International Agreements Database mining

Tomasz Siudalski, Weronika Plichta, Michał Taczała

Introduction

This project realized in cooperation with faculty members from the University of Lodz aims to explore the application of Natural Language Processing (NLP) techniques to analyze a vast collection of international agreements from U.S. states and municipalities. The project will focus on automating the extraction of 13 key attributes proposed by the faculty members, such as areas of cooperation, parties involved, agreement types, and recurring clauses.

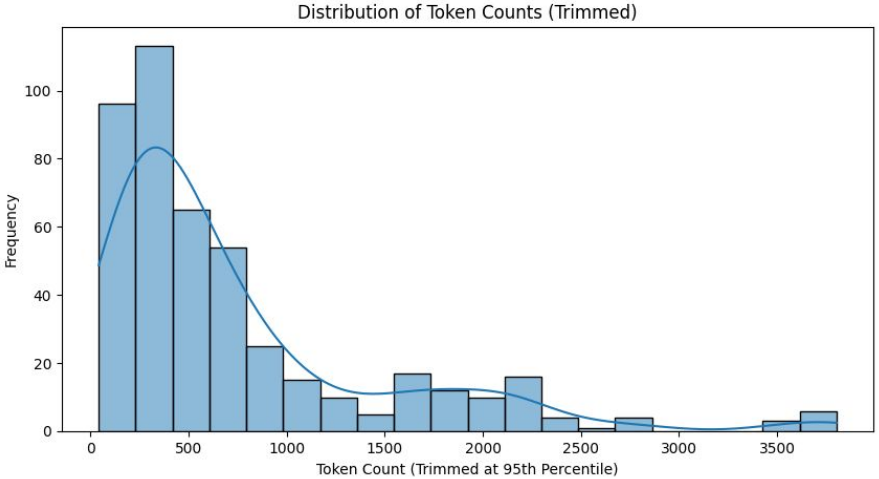
Dataset

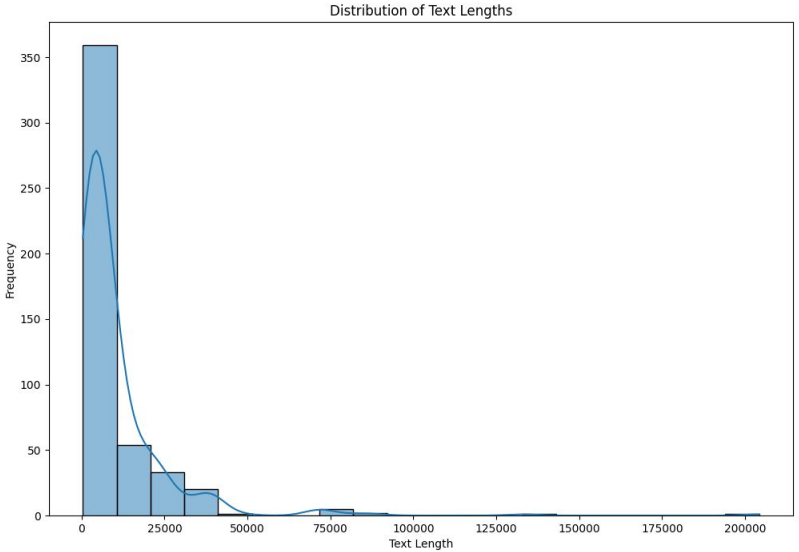
- 180 PDF documents
- Images in PDF format of agreements between US States
- Images are of a different quality(some are quite blurry)
- Confidential (As agreements are from private paid dataset)

