August
2021

Finding Semantically Similar Legislation Across Jurisdictions

Transjurisdictional Transformers

Neil Yap

Outline

1. Problem Statement
2. Project Workflow
3. Embedding Methods
4. Results Analysis
5. Moving Forward
6. Reflections and Lessons

Glossary

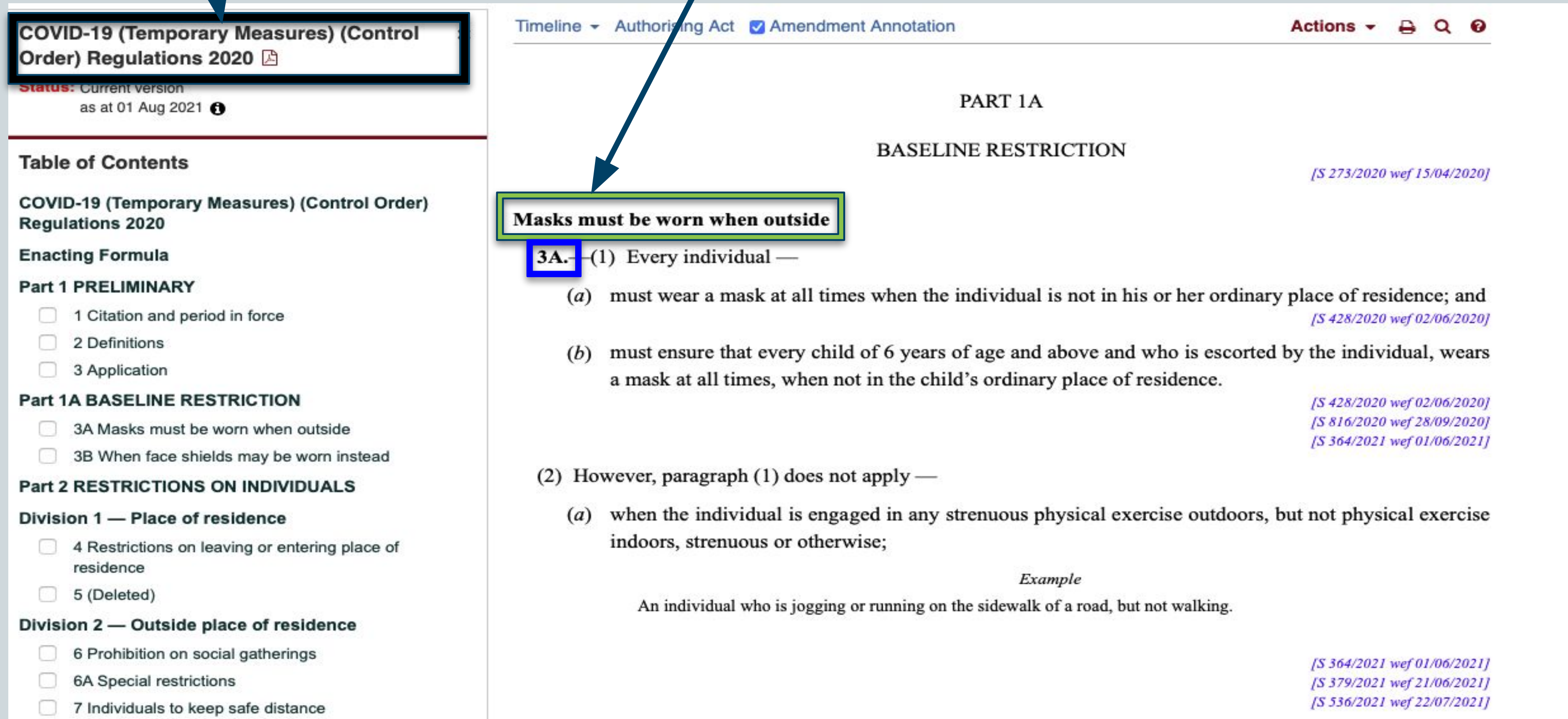
Legislation/Regulation/Act ≈ Piece of Law

Provision/Section ≈ Smaller Piece of the Piece of Law

Jurisdiction ≈ Country

Legislation/Act/Regulation

Title (green box),
Section/Provision Number (blue box)



COVID-19 (Temporary Measures) (Control Order) Regulations 2020

Status: Current version
as at 01 Aug 2021

Table of Contents

PART 1A

BASELINE RESTRICTION

[S 273/2020 wef 15/04/2020]

Masks must be worn when outside

3A.—(1) Every individual —

(*a*) must wear a mask at all times when the individual is not in his or her ordinary place of residence; and

[S 428/2020 wef 02/06/2020]

(*b*) must ensure that every child of 6 years of age and above and who is escorted by the individual, wears a mask at all times, when not in the child's ordinary place of residence.

[S 428/2020 wef 02/06/2020]
[S 816/2020 wef 28/09/2020]
[S 364/2021 wef 01/06/2021]

(2) However, paragraph (1) does not apply —

(*a*) when the individual is engaged in any strenuous physical exercise outdoors, but not physical exercise indoors, strenuous or otherwise;

