POLYTECH NANTES

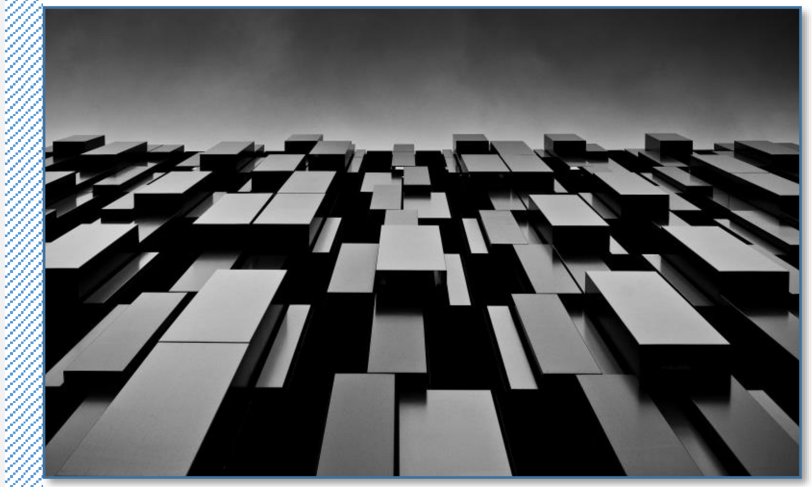
Pôle Sciences et technologie
Nantes Université

Regulation of High-risk AI systems

Ontology Building Defense

Yingqing CHEN & Robin Pourtaud

GOAL



Based on a document for regulation laying down harmonised rules on artificial intelligence and amending certain Union legislative acts, extract relevant terms of a domain from a corpus and classify or cluster them regarding core concepts, in order to build then an ontology of this domain. A cluster or class should be associated to a core concept and should include the terms that specialis[redacted] this core concept.

STEP 1 : Corpus Domain, Source

The domain of the corpus is the regulation of the use of Artificial Intelligence (AI) in the European Union.

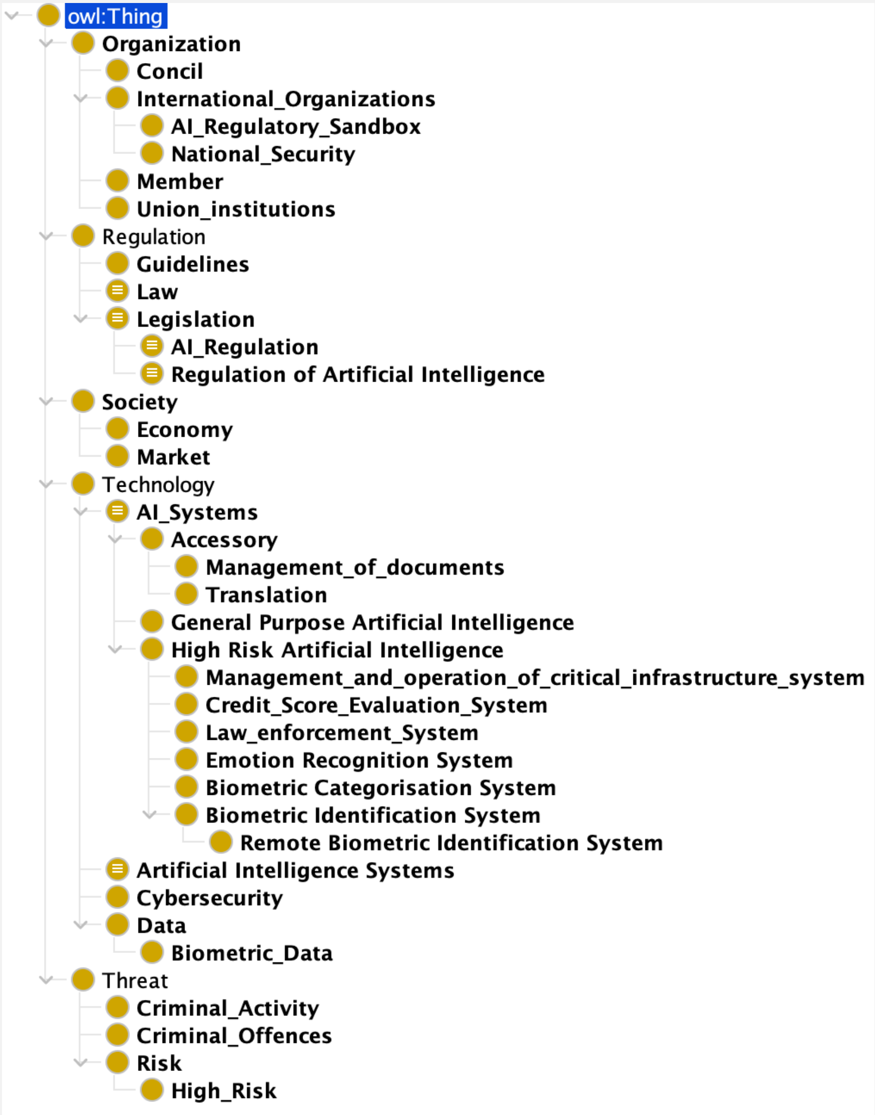
Council of the European Union

The corpus covers various aspects of the development, marketing, and use of AI, such as safety and fundamental rights, the internal market, and harmonized rules on the placing on the market, putting into service, and use of AI systems. It lays out specific prohibited practices. The corpus consists mainly of legislative documents, regulations, and guidelines that govern the use of AI in the European Union. The whole document has a size of around : 28325 words, 153559 characters, 98-99 pages.

STEP 1 : Core Concepts

We kept as Core concepts what we thought was very important:

•Regulation (superclass of law, rules…)
•Technologie (superclass of AI Systems, Cybersecurity…)
•Organization (superclass of institutions…)
•Threats (Criminal Activities…)



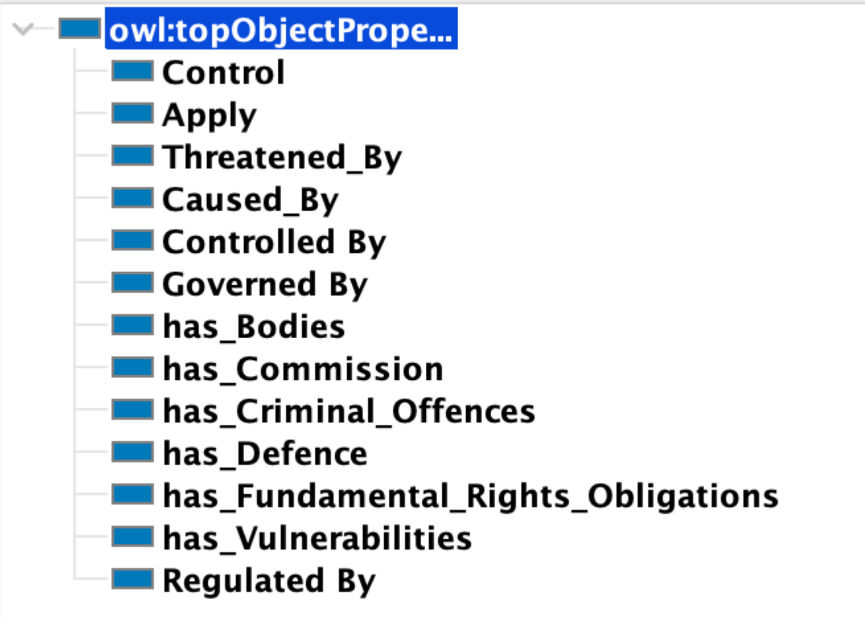
- owl:Thing
  - Organization
    - Concil
    - International_Organizations
      - AI_Regulatory_Sandbox
      - National_Security
    - Member
    - Union_institutions
  - Regulation
    - Guidelines
    - Law
    - Legislation
      - AI_Regulation
      - Regulation of Artificial Intelligence
  - Society
    - Economy
    - Market
  - Technology
    - AI_Systems
      - Accessory
        - Management_of_documents
        - Translation
      - General Purpose Artificial Intelligence
      - High Risk Artificial Intelligence
        - Management_and_operation_of_critical_infrastructure_system
        - Credit_Score_Evaluation_System
        - Law_enforcement_System
        - Emotion Recognition System
        - Biometric Categorisation System
        - Biometric Identification System
          - Remote Biometric Identification System
    - Artificial Intelligence Systems
    - Cybersecurity
    - Data
      - Biometric_Data
  - Threat
    - Criminal_Activity
    - Criminal_Offences
    - Risk
      - High_Risk

STEP 1 : Core Relations

•Core relations

•**Technologies** can be "Regulated By" **Organization**
•**Threat** can be "caused By" **Technologies**
•**Society** can be "threatened" by **Threat**
•**Organization** can "apply" **Regulation**

Regulation can "control" **Technologies, Society**



- owl:topObjectPrope...
  - Control
  - Apply
  - Threatened_By
  - Caused_By
  - Controlled By
  - Governed By
  - has_Bodies
  - has_Commission
  - has_Criminal_Offences
  - has_Defence
  - has_Fundamental_Rights_Obligations
  - has_Vulnerabilities
  - Regulated By

STEP 2 : Domain Ontology Competences

- The ability to understand the concepts and relationships defined in the ontology related to the regulation of the use of AI in the European Union
- The ability to use the ontology to understand and analyze legislative documents, regulations, and guidelines that govern the use of AI in the European Union

- Knowledge of the legal, technical and ethical aspects related to AI governance and the use of AI in the EU context.

