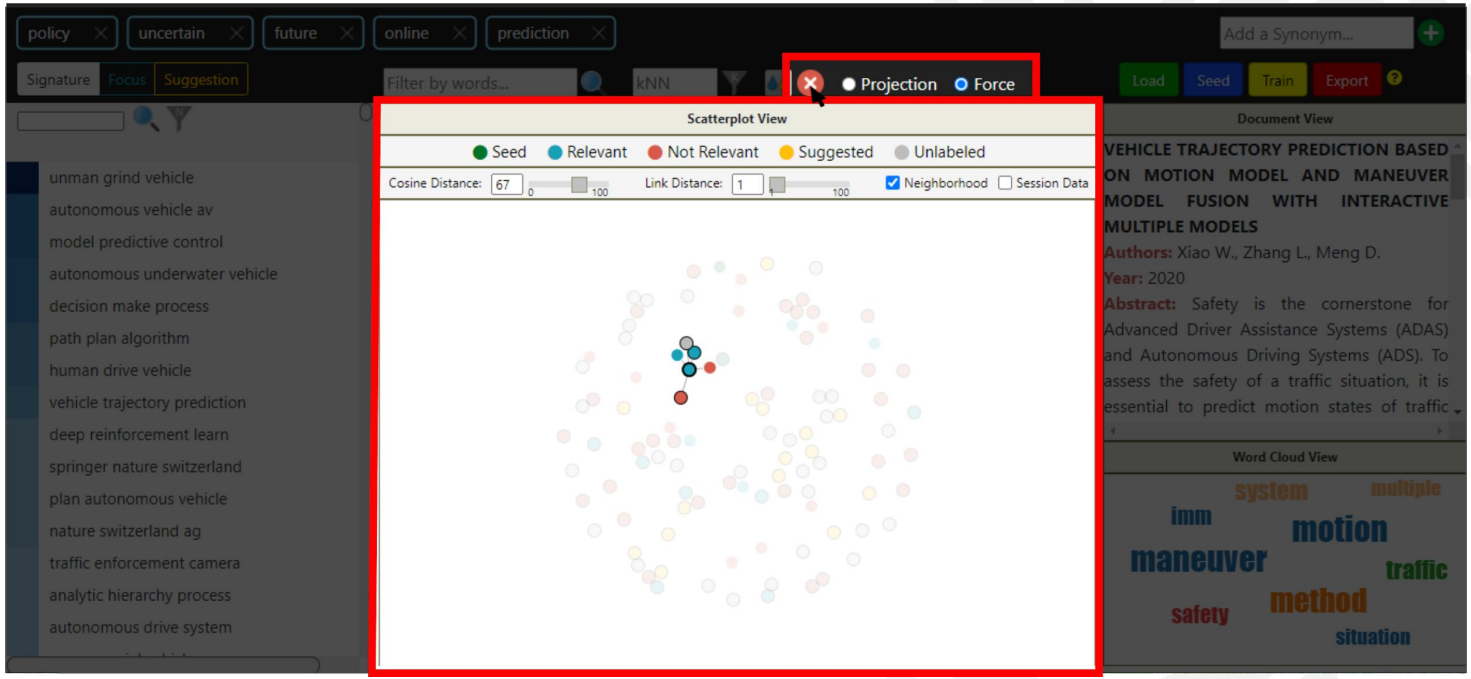
TRIVIR



Parameters

Neighborhood: Show ‘closest neighbors’ to a document, i.e., its most similar documents.

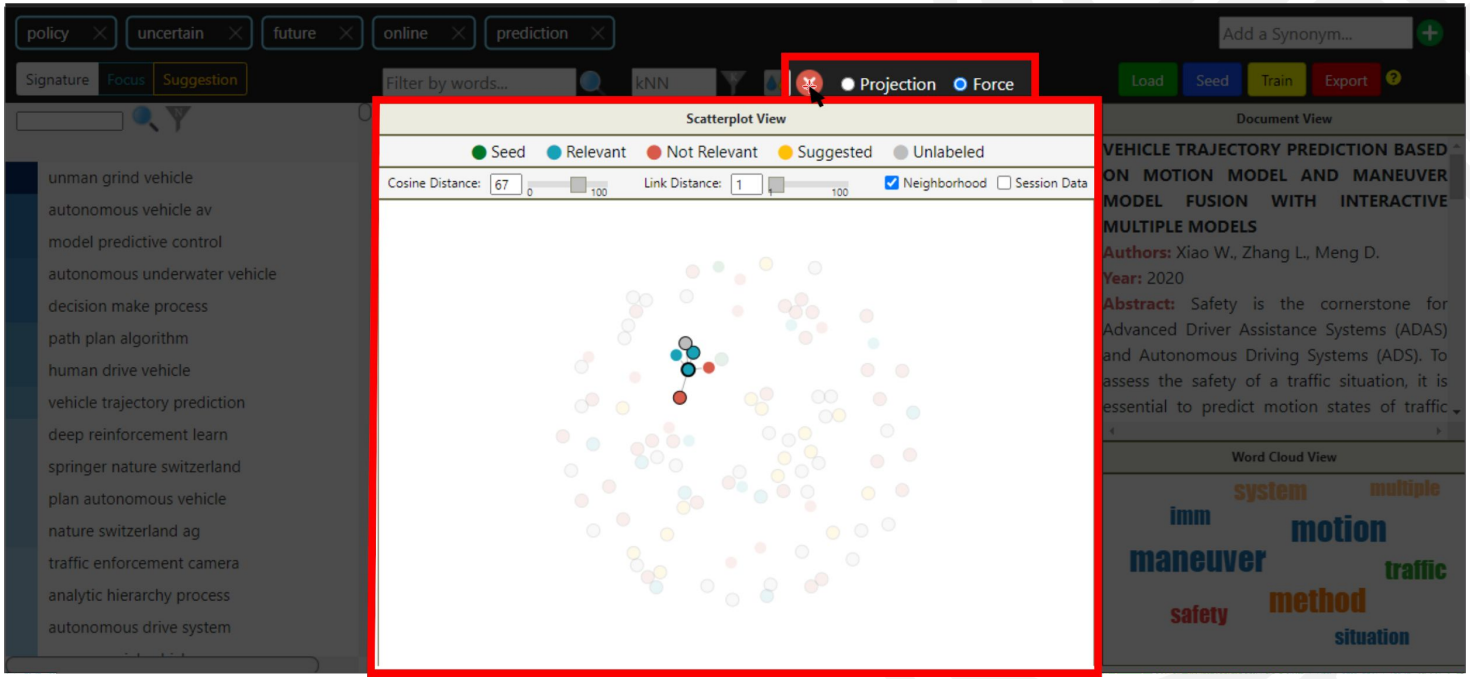
TRIVIR



Parameters

Neighborhood: Show ‘closest neighbors’ to a document, i.e., its most similar documents.

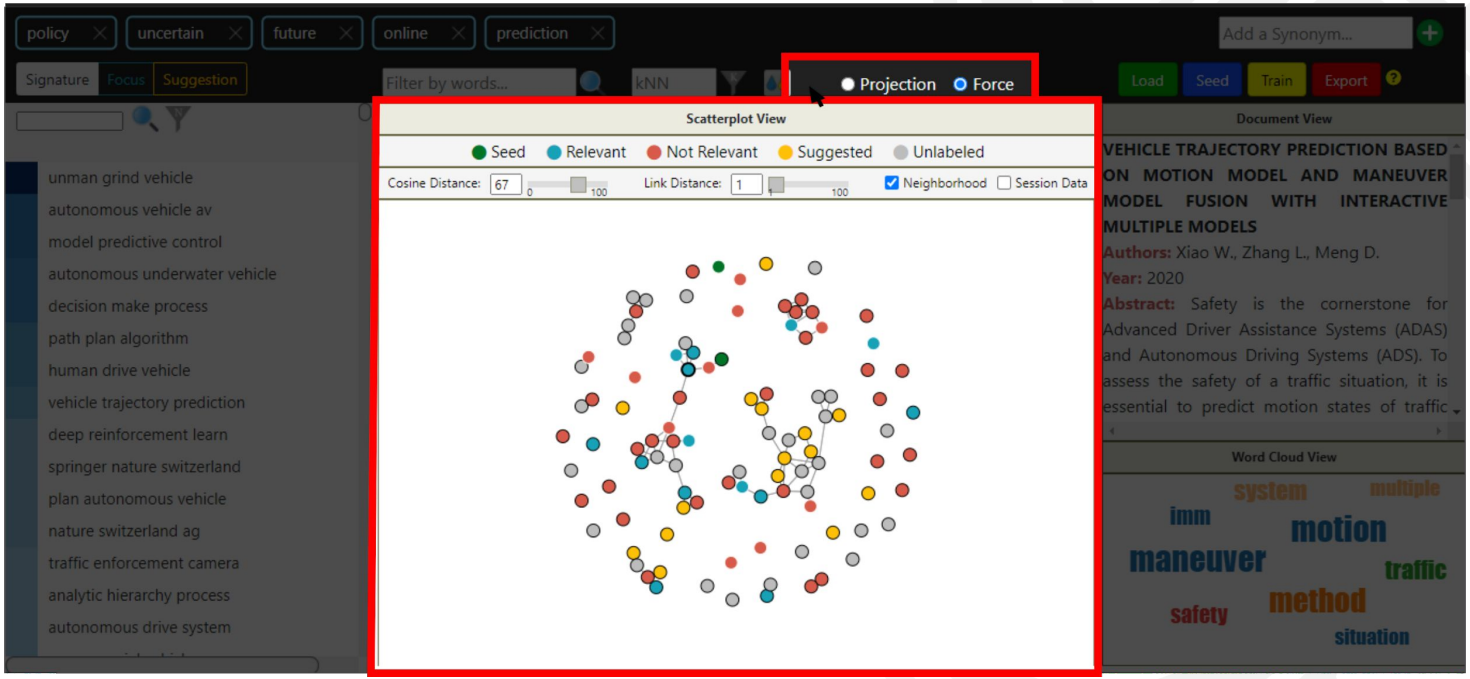
TRIVIR



Parameters

Neighborhood: Show ‘closest neighbors’ to a document, i.e., its most similar documents.

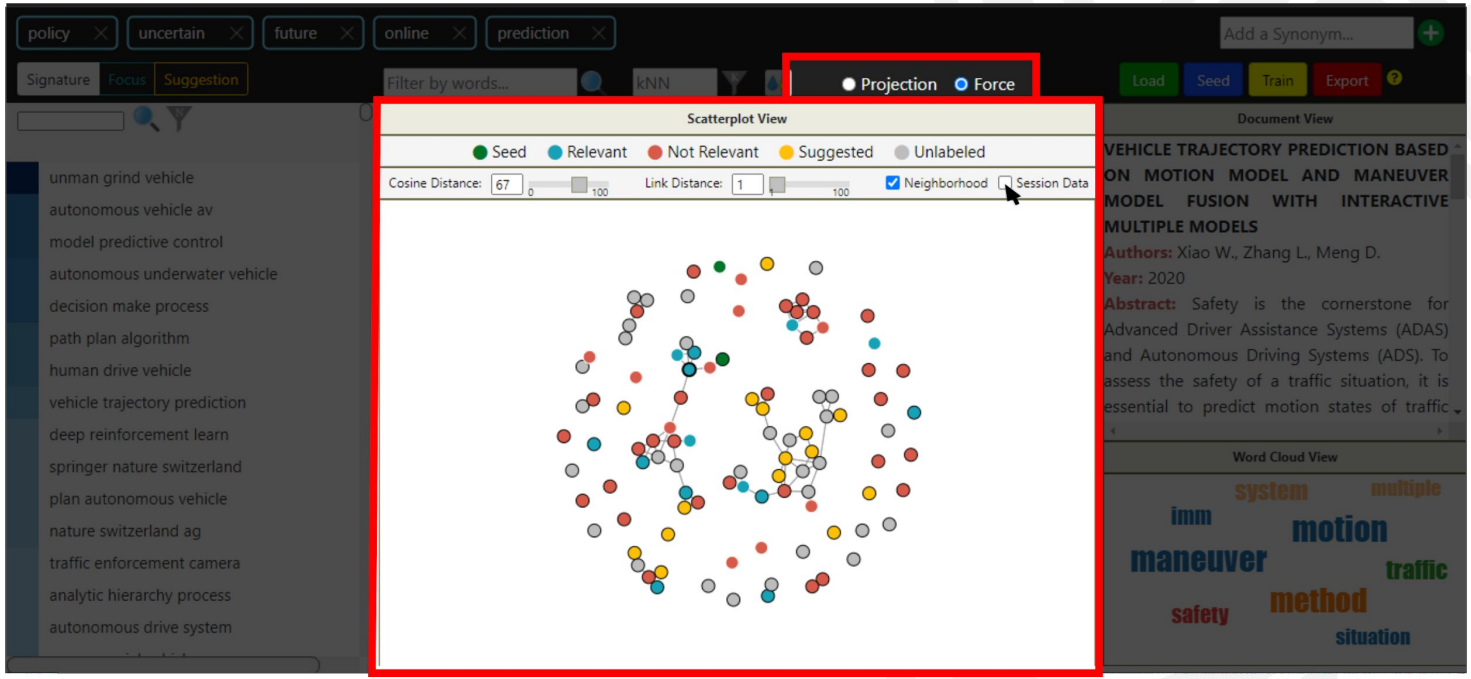
TRIVIR



Parameters

Neighborhood: Show 'closest neighbors' to a document, i.e., its most similar documents.

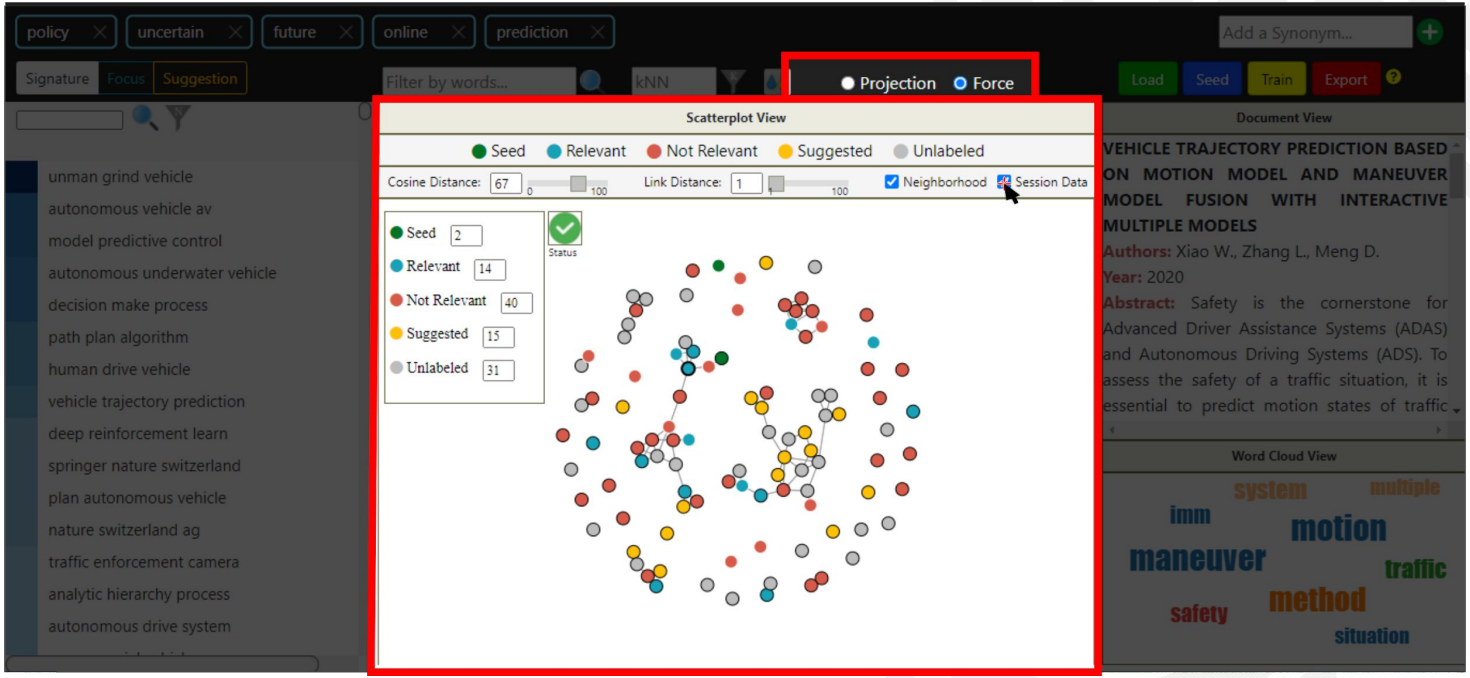
TRIVIR



Parameters

Session Data: Show how many documents per label, and the system status.