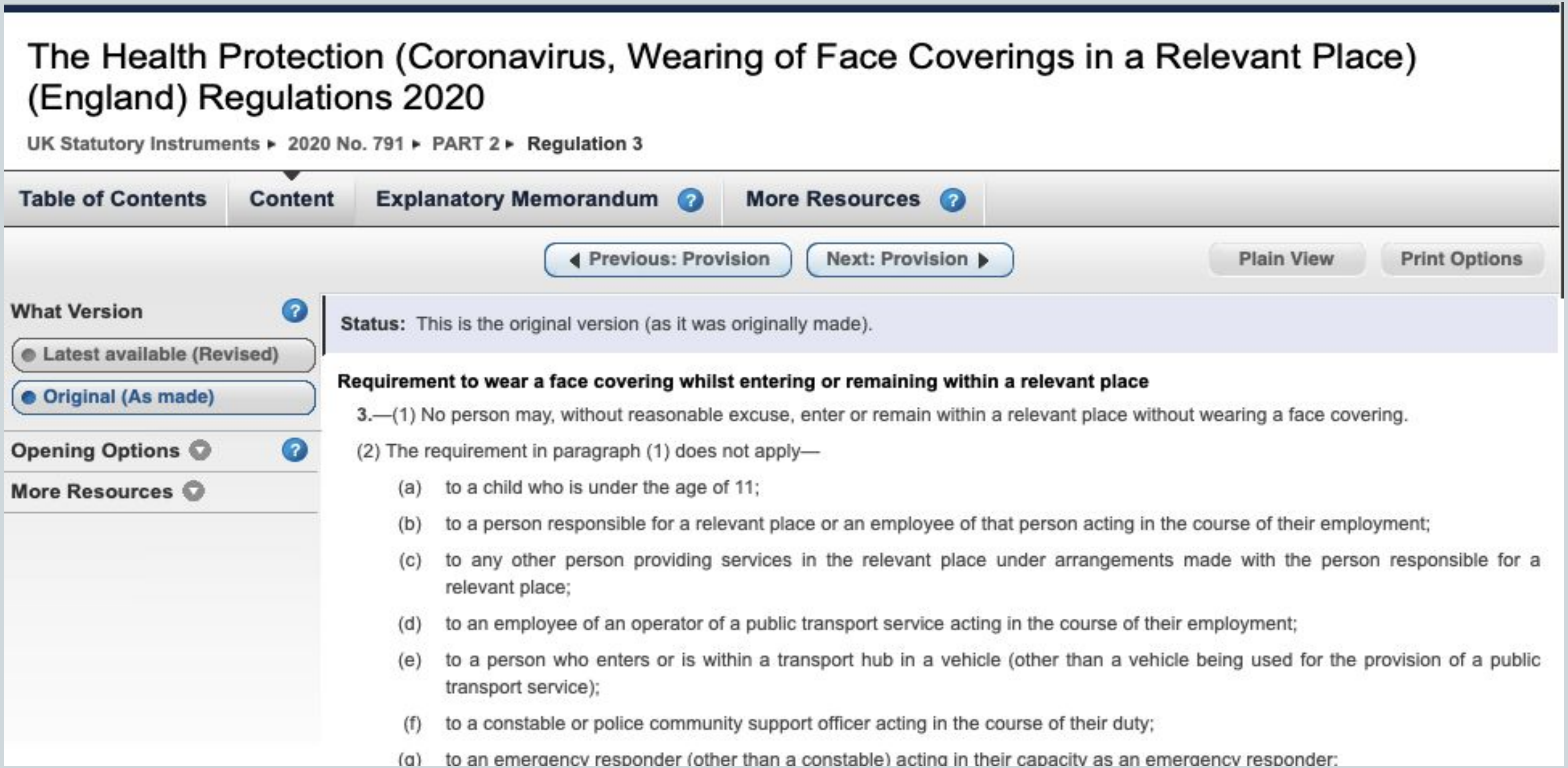
Example

An individual who is jogging or running on the sidewalk of a road, but not walking.

[S 364/2021 wef 01/06/2021]
[S 379/2021 wef 21/06/2021]
[S 536/2021 wef 22/07/2021]

Jurisdiction = Singapore

Example of Closest UK Equivalent Provision



The Health Protection (Coronavirus, Wearing of Face Coverings in a Relevant Place) (England) Regulations 2020

UK Statutory Instruments ▸ 2020 No. 791 ▸ PART 2 ▸ **Regulation 3**

Status: This is the original version (as it was originally made).

Requirement to wear a face covering whilst entering or remaining within a relevant place

3.—(1) No person may, without reasonable excuse, enter or remain within a relevant place without wearing a face covering.

(2) The requirement in paragraph (1) does not apply—

(a) to a child who is under the age of 11;

(b) to a person responsible for a relevant place or an employee of that person acting in the course of their employment;

(c) to any other person providing services in the relevant place under arrangements made with the person responsible for a relevant place;

(d) to an employee of an operator of a public transport service acting in the course of their employment;

(e) to a person who enters or is within a transport hub in a vehicle (other than a vehicle being used for the provision of a public transport service);

(f) to a constable or police community support officer acting in the course of their duty;

(g) to an emergency responder (other than a constable) acting in their capacity as an emergency responder;

Jurisdiction = United Kingdom

Problem Statement

Given a Singapore legislation provision, how might a lawyer quickly view the *closest equivalent provision* of another jurisdiction?