STEP 2 : NLP Tools

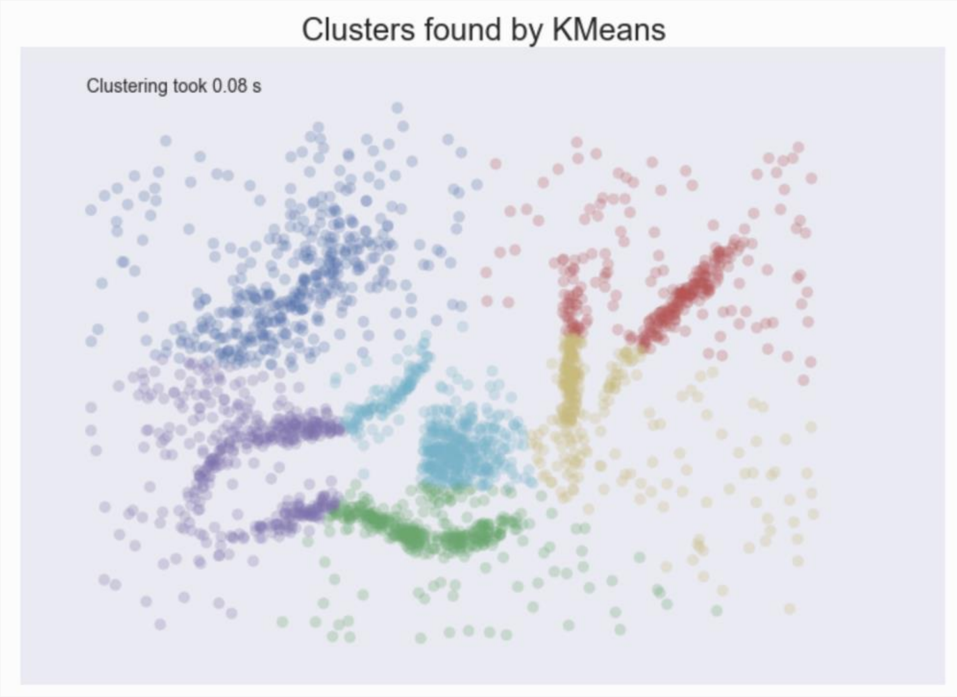
For this project, we will use Python3.11 and the following packages:

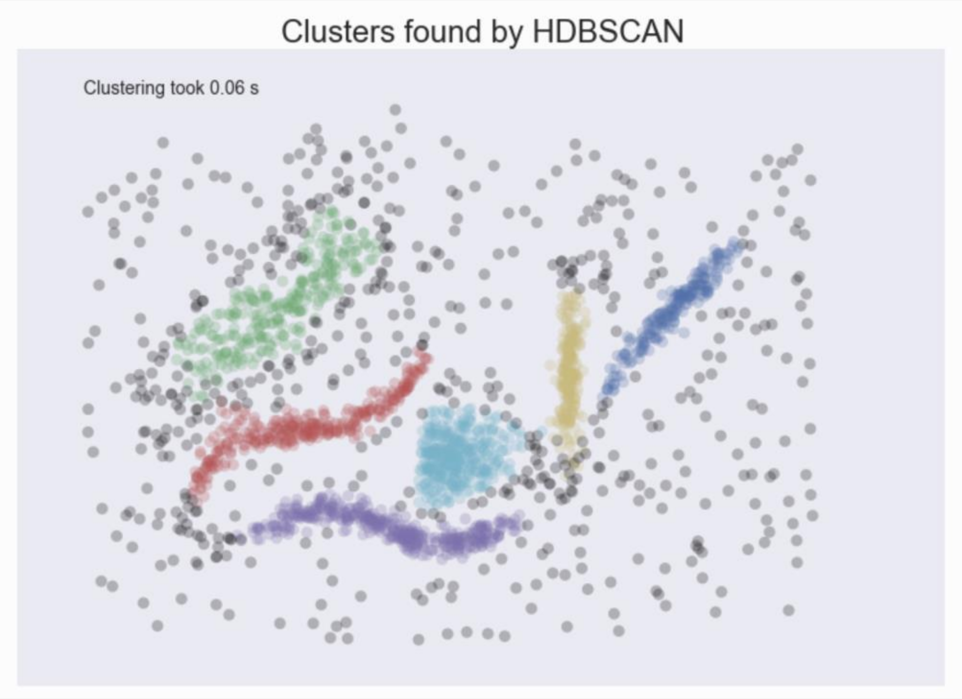
- Pandas
- Numpy
- Plotly
- Sklearn
- PDFminer
- Hdbscan

And the following natural language processing package:

- Spacy
- Gensim
- OpenAI API

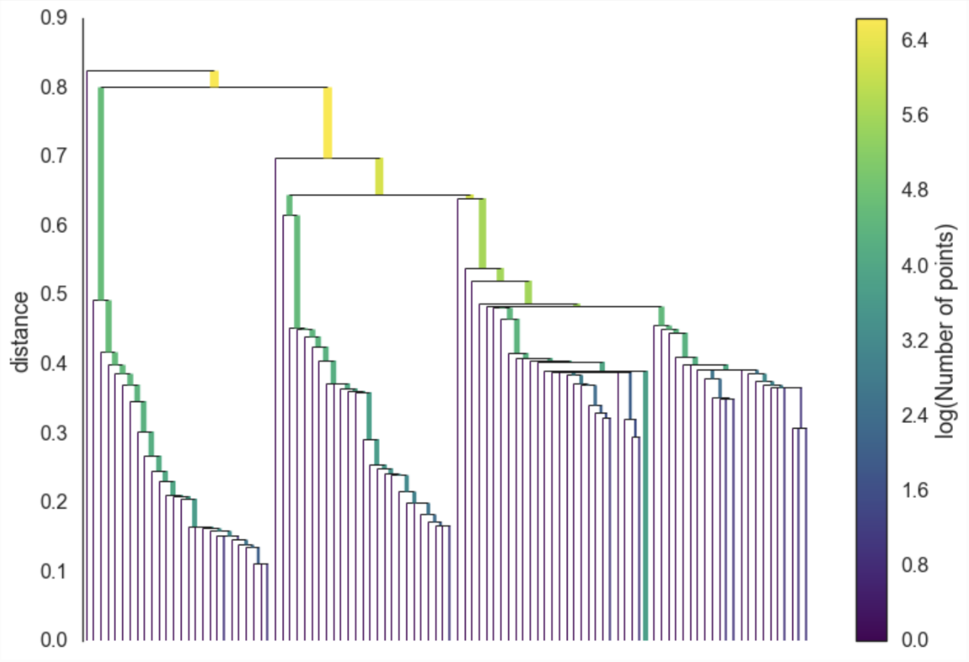
STEP 2 : Unsupervised Learning (HDBSCAN) - 1