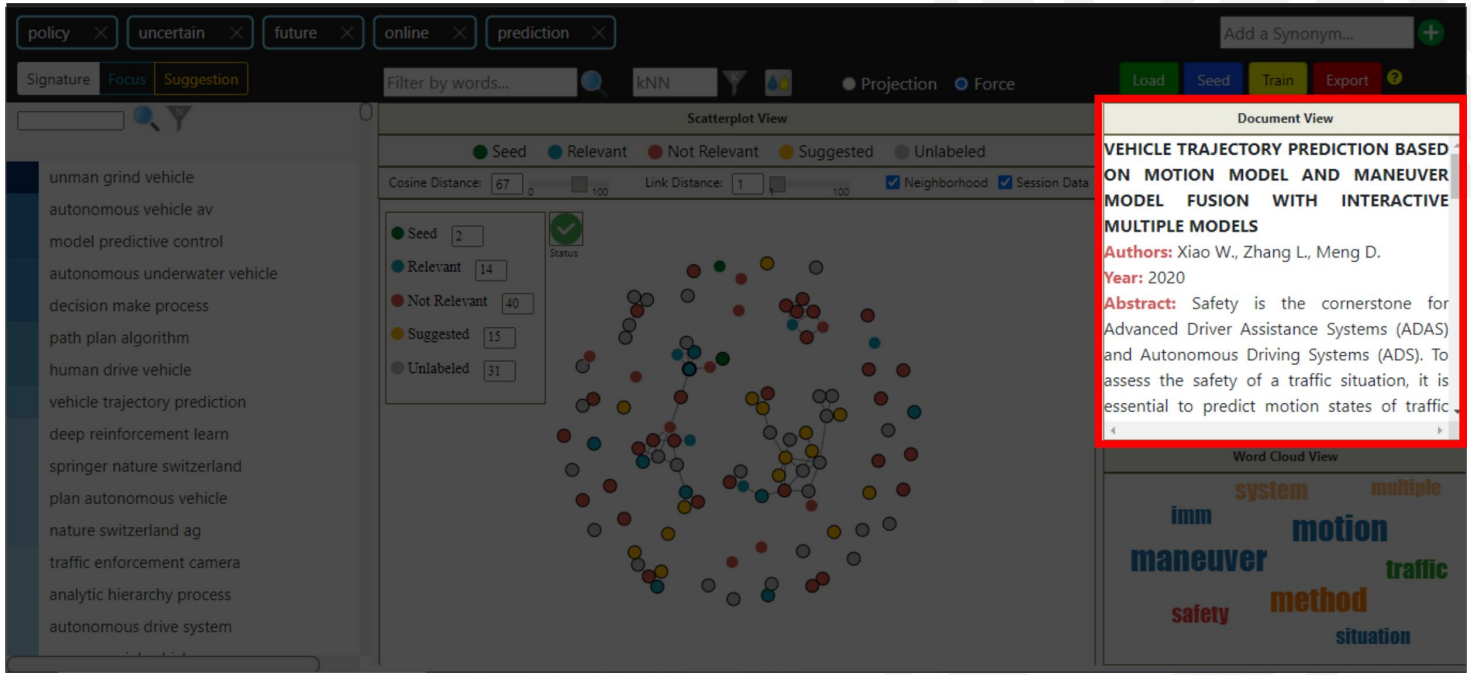
TRIVIR



Parameters

Session Data: Show how many documents per label, and the system status.

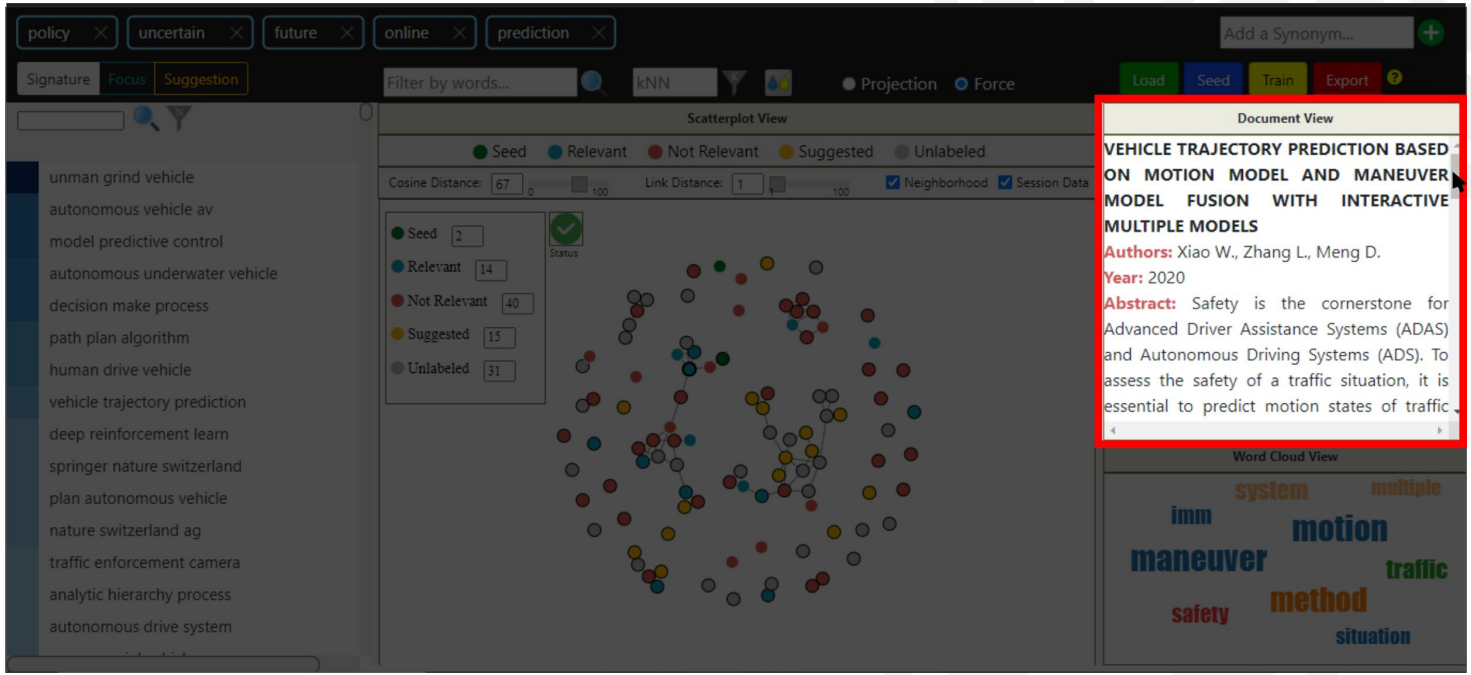
TRIVIR



Query document shown:

Clicking on a circle, the selected document's content is displayed in the Document view.

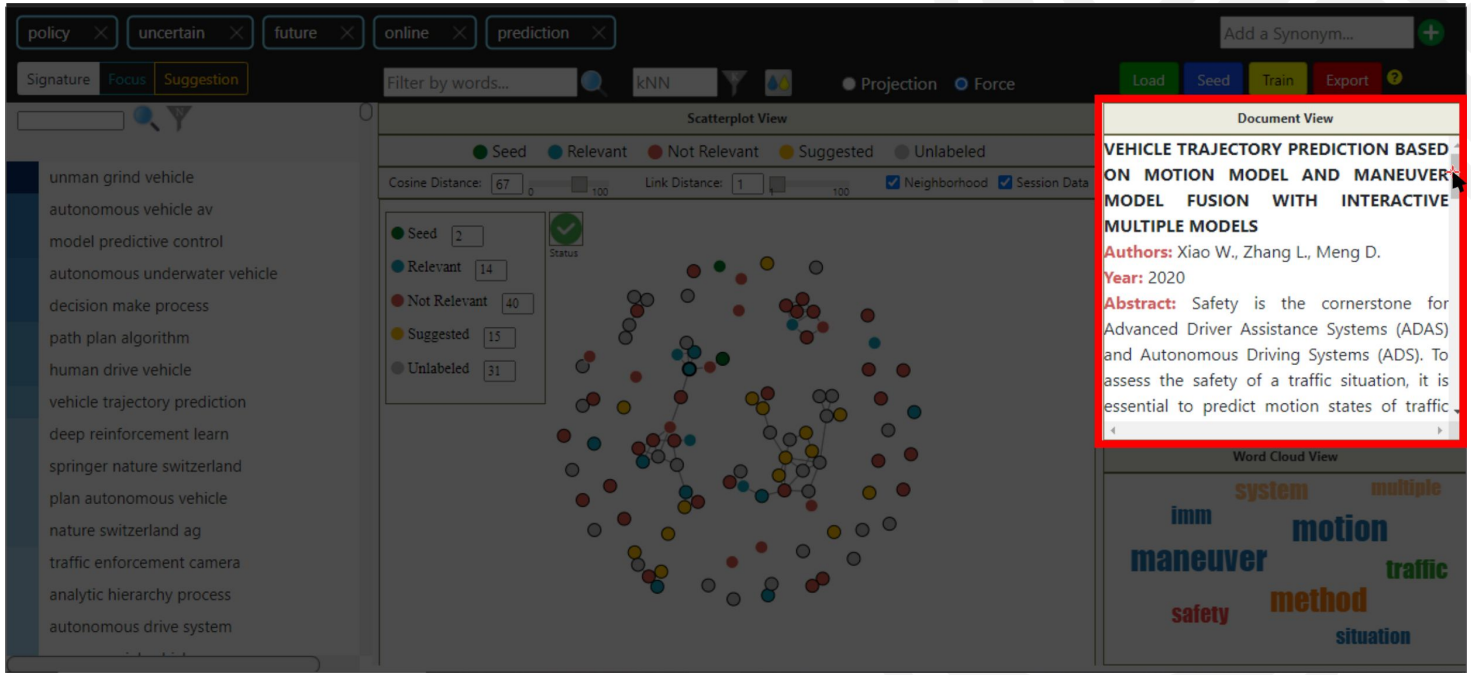
TRIVIR



Query document shown:

Clicking on a circle, the selected document's content is displayed in the Document view.

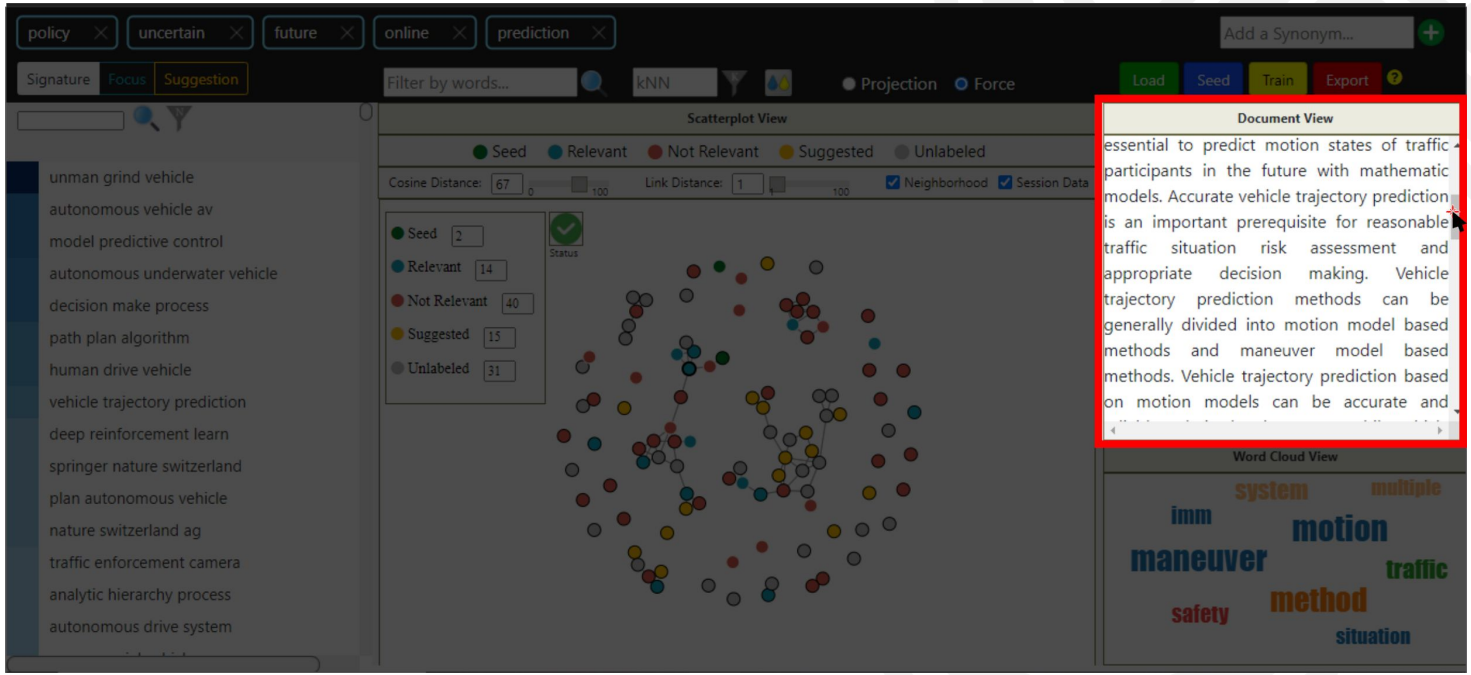
TRIVIR



Query document shown:

Clicking on a circle, the selected document's content is displayed in the Document view.

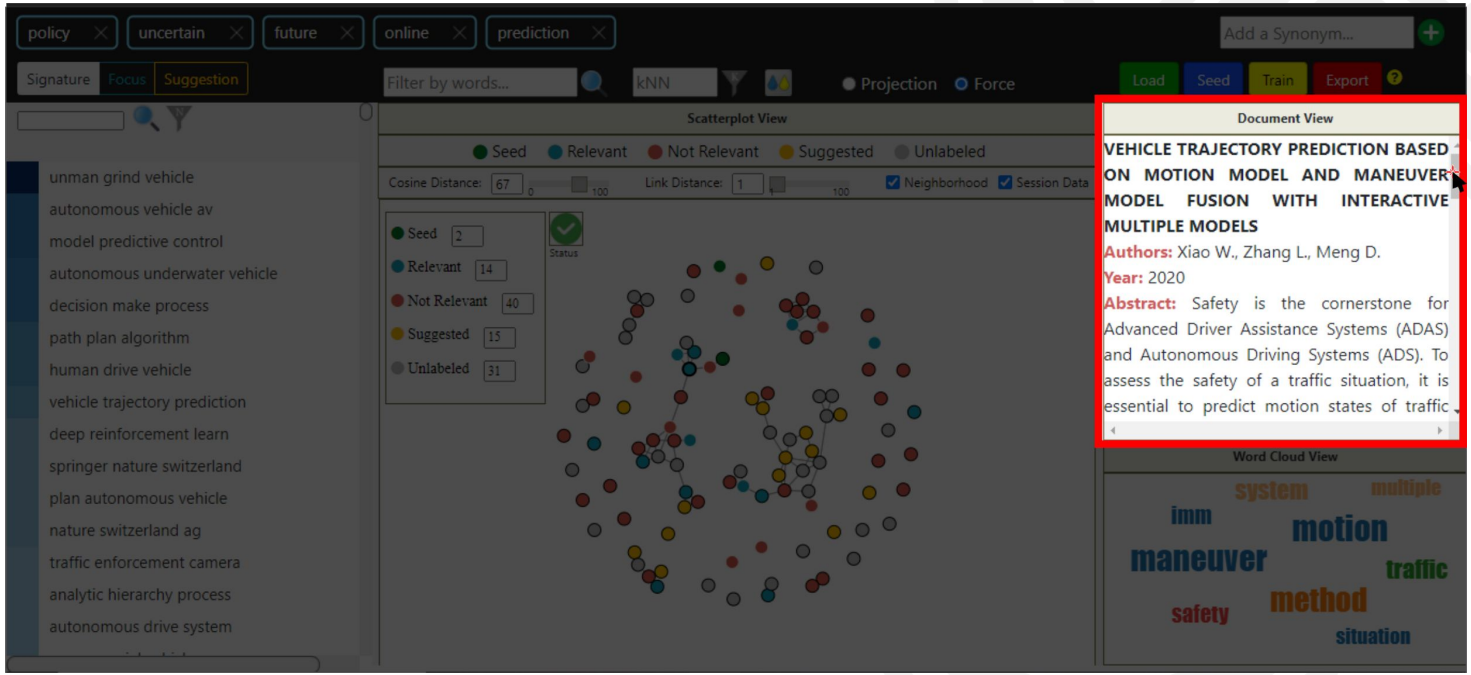
TRIVIR



Query document shown:

Clicking on a circle, the selected document's content is displayed in the Document view.