Task

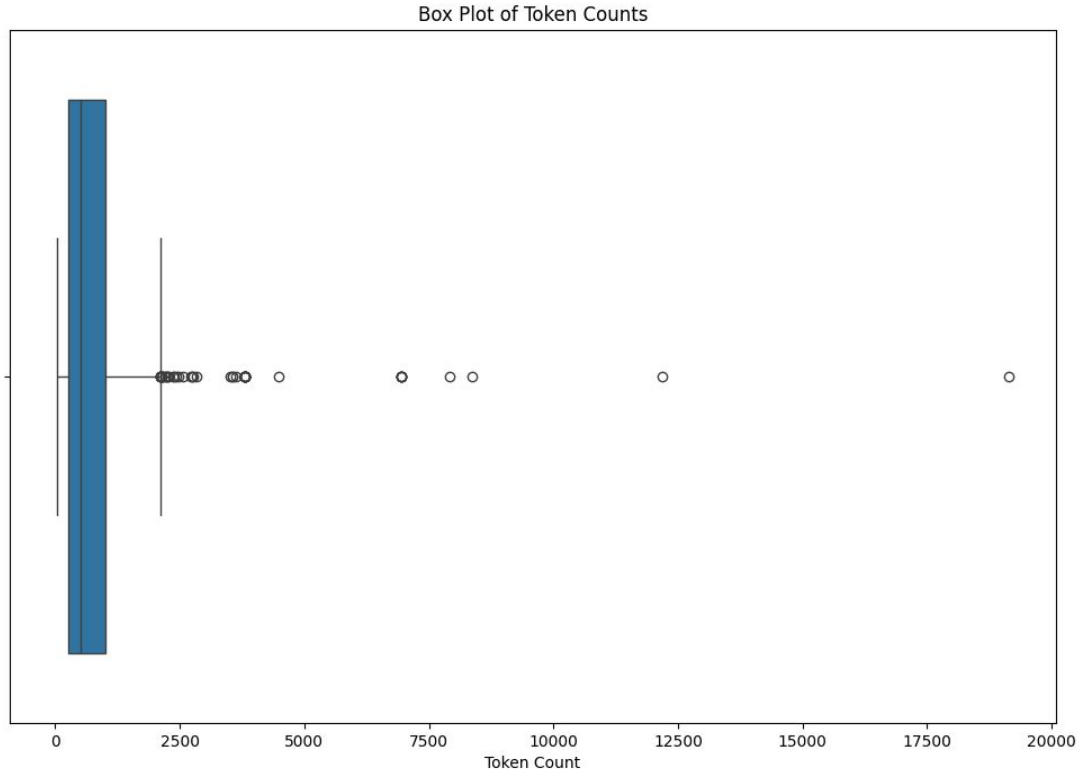
1. Identification of areas of cooperation mentioned in the agreements.
2. Identification of the parties involved (are any institutions or local partners mentioned besides the states?).
3. Identification of the types of agreements (e.g., Memorandum of Understanding, Sister Cities Agreement, etc.).
4. Determination of the percentage of agreements under the patronage of Sister Cities International.
5. Identification of international organizations mentioned in the agreements.
6. Determination of the terms of validity for each agreement (until when?).
7. Identification of the length of each agreement (number of pages or words).
8. Determination of the conditions for extending each agreement (automatic or by decision?).
9. Analysis of the frequency of recurring clauses in the agreements (always, often, rarely) – the level of detail in the agreements.
10. Identification of the partners with whom the agreements tend to be more detailed.
11. Indication of whether the agreement includes an evaluation of its implementation.
12. Identification of whether the agreement mentions any coordination of activities with other entities (e.g., government, other cities/states, international organizations).
13. Identification of whether the agreement refers to other legal documents.

EDA


Distribution of Token Counts (Trimmed)
Frequency
Token Count (Trimmed at 95th Percentile)
0
20
40
60
80
100
0
500
1000
1500
2000
2500
3000
3500


Distribution of Text Lengths
Frequency
Text Length
0
50
100
150
200
250
300
350
0
25000
50000
75000
100000
125000
150000
175000
200000