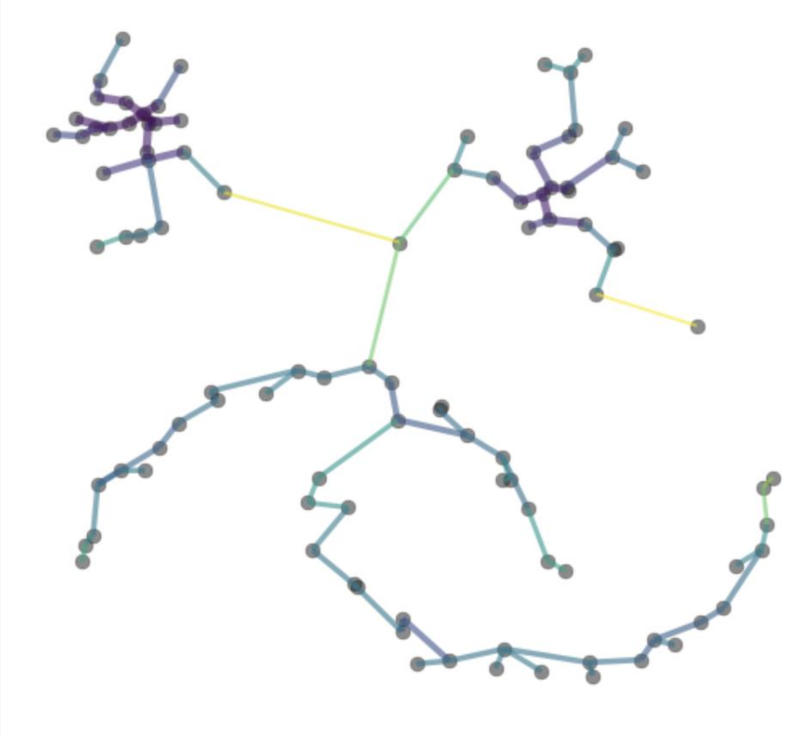




https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

STEP 2 : Unsupervised Learning (HDBSCAN) - 2

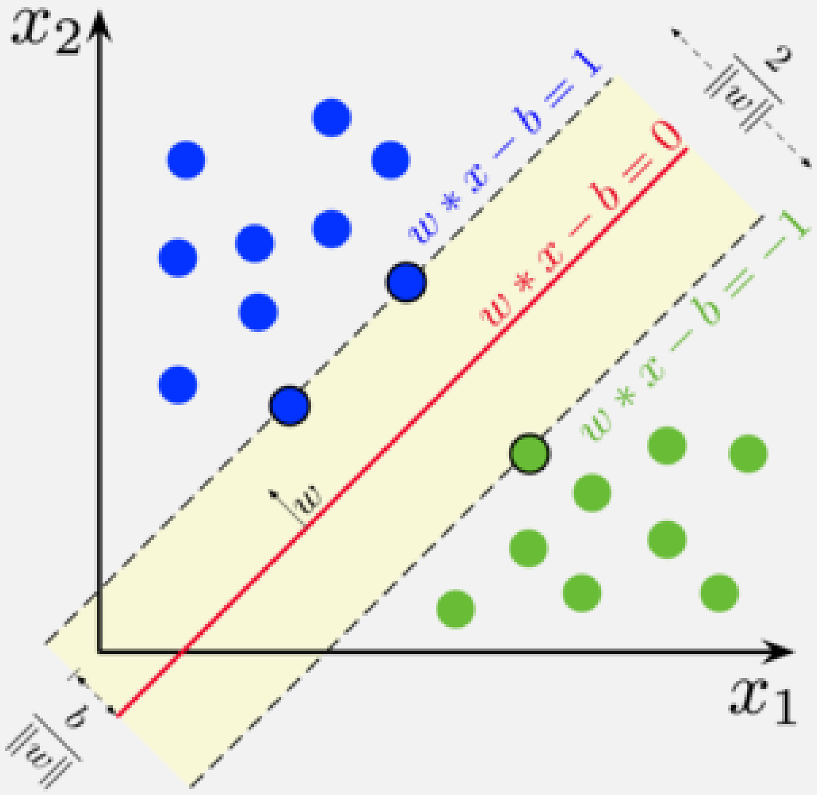




https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

[redacted]

STEP 2 : Supervised Learning (SVM)



Train Test Split : 0.8, 0.2

https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

STEP 2 : Lexical Units, Words, Terms

1. Nouns: Representing classes or concepts.

2. Verbs: Representing relations or properties between classes or concepts.

3. Adjectives: Representing attributes of classes or concepts.

4. Adverbs: Representing qualifiers or modifiers of relations or properties.

5. Name: Representing instances or individuals of classes or concepts.

6. **Prepositions:** Representing the relationships between instances or individuals of classes or concepts in the ontology.

STEP 2 : Feature Selection

Obviously, we don’t want to consider the whole corpus. We want to keep only words and sentences that respect some metric and pattern constraints:

Sentences that are related to core concepts: One approach could be to use keyword matching or text similarity methods to identify sentences that contain key terms related to our core concept. Some verbs, or words like “as”, “like” etc can give us more information.

Core concepts co-occurrences: Identifying the co-occurrences of core concepts in the corpus can help to identify patterns of association between concepts and to extract important relationships.

STEP 3 : Processing the corpus

Preprocessing of the corpus by using NLP techniques.

- Using spacy: parsedSent = nlp(sent).
- Obtain lemma, POS tag, dependency word, and dependency relation for each word.
- Extract frequent terms.
- Transform the corpus into a special format.

STEP 3 : NPs (or terms) extraction

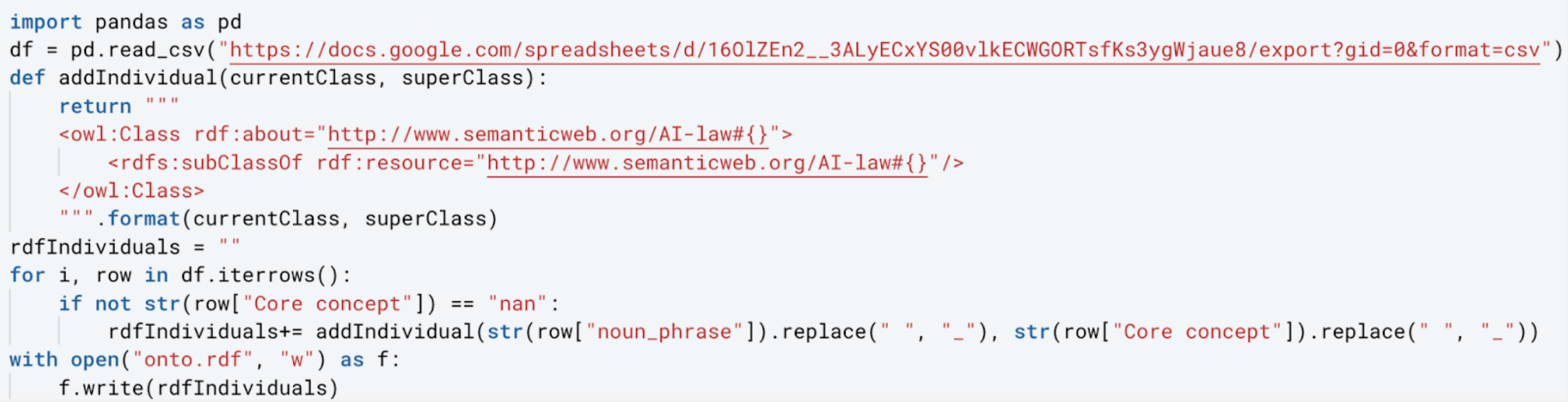
Extract frequent Noun phrases (NP) that occur above minimum frequency

- Pass over each sentence of the corpus
- Use spacy NP chunker to extract noun phrases
- Remove not interesting NPs **if they exist in the list of StopWords**
- If NP exists in the dictionary, increase its frequency value by 1
- Else add NP to the dictionary with value = 1
- Finally, save all NPs in the dictionary with frequency value > threshold into a file

STEP 4 : Gold standard ontology building

Extract frequent Noun phrases (NP) that occur above minimum frequency

- We classify manually selected NPs according to core concepts and generate the ontology.



```
import pandas as pd
df = pd.read_csv("https://docs.google.com/spreadsheets/d/16OlZEn2__3ALyECxYS00vlkECWGORTsfKs3ygWjaue8/export?gid=0&format=csv")
def addIndividual(currentClass, superClass):
    return """
    <owl:Class rdf:about="http://www.semanticweb.org/AI-law#{}">
        <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/AI-law#{}"/>
    </owl:Class>
    """.format(currentClass, superClass)
rdfIndividuals = ""
for i, row in df.iterrows():
    if not str(row["Core concept"]) == "nan":
        rdfIndividuals+= addIndividual(str(row["noun_phrase"]).replace(" ", "_"), str(row["Core concept"]).replace(" ", "_"))
with open("onto.rdf", "w") as f:
    f.write(rdfIndividuals)
```

STEP 5 : NP-Parsing and Word Embedding

Multiples Embeddings Possible:

1 – Word2Vec on NP (without context)

2 – Word2Vec on NP + Corpus (contextual)

3 – Word2Vec on Google News (3gb)

4 – Word2Vec on Google News + Corpus

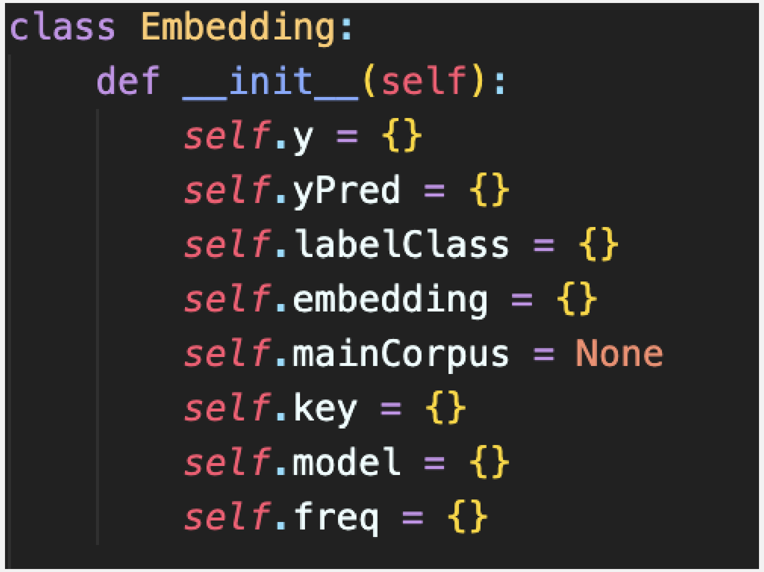
5 – Word2Vec on WikipediaEN (17gb...)

6 – GPT3 Embedding (OpenAi)

(Not tried: Bert Embedding, FastText, GloVe......)