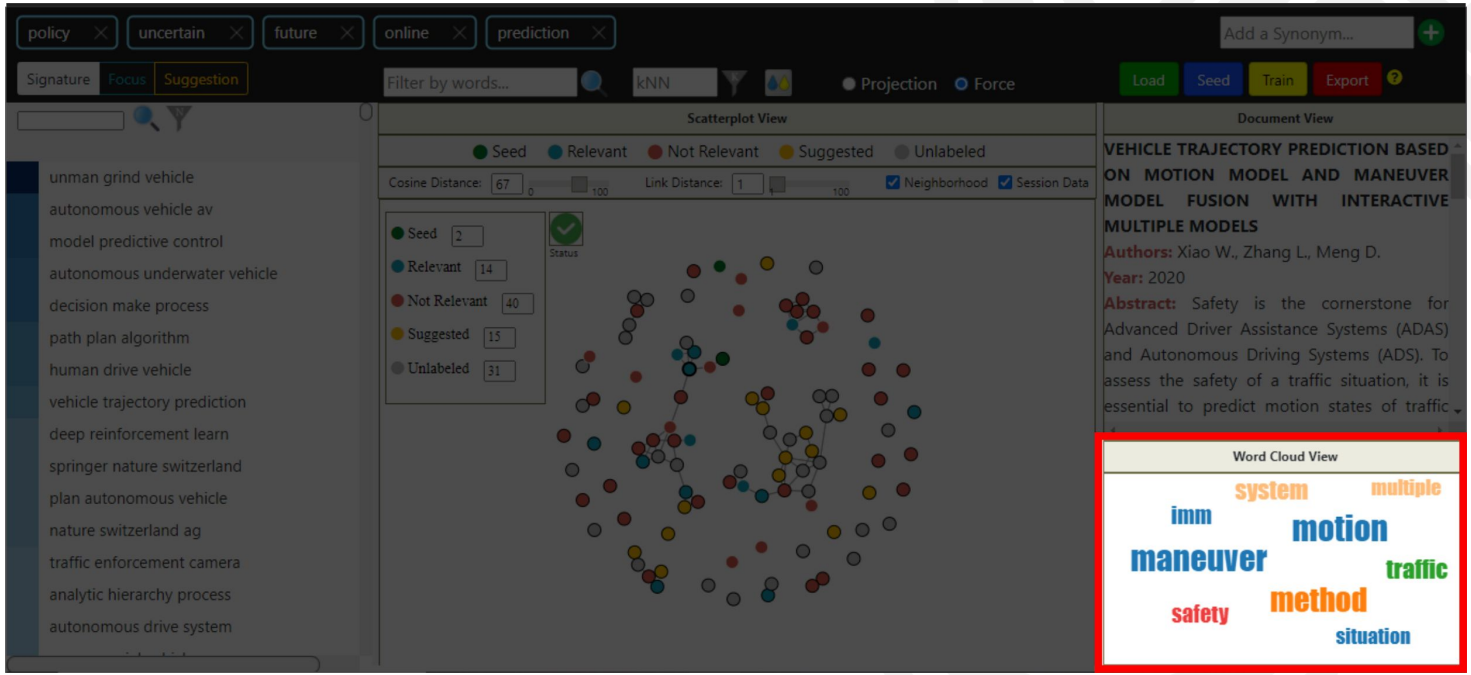
TRIVIR



Query document shown:

Clicking on a circle, the selected document's content is displayed in the Document view.

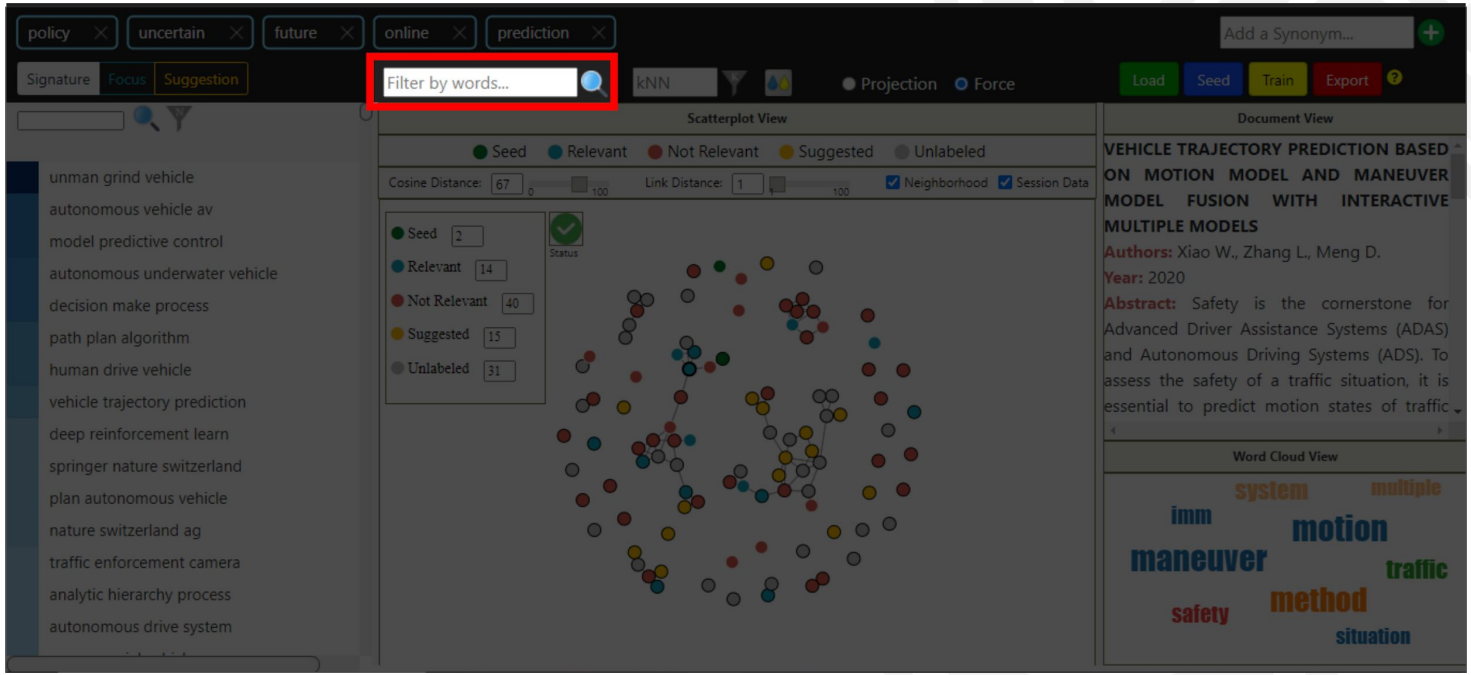
TRIVIR



Word Cloud View:

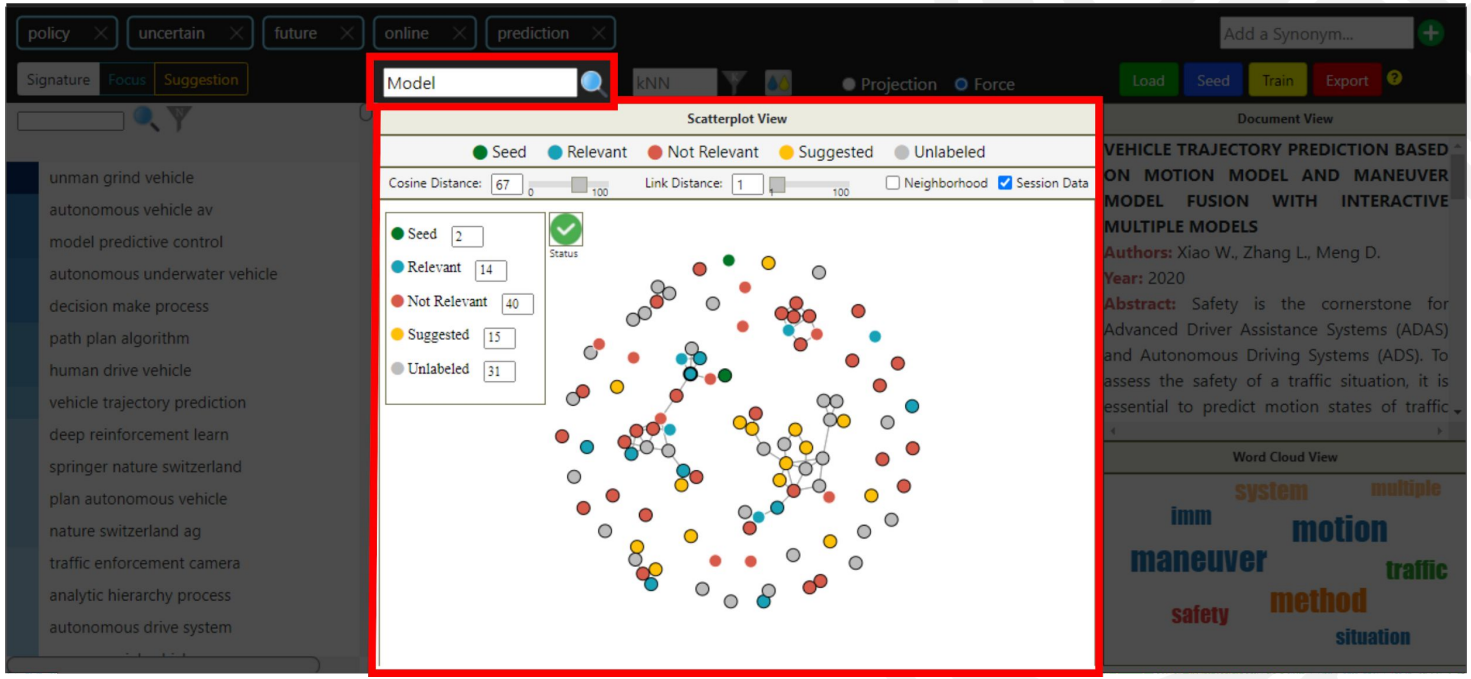
Show most frequent words in the selected document.

TRIVIR



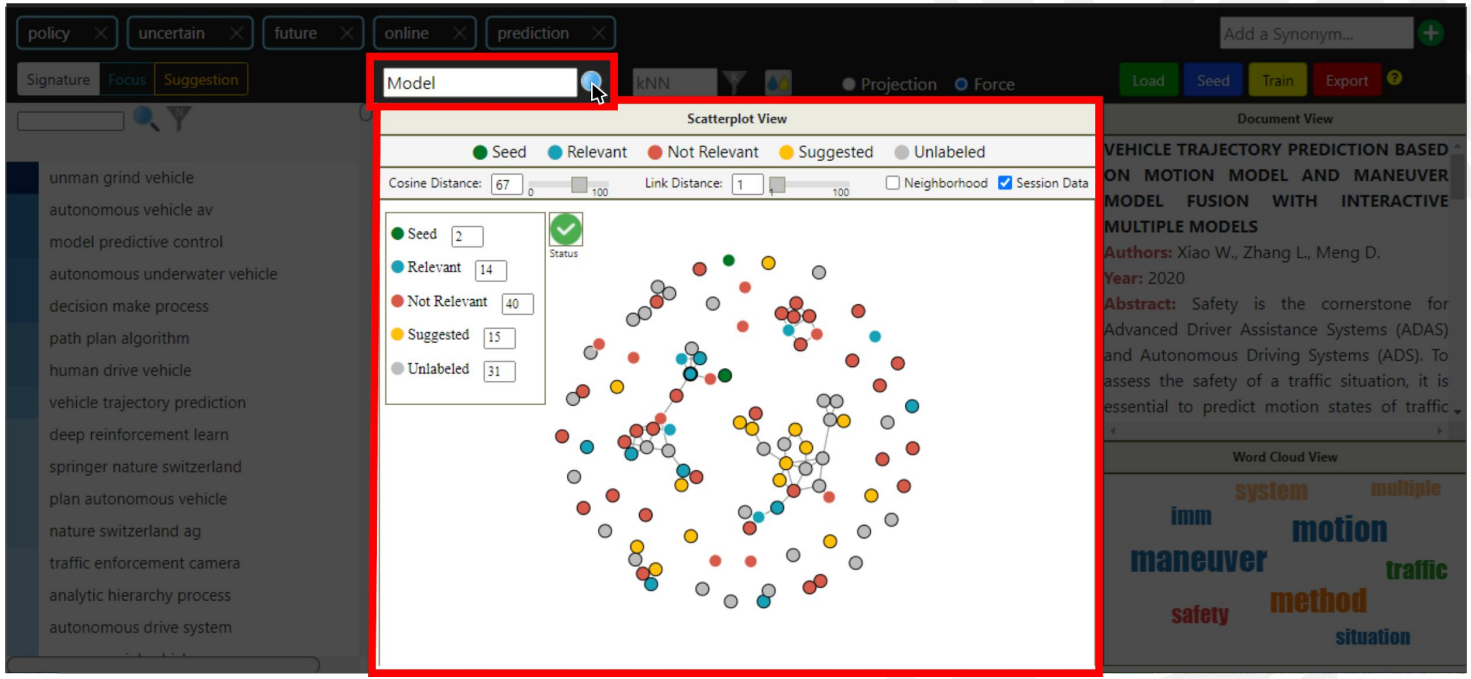
Show only documents that include a given term

TRIVIR



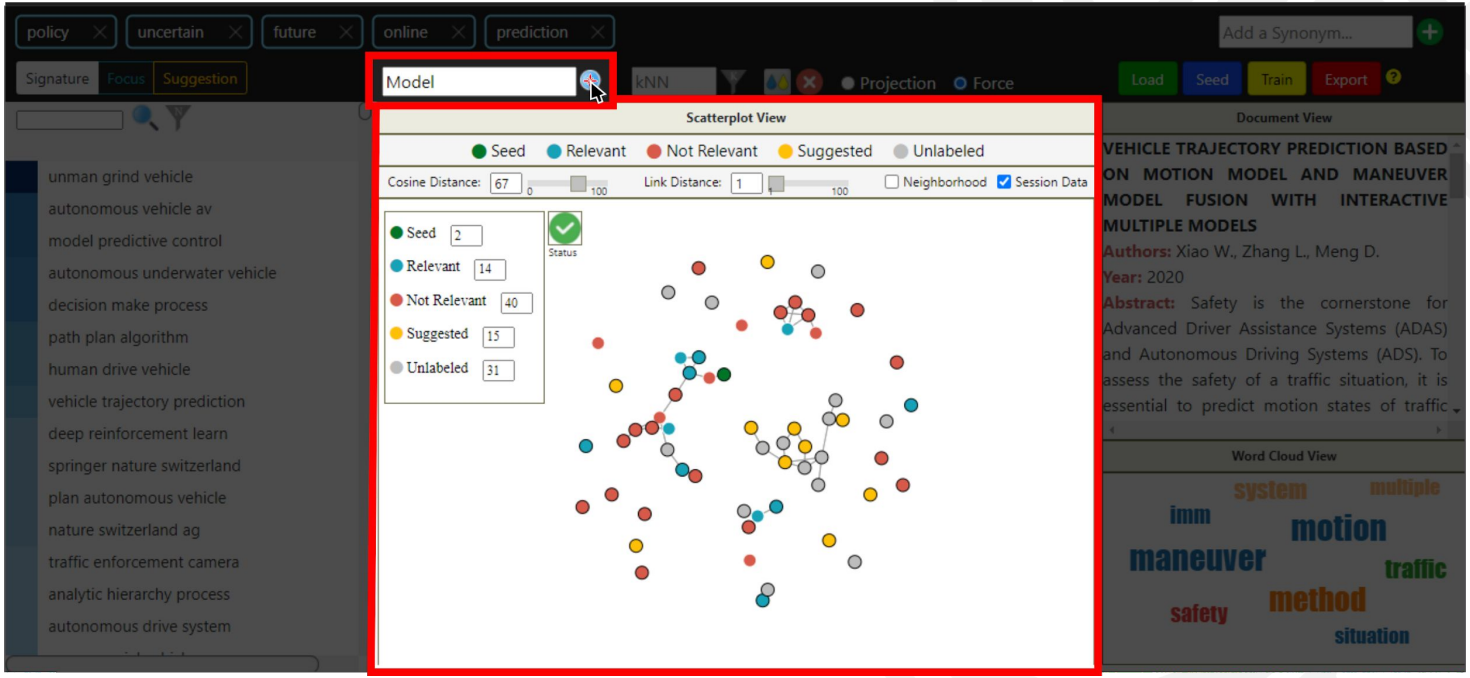
Show only documents that include a given term

