Topic Modelling of Canadian Regulations

By: Xin Zhao
Josephine Kamei

Abstract

Canadian regulations cover a wide range of departments and regulatory agencies and with the immense increases in regulations, topic modelling with Latent Dirichlet Allocation (LDA) was utilized in order to go through the regulations and find topic words that commonly occurred together with the topic. After preprocessing the data, topic modelling with LDA was applied to the Canadian regulations' dataset from The Justice Canada website. It was observed that there were 60 different topics out of the 2062 Canadian regulations. Topic modelling using LDA helped the reader get a quick overview of the entire dataset of Canadian regulations. From this, we were able to establish the topic labels using the commonly occurs words that frequent the topic label. Visualizations were generated in presenting the results because they helped to show the same results in different ways. This further enhanced the comprehension of regulations summary. The next proposed step for this project would be applying trend analysis. It could help to determine future models by finding patterns within historical and current data which would be relevant towards forecasting topic trends in future Canadian regulations.

Introduction

Canadian regulations cover a wide range of departments and regulatory agencies which include Transport Canada (regulations for air, marine and rail road transportation), Fisheries and Oceans Canada (regulations regarding fisheries and aquatic species) [1], and Health Canada which established The Food and Drugs Act (FDA) where the primary regulations governs the rules and regulations for the sale, health and safety of food, drugs and goods within Canada. [1,2,3].

With the immense increase in regulations, data mining has the ability identify patterns throughout massive amounts of documents and or datasets from which predictions could be made from the outcomes [4]. The use of data analytics and data mining continues to be prevalent within many government agencies throughout Canada where many sectors within the Government of Canada have utilized data mining and data analytics in order to ensure the importance of reporting and maintaining data quality [5]. In the United States (U.S.), the U.S. Food and Drug Administration (FDA) which is comparable to Health Canada and the Canadian Food Inspection Agency [6], the use of data mining has been increasing and expanding due to the massive surges in the report databases (at two million reports as of 2018) which not only limited to including the misuse, adverse effects, consumer and health care complaints regarding the product, to be quicker and more efficient regarding identifying potential safety issue of the products and to publicize them to the public [7]. Not only has data mining become more prevalent within government agencies, it has been used in the legal sector such as the High Court of Australia [8], used extensively in scientific publications, humanities and within the financial industries [9,10].

There are many techniques of data mining that can be utilized in order to analyze large amounts of documents to get topics extraction and topic modelling is included as one of those examples [11]. Topic modelling is able to go through vast amounts of data or documents and find topic words that commonly occur together with the topic. [12]. Latent Dirichlet Allocation

(LDA) tells us what topics are present in any given document by observing all the words in it and producing a topic distribution. From figure 1, **m** denotes the number of documents, **n** is the number of words, words in the document is shown as **w**, and **z** is the topic of the word in the document. Θ represents the topic distribution for each document, α controls per document topic distribution, and β controls per topic word distribution. The final result of the LDA model will be all the topics made of all the words with the probabilities that belongs to the topic [13].

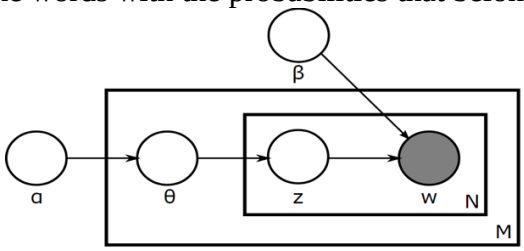

Figure 1: Latent Dirichlet Allocation (LDA) [13]

Applying the concept of topic modelling through LDA, the objective of this project is to show the different groups of regulations in Canada using the document dataset from Justice Canada [14]. Firstly, an analysis of all regulations will be implemented in order to create topic labels by using topic modelling. Once the regulations of the topics are identified, each document will be analyzed to find out which topics the document belongs to. Then the use of graphs and multiple visualizations to illustrate the regulations within each topic for easier, faster comprehension.

Related Work

There have been no past shortages of literature regarding the applications of topic modelling to get topics extraction from large amounts of documents. Duggirala *et al.* [7] described the different uses of data mining techniques used in order identify safety of goods, food and drugs/medicine regulated by the FDA. The authors chronicled the advantages and challenges of using data mining at the FDA which included how topic modelling was going to be applied with their tobacco documents. Despite this task being at the developmental stage as of 2018, the authors were planning to use this methodology to create a software program that uses topic modelling and implementing LDA to recognize topics that are located in distinct texts are currently in the works [7].