[redacted]

STEP 5 : Embeding Module



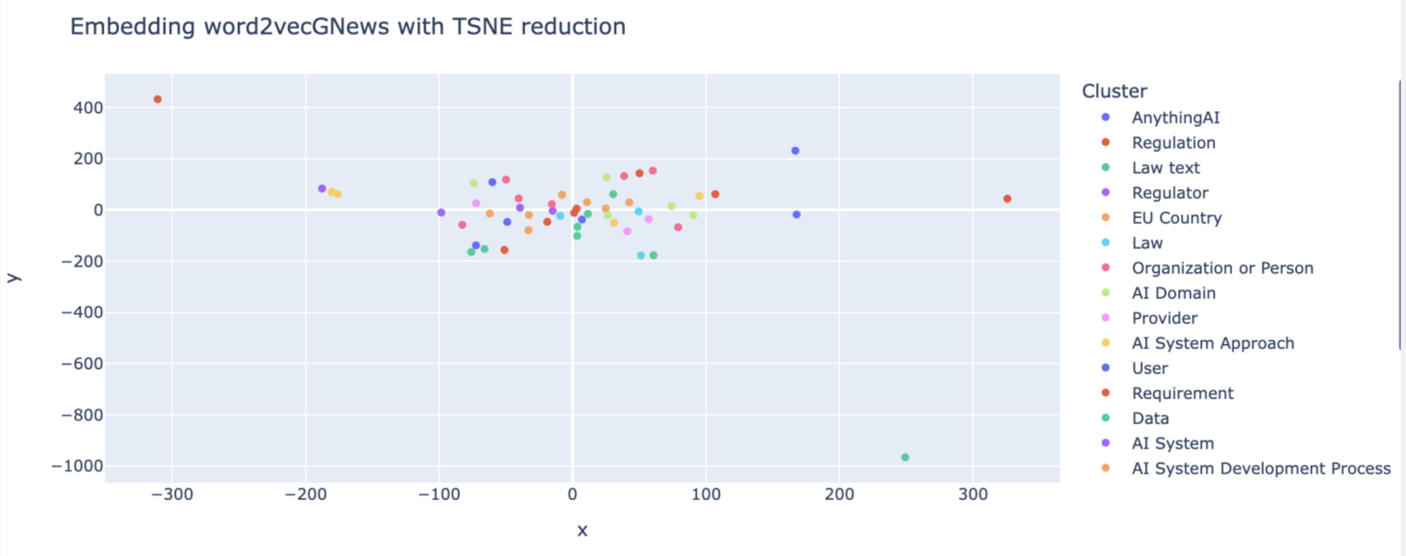
```python
class Embedding:
    def __init__(self):
        self.y = {}
        self.yPred = {}
        self.labelClass = {}
        self.embedding = {}
        self.mainCorpus = None
        self.key = {}
        self.model = {}
        self.freq = {}
```

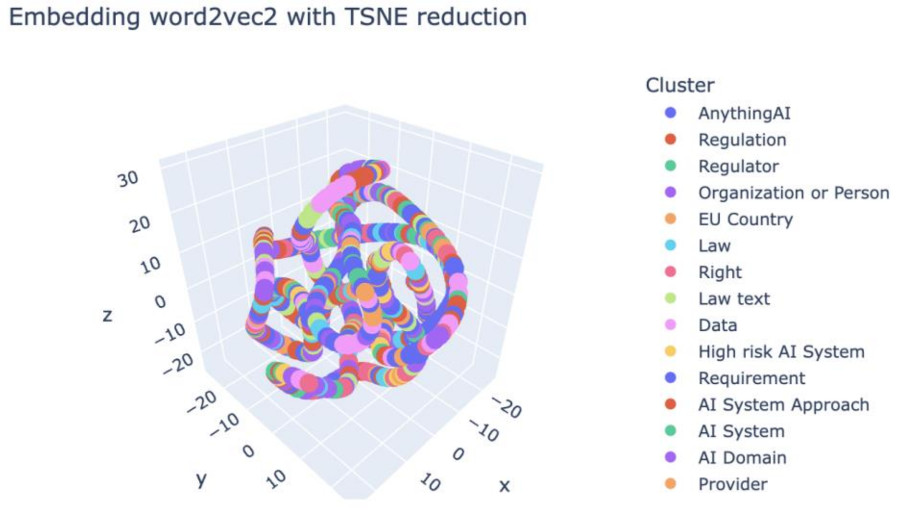
Extensible Class that can handle multiple operations:

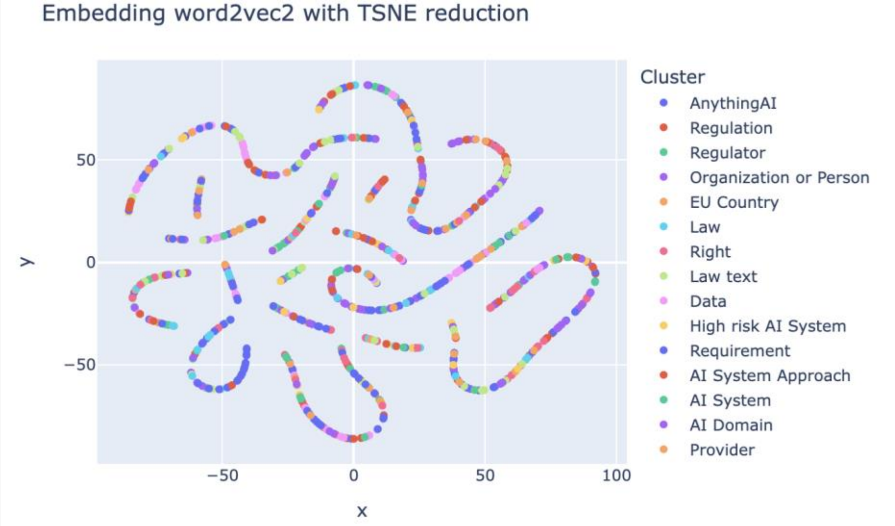
- Reduction methods
- Embedding methods
- Plots (dendrograms...)
- ML/DM methods...
- Export
- ....

[redacted]