Project Workflow

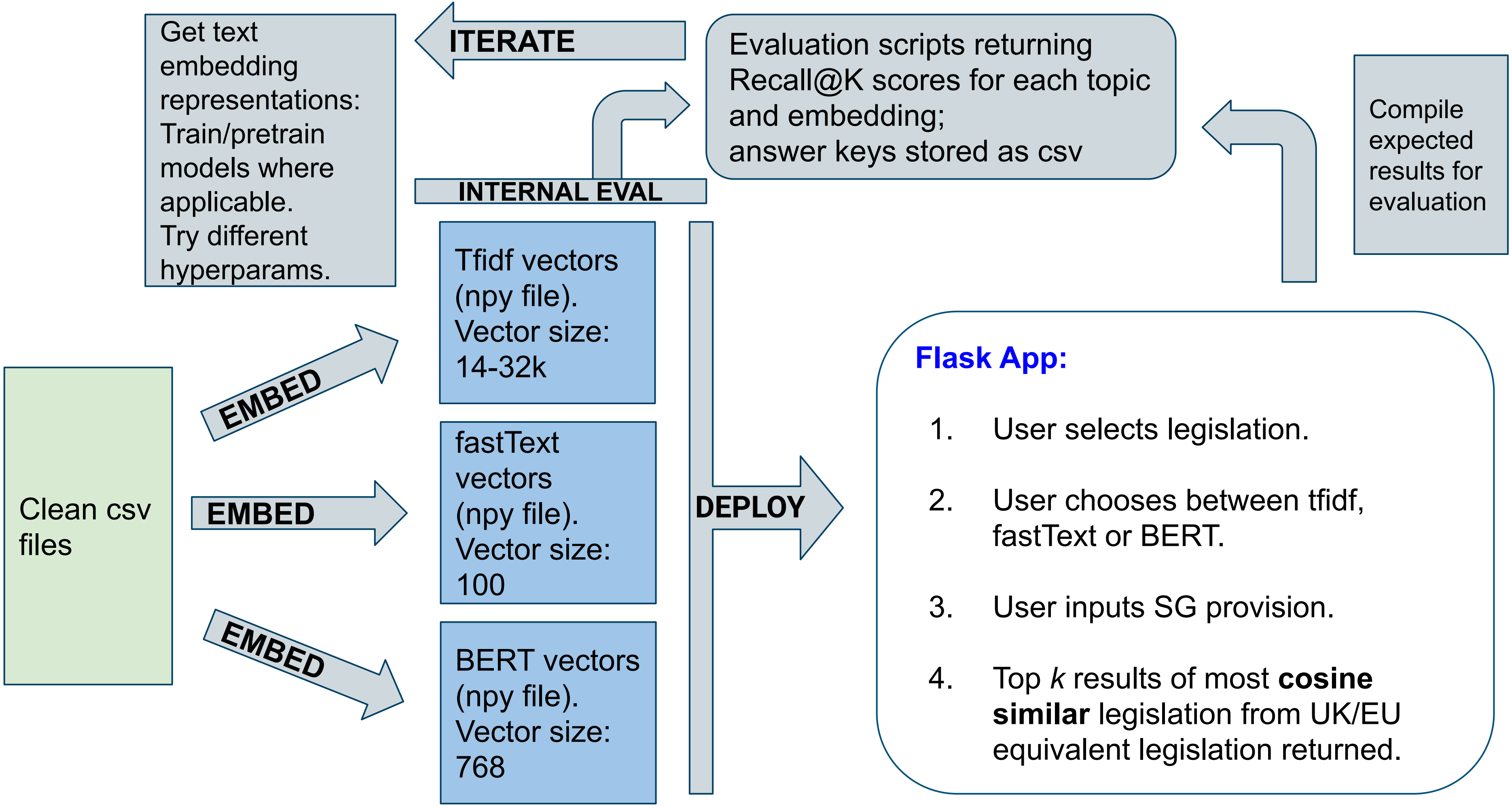
Workflow Outline

1. Crawl and clean data for these 3 legal topics:
    a. Copyright Legislation (SG & UK): 700+ entries
    b. Trade Mark Legislation (SG & UK): 200+ entries
    c. Data Protection Legislation (SG & EU): 180+ entries

2. Get embeddings (vectors saved in npy file). For each legal topic get:
    a. Tfidf
    b. fastText
    c. BERT

3. Get top k provision matches based on cosine similarity

4. Evaluate based on Recall@3 with: self-assembled answer key; answers from user testing



Get text embedding representations: Train/pretrain models where applicable. Try different hyperparams.

ITERATE

Evaluation scripts returning Recall@K scores for each topic and embedding; answer keys stored as csv

Compile expected results for evaluation

INTERNAL EVAL

Clean csv files

EMBED

EMBED

EMBED

Tfidf vectors (npy file). Vector size: 14-32k

fastText vectors (npy file). Vector size: 100

BERT vectors (npy file). Vector size: 768

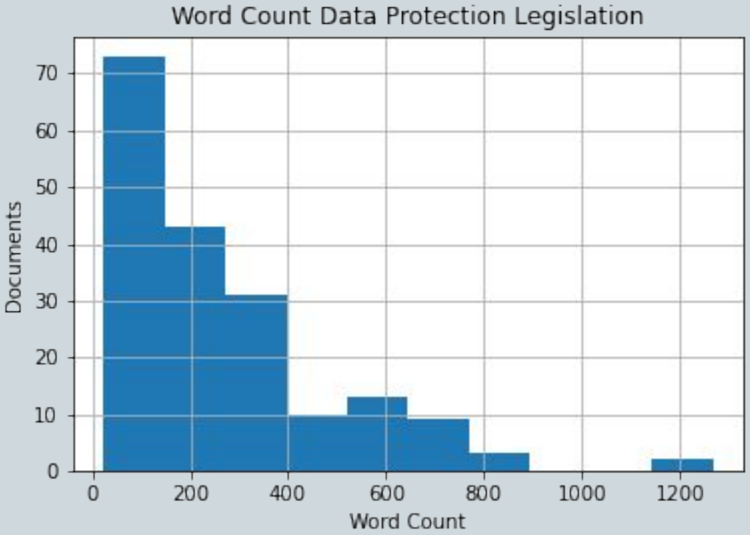
DEPLOY

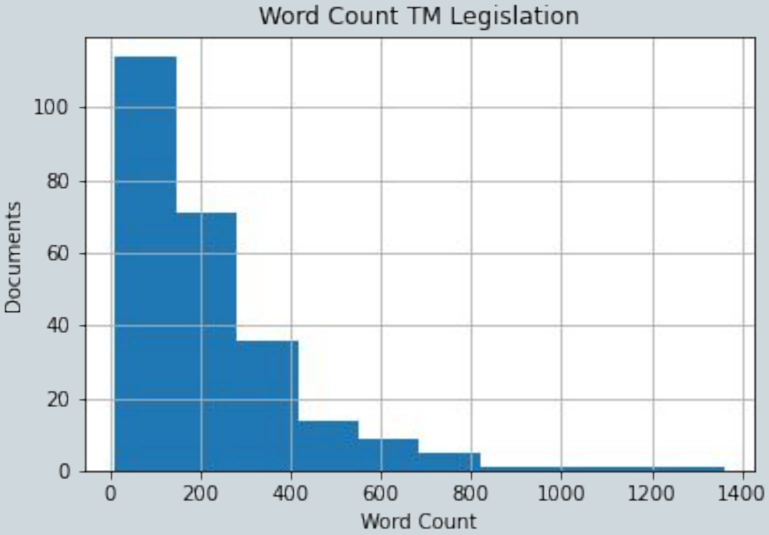
Flask App:

1. User selects legislation.
2. User chooses between tfidf, fastText or BERT.
3. User inputs SG provision.
4. Top *k* results of most **cosine similar** legislation from UK/EU equivalent legislation returned.

Embedding Methods

General Sense of Word Counts





Not short snippets, but not super long either. BERT max token window is 512.