Topic modelling has been instrumental in being employed to examine an entire legal text corpus regarding the High Court of Australia dated from 1903-2015 [8]. The authors Carter *et al.* documented that to their knowledge that they were the first to apply topic modelling towards an entire legal textual corpus. The researchers noticed that there were fluctuations in the number of documents presented to the Court throughout the years and correlated that specific events had an immense impact on the workload of the judicial High Court. For example, when the commencement of the High Court was during the early stages, there was an increase in volume of texts, however during World War 2, there was a visible decrease in documents presumably due to diversion of attention to war related issues. Carter *et al.* focused on the viability, methodology and the value topic modelling provided when dissecting the substantial legal corpus. The authors concluded topic modelling showed the various subject matters that occurred throughout its timeframe. The results provided a predictive model for future legal textual corpus

and anticipated that their findings could be extrapolated towards other legal knowledge and or sectors.

Not only has application of topic modelling been applied in the legal text corpus, it has also been used in United Kingdom legislative documents. With the dramatic increases in legislative texts, lawmakers need to selectively choose which parliamentary legislations they need to attend while missing many other important sessions in the process. O'Neill *et al.* focused their research on using topic modelling and LDA for topic extraction from all the legislative documents and providing visualizations to make it easier to browse and identify specific legal topics making it much easier to for the lawmakers.

This project will derive a similar approach using the above authors where we will be using the dataset of regulations obtained from the Justice Laws website [1] where topic modelling using LDA will be accomplished. This will show all the different kinds topics from the regulations from which we will create topic labels from the topic words, and illustrate the findings using associated visualization.

Dataset

The dataset we used for this project contains a total of 2062 Canadian regulations downloaded from The Justice Canada FTP server (ftp://205.193.86.89/) [15]. Each regulation file was originally in XML format, and by using the parser created by Dr. Shariyar [16], we converted each regulation into a dictionary with instrument number, short title, long title, modified year, registration year, consolidated year and content.

Methodology

Data Preprocessing

In order to proceed with topic modelling, we first stored each regulation's content into a list. Then we tokenized each regulation content and get rid of the stop words, punctuations and special characters. For tokenization, we assigned part of speech (POS) tag to each word in the regulations by using Penn Treebank tags then converted all Adjectives, Verbs, Nouns, and Adverbs to WordNet Tags. This step is crucial for lemmatization because many words can have different meanings, for example the word tear can be a noun as the meaning of the tear in the eye and it can also be a verb as the meaning of ripping off something. By having different POS tags, the word will have different lemmas. After tokenization, we created a document term matrix for words occurred less than 7000 times, and more than 10 times. The document term matrix produced is an encoded vector with a length of 4290 words and each entry is the number of times each word appeared in each regulation.

LDA for Topic Modelling

The topic modelling technique that we used for this project was Latent Dirichlet Allocation (LDA). LDA assumes every document is a probability distribution of a topic, and every topic is a probability distribution of words. LDA looks at all the words in the document, and randomly assign each word in each document to a topic. Then LDA will go through every word and its topic assignment in each regulation, and look at how often the topic occurs in the document, and how often the word occurs in the topic. Based on this info, the model will create a new topic assignment for the word. The model will go through multiple iterations to find out the optimal topics. The output of the LDA model will be all the topics made of all the words with the probabilities that belongs to the topic [17].

The first thing we had to figure out was to find out the optimal number of topics. There are multiple ways to identify the number of topics. The way we used was to run LDA models with different number of topics, 40, 60, 80 and 100. For each model, its perplexity are calculated. More details about how the perplexity was calculated will be provided in Results & Evaluation section. By rule of thumb, if the model with the highest log likelihood and lowest perplexity before the changes flatten out is the best model [18]. After running all 4 models, we found that the model with 60 topics had the best result.

After the LDA model has trained on the regulation contents, we looped through all words in each topic to find out the topic 10 words with the highest probability in each model. Then we assigned each regulation with a topic by finding the index of the highest topic probability of the regulation. We also wanted to find out the top 20 regulations that belongs to each topic. We created a data frame to combine the regulation title, the topic, and the probability of the topic that belongs to the regulation, then used groupby function on topic column to find out the top 20 regulations with highest probabilities that belong to the topic.