STEP 5 : Views of Embedding

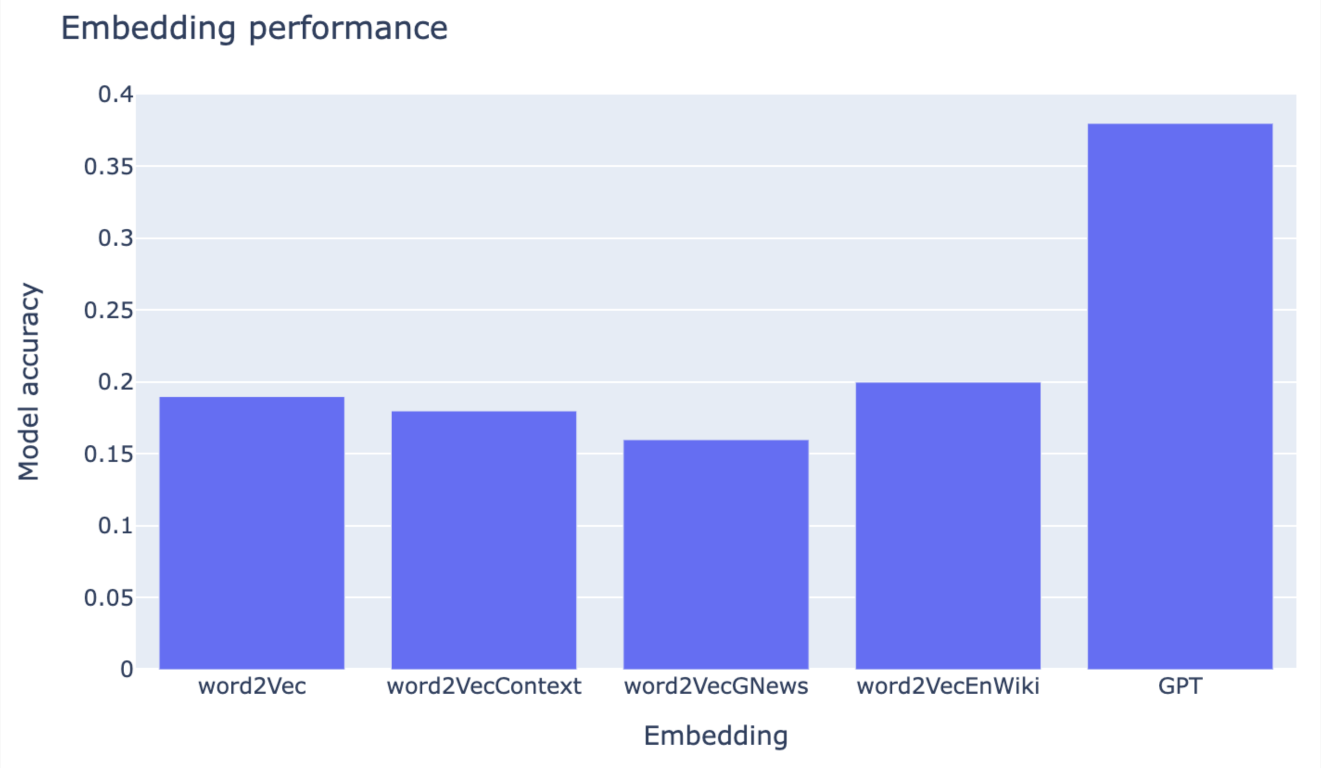


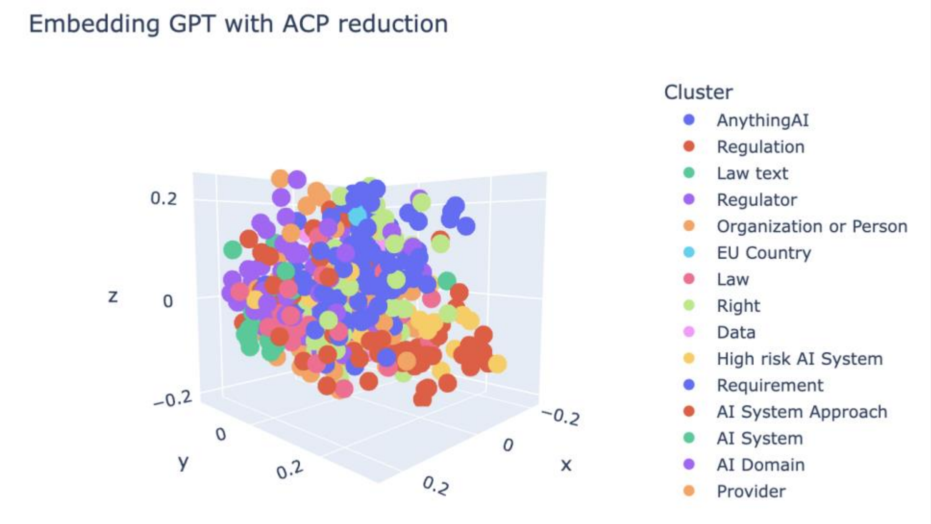


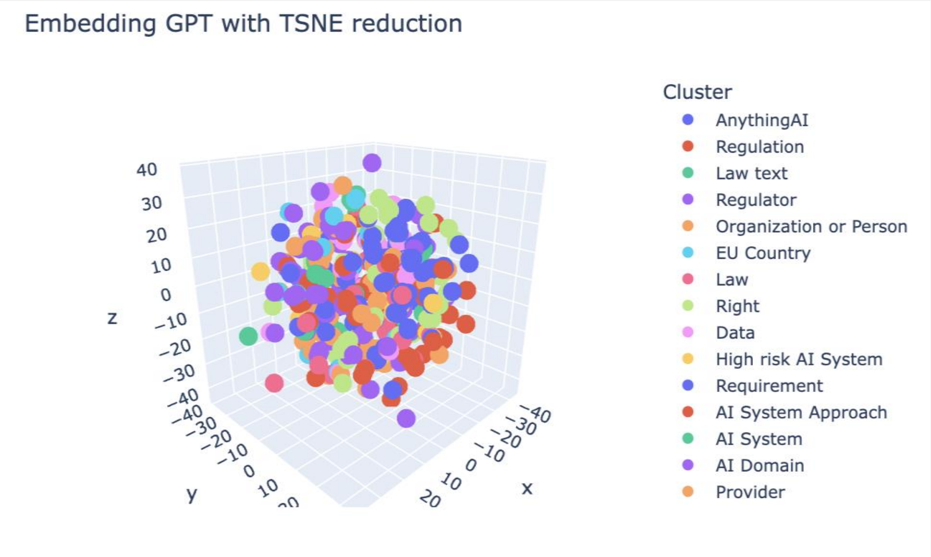


[redacted]

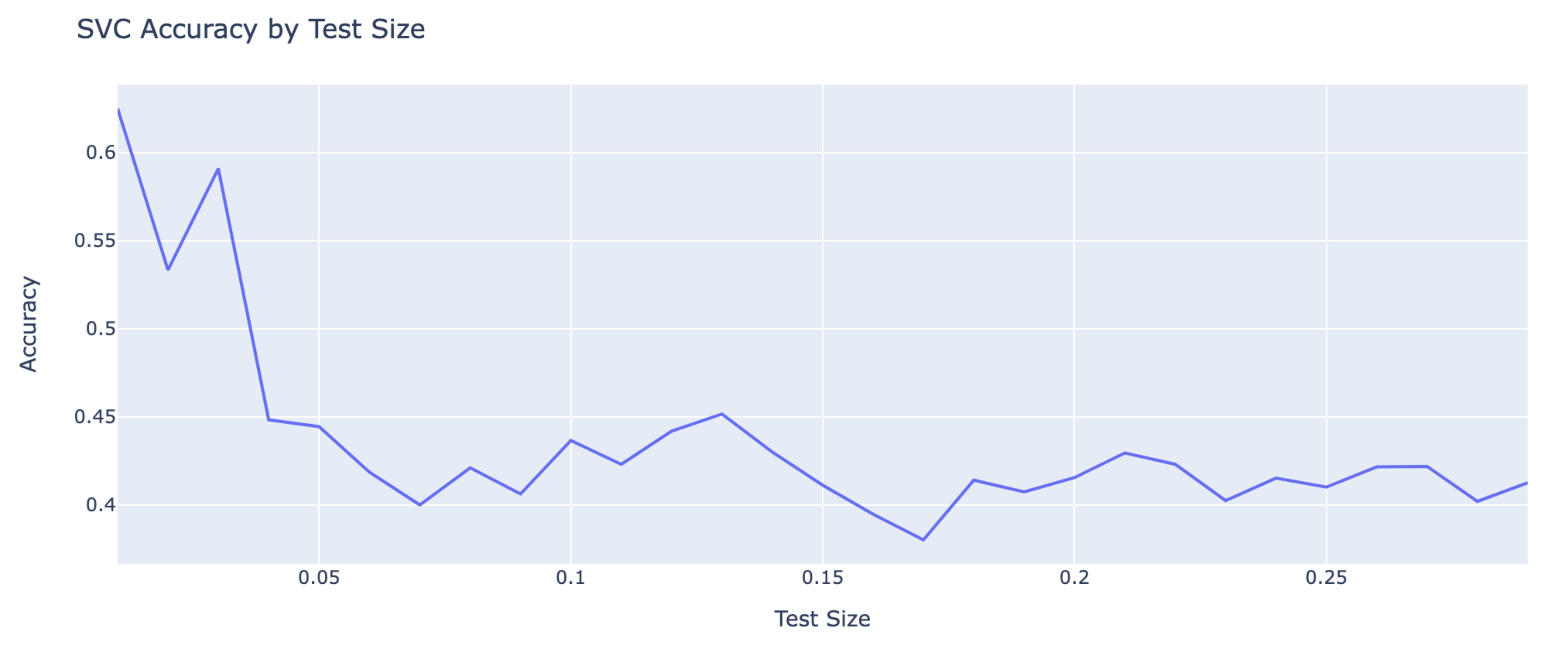
STEP 5,6,7 : Which Embedding to keep?


Embedding performance


Embedding GPT with ACP reduction


Embedding GPT with TSNE reduction

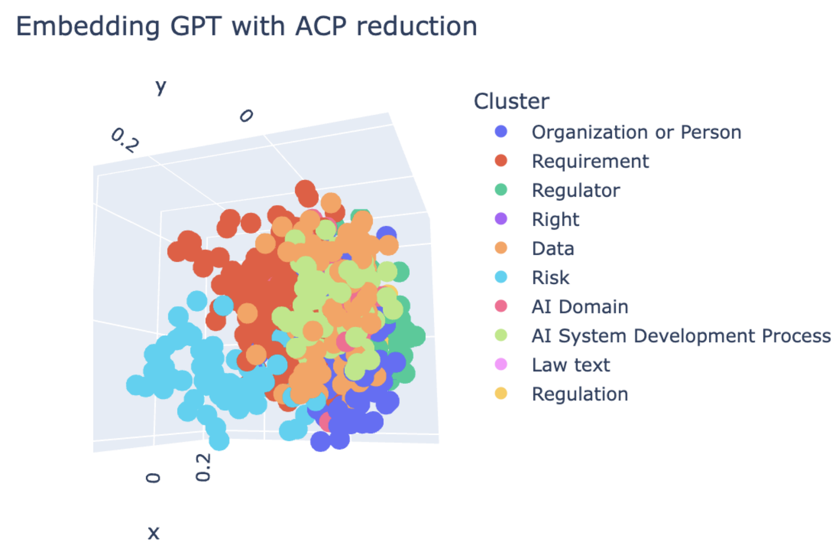
STEP 5,6,7: SVM Classification on GPT - 1