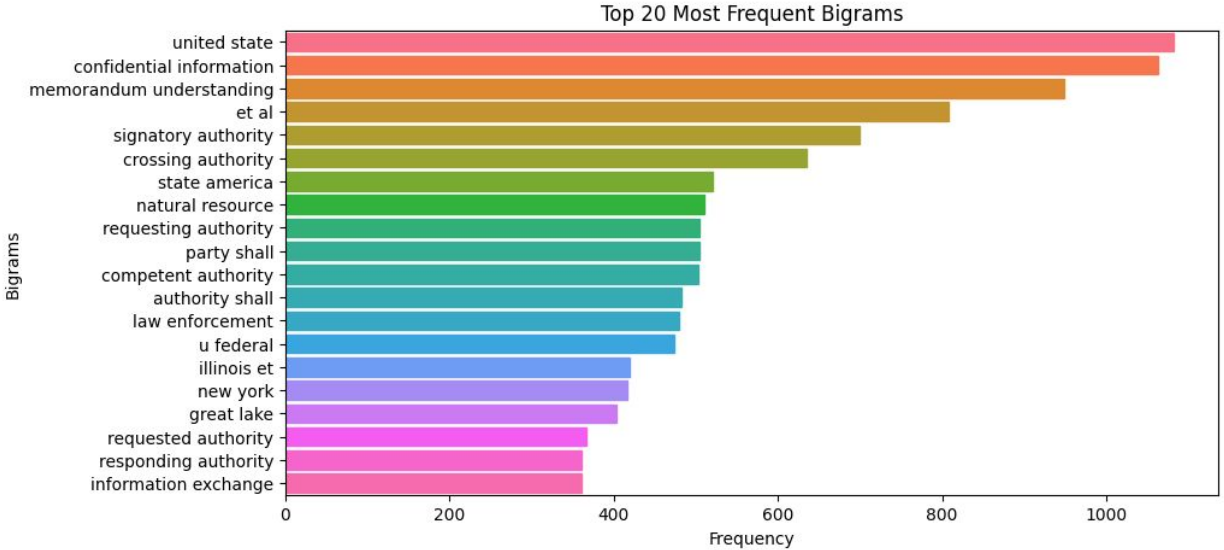
EDA

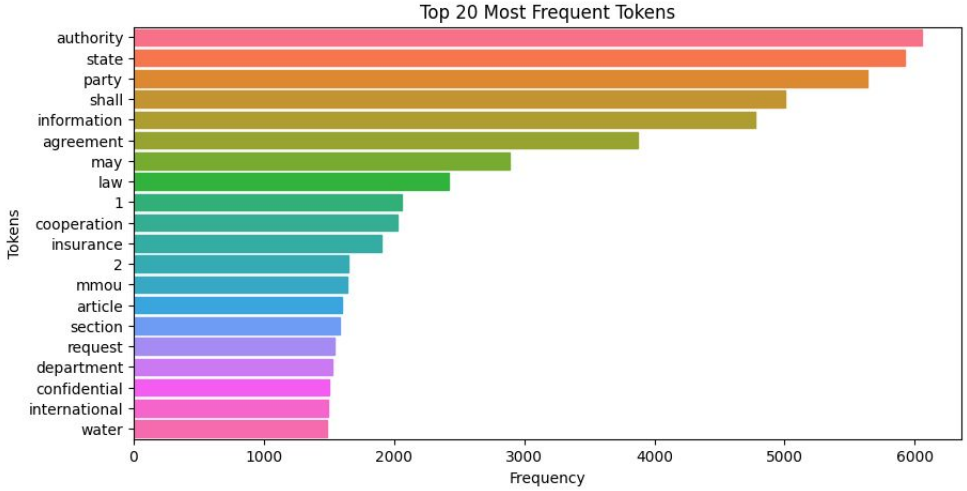

Box Plot of Token Counts
0
2500
5000
7500
10000
12500
15000
17500
20000
Token Count