Results & Evaluation

Since we wanted to provide accurate topic labels for the regulations, we did a lot of research by search the topic words in each topic and come up with 60 different topics for the 2062 Canadian regulations. The 60 topics we obtained from the LDA model is described in appendix 1.

In order to understand the output of LDA output, we used various visualizations including bar charts, word clouds and tables to provide insights of different topics for all regulations. The first visualization is a bar graph that shows the total number of regulations belong to each topic (Figure 2). As we can see most regulations belong to topic 31 (Canadian national service fees), topic 17 (bank insurance), topic 1(consumer goods), topic 51 (agricultural trade), topic 53 (pension plan) and topic 57(custom tariff).

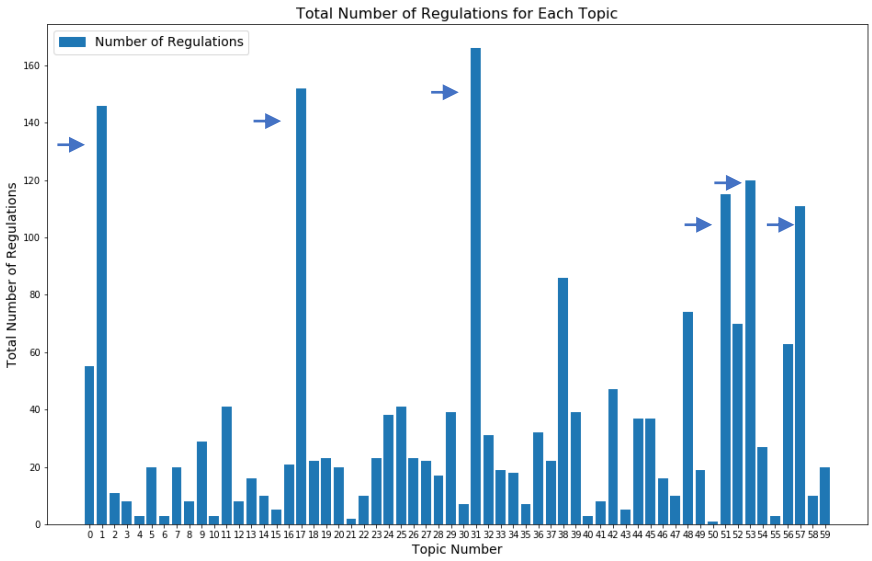

Figure 2: Bar graph depicting the total number of regulations belonging to each topic

Another effective visualization tool we used was word cloud. The word cloud tool will display the words in each topic and the higher the probability the word belongs to the topic the

bigger size the word will be shown. For illustration we chose 3 topics and by using word clouds to illustrate the topic words in each topic. As we can see for topic 17 bank insurance, some of the words with the highest probabilities are company, insurance, bank, and loan. For topic 28 cannabis licence application, the words with the highest probabilities are cannabis, licence, application, person, and holder. For topic 38 airport zoning, the words with the highest probabilities are airport, runway, zone, and surface, imaginary.