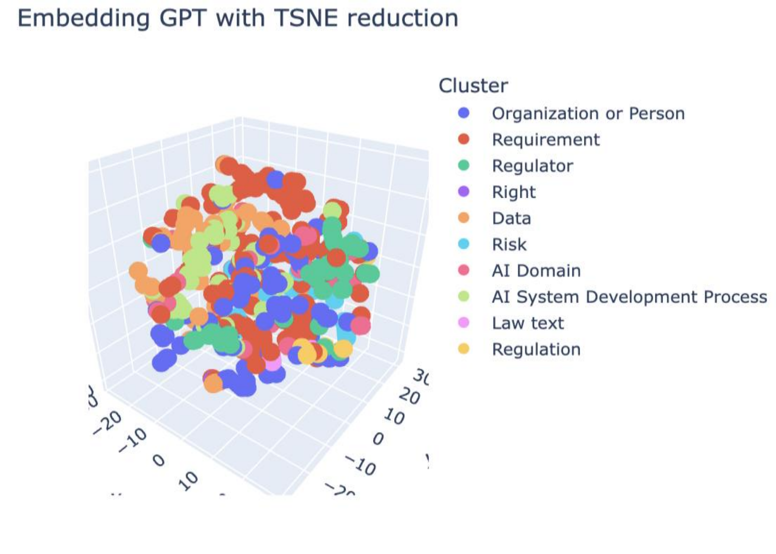


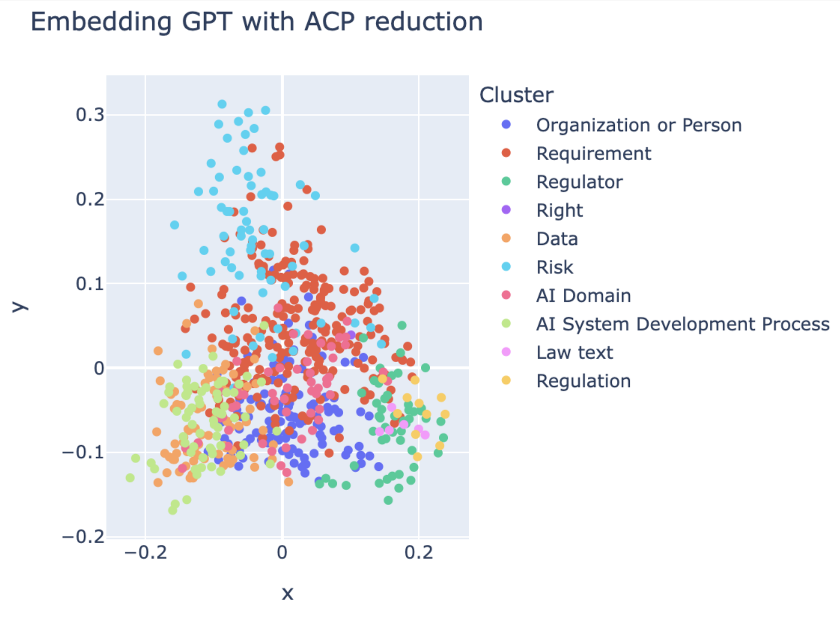
We decided on using 0.13 with an accuracy of 0.4516

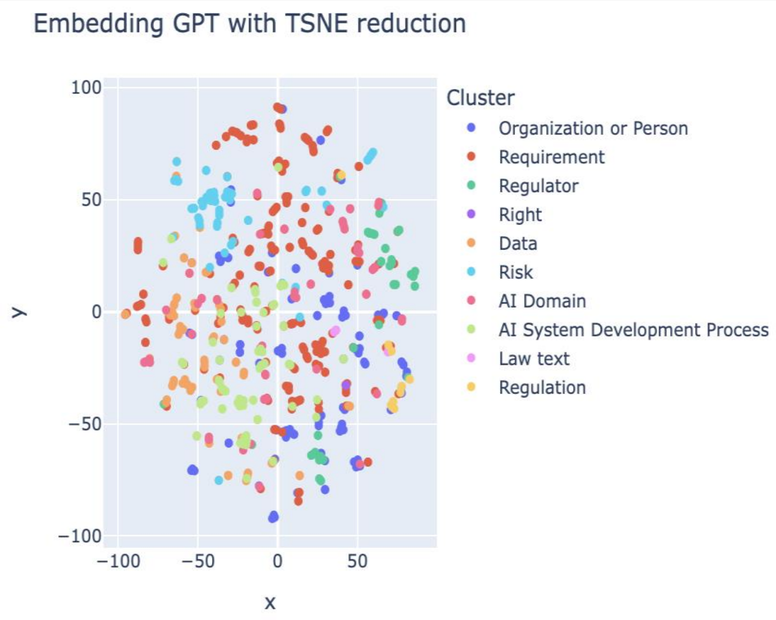
[redacted]