Implementation

Embedding	Python Library	Hyperparams Chosen	Train+Embed Time 200-800 examples on CPU	Embedding Size per Entry
tfidf	sklearn	ngram_range=(1,2), max_df=0.95	< 1 min	About 14-36k (vocab size)
fastText	fasttext	skipgram training, 2 wordngrams*	< 1 min	100*
BERT	transformers by huggingface; pytorch	MLM pre-training, 512 token window, 4 epochs, lr of 3e-4, batch size 8	Pretrain model to dataset of 200-800 entries: 2-3 hours	768

* different hyerparams for data protections set: (**dim=50**, lr=0.0001, epoch=50, minn=6, minCount=3, ws=10, model='skipgram', wordNgrams=3)

Implementation

Embedding	Python Library	Hyperparams Chosen	**Train+Embed Time** **200-800 examples on CPU**	Embedding Size per Entry
tfidf	sklearn	ngram_range=(1,2), max_df=0.95	< 1 min	About 14-36k (vocab size)
fastText	fasttext	skipgram training, 2 wordngrams*	< 1 min	100*
BERT	transformers by huggingface; pytorch	MLM pre-training, 512 token window, 4 epochs, lr of 3e-4, batch size 8	Pretrain model to dataset of 200-800 entries: 2-3 hours	768

* different hyerparams for data protection set: (**dim=50**, lr=0.0001, epoch=50, minn=6, minCount=3, ws=10, model='skipgram', wordNgrams=3)

fastText

- Similar to word2vec, learn representations with
  - CBOW
  - Skipgram
- Character level ngram representations: preserves subword information
  - data → <da, dat, ata, ta>
- Relatively fast but can be memory intensive
- Known to need a lot of training data to see results (1000 docs is little)

Enriching Word Vectors with Subword Information (Bojanowski et al., 2016)

BERT (Bidirectional Encoder Representations from Transformers)

- Contextual representations
  - Positional encodings
- Transformers: Self-attention
- Pretrain to adjust to domain. Methods:
  - MLM (masked language model)
  - NSP (next sentence prediction)
- BERT base has 12 hidden layers
  - Used last hidden layer to get word representations

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)