EDA

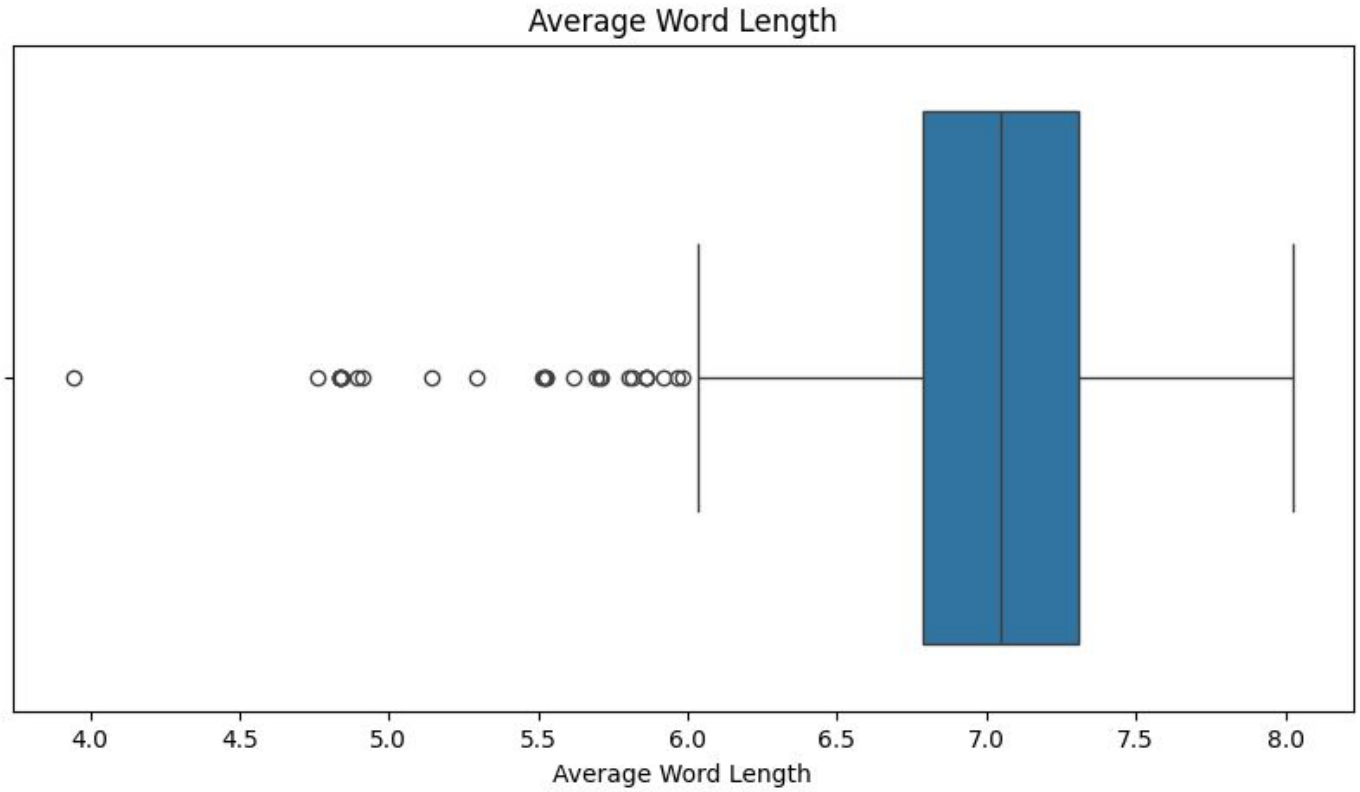

Word Cloud of Tokens
party
crossing authority
et al
requested authority
law regulation
information exchange
great lake
competent authority
requesting authority
law enforcement
state america
memorandum understanding
confidential information
signatory authority
agreement
united state
federal
natural resource
illinois et
new york
support
financial service
iais multilateral
environment natural
provide
information provided
multilateral memorandum
driver license
new brunswick
economic development
international authority
supervisory college
cooperation information
insurance california
applicable law
contracting state
california et
party may
party agree
michigan party
february revised
provision agreement
international crossing
regulated entity
regional body
information received
authority may
originating party
revised july
understanding cooperation
exchange information
iais mmou
mmou february
responding authority
department insurance
request assistance
third party
program
including
federal plaza
state maryland
development
exchange iais
project
work
activity
july page
extent possible
new increased
working group
necessary
new jersey
state maine
republic china
supervisory authority
collaboration
mutual interest
forest fire
louisiana
develop
area
side
set forth
responsible
water quality
professional secrecy
regional review
relationship
required
contact person
effective date
party agreement
either party
trade investment
consumptive use
withdrawal consumptive
hereinafter referred
plan
state new
one
york state
bridge
meeting
state province
commercial vehicle
emergency management
share information
relation
section
written notice
party jurisdiction
member state
event
school district
character
fnmb team
provide information
authority mean
governor state
motor vehicle
de la
fax email
agreed upon
business day
facility
following
appropriate
state illinois
use reasonable
establish
resource illinois
representative
best practice
vehicle inspection
water resource
united kingdom
page insurance
public private agreement
federal agency
michigan crossing
partnership
entity person
regulation requirement
enforcement illinois
michigan interchange
insurance company
consumptive use
state michigan
name title
annex b
well
term agreement
water conservation
attorney general
exception standard
tml fax
sister state
college member
lawrence river
additional
share
based
c
u
e

EDA


Top 20 Most Frequent Bigrams
Bigrams
united state
confidential information
memorandum understanding
et al
signatory authority
crossing authority
state america
natural resource
requesting authority
party shall
competent authority
authority shall
law enforcement
u federal
illinois et
new york
great lake
requested authority
responding authority
information exchange
0
200
400
600
800
1000
Frequency


Top 20 Most Frequent Tokens
Tokens
authority
state
party
shall
information
agreement
may
law
1
cooperation
insurance
2
mmou
article
section
request
department
confidential
international
water
0
1000
2000
3000
4000
5000
6000
Frequency

EDA


Average Word Length
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Average Word Length

3. Identifying agreement type

For identifying the agreement type we used “facebook/bart-large-mnli” model for “zero-shot-classification” task.