TRIVIR



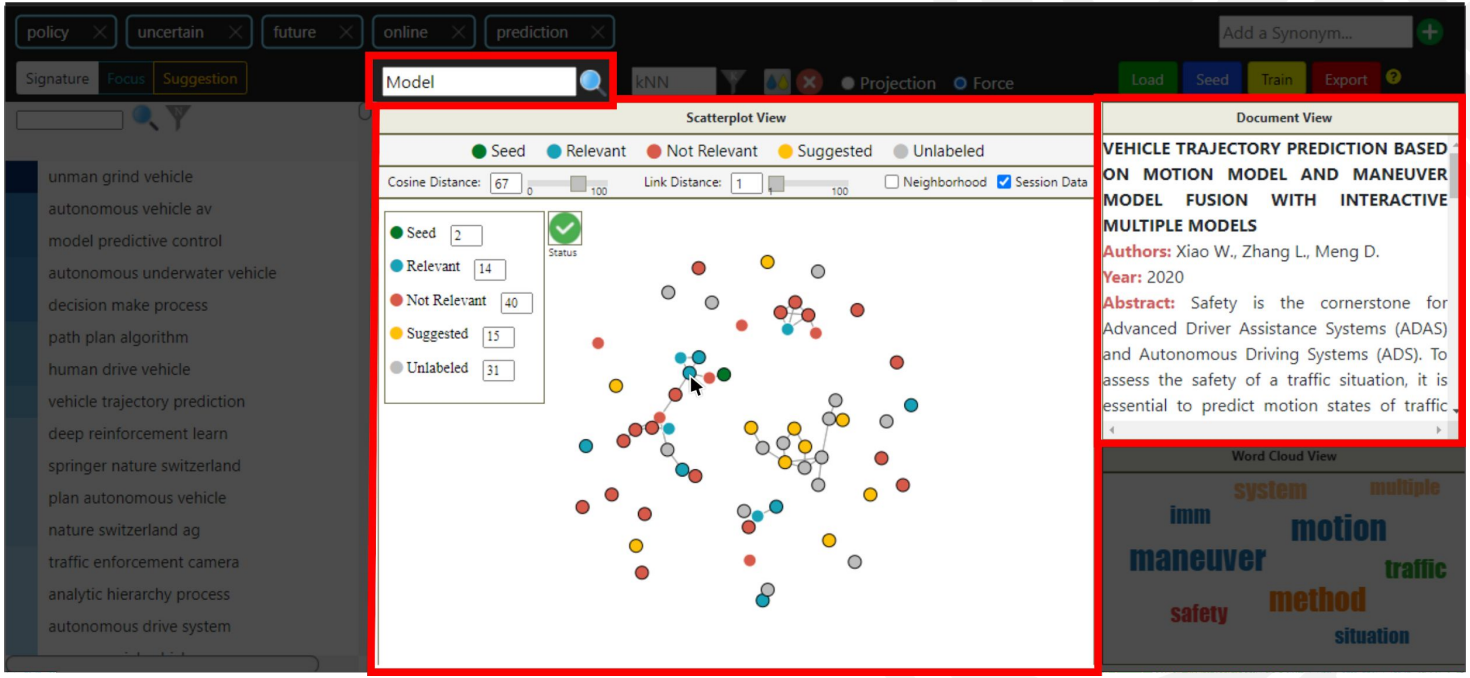
Show only documents that include a given term

TRIVIR



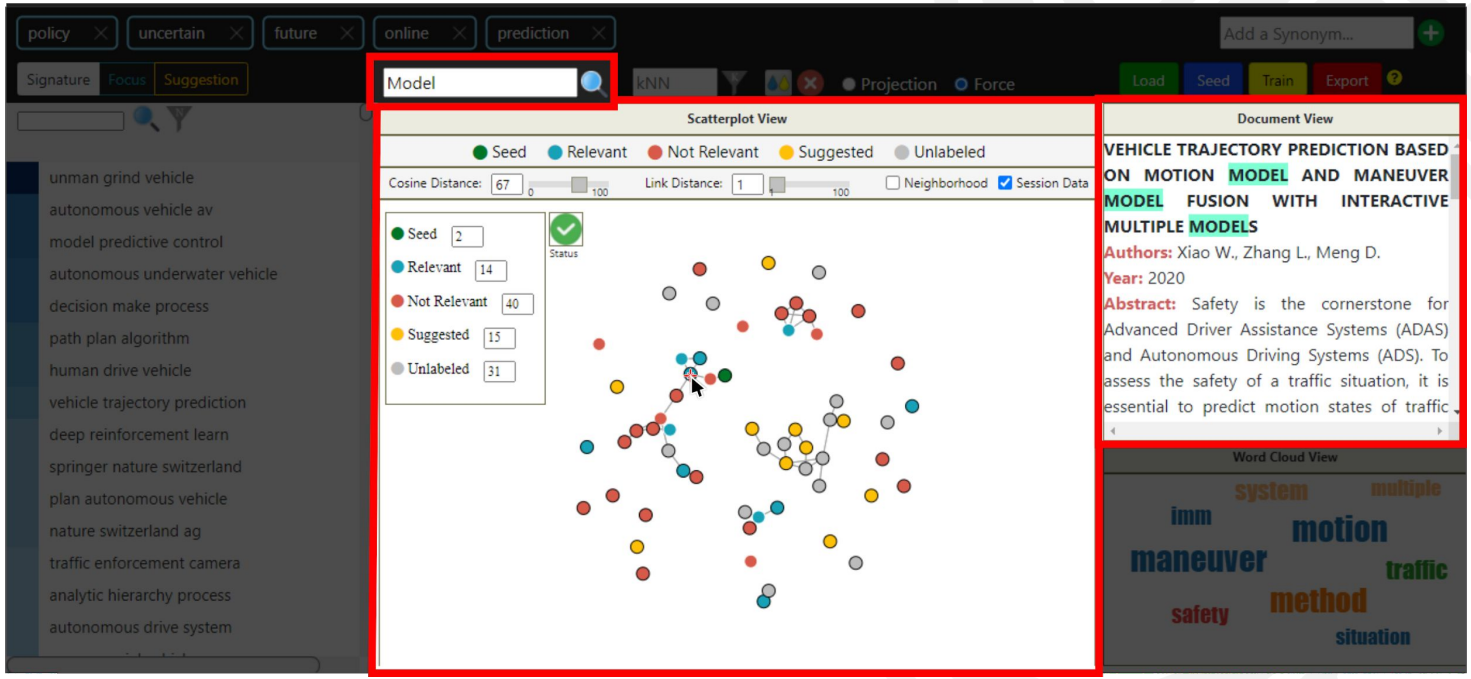
Show only documents that include a given term

TRIVIR



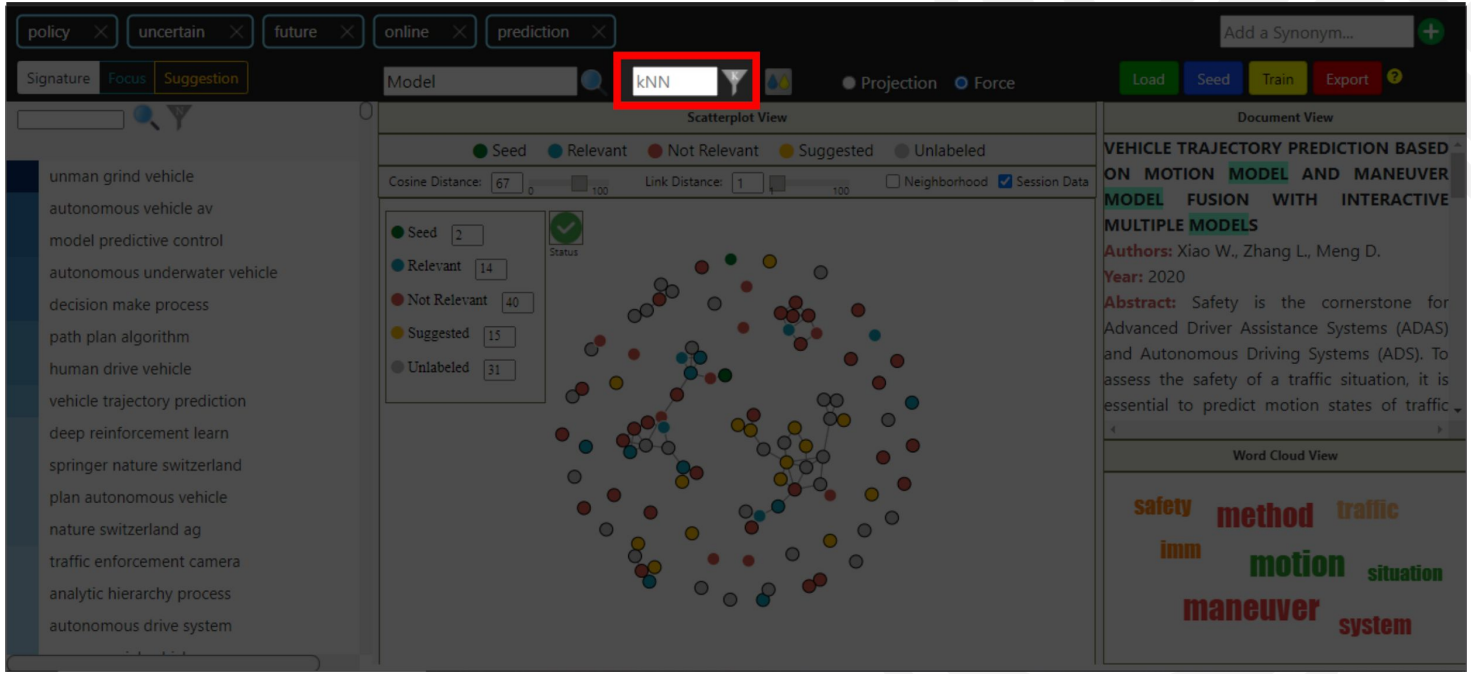
Show only documents that include a given term

TRIVIR



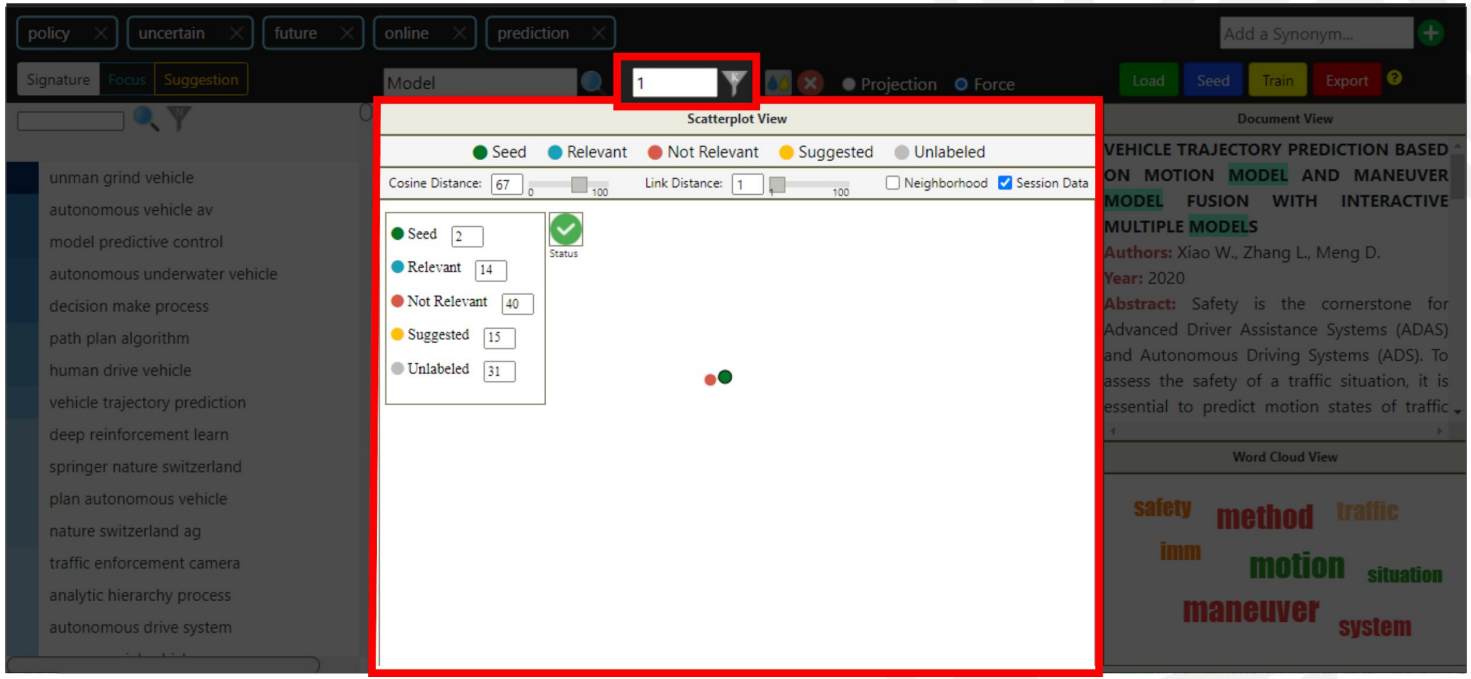
Highlight the given term in Document View

TRIVIR



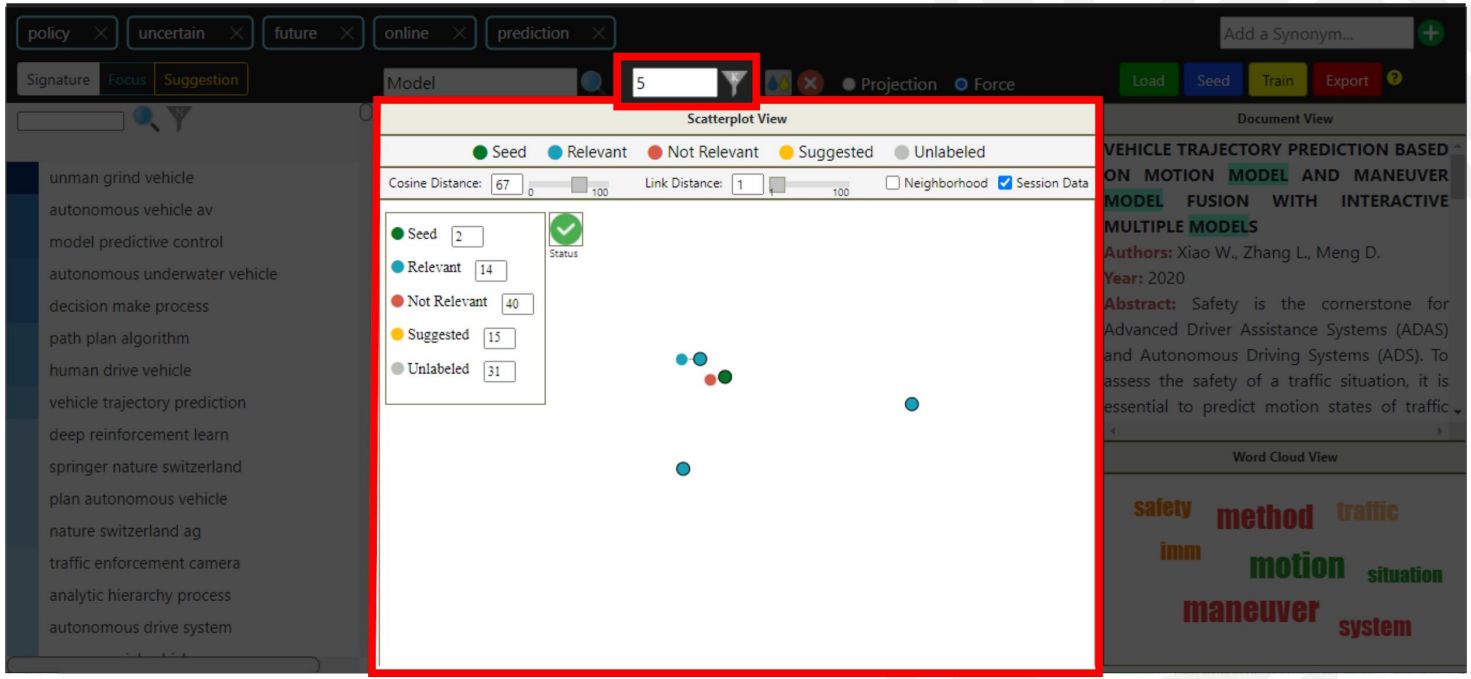
Filter the K documents most similar to the initial query document