Masked Lang Model

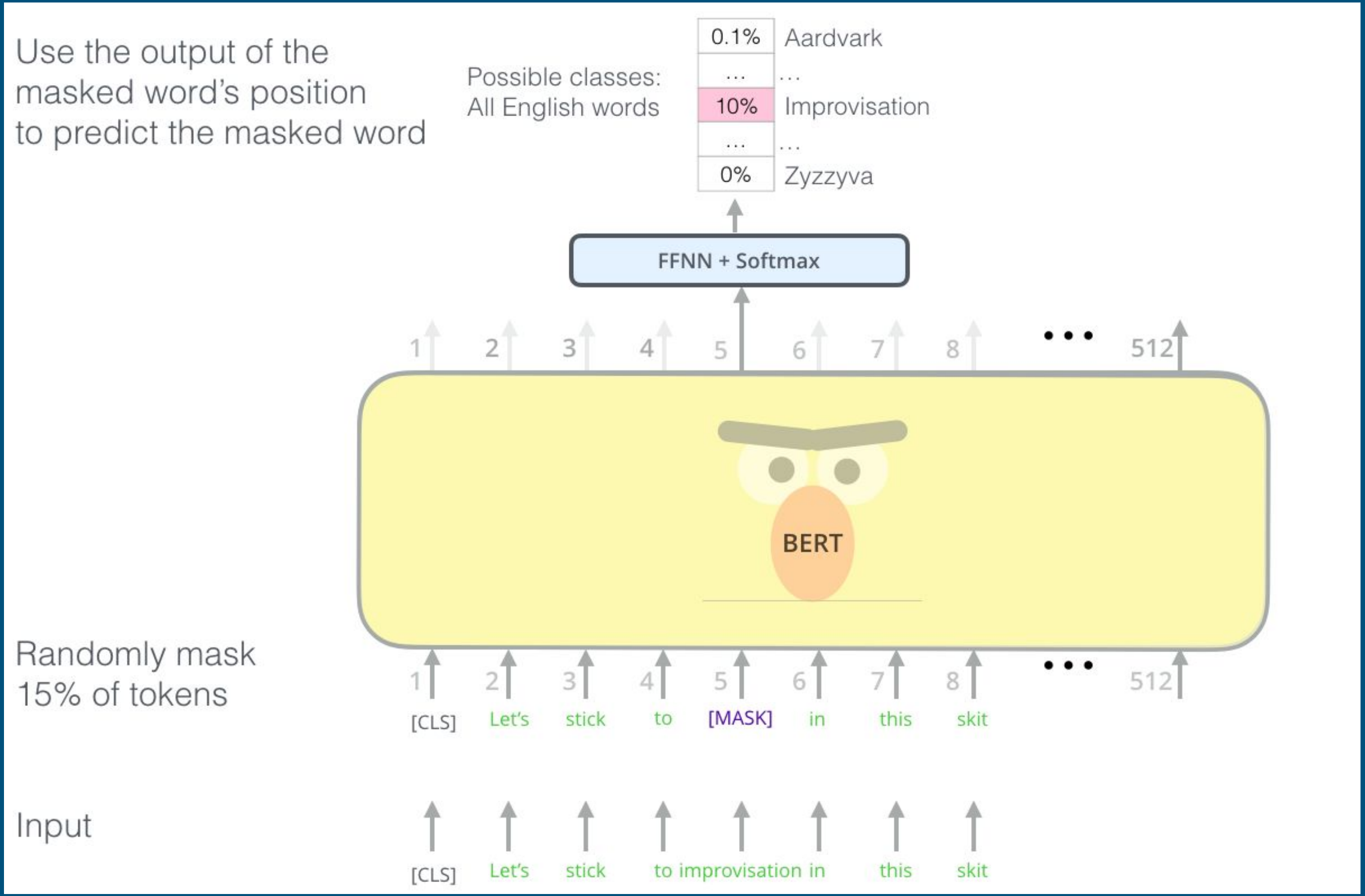


Image from: https://jalammar.github.io/illustrated-bert/

Results & Analysis

Performance

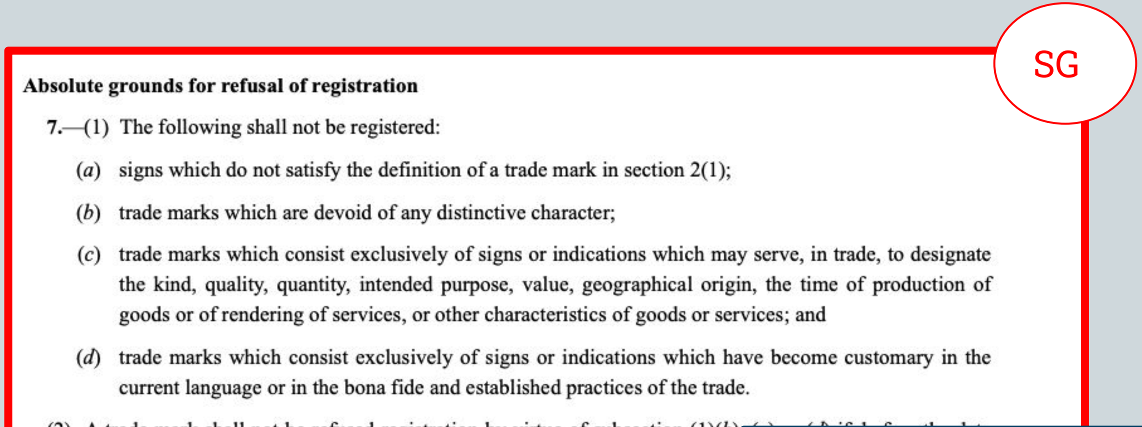
Method	Copyright Recall @ 3 20 Examples	Trade Mark Recall @ 3 12 Examples	Data Protection Recall @ 3 12 Examples
Title Edit Distance (Baseline)	0.35	0.83	0.17
tfidf (cos sim)	0.8	1	0.33
fastText (cos sim)	0.4	0.33	0.25
BERT (cos sim)	0.55	0.83	0.25

Analysis

- tfidf performs best

- However, expectedly, tfidf only works well with very similarly worded input and targets

- In 1/12 of the less similarly worded examples, BERT got the correct result over tfidf
    - some potential for BERT to catch ‘harder cases’
    - but need more data and tests to see if this is really the case



SG

Absolute grounds for refusal of registration

7.—(1) The following shall not be registered:

(*a*) signs which do not satisfy the definition of a trade mark in section 2(1);

(*b*) trade marks which are devoid of any distinctive character;

(*c*) trade marks which consist exclusively of signs or indications which may serve, in trade, to designate the kind, quality, quantity, intended purpose, value, geographical origin, the time of production of goods or of rendering of services, or other characteristics of goods or services; and

(*d*) trade marks which consist exclusively of signs or indications which have become customary in the current language or in the bona fide and established practices of the trade.

(2) A trade mark shall not be refused registration by virtue of subsection (1)(*b*)

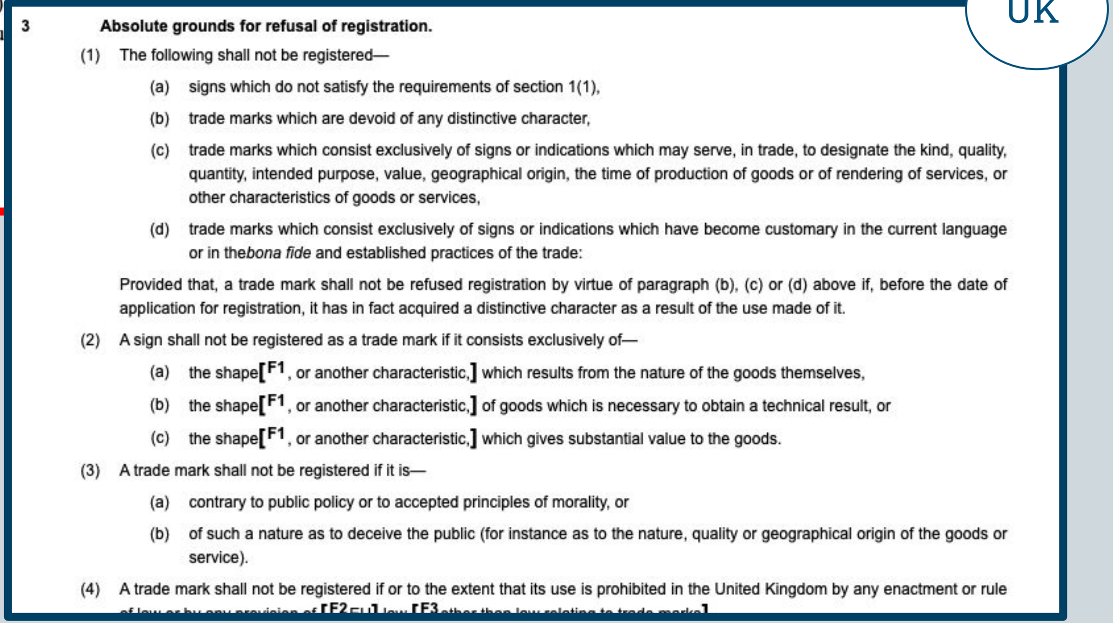
of application for registration, it has in fact acquired a distinctive character as a resu

(3) A sign shall not be registered as a trade mark if it consists exclusively of —

(*a*) the shape which results from the nature of the goods themselves;

(*b*) the shape of goods which is necessary to obtain a technical result; or

(*c*) the shape which gives substantial value to the goods.



UK

3 Absolute grounds for refusal of registration.

(1) The following shall not be registered—

(a) signs which do not satisfy the requirements of section 1(1),

(b) trade marks which are devoid of any distinctive character,

(c) trade marks which consist exclusively of signs or indications which may serve, in trade, to designate the kind, quality, quantity, intended purpose, value, geographical origin, the time of production of goods or of rendering of services, or other characteristics of goods or services,

(d) trade marks which consist exclusively of signs or indications which have become customary in the current language or in thebona *fide* and established practices of the trade:

Provided that, a trade mark shall not be refused registration by virtue of paragraph (b), (c) or (d) above if, before the date of application for registration, it has in fact acquired a distinctive character as a result of the use made of it.

(2) A sign shall not be registered as a trade mark if it consists exclusively of—

(a) the shape[F1, or another characteristic,] which results from the nature of the goods themselves,

(b) the shape[F1, or another characteristic,] of goods which is necessary to obtain a technical result, or

(c) the shape[F1, or another characteristic,] which gives substantial value to the goods.