Insights after manual testing:

- Model has some flaws and is not always 100% accurate
- POC works, but we will try to improve its performance for the final version

```
file_name,Agreement Type
1June20.pdf,agreement
18102023.pdf,memorandum of understanding
1August12.pdf,Unknown
```

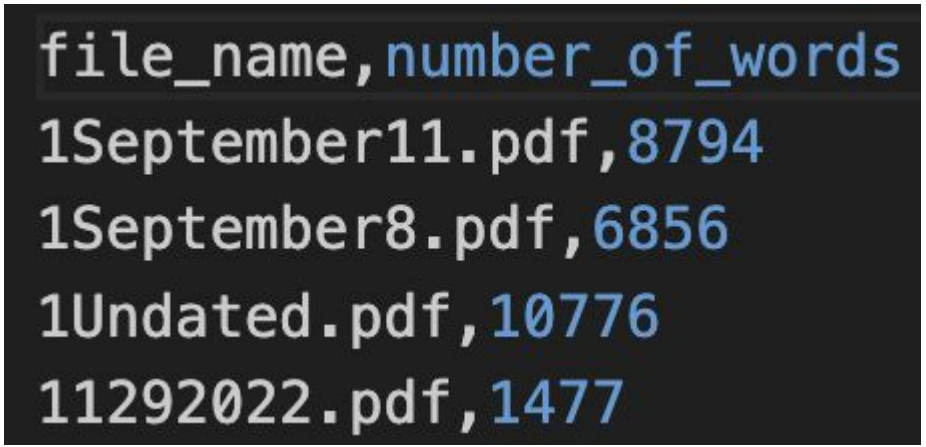
4. Determination of percentage of agreements under the patronage of Sister Cities International

- We used “facebook/bart-large-mnli” model for “zero-shot-classification” task.
- Model yields 22% of agreements as under the patronage of SSI, which is more or less matching the real value
- Also to be improved for the final version

```
file_name,Is Sister Cities International
1September11.pdf,False
1September8.pdf,False
1Undated.pdf,False
11292022.pdf,False
```

7. Identification of the length of each agreement

Because we had had a list of words for each agreement after data preprocessing, we haven’t used any ML model for this task, but simply counted the number of words



Phi-3-Mini-4K-Instruct

- lightweight, fast text generation model
- only 3.8B parameters, 2.18 GB
- trained on synthetic data and the filtered publicly available websites data with a focus on high-quality and reasoning dense properties.

Used for 4 tasks:

- identifying areas of cooperation
- extracting local and international organizations
- determining validity of the agreement
- identifying of whether the agreement includes an evaluation of its implementation

Spacy and KeyBert

Spacy:

- free, open-source library for advanced Natural Language Processing (NLP) in Python.
- suited for Named Entity Recognition
- fast but the results are far from perfect

KeyBert:

- minimal and easy-to-use keyword extraction technique
- leverages BERT embeddings to create keywords and keyphrases that are most similar to a document.

1. Identifying areas of cooperation

- keywords extracted with KeyBert
- LLM asked with identifying possible cooperation areas based on those keywords
- high variety of returned areas, preparing valid areas and classification would help standardize the results

cooperation_areas
['Energy Management', 'Clean Energy Innovation', 'New Energy Development']
['Great Lakes Environmental Protection', 'Basin Research Collaboration', 'Lake States Governance Coordination']
['Environmental Protection in Shanghai', 'Climate Environment Cooperation Shanghai-California', 'Ecological Collaboration between Shanghai and California']
['Inter-Border Coordination', 'Bilateral Conference', 'State Participation', 'Joint Meetings', 'Mexico States Collaboration']
['Goodwill Delegation Meetings', 'Joint Economic Cooperation', 'State-to-Province Collaboration']
['Technology Innovation', 'Energy Resources Development', 'Clean Technology Partnership']
['Cross-border Coordination', 'Border Infrastructure Development', 'Port Management and Entry Regulation']