STEP 5,6,7: SVM Classification on GPT - 1

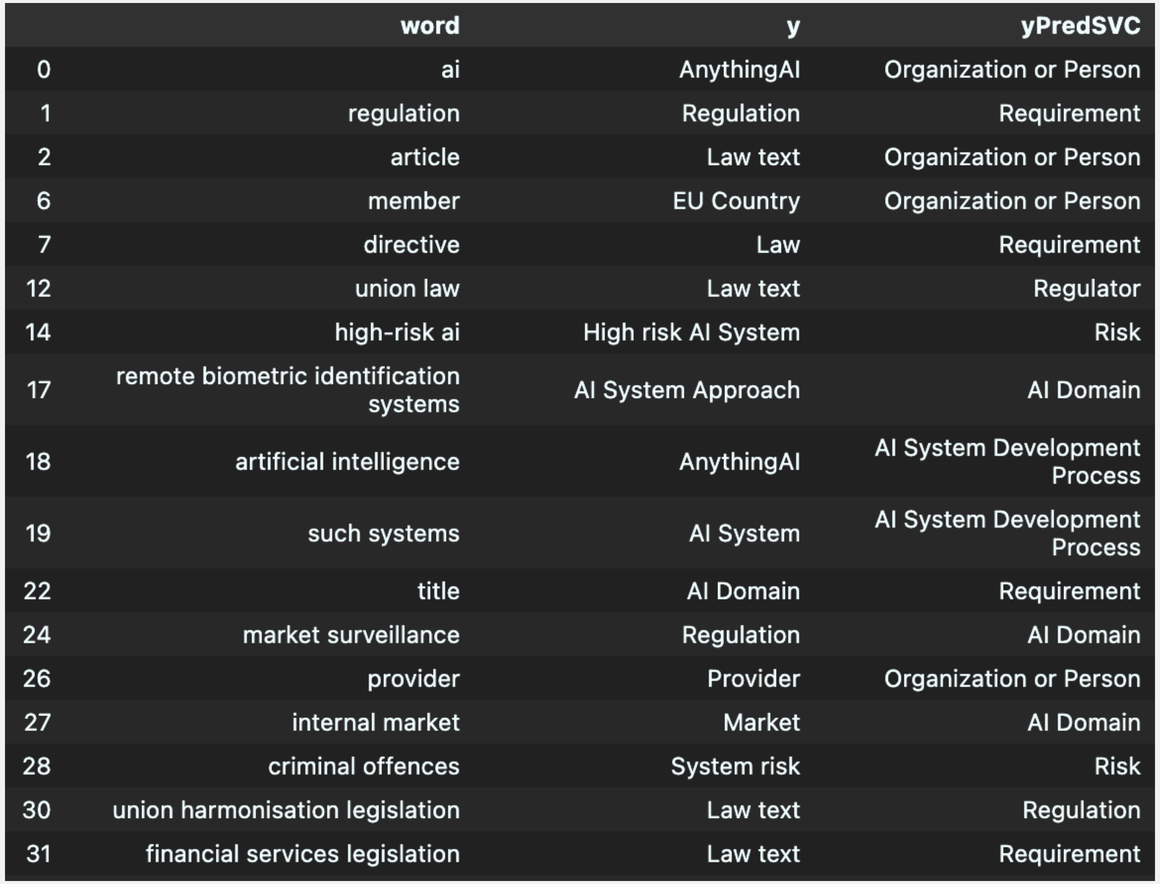

Embedding GPT with ACP reduction


Embedding GPT with TSNE reduction

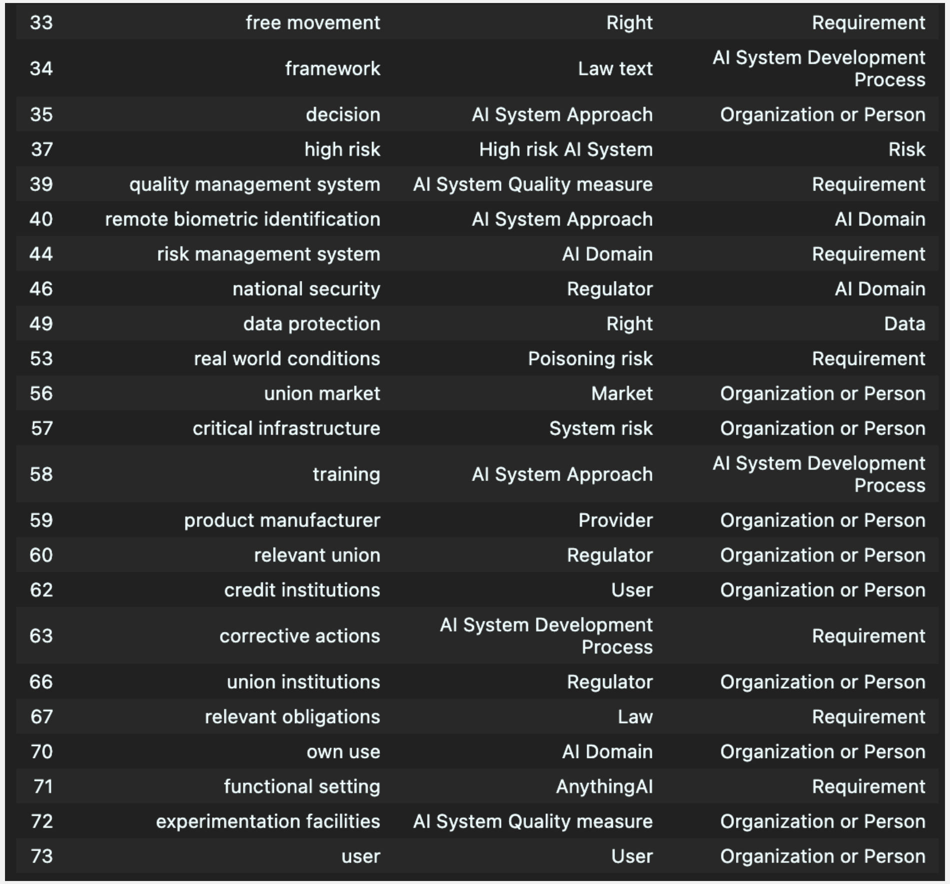

Embedding GPT with ACP reduction


Embedding GPT with TSNE reduction

STEP 5,6,7: SVM Classification on GPT - 2

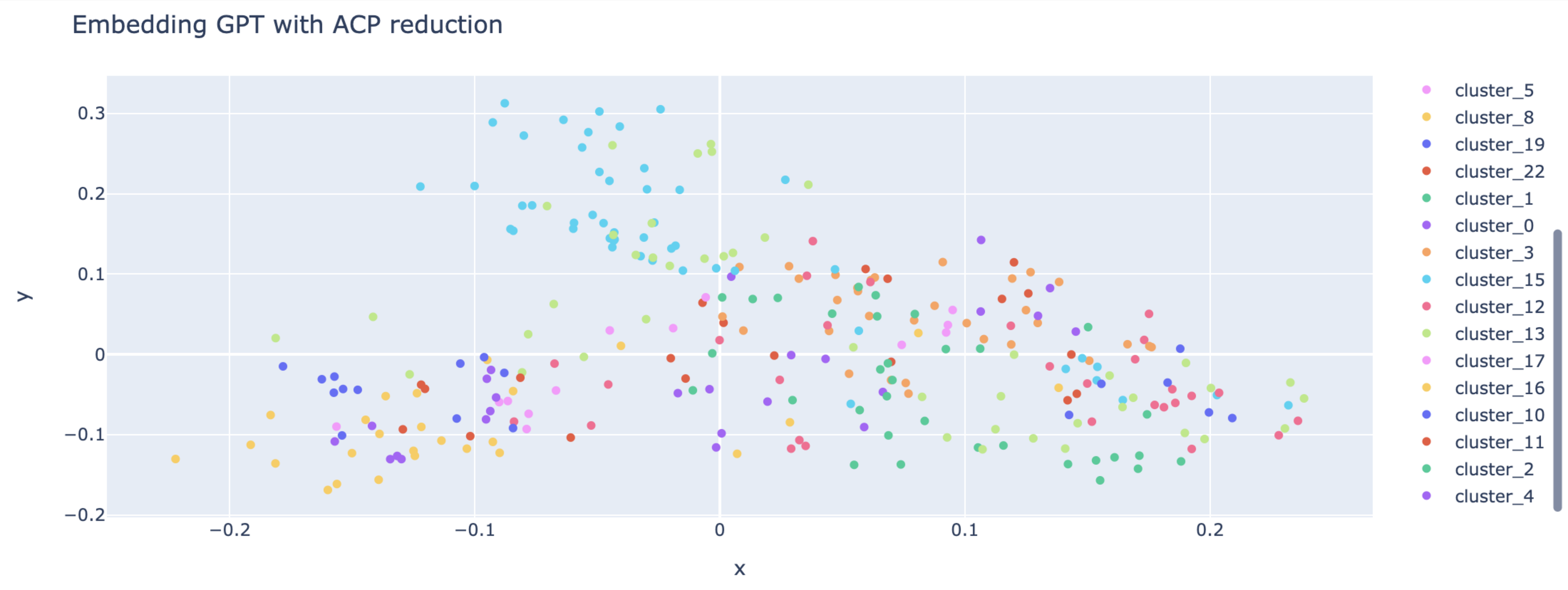


	word	y	yPredSVC
0	ai	AnythingAI	Organization or Person
1	regulation	Regulation	Requirement
2	article	Law text	Organization or Person
6	member	EU Country	Organization or Person
7	directive	Law	Requirement
12	union law	Law text	Regulator
14	high-risk ai	High risk AI System	Risk
17	remote biometric identification systems	AI System Approach	AI Domain
18	artificial intelligence	AnythingAI	AI System Development Process
19	such systems	AI System	AI System Development Process
22	title	AI Domain	Requirement
24	market surveillance	Regulation	AI Domain
26	provider	Provider	Organization or Person
27	internal market	Market	AI Domain
28	criminal offences	System risk	Risk
30	union harmonisation legislation	Law text	Regulation
31	financial services legislation	Law text	Requirement



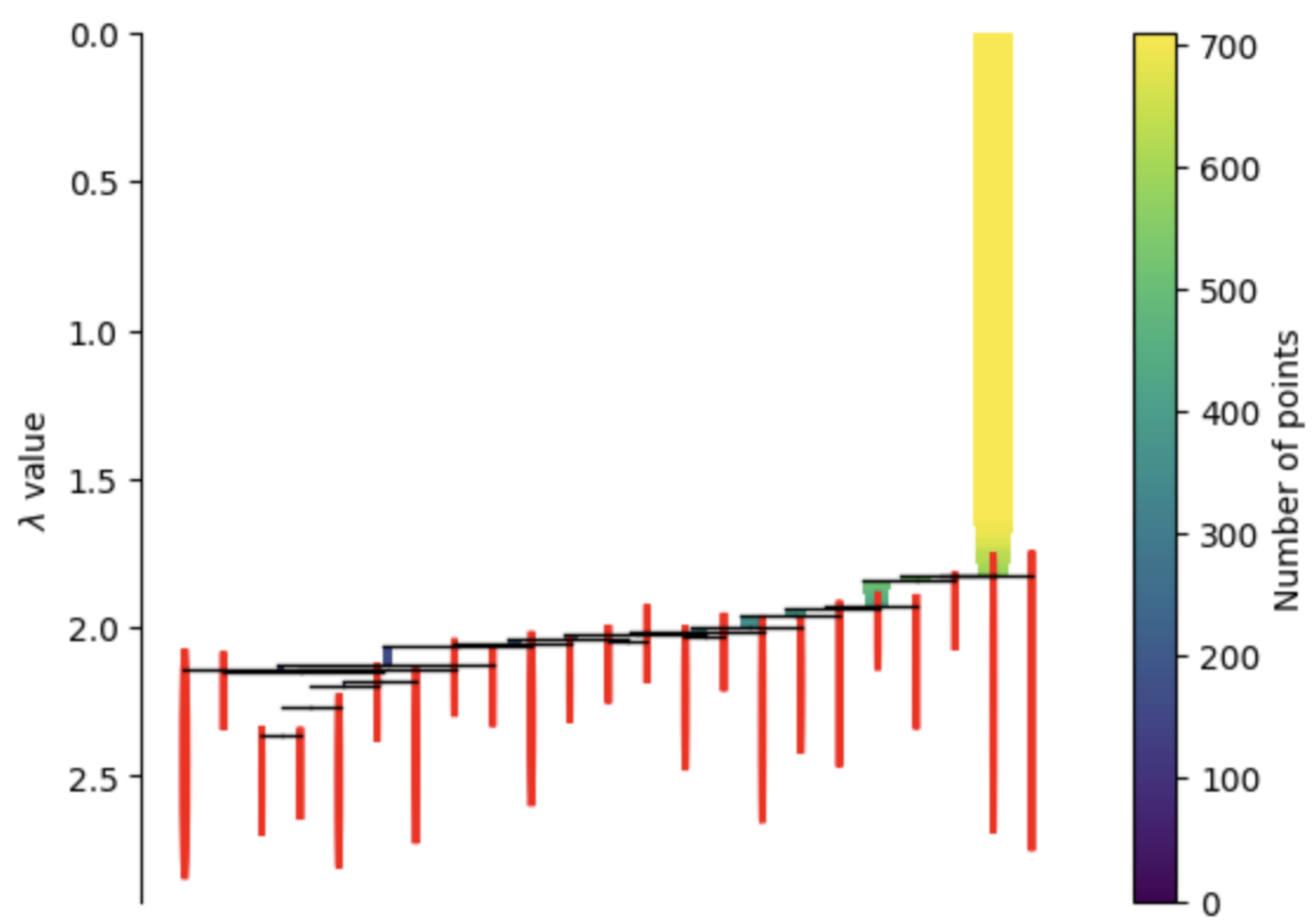
33	free movement	Right	Requirement
34	framework	Law text	AI System Development Process
35	decision	AI System Approach	Organization or Person
37	high risk	High risk AI System	Risk
39	quality management system	AI System Quality measure	Requirement
40	remote biometric identification	AI System Approach	AI Domain
44	risk management system	AI Domain	Requirement
46	national security	Regulator	AI Domain
49	data protection	Right	Data
53	real world conditions	Poisoning risk	Requirement
56	union market	Market	Organization or Person
57	critical infrastructure	System risk	Organization or Person
58	training	AI System Approach	AI System Development Process
59	product manufacturer	Provider	Organization or Person
60	relevant union	Regulator	Organization or Person
62	credit institutions	User	Organization or Person
63	corrective actions	AI System Development Process	Requirement
66	union institutions	Regulator	Organization or Person
67	relevant obligations	Law	Requirement
70	own use	AI Domain	Organization or Person
71	functional setting	AnythingAI	Requirement
72	experimentation facilities	AI System Quality measure	Organization or Person
73	user	User	Organization or Person