(3) A trade mark shall not be registered if it is—

(a) contrary to public policy or to accepted principles of morality, or

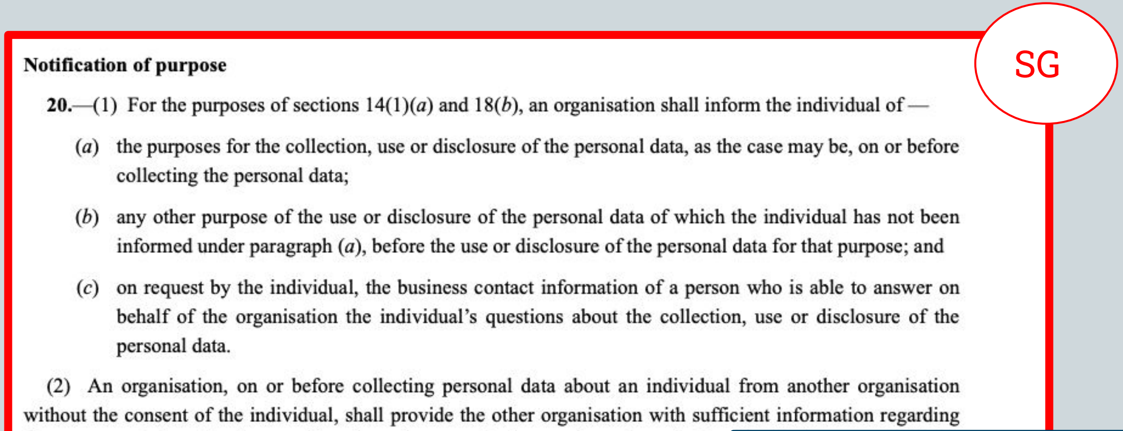
(b) of such a nature as to deceive the public (for instance as to the nature, quality or geographical origin of the goods or service).

(4) A trade mark shall not be registered if or to the extent that its use is prohibited in the United Kingdom by any enactment or rule of law or by any provision of [F2EU] law [F3other than law relating to trade marks]

Easy!

Baseline can catch

tfidf very strong



SG

Notification of purpose

20.—(1) For the purposes of sections 14(1)(*a*) and 18(*b*), an organisation shall inform the individual of —

(*a*) the purposes for the collection, use or disclosure of the personal data, as the case may be, on or before collecting the personal data;

(*b*) any other purpose of the use or disclosure of the personal data of which the individual has not been informed under paragraph (*a*), before the use or disclosure of the personal data for that purpose; and

(*c*) on request by the individual, the business contact information of a person who is able to answer on behalf of the organisation the individual's questions about the collection, use or disclosure of the personal data.

(2) An organisation, on or before collecting personal data about an individual from another organisation without the consent of the individual, shall provide the other organisation with sufficient information regarding the purpose of the collection to allow that other organisation to determine whether

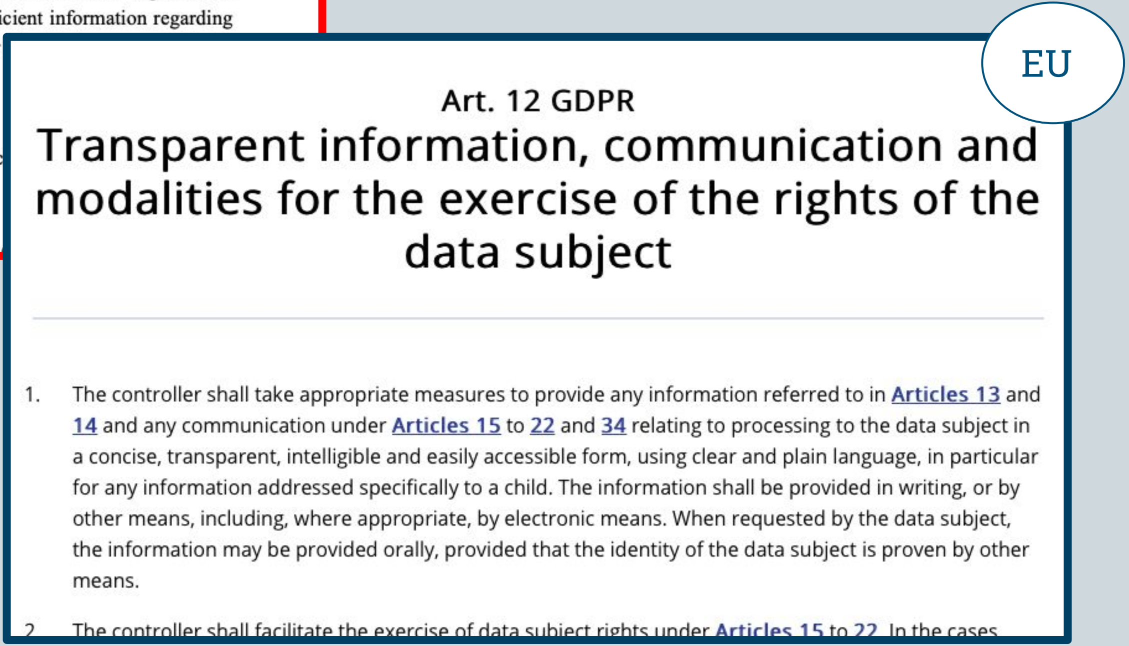
accordance with this Act.

(3) Subsection (1) shall not apply if —

(*a*) the individual is deemed to have consented to the collection, use or disc

under section 15 or 15A; or

(*b*) the organisation collects, uses or discloses the personal data without the



EU

Art. 12 GDPR

Transparent information, communication and modalities for the exercise of the rights of the data subject

1. The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means. When requested by the data subject, the information may be provided orally, provided that the identity of the data subject is proven by other means.

2. The controller shall facilitate the exercise of data subject rights under Articles 15 to 22. In the cases

Hard!

BERT right

tfidf wrong

Conclusion as of August 2021

- This tool can benefit lawyers needing to quickly map very similarly worded legislation, and can be rapidly built with tfidf.

- For more challenging situations (less similarly worded legislation), more work needs to be done to explore the potential of more sophisticated embeddings, but they do not work so well “out of the box” and with less than 1000 examples of training data.

Moving Forward

Next Steps

- Try out more fastText hyperparams. Try pretraining with more external data of related legal topics.

- Dive deeper into error analysis. Look at wrong examples and examine examples that models predicted differently.