TRIVIR



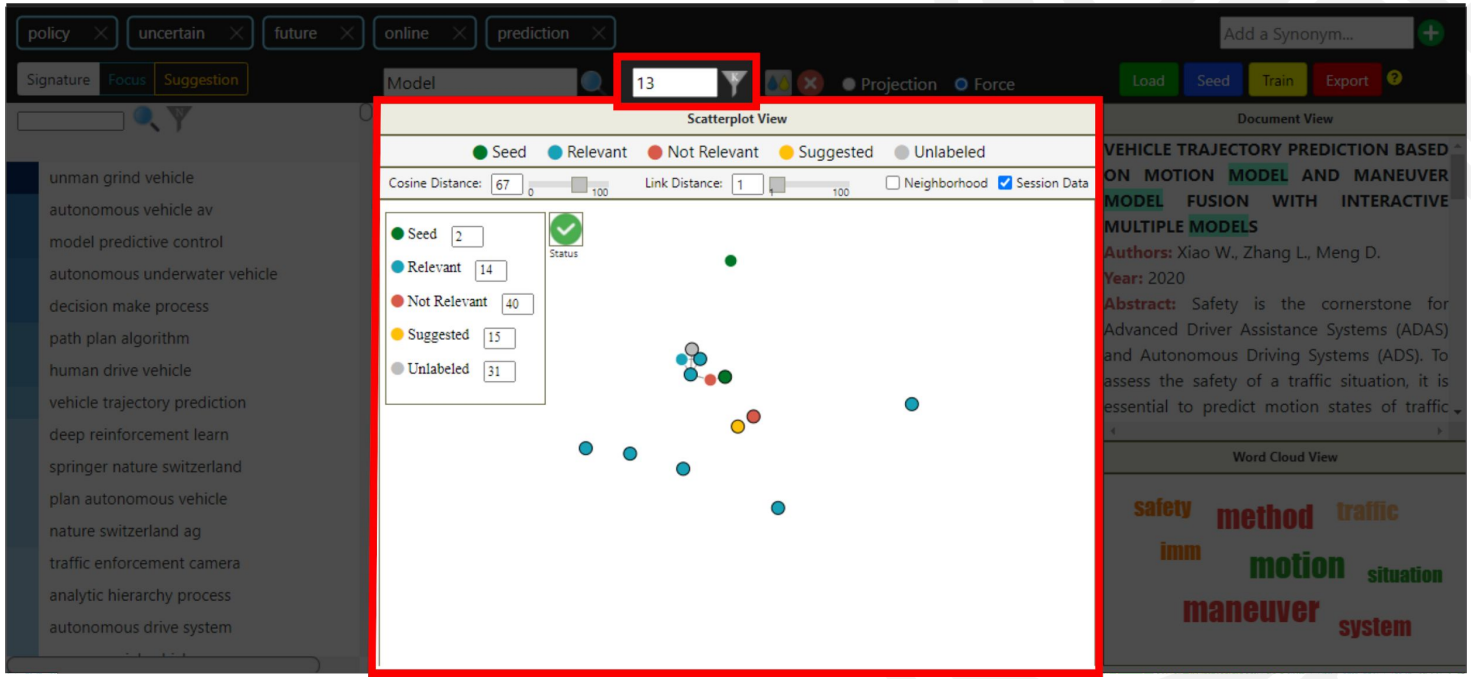
Filter the K = 1 documents most similar to the initial query document

TRIVIR



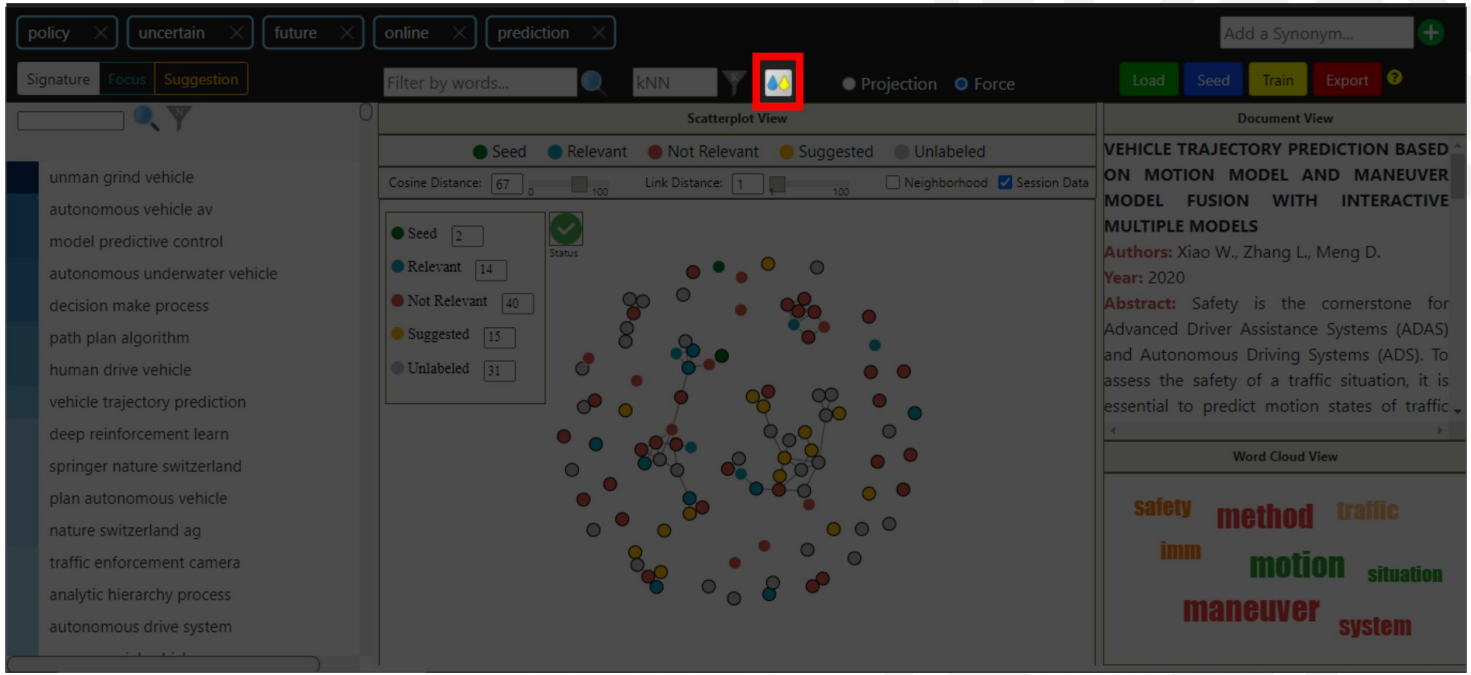
Filter the K = 5 documents most similar to the initial query document

TRIVIR



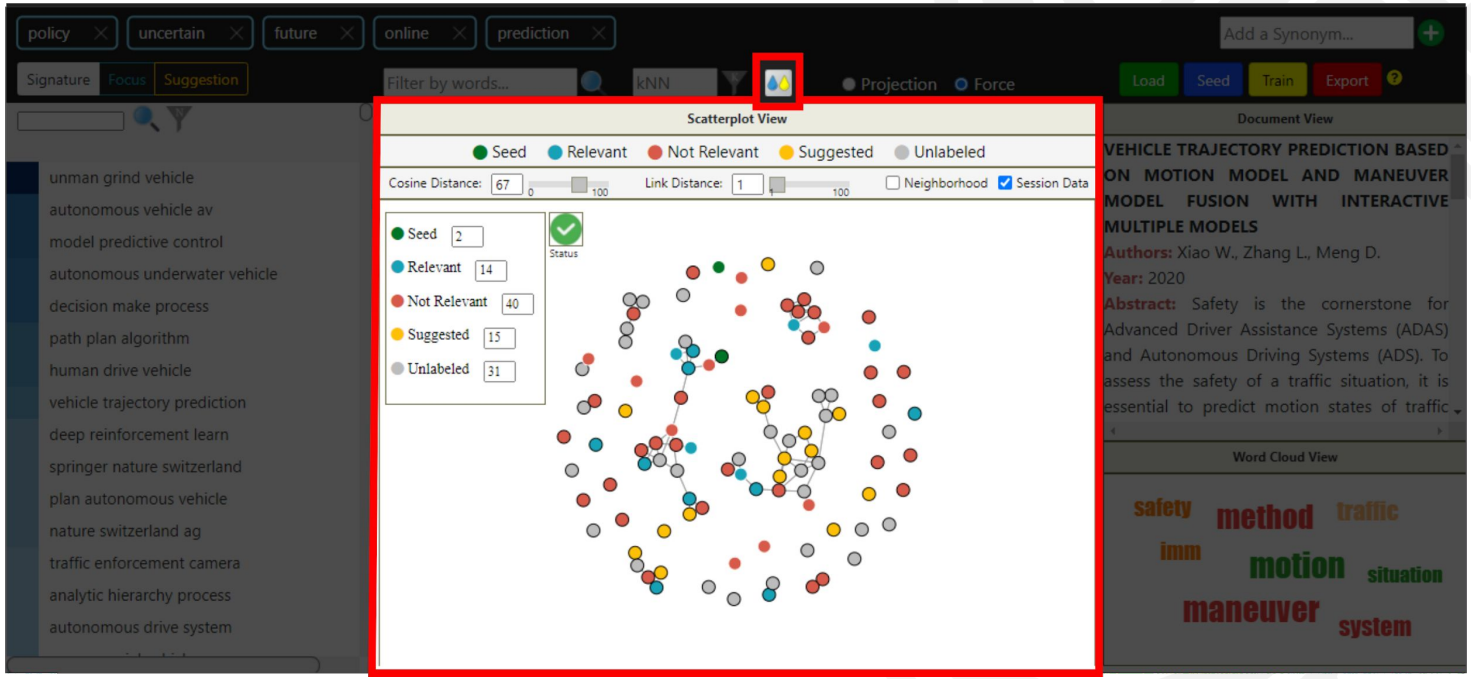
Filter the K = 13 documents most similar to the initial query document

TRIVIR



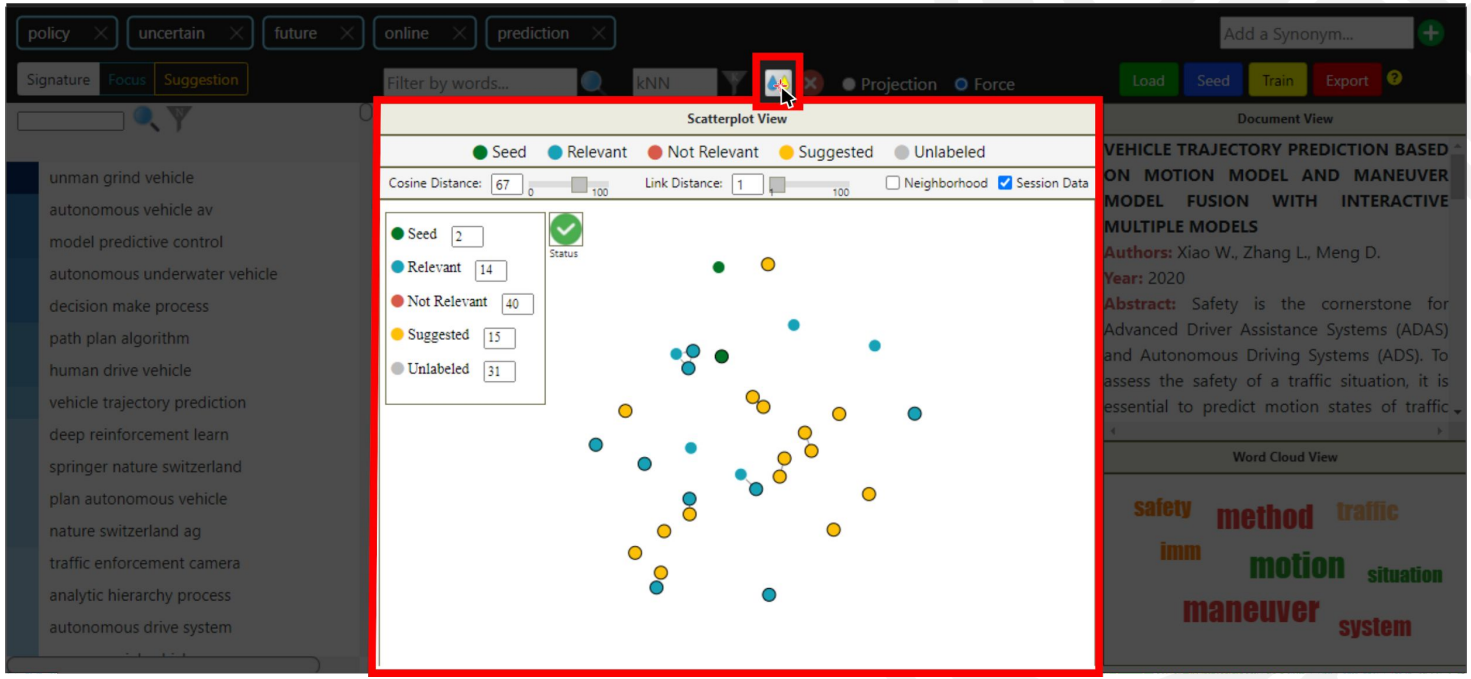
Reduce point clutter by displaying only the seed, relevant and suggested documents