STEP 5,6,7: HDBSCAN Clustering on GPT - 1

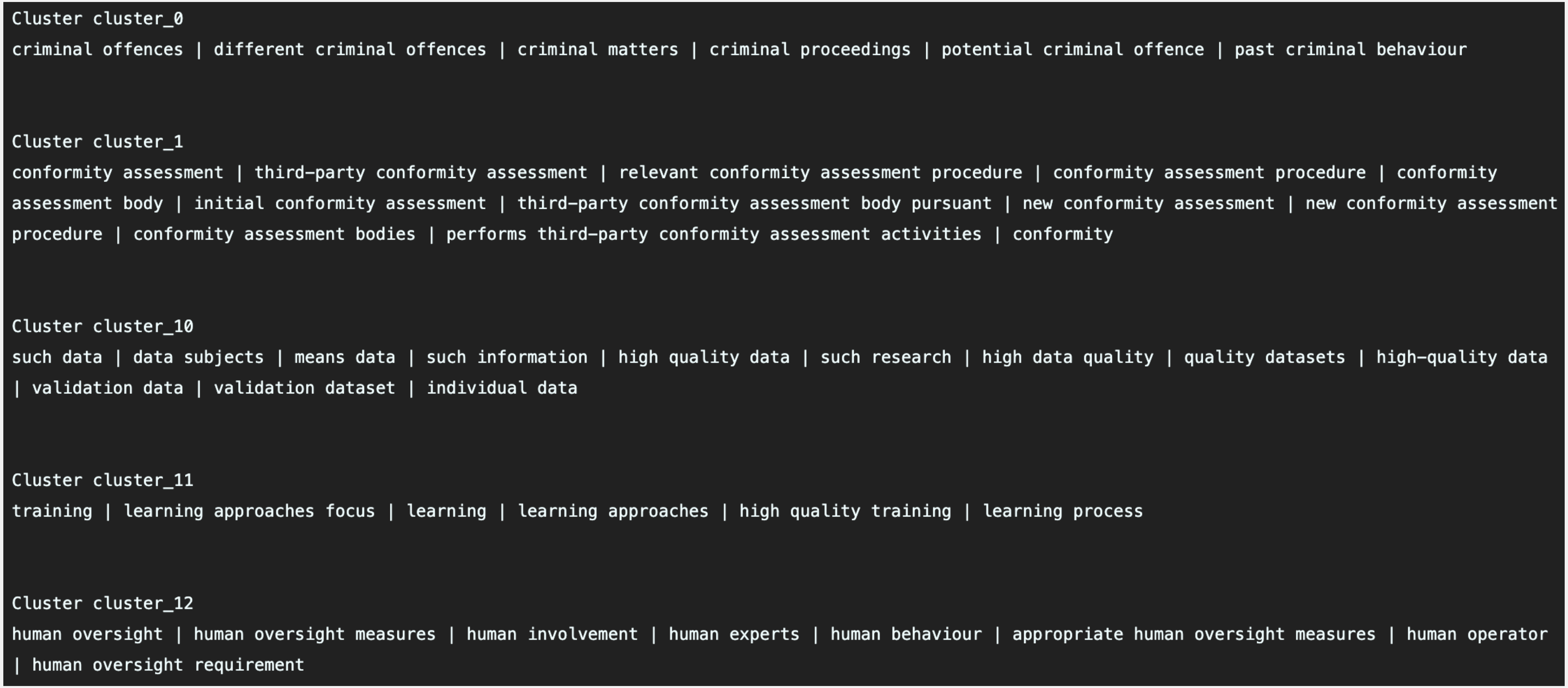


The noise is hidden !

STEP 5,6,7: HDBSCAN Clustering on GPT - 2

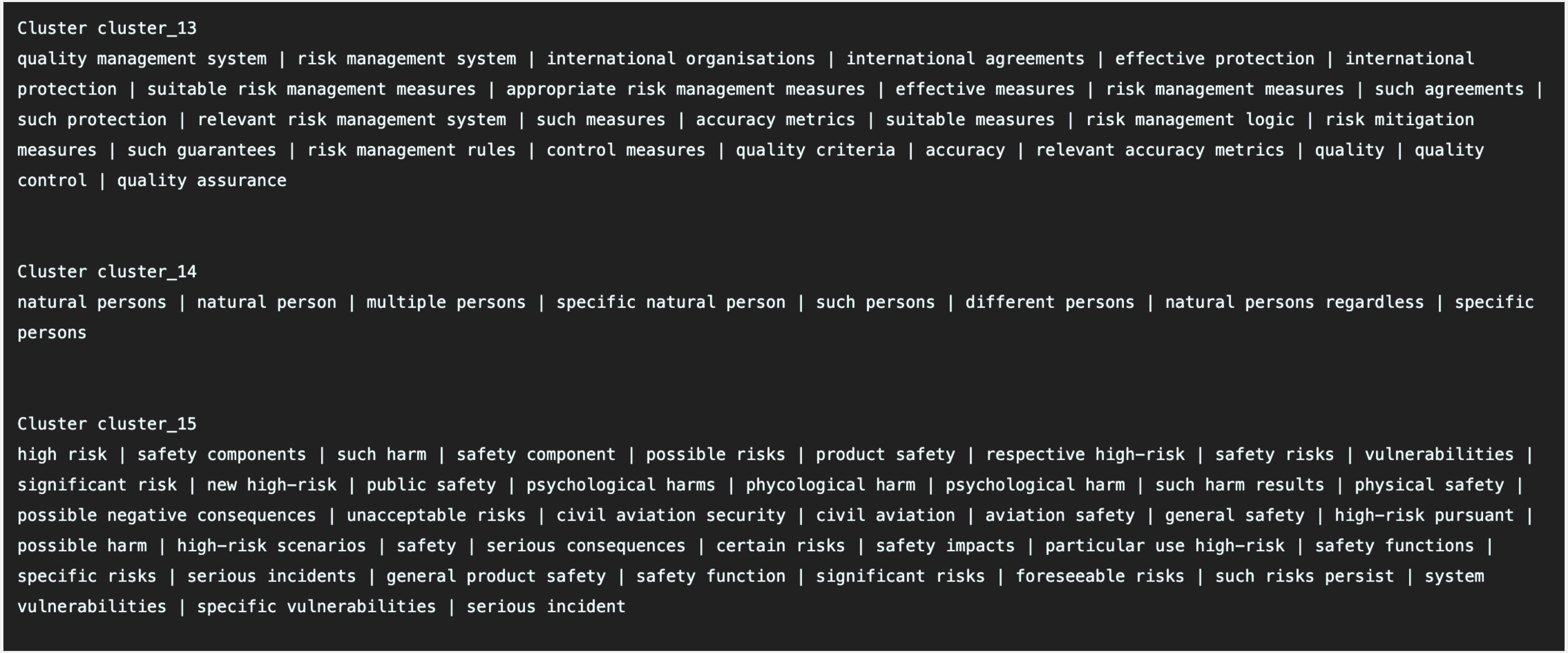


STEP 5,6,7: HDBSCAN Classification on GPT - 3



```
Cluster cluster_0
criminal offences | different criminal offences | criminal matters | criminal proceedings | potential criminal offence | past criminal behaviour


Cluster cluster_1
conformity assessment | third-party conformity assessment | relevant conformity assessment procedure | conformity assessment procedure | conformity
assessment body | initial conformity assessment | third-party conformity assessment body pursuant | new conformity assessment | new conformity assessment
procedure | conformity assessment bodies | performs third-party conformity assessment activities | conformity


Cluster cluster_10
such data | data subjects | means data | such information | high quality data | such research | high data quality | quality datasets | high-quality data
| validation data | validation dataset | individual data


Cluster cluster_11
training | learning approaches focus | learning | learning approaches | high quality training | learning process


Cluster cluster_12
human oversight | human oversight measures | human involvement | human experts | human behaviour | appropriate human oversight measures | human operator
| human oversight requirement
```

STEP 5,6,7: HDBSCAN Classification on GPT - 4



```
Cluster cluster_13
quality management system | risk management system | international organisations | international agreements | effective protection | international
protection | suitable risk management measures | appropriate risk management measures | effective measures | risk management measures | such agreements |
such protection | relevant risk management system | such measures | accuracy metrics | suitable measures | risk management logic | risk mitigation
measures | such guarantees | risk management rules | control measures | quality criteria | accuracy | relevant accuracy metrics | quality | quality
control | quality assurance


Cluster cluster_14
natural persons | natural person | multiple persons | specific natural person | such persons | different persons | natural persons regardless | specific
persons


Cluster cluster_15
high risk | safety components | such harm | safety component | possible risks | product safety | respective high-risk | safety risks | vulnerabilities |
significant risk | new high-risk | public safety | psychological harms | phycological harm | psychological harm | such harm results | physical safety |
possible negative consequences | unacceptable risks | civil aviation security | civil aviation | aviation safety | general safety | high-risk pursuant |
possible harm | high-risk scenarios | safety | serious consequences | certain risks | safety impacts | particular use high-risk | safety functions |
specific risks | serious incidents | general product safety | safety function | significant risks | foreseeable risks | such risks persist | system
vulnerabilities | specific vulnerabilities | serious incident
```

POLYTECH NANTES | Pôle Sciences et technologie Nantes Université

THANKS FOR YOUR ATTENTION