- Try on more examples of less similarly worded legislation.
  - Less ‘keyword’ based
  - More challenging for tfidf, see if BERT can outperform in those situations

- Consider stacking models with some rules based hierarchy
  - E.g. tfidf -> BERT

Lessons

Data Science Workflow Learnings

- Much better sense of of how ‘out of the box’ the NLP models are for a specific domain. Transfer learning advancements in NLP are promising, but not yet a magic pill.

- Data labelling is the most expensive both time wise and expertise wise, likewise for qualitative error analysis. (Duh! But worth repeating.)

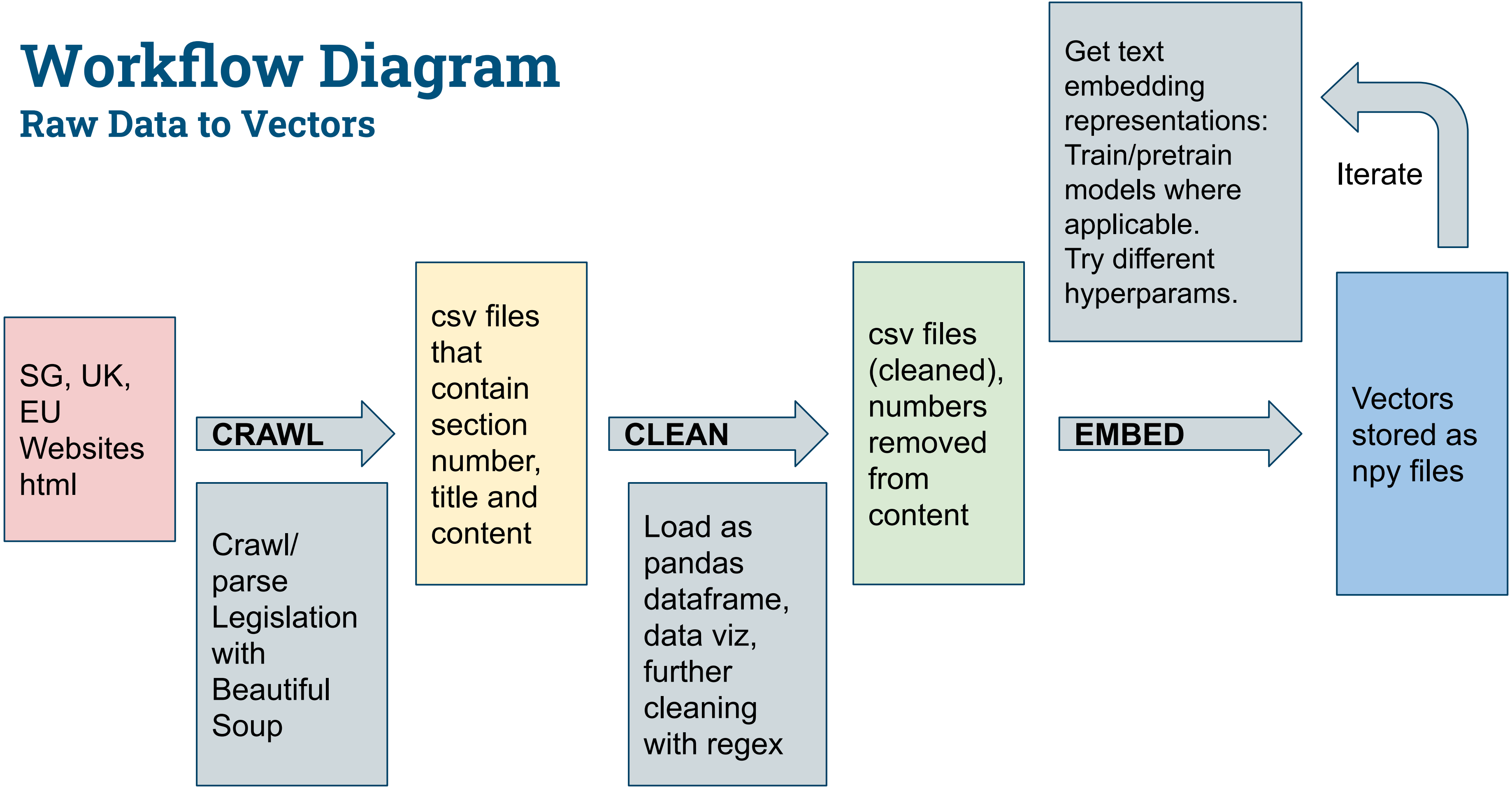
- Do not get discouraged by low numbers, the real world is not Kaggle, always compare to some realistic baseline to see the potential in your solution.

Thank You!
Congrats DSI 22!