TRIVIR



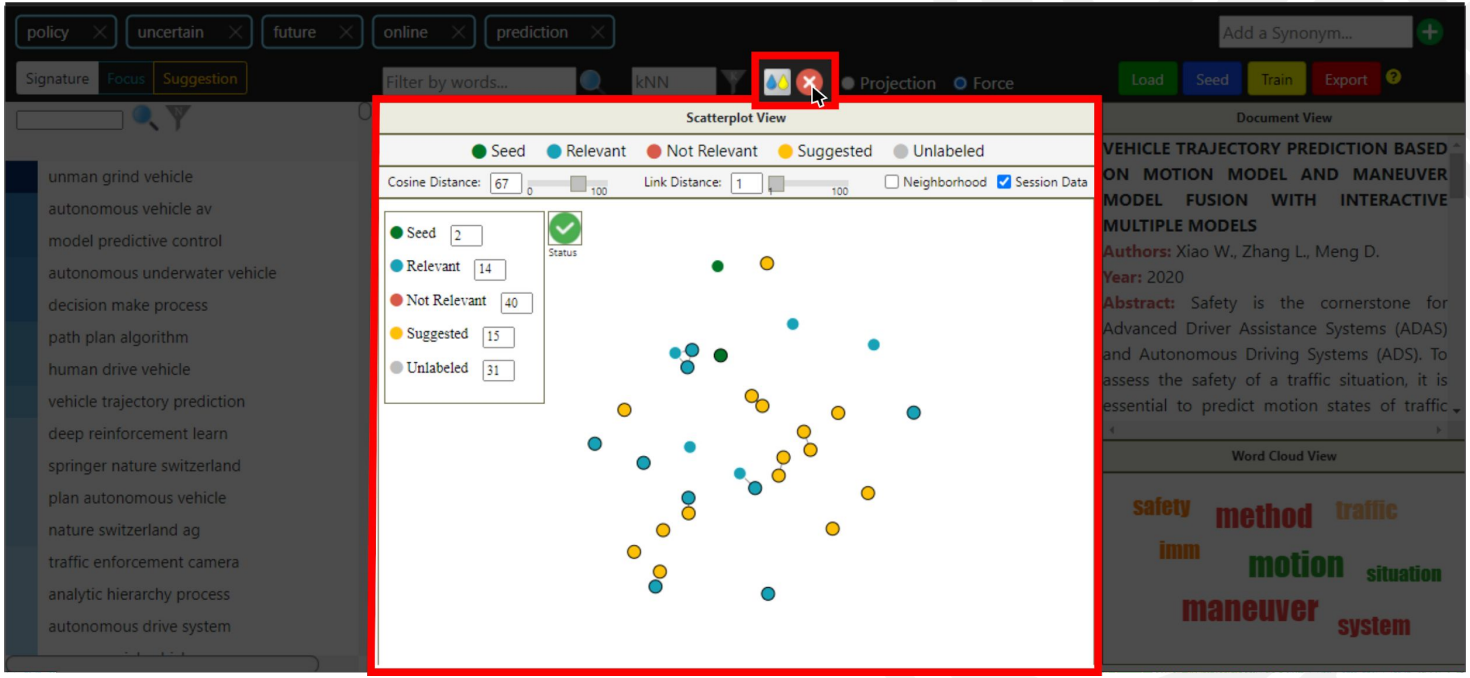
Reduce point clutter by displaying only the seed, relevant and suggested documents

TRIVIR



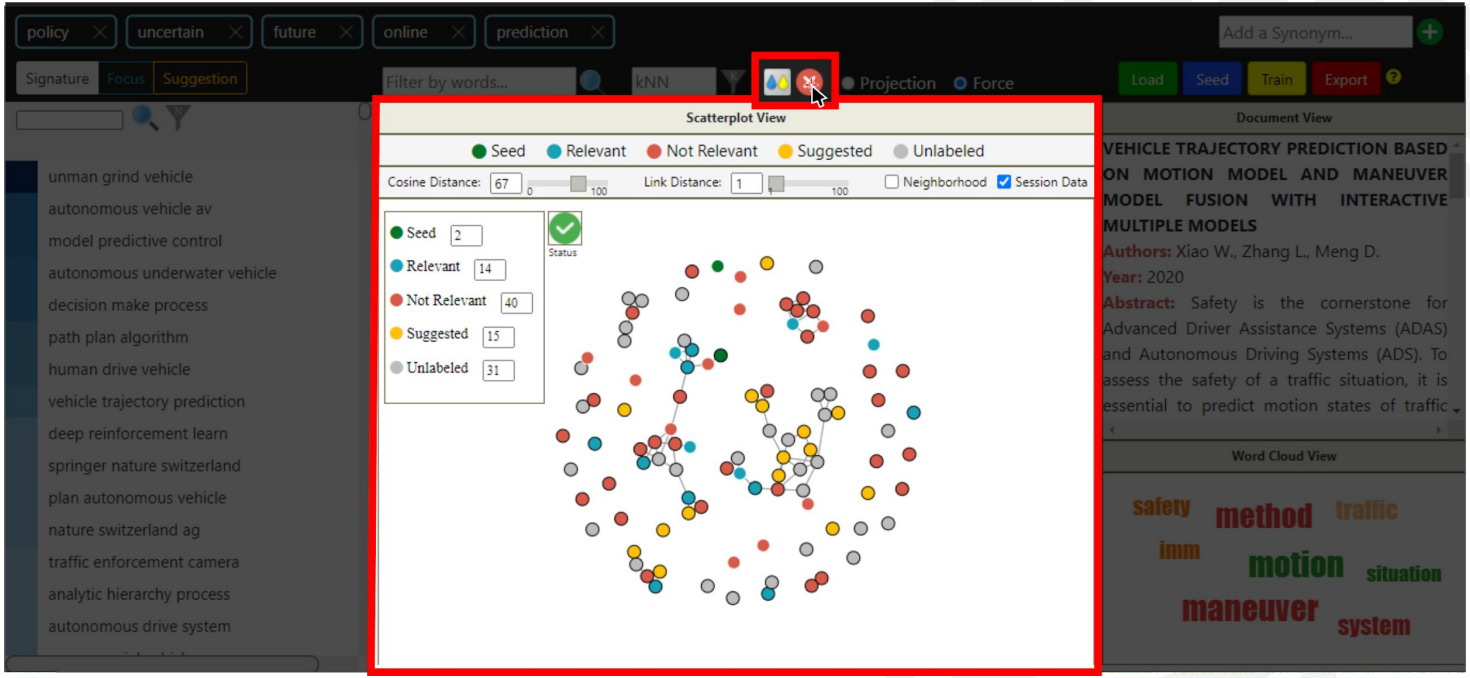
Reduce point clutter by displaying only the seed, relevant and suggested documents

TRIVIR



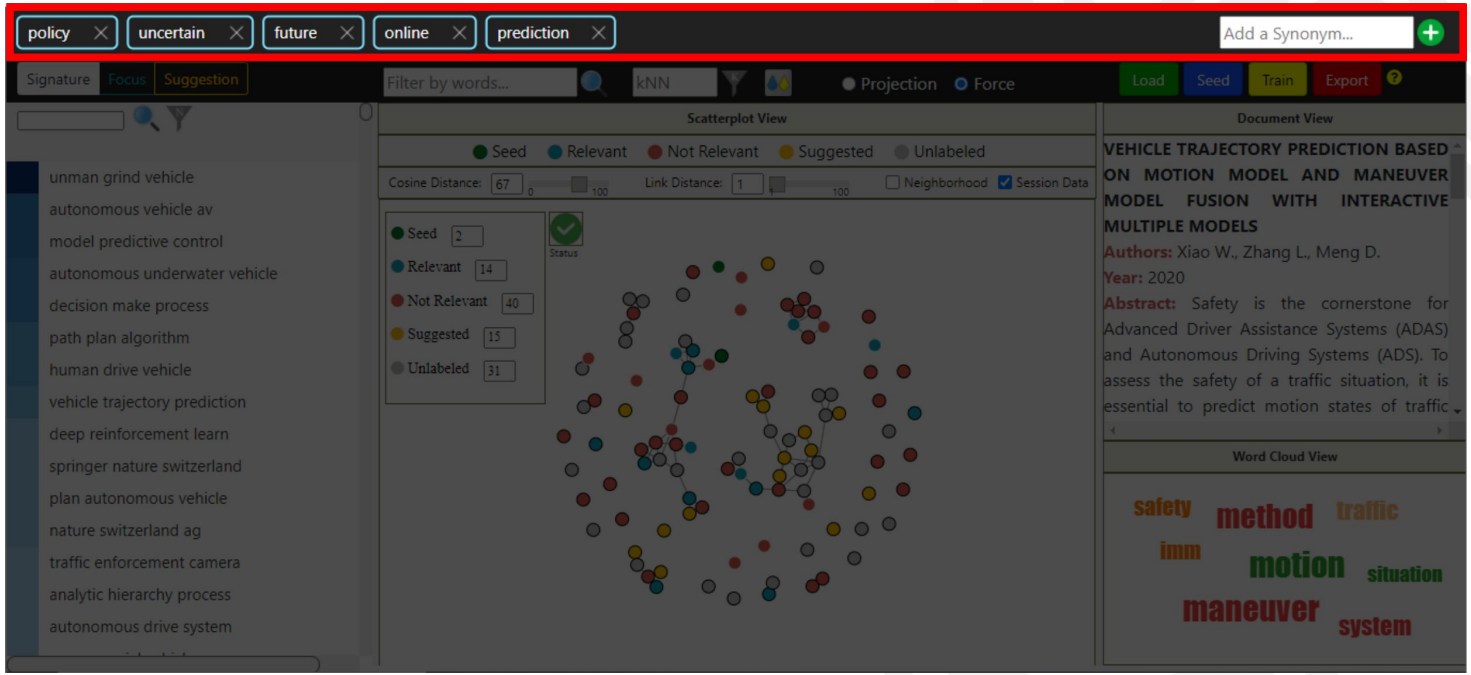
Reduce point clutter by displaying only the seed, relevant and suggested documents

TRIVIR



Reduce point clutter by displaying only the seed, relevant and suggested documents

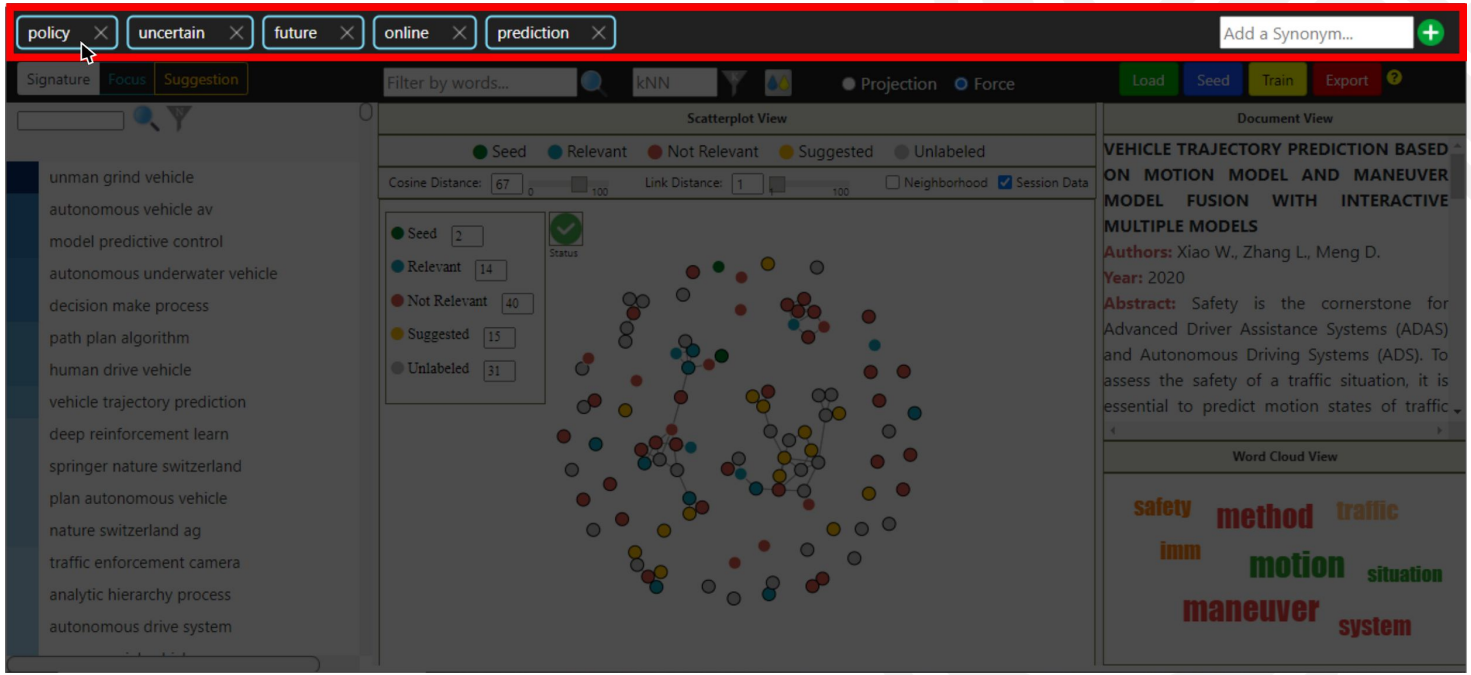
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

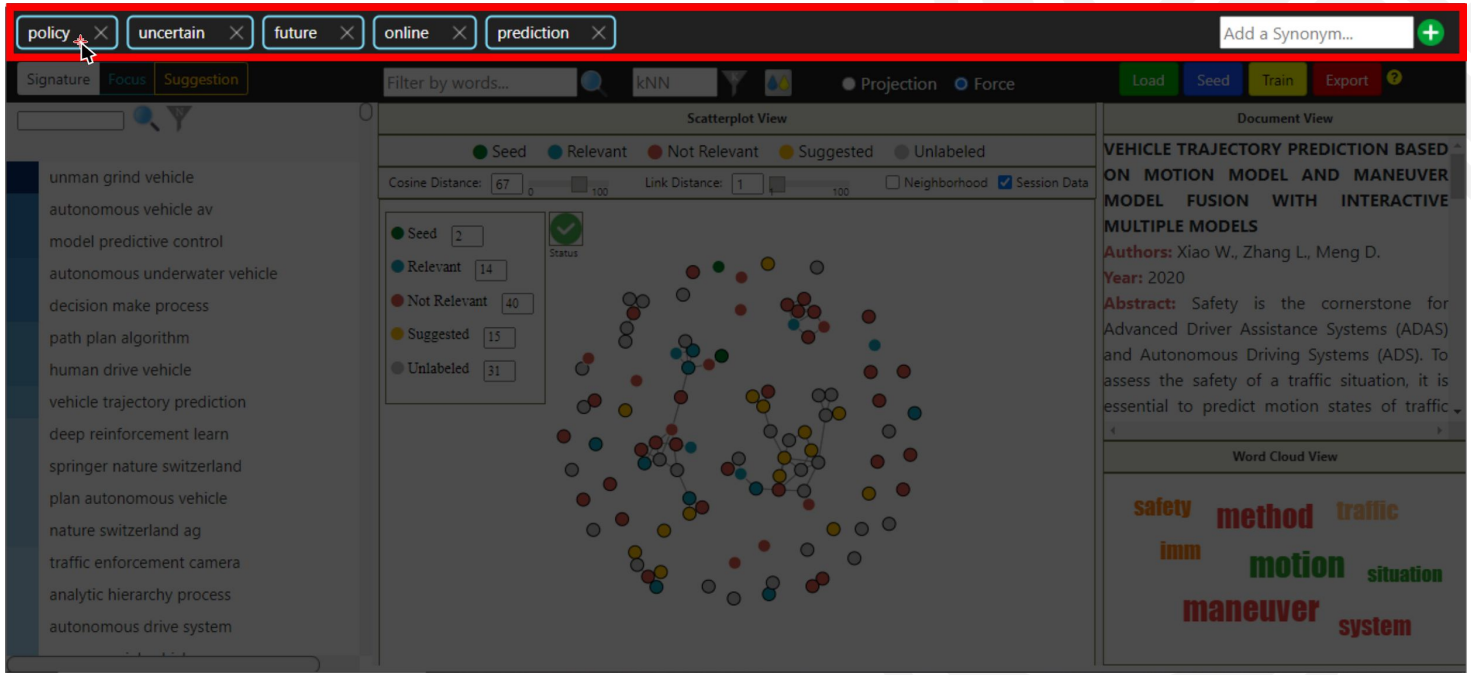
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