2. Identifying organizations

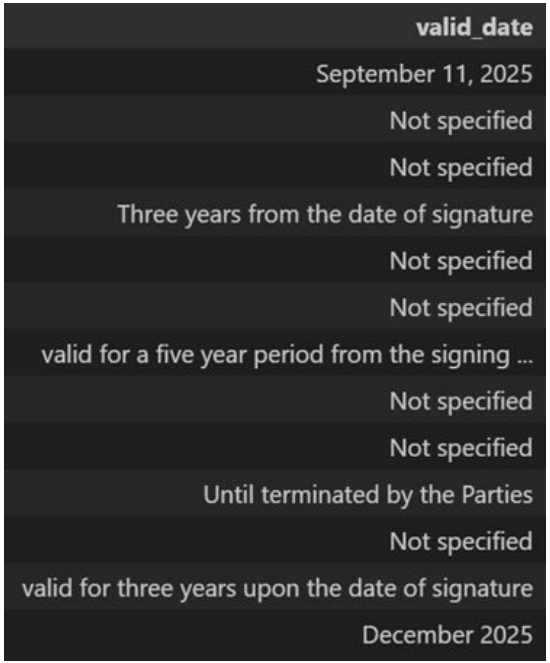
- Named entities identified with spacy
- many duplicates and lots of random noise
- LLM asked to recognize real organizations and discard noise
- output in JSON, sometimes incorrect format returned by LLM

[{'name': 'Govern of the State of California', 'abbreviation': 'GOV'}, {'name': 'Californias Office of Business and Economic Development', 'abbreviation': None}, {'name': 'Generalitat de Catalunya', 'abbreviation': 'GCAT'}, {'name': 'Catalonia Agency for Competitiveness Industry, Tourism, Knowledge and Innovation AGCC', 'abbreviation': 'AGCC'}, {'name': 'European Union Affairs Department of the Government of Catalonia', 'abbreviation': None}, {'name': 'Mobile World Congress', 'abbreviation': 'MWC'}]

[{'name': 'California Department of Food and Agriculture', 'abbreviation': 'CDFA'}, {'name': 'University of California, Division of Agricultural Sciences', 'abbreviation': None}, {'name': 'United States Department of Agriculture USDA', 'abbreviation': 'USDA'}, {'name': 'TriNational Agricultural Accord Meeting Organization', 'abbreviation': None}, {'name': 'Ministry of Natural Resources and Environmental Protection, Mexico', 'abbreviation': None}, {'name': 'Casa de California Foundation', 'abbreviation': None}]

6. Identifying extraction date

- all dates identified with spacy
- LLM fed with chunks containing the date and asked to recognize if the date is referring to agreement expiration date
- responses in both date and text format, require standardization



How we could measure level of detail in the formal agreements?

- Determining if all necessary legal elements and clauses are present requires deep domain knowledge
- Delicate diplomatic language that preserves national sovereignty
- Evaluating if timing-related details (such as specific plans for the cooperation) are sufficiently specified -> That’s why we’ve merged it with next task

Indication of whether the agreement includes an evaluation of its implementation

	evaluation_yes_no
0	Yes, there are specific actions mentioned in the document: 1. The Parties will meet from time to time (implied timeline) - "The present Memorandum of Understanding shall be terminated upon thirty days prior written notice by either Party." 2. Termination with 30-day advance notice - "either Party may at any time terminate this Memorandum of Understanding by giving thirty days prior notice in writing to the other Party." 3. Renewal for a maximum period (implied timeline) - "Notwithfering the above, either Party may at any time renew this Agreement...for up to five years from such expiration."
1	Yes. The agreement text includes specific provisions regarding the implementation evaluation through: - ITEM 1 MEETINGS AND DISCUSSION: "The Participants will meet from time to time in order to exchange information or views with respect to potential resource development projects." This implies regular evaluations and discussions. - DURATION: The agreement has a duration of two years, which can be renewed every 24 months after giving notice. It also allows for termination by either party providing prior written notice. These provisions ensure that the implementation is regularly reviewed and evaluated over time.
3	NO. There are no specific provisions for evaluating the implementation mentioned in this agreement text, such as timelines of meetings or events, review periods, or assessment criteria. The document focuses on reaffirming cooperation and expanding collaboration between Alberta and Alaska through a bilateral council without detailing evaluation mechanisms.
6	Yes, the agreement text includes specific provisions for evaluating its implementation. The actions mentioned in the document are: 1) Establishing a Bilateral Working Group on the Protection of Transboundary Waters with reporting and oversight responsibilities to Alaska's Lieutenant Governor (state level). 2) Developing reciprocal procedures for inviting interested government representatives, scientists in environmental assessments under provincial/state law. 3) Facilitating participation in permitting processes triggered by federal laws. 4) Sharing best practices on workforce development and training between Alaska and British Columbia. 5) Collaboration to promote marine transportation reliability and safety, including measures for accident prevention and spill consequences reduction. 6) Continued cooperation in emergency management mutual aid through the Pacific Northwest Emergency Management Arrangement (PNEMA). 7) Sharing information about infrastructure development and promoting increased travel/shipping between Alaska and British Columbia. 8) Exploring other areas for joint action, including natural resource development, fisheries, border management, trade & investment, climate change adaptation etc.