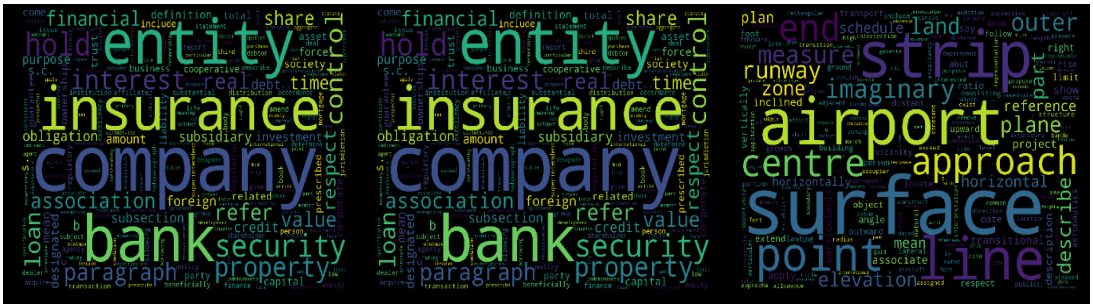

Figure 3: Word cloud displaying topics: 17,28,38

In Table 1, we illustrated how we can organize the result into a table by combining topic number& label, topic words and the top 3 regulations that belong to the topic. For topic 17 bank insurance, the topic words are company, bank, entity, insurance, security, hold, property, interest. And some of the regulations belong to this topic include Minority Investment, Regulatory Capital, Going-Private Transaction. For topic 28 Cannabis Licence Application**,** the topic words include cannabis, licence, person, holder, application, and some of the regulations belong to the topic include Cannabis Regulations, Access to Cannabis for medical purposes, and new classes of practitioners Regulation. For topic 38 Airport Zoning, the topic words include surface, strip, airport, line, point, approach, centre, end, imaginary and measure. The top 3 regulations belong to the topic are Regulations Respecting Zoning at Flin Flon Airport, Regulations Respecting Zoning at Hall Beach Airport, Regulations Respecting Zoning at Aklavik Airport.

Topic Number & Label	**Topic Words:**	**Regulation Title**
Topic 17: **Bank Insurance**	company, bank, entity, insurance, security, hold, property, interest, control, paragraph	• Minority Investment (Bank Holding Companies) Regulations • Regulatory Capital (Cooperative Credit Associations) Regulations • Going-Private Transaction (Banks and Bank Holding Companies) Regulations
Topic 28: **Cannabis Licence Application**	cannabis, licence, person, holder, name, refer, subsection, information, application, provide	• Cannabis Regulations • Access to Cannabis for Medical Purposes Regulations • New Classes of Practitioners Regulations
Topic 38 **Airport Zoning:**	surface, strip, airport, line, point, approach, centre, end, imaginary, measure	• Regulations Respecting Zoning at Flin Flon Airport • Regulations Respecting Zoning at Hall Beach Airport • Regulations Respecting Zoning at Aklavik Airport

Table 1: Selected topic number, labels and words with their regulation title

The bar graphs shown in Figure 4 provide insights of the topic changes from the time period of 2010 to 2018. In 2010 most of the topic were topic 17 bank insurance, because at that time, in the U.S, there was just a huge housing financial crisis. And on May 27th 2010, Minister of Finance proposed tighter regulation of insurance promotion by banks which further backs the result [19]. In 2018, we can see the most regulations were created in topic 28 cannabis licence application, and topic 45 food package labeling. Starting from October 17th, 2018 Cannabis act came into effect and cannabis became legalized [20], so there were a lot of regulation created for possession, acquisition and consumption of cannabis.