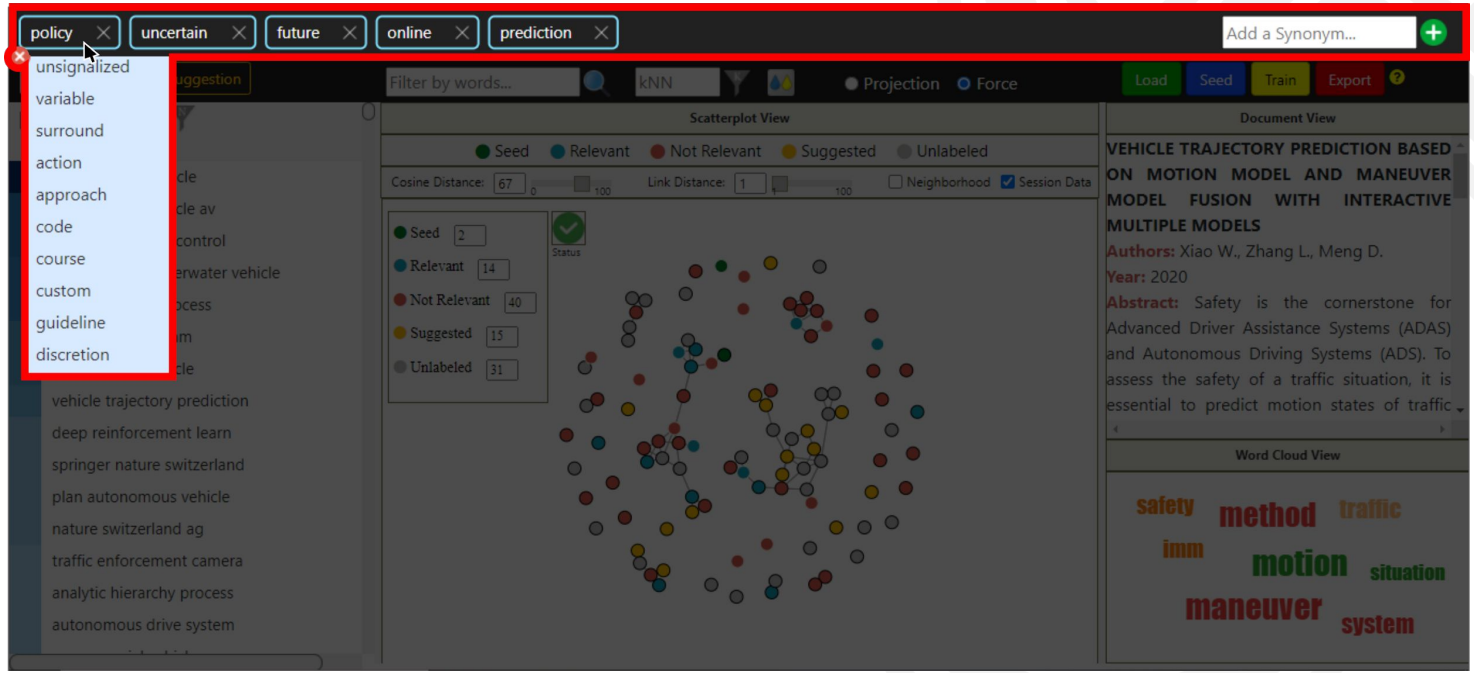
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

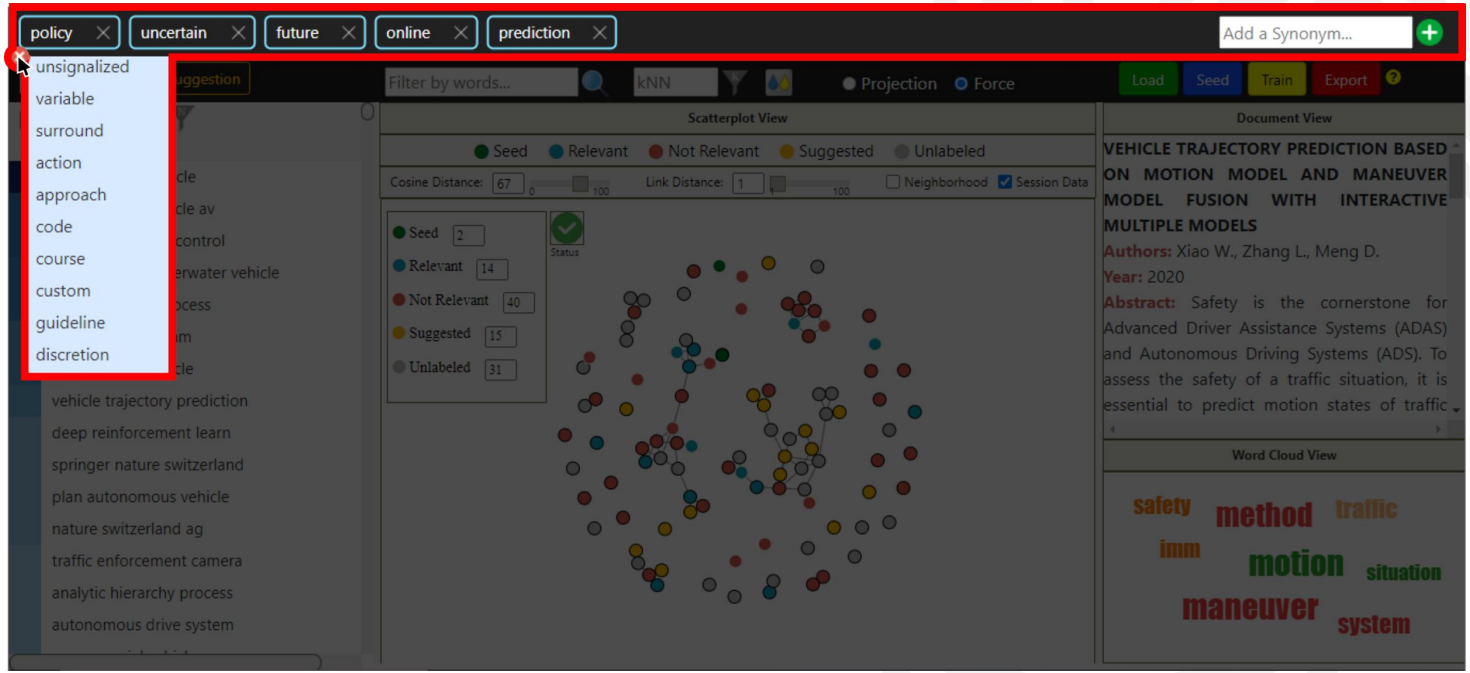
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

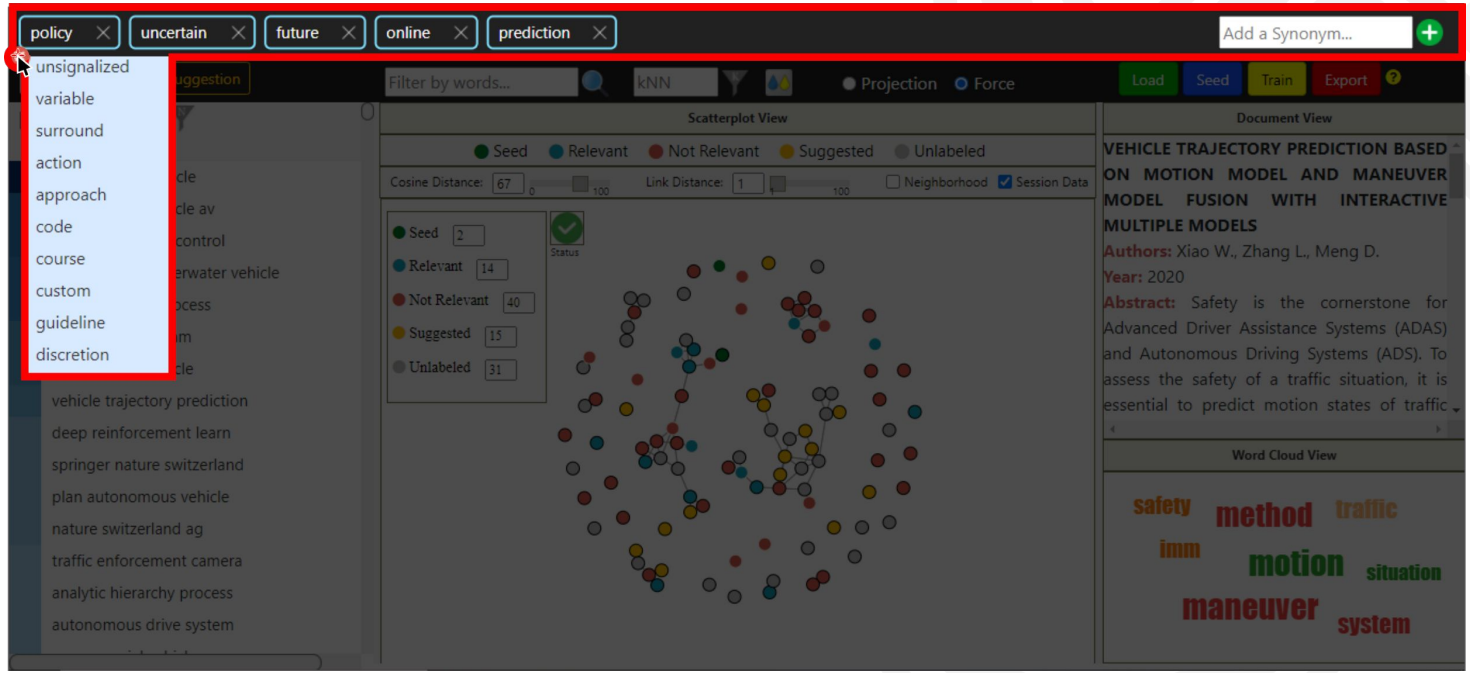
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

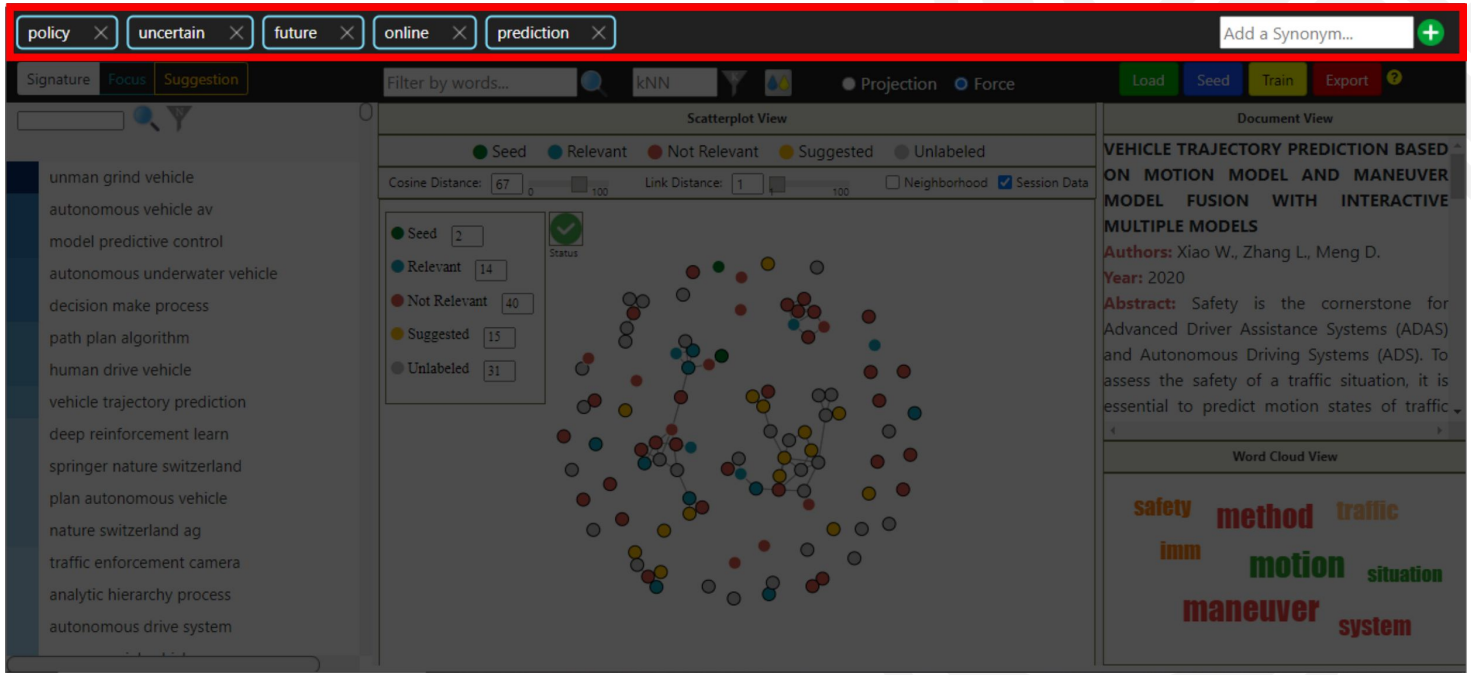
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

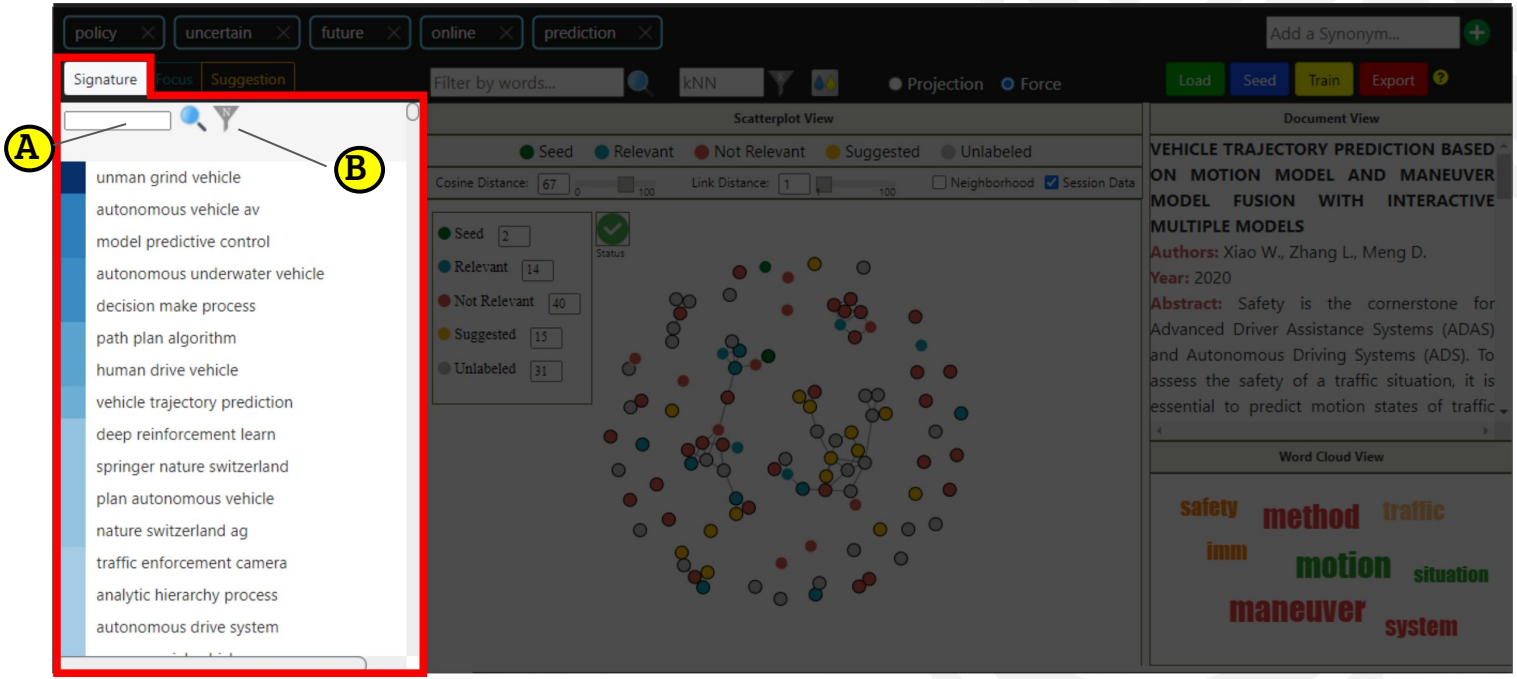
TRIVIR



The Terms view:

Show the five most frequent terms in the initial query document.
It is possible to remove or to add terms, and also find synonyms.