Checking if prolongation of agreement will be automatic or some actions needs to be taken

6	The extension of the agreement will require active approval from both parties involved. According to section II, paragraph 2(a) and (b), it is stated that "The Governors Office" in Alaska and "Intergovernmental Relations Secret As per section II, paragraph 2(c), "Officials appointed by the Governor and the Premier may negotiate jointly," which implies that for an extension of this agreement to take place without further decision-making, it would requi There are no specific mechanisms mentioned for determining when an extension is needed, but given that this MOU includes several commitments across various areas like transboundary waters protection, workforce develo In conclusion, while there are no explicit mechanisms for deciding when an extension is needed within the MOU itself, it can be inferred that a combination of regular assessments and approvals from designated officials wou
7	The extension of the agreement will require a simple majority vote from all members involved in the Northwest Wildland Fire Protection Agreement to renew or extend it beyond its current term. This is stated under Article X (S In case of a significant alteration that requires Congr fearing the agreement's continuity or changes, it would be necessary for Congress approval according to its terms stated under SEC.2 which allows any state to become p In conclusion, while there's no formal approval process required for an extension as per the given text (unless it involves significant changes), regular review and potential amendments through majority voting among member
8	The extension of the Memorandum of Understanding (MOU) between Japan Bank for International Cooperation (JBIC) and Alaska Department of Natural Resources will require active approval from both parties involved. This To make this determination, a decision-making mechanism should be established to review and discuss potential extensions before submitting an official request for approval from both parties involved: JBIC and Alaska DNR 1. Regular meetings between representatives of Japan Bank for International Cooperation (JBIC) and Alaska Department of Natural Resources to review progress, assess current needs, and discuss potential future collaborat 2. Formal reviews: Both parties should conduct formal evaluations every few years (e.g., 3 to 5) before deciding whether an extension is necessary and appropriate, based on factors such as project outcomes, financial consi 3. MOU extension requests: Once the decision is made that an extension would benefit ongoing projects or future collaborations, both parties should submit formal requests for approval through their respective channels (e.g 4. Acknowledgment: Upon receiving the extension request, both parties should acknowledge receipt promptly to avoid any misunderstandings or delays in processing requests for MOU extensions and terminations as demo In summary, an automatic extension of the agreement is not indicated; instead, active approval from both parties involved will be required before any renewal or termination decisions are made regarding their Memorandum o

Blackstone with spacy - Identification of whether the agreement refers to other legal documents.

- trained on UK law documents

Named-Entity Recogniser

The NER component of the Blackstone model has been trained to detect the following entity types:

Ent	Name	Examples
CASENAME	Case names	e.g. *Smith v Jones*, *In re Jones*, In *Jones'* case
CITATION	Citations (unique identifiers for reported and unreported cases)	e.g. (2002) 2 Cr App R 123
INSTRUMENT	Written legal instruments	e.g. Theft Act 1968, European Convention on Human Rights, CPR
PROVISION	Unit within a written legal instrument	e.g. section 1, art 2(3)
COURT	Court or tribunal	e.g. Court of Appeal, Upper Tribunal
JUDGE	References to judges	e.g. Eady J, Lord Bingham of Cornhill

Blackstone with spacy - Identification of whether the agreement refers to other legal documents.

- It’s detecting legal documents with specific syntax
  <year> <code/name>
- We should test different model for detecting full names for such documents

```
2012 CEAA 2012.,1
NA,0
NA,0
112 STAT.,1
NA,0
4 MK00024 .;10 MK00030 ELARM Sept. 30,2
NA,0
1949 INTERSTATE FOREST FIRE PROTECTION COMPACT,1
NA,0
1949 INTERSTATE FOREST FIRE PROTECTION COMPACT,1
```

Thank you for your attention