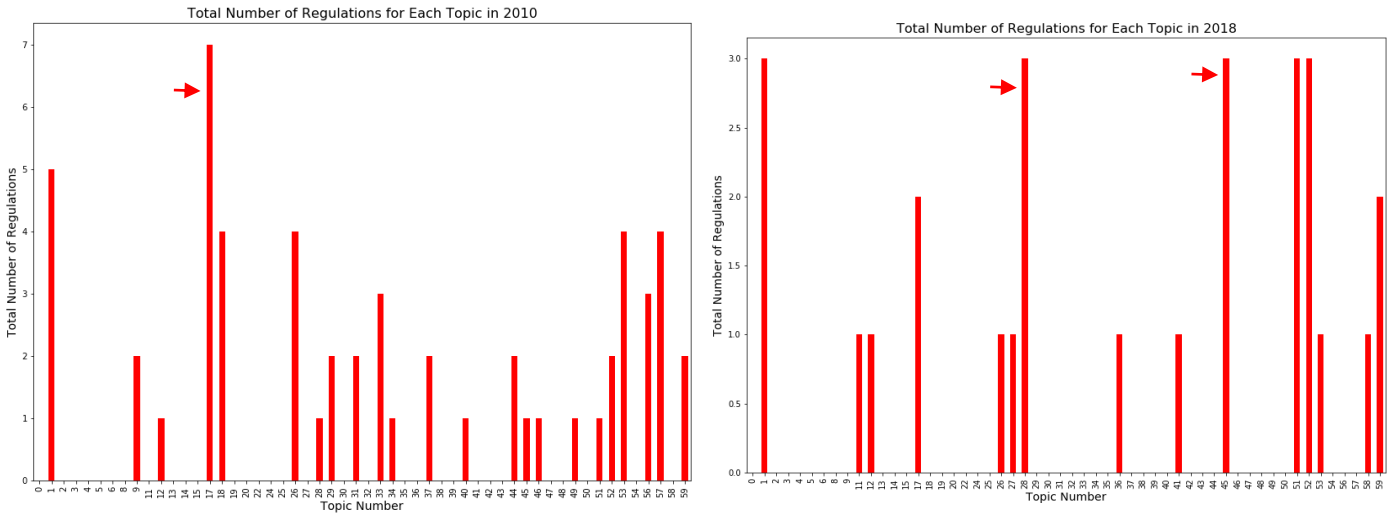

Figure 4: Bar graphs displaying the total number of regulations for each topic for 2010 and 2018.

In order to evaluate our LDA models, we calculated perplexity for each model. Perplexity measures the performance of a probability model [18]. For LDA topic modelling evaluation, the equation of how each model's perplexity is calculated is shown in figure 7. N denotes the total number of words in the model, p(xi) is the probability of each word belong to each topic multiply by the probability of the regulation of the word that belongs to each topic. For our LDA models, the perplexity of different topics are shown in table 2. The decrease of perplexity from 40 topics to 60 topics is significant and when the number of topics changed from 60 to 80 and 100, the decrease of perplexity seems to flatten out.

$$\exp\left(-\frac{1}{N}\sum_i \log p(x_i)\right)$$

Figure 5: Equation of perplexity

Topics	Perplexity	Difference
40	454	
60	417	-8.14%
80	397	-4.80%
100	386	-2.77%

Table 2: Evaluating models with different number of topics

Conclusion

Topic modelling using LDA helped the reader get a quick overview of the entire dataset of Canadian regulations. From this, we were able to establish topic labels using the commonly occurs words that frequent the topic label. However, we learned that in some instances, forming topic labels with the collection of words that was not easy to determine. For the most part, it was the difficulty in creating a topic label was due to the keyword being not present in the topic words. For topic 21, it the regulation was discussing railway safety, but the word train, railway was absent. Perhaps if we were familiar with the various kinds of regulations that are current via self-knowledge through the news, studied government and or Canadian law, a topic label might be easier to decipher. Visualizations were important in presenting the results because they helped show the same results in different ways through utilizing word clouds, bar graphs, and tables. This further enhanced the comprehension of regulations summary even easier and faster.

Future Direction of this Project

Since topic extractions was completed through topic modelling, the next proposed step for this project would be applying trend analysis. It could help determine future models by finding patterns within historical and current data [21,22] which would be relevant towards forecasting topic trends in future Canadian regulations.

References

[1] Consolidated federal laws of Canada, Consolidated Acts.
https://laws-lois.justice.gc.ca/eng/FAQ/#g1

[2] Policies, Regulations and Laws by Department or Agency
https://www.canada.ca/en/government/policy/dept.html

[3] Canada's Food and Drugs Act and Regulations
https://www.canada.ca/en/health-canada/services/food-nutrition/legislation-guidelines/acts-regulations/canada-food-drugs.html

[4] Data Mining Concepts
https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON002

[5] Health Canada Detailed Action Plan
http://www.ourcommons.ca/content/Committee/421/PACP/WebDoc/WD8148750/Action_Plans/51-HealthCanada-e.pdf

[6] U.S. & Canada have comparable food safety systems: CFIA and FDA
https://www.foodincanada.com/food-in-canada/u-s-canada-comparable-food-safety-systems-cfia-fda-133960/

[7] Duggirala, H.J., Tonning, J.M., Smith, E., Bright, R.A., Baker, J.D., Ball, R., Bell,C., Bouri, K., Bright-Ponte, S.J., Botsis, T., Boyer, M., Burkhart, K., Condrey, G.S., Chen, J.J., Chirtel, S., Filice, R.W., Francis, H., Jiang, H., Levine, J., Martin, D., Oladipo, T., O'Neill, R., Palmer, L.M., Paredes, A., Rochester, G., Sholtes, D.,Wong, H., Xu, Z., Szarfman, A., and T. Kass-Hout. 2018. Data Mining at FDA. U.S. Food and Drug Administration.
https://www.fda.gov/scienceresearch/dataminingatfda/ucm446239.htm

[8] Carter, D.J., Brown, J., and A. Rahmani. 2016. Reading the high court at a distance: topic modelling the legal subject matter and judicial activity of the high court of Australia, 1903-2015. UNSW Law Journal. Vol.39, No.4

[9] Boyd-Graber, J., Hu, Y., and D. Minmo. 2017. Applications of topic models. Founds and Trends in Information Retrieval. Vol. 11, No.2 pp.1-158.