After Credits

Workflow Diagram

Raw Data to Vectors



SG, UK, EU Websites html

CRAWL

Crawl/ parse Legislation with Beautiful Soup

csv files that contain section number, title and content

CLEAN

Load as pandas dataframe, data viz, further cleaning with regex

csv files (cleaned), numbers removed from content

EMBED

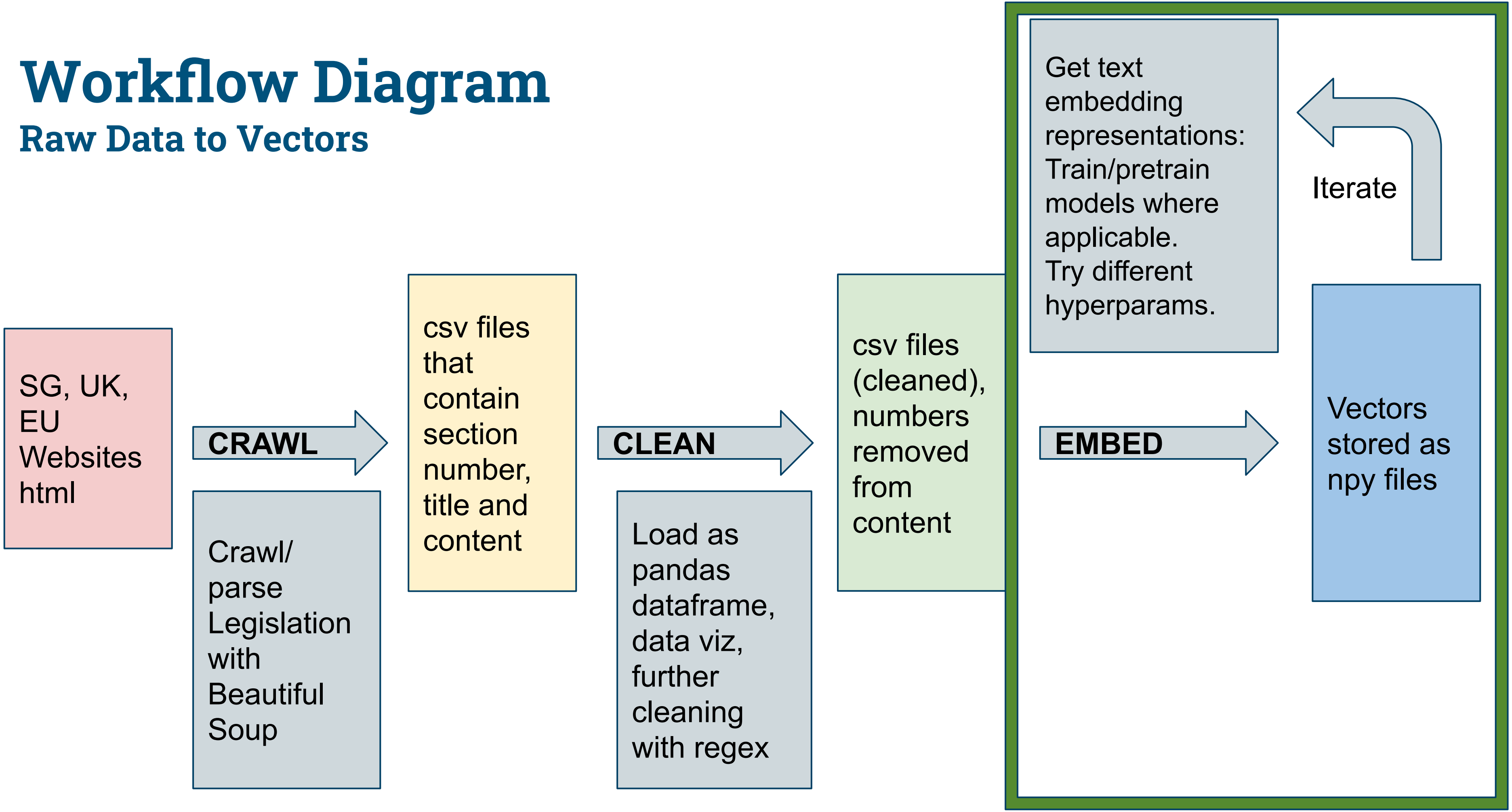
Get text embedding representations: Train/pretrain models where applicable.
Try different hyperparams.

Vectors stored as npy files

Iterate

Workflow Diagram

Raw Data to Vectors



SG, UK, EU Websites html

CRAWL

Crawl/ parse Legislation with Beautiful Soup

csv files that contain section number, title and content

CLEAN

Load as pandas dataframe, data viz, further cleaning with regex

csv files (cleaned), numbers removed from content

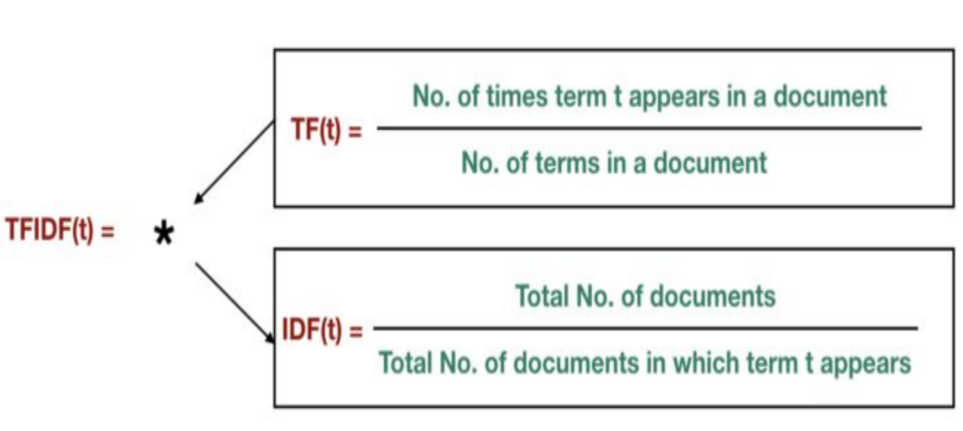
EMBED

Get text embedding representations: Train/pretrain models where applicable. Try different hyperparams.

Iterate

Vectors stored as npy files

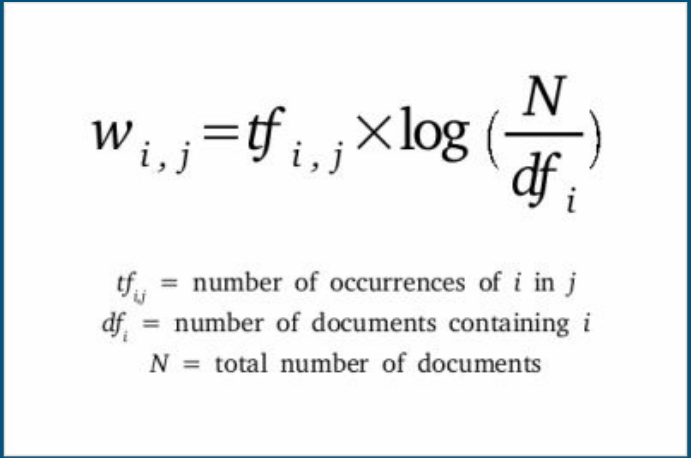
tfidf



$$\text{TFIDF(t)} = \text{TF(t)} * \text{IDF(t)}$$

$$\text{TF(t)} = \frac{\text{No. of times term t appears in a document}}{\text{No. of terms in a document}}$$

$$\text{IDF(t)} = \frac{\text{Total No. of documents}}{\text{Total No. of documents in which term t appears}}$$



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Image from:
https://medium.com/swlh/sentiment-classification-for-restaurant-reviews-using-tf-idf-42f707bfe44d

Image from:
http://www.digitalmarketingchef.org/what-is-tfidf-and-why-does-it-matter-for-seo/