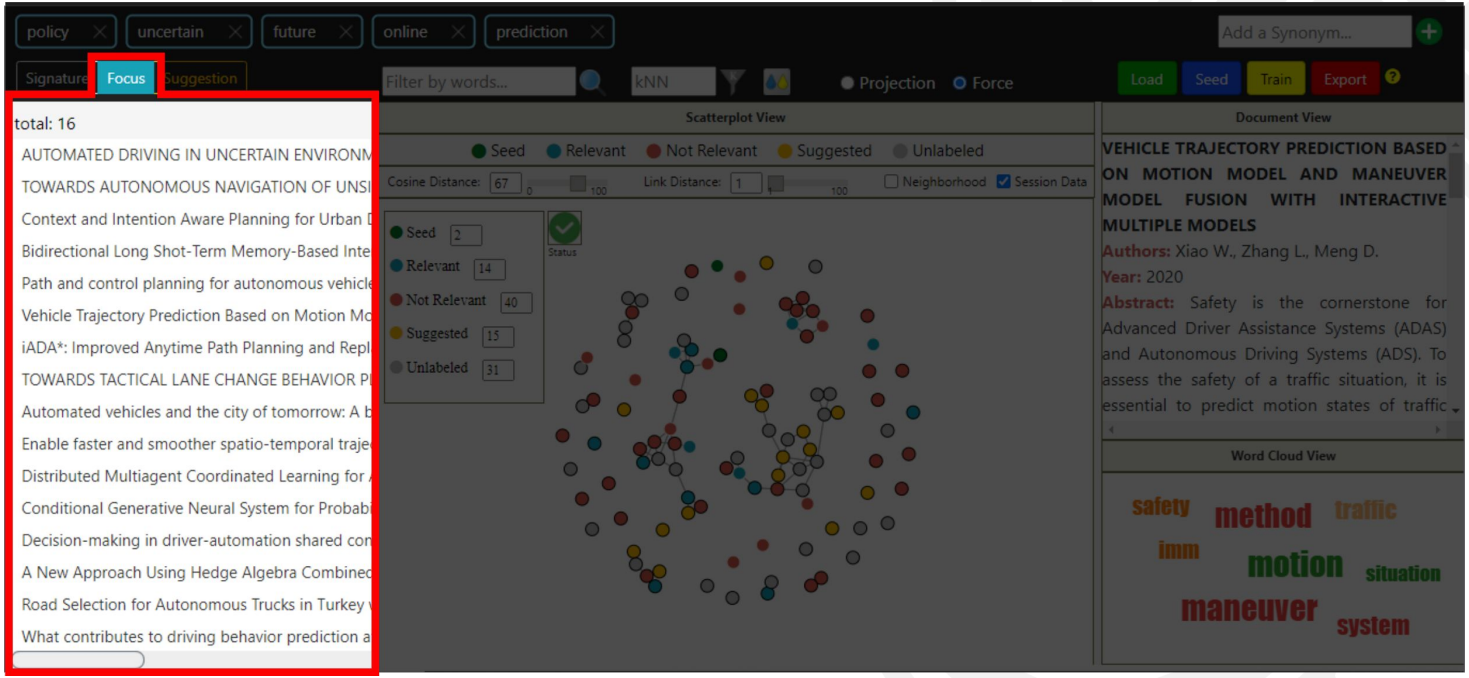
TRIVIR



Signature List view:

Show a rank of relevant 3-grams in the corpus. The user can filter 3-grams with a particular term (A) or with terms from the Terms view (B).

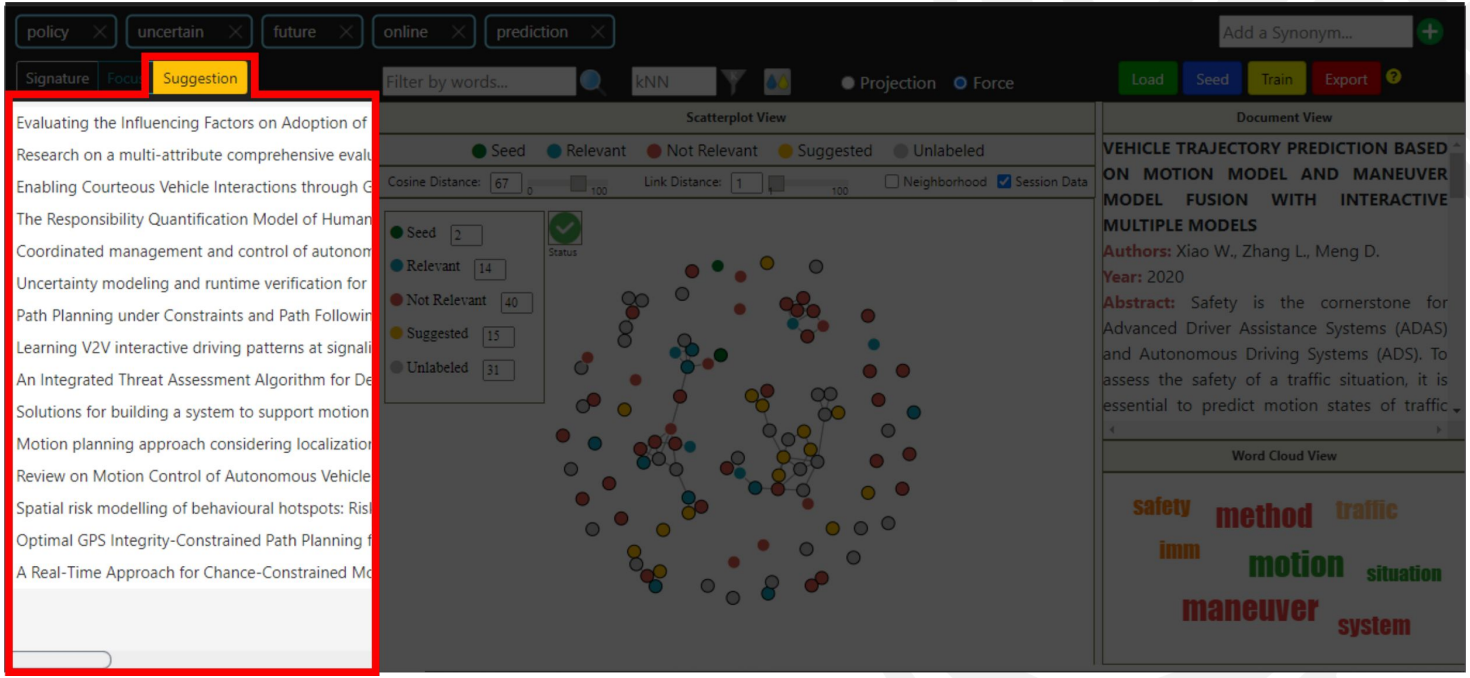
TRIVIR



Focus List view:

List all documents currently labeled as seed or relevant.

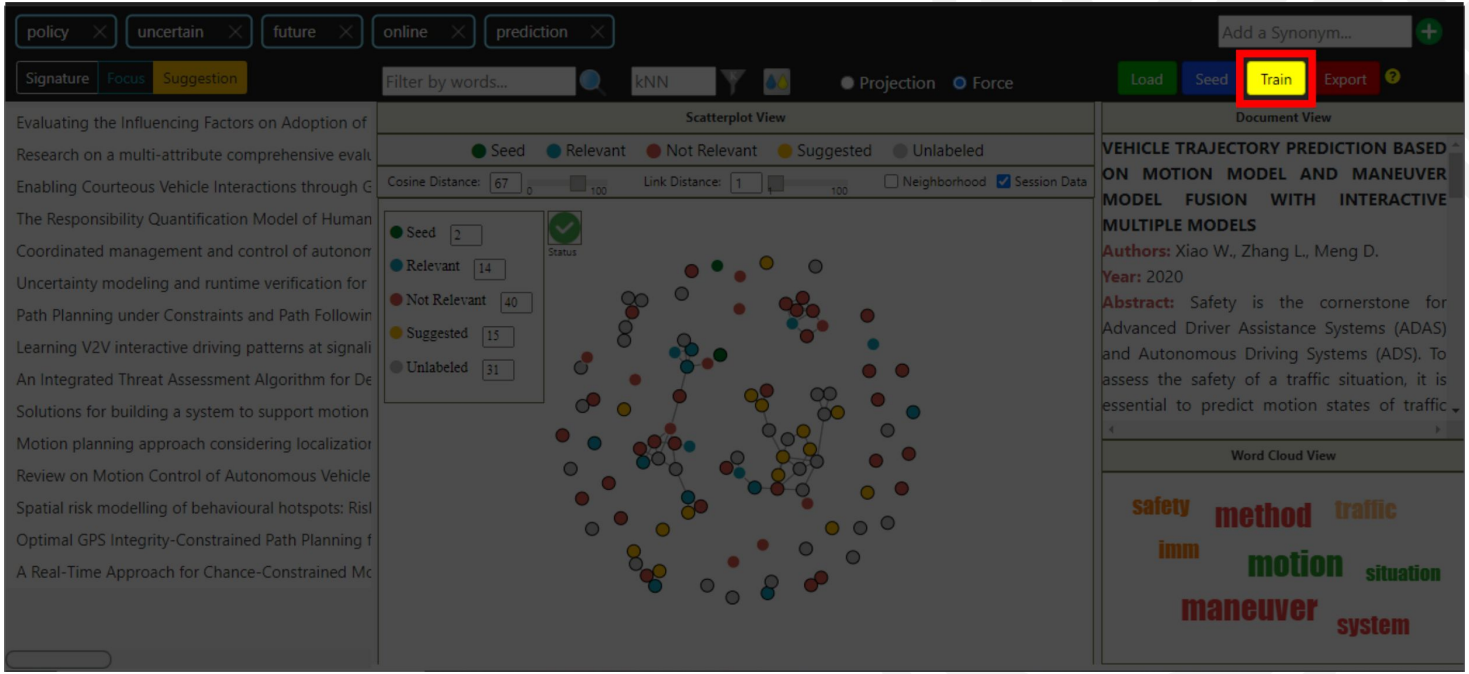
TRIVIR



Suggestion List view:

List the documents currently suggested as relevant by the Machine Learning classifier.

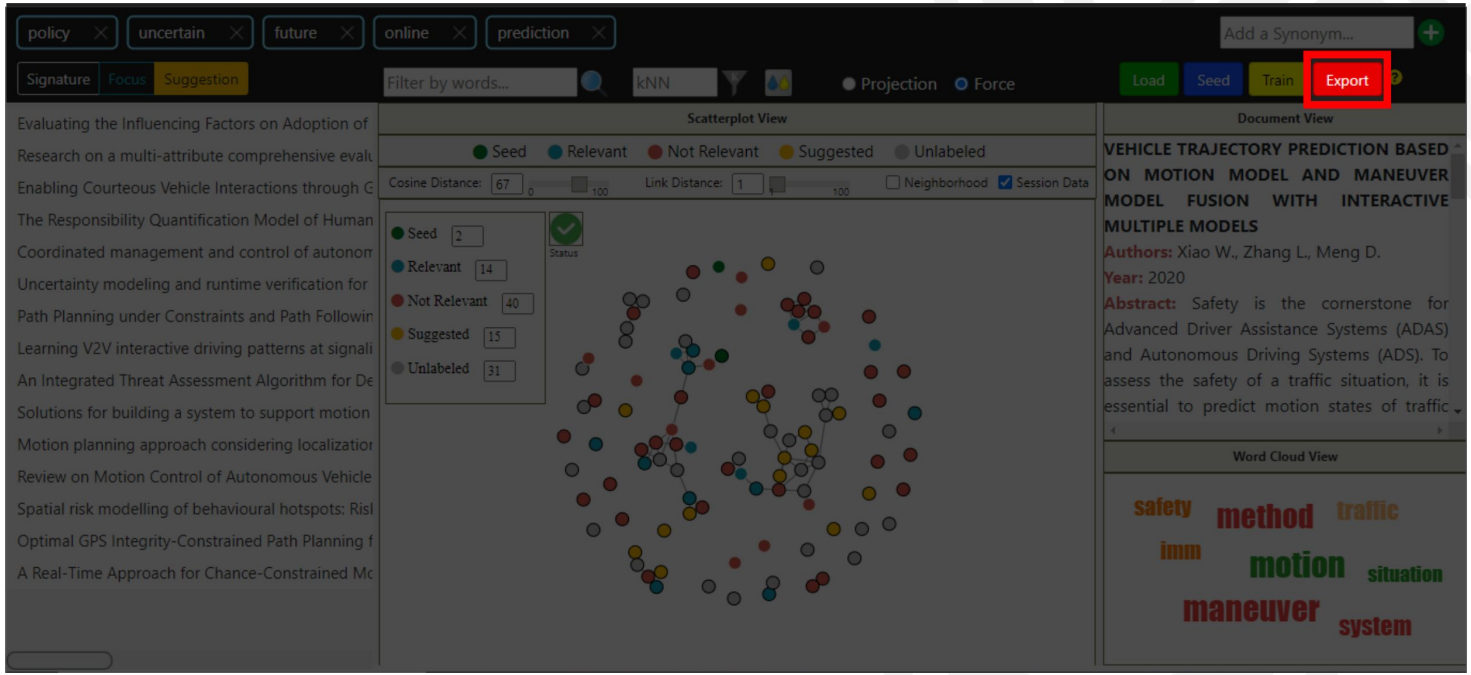
TRIVIR



Train Button

Retrain the ML algorithm considering previous user feedback.
This step can improve the recommendation of documents.

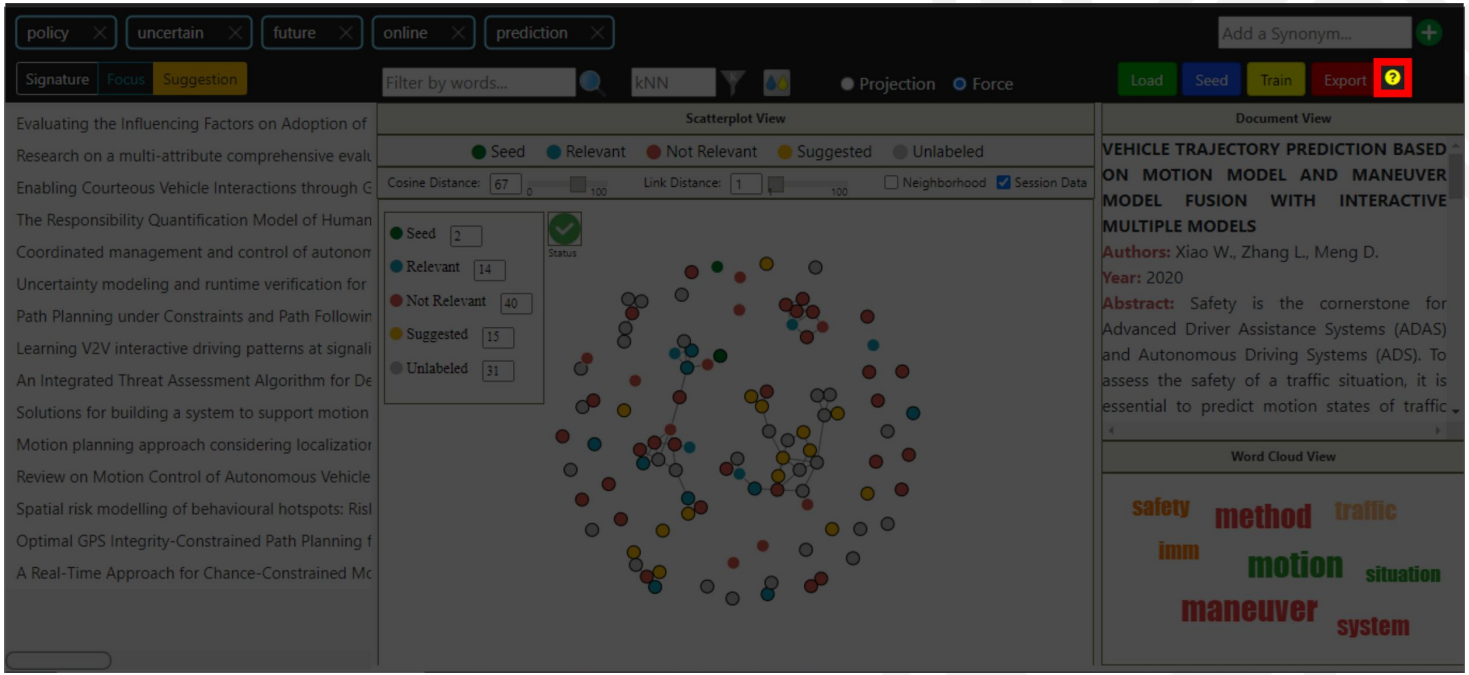
TRIVIR



Export Button

Download from the server a file containing all session data for further exploration.

TRIVIR



Tips Button (Help)

The tips button shows a Guided Tour through the system functionalities.

User Study

Hands On!

User Study

Thinking aloud:

- Describe your thoughts as you go along

Make notes:

- Strengths
- Weaknesses
- General Considerations