[10] Hurtado, J.L., Agarwal, A., and X. Zhu. 2016. Topic discovery and future trend forecasting for texts. Journal of Big Data. Vol. 3, No.1

[11] Karl, A., Wisnowski, J., and W.H. Rushing. 2015. A practical guide to text mining with topic extraction. WIRESs Computational Statistics. Vol. 7, No. 5

[12] Alghamdi, R., and K. Alfalqi. 2015. A survey of topic modeling in text mining. (IJACSA) International Journal of Advanced Computer Science and Applications. Vol.6, No.1

[13] Latent Dirichlet Allocation (LDA) from Wikipedia
https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[14] Food and Drug Regulations.
https://laws-lois.justice.gc.ca/eng/regulations/C.R.C.,_c._870/index.html

[15] Consolidated federal laws of Canada, Consolidated Regulations. Retrieved from
https://laws-lois.justice.gc.ca/eng/regulations/

[16] XML Python Parser downloaded from D2L

[17] Blei, D., Ng, A., Jordan, M. 2003. Latent dirichlet allocation. Journal of machine Learning research, pp.993-1022.

[18] Goodman, J. T. 2001. A bit of progress in language modeling. Computer Speech & Language, Vol. 15, No. 4 pp.403-434

[19] Department of Finance. 2010. Archived - Minister of Finance Proposes Tighter Regulation of Insurance Promotion by Banks. https://www.fin.gc.ca/n10/10-052-eng.asp

[20] Consolidated federal laws of Canada, Cannabis Act.
https://laws-lois.justice.gc.ca/eng/acts/C-24.5/

[21] Hwang, M-H., Ha, S., In, M., and K. Lee. 2018. A method of trend analysis using Latent Dirichlet Allocation. International Journal of Control and Automation. Vol. 11, No. 5 pp.173-182

[21] O'Neill, J.O., O'Brien, L., Robin, C., and P. Buitelaar. 1999. An analysis of Topic Modelling for Legislative Texts. In Proceedings of 2nd Work on Automated Semantic Analysis of Information in Legal Texts (ASAIL'17)

Appendix 1

Topic number, words and labels

Topic	Topic Words	Topic Labels
0	loan, amount, agreement, cost, board, veteran, interest, director, respect, rate	Veteran's land regulations
1	force, come, day, subsection, good, b, register, purpose	Consumer goods
2	system, installation, fire, platform, space, equipment, area, mean, standard, production	Offshore platform safety
3	type, column, failure, violation, set, item, prescribed, provision	Monetary penalties
4	explosive, unit, storage, use, licence, store, person, site, magazine, certificate	Explosives license
5	subsection, latitude, longitude, thence, north, west, point, area, marine	Marine conservation
6	diving, operation, dive, use, person, diver, supervisor, certificate, safety, category	Oil and gas operation
7	free, item, tariff, january, part, good,article, schedule, following	Tariff on goods
8	vehicle, engine, emission, year, model, subsection, class, standard, company, case	Vehicle emissions
9	person, security, property, designated, resolution, canadian, mean, date, prohibit, paragraph	United nation resolution
10	southern, northern, western, eastern, concession, boundary, road, front, southeastern, northwestern	Concession boundary
11	application, land, territory, indian, lease, northwest, subsection, area, order, respect	Territorial lands application
12	well, oil, gas, operator, production, ensure, equipment, operation, installation, mean	Oil and gas production
13	metre, lot, direction, limit, part, measure, hundred, food, drug, parcel	Food additives
14	standard, product, column, csa, manufacture, set, test, table, mean, information	CSA standards and information
15	claim, mine, mining, work, cost, year, mineral, subsection, permit, respect	Mining permits
16	tank, storage, foot, use, ship, less, track, building, subsection, person	Ship storage regulations
17	company, bank, entity, insurance, security, hold, property, interest, control, paragraph	Bank insurance
18	vessel, certificate, master, voyage, column, board, requirement, set, respect, person	Vessel certificate requirements

19	good, verification, origin, officer, law, person, custom, reference, paragraph, visit	Verification of goods origin
20	project, authority, assessment, review, decision, environmental, subsection, member, conduct, report	Environmental assessment report
21	end, inch, car, side, less, specify, schedule, brake, minimum, ladder	Workforce car usage
22	railway, company, employee, safety, local, service, work, include, refer, process	Railway safety process
23	product, board, person, agency, commodity, sale, order, regulated, sell, transport	Sale and transport of products
24	security, ferry, facility, number, domestic, member, force, police, service, person	Domestic ferry security
25	party, application, hearing, person, document, appeal, provide, division, notice, rule	Immigration and refugee
26	information, document, electronic, use, purpose, provide, form, record, name, consent	Electronic information of consent document
27	firearm, know, individual, design, version, code, prohibit, model, modify, criminal	Firearm prohibitions for criminals
28	cannabis, licence, person, holder, name, refer, subsection, information, application, provide	Applications for cannabis license
29	water, specie, use, waste, fish, activity, deposit, park, fishery, substance	Fisheries activity
30	point, island, water, follow, line, begin, beginning, true, northerly, light	Sea geographical orders
31	canadian, order, service, fee, respect, pay, national, administration, claim, title	Canadian national service fees
32	person, port, transportation, transport, nuclear, failure, authority, safety, material, carrier	Transport of nuclear materials
33	tax, property, supply, subsection, person, respect, amount, province, part, service	Property tax
34	land, first, nation, right, document, register, interest, registration, record, registry	First nation land registration
35	subheading, change, heading, good, value, material, head, chapter, use, method	Rules of origin
36	corporation, share, facility, canadian, carrier, subsection, voting, hold, office, development	Telecommunication development
37	licence, licensee, province, power, time, provincial, work, designate, crown, purpose	Loan protection
38	surface, strip, airport, line, point, approach, centre, end, imaginary, measure	Airport zoning

39	lot, thence, corner, limit, surface, say, boundary, concession, westerly, li ne	International Airport zoning
40	particular, plan, institution, financial, year, period, investment, amount, fiscal, day	Corporate tax
41	subsection, refer, emission, x, establishment, year, method, determine, u nit	Method of emissions
42	export, permit, import, country, good, exporter, state, united, republic	Permit of import and export and of goods
43	category, —, mixture, data, substance, hazard, classify, product, chemica l, concentration	Classify hazardous chemical product concentration
44	service, public, group, period, employment, department, person, support, month, category	Public service employment support
45	product, food, case, use, set, package, mm, label, apply	Apply product labels to food
46	commission, request, file, person, complaint, address, name, day, docum ent, tobacco	Commission of tobacco
47	day, annuity, spouse, judge, partner, benefit, former, common-law, electi on, period	Judges annuity benefits
48	applicant, officer, member, application, subsection, day, director, person, notice, name	Council member election
49	export, product, allocation, primary, quebec, producer, volume, quantity, softwood, lumber	Softwood lumber export
50	subdivision, legal, road, allowance, order	Species at Risk
51	board, producer, order, market, marketing, mean, levy, interprovincial, c ommodity, trade	Agricultural trade
52	bay, organization, immunity, thunder, privilege, convention, internationa l, vancouver, article, order	Privileges and immunities of foreign organizations
53	amount, year, pay, service, subsection, pension, person, payment, period, day	Yearly pension plan payment
54	thence, approximate, longitude, latitude, straight, tissue, organ, line, proc lamation, foot	Designated states
55	system, pump, ship, inspection, item, machinery, pipe, control, space, pre ssure	Ship machinery inspections
56	bank, deposit, association, institution, member, person, notice, account, b ranch, service	Banking services
57	january, custom, remission, duty, import, order, tariff, good, respect, dec ember	Tariff on import of goods

58	operation, activity, area, operator, geophysical, animal, habitat, officer, person, data	Geophysical activities and operations
59	person, pipeline, licence, board, number, locomotive, report, activity, fac ility, name	Energy board pipeline