N-GrAM: New Groningen Author-profiling Model

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim

Overview

Meet the Team

Task and Data

Our approach

Data insights

Conclusion

MEET THE TEAM

During and after writing

Malvina Nissim

(Head honcho)

Hessel Haagsma
Masha Medvedeva

(PAN Veterans)

Gareth Dwyer
Josine Rawee
Angelo Basile

(PAN Newbies)

TASK AND DATA

Task and data

Twitter data:

- ~100 tweets/ author
- 600 authors / variety

Language	Varieties	Authors
Arabic	4	2400
English	6	3600
Portuguese	2	1200
Spanish	7	4200

Task and data

Twitter data:

- ~100 tweets/ author
- 600 authors / variety

Language	**Varieties**	**Authors**
Arabic	4	2400
English	6	3600
Portuguese	2	1200
Spanish	7	4200

Gender and **Language Variety** profiling

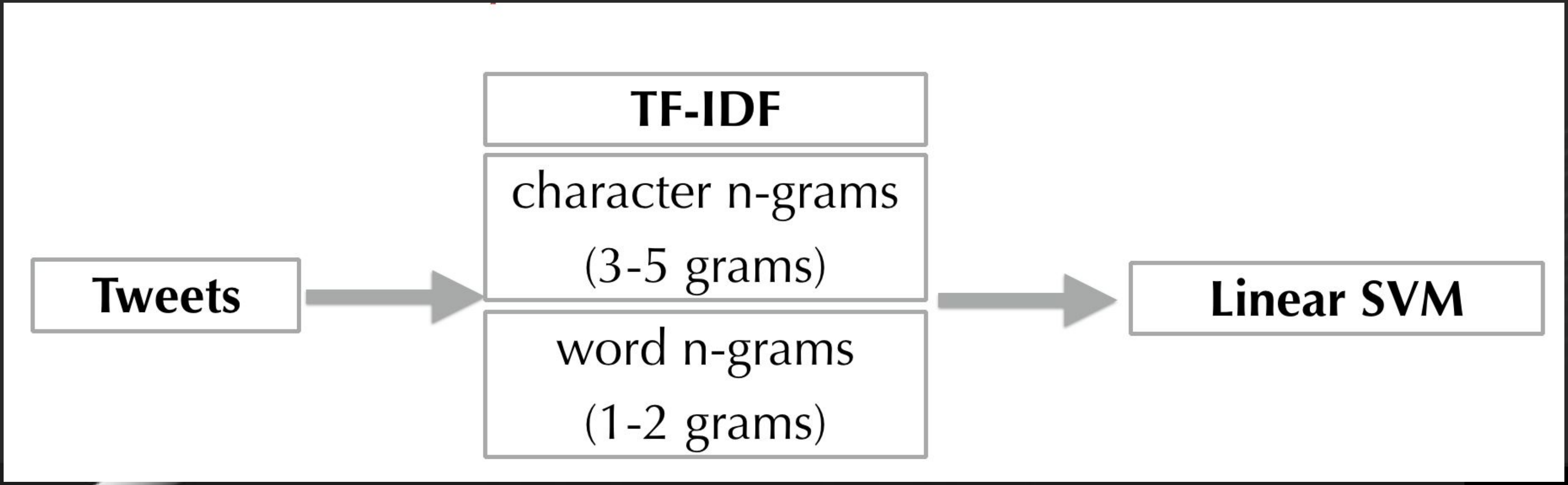
- Is the author Male or Female?
- What language variety are they using?

OUR APPROACH

N-grams + SVM

Start with basic system

- Word and Character n-grams
- TF-IDF
- Linear Support Vector Machine



Gotta use it all

Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset

Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset

More features

- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset

More features

- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers

- Fast Text, Decision Trees, Neural Networks

More data is better data!

Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset

More features

- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers

- Fast Text, Decision Trees, Neural Networks

More Data

Adding data from previous pan years

- Train on 2016, test on 2017
- Vice versa
- :(

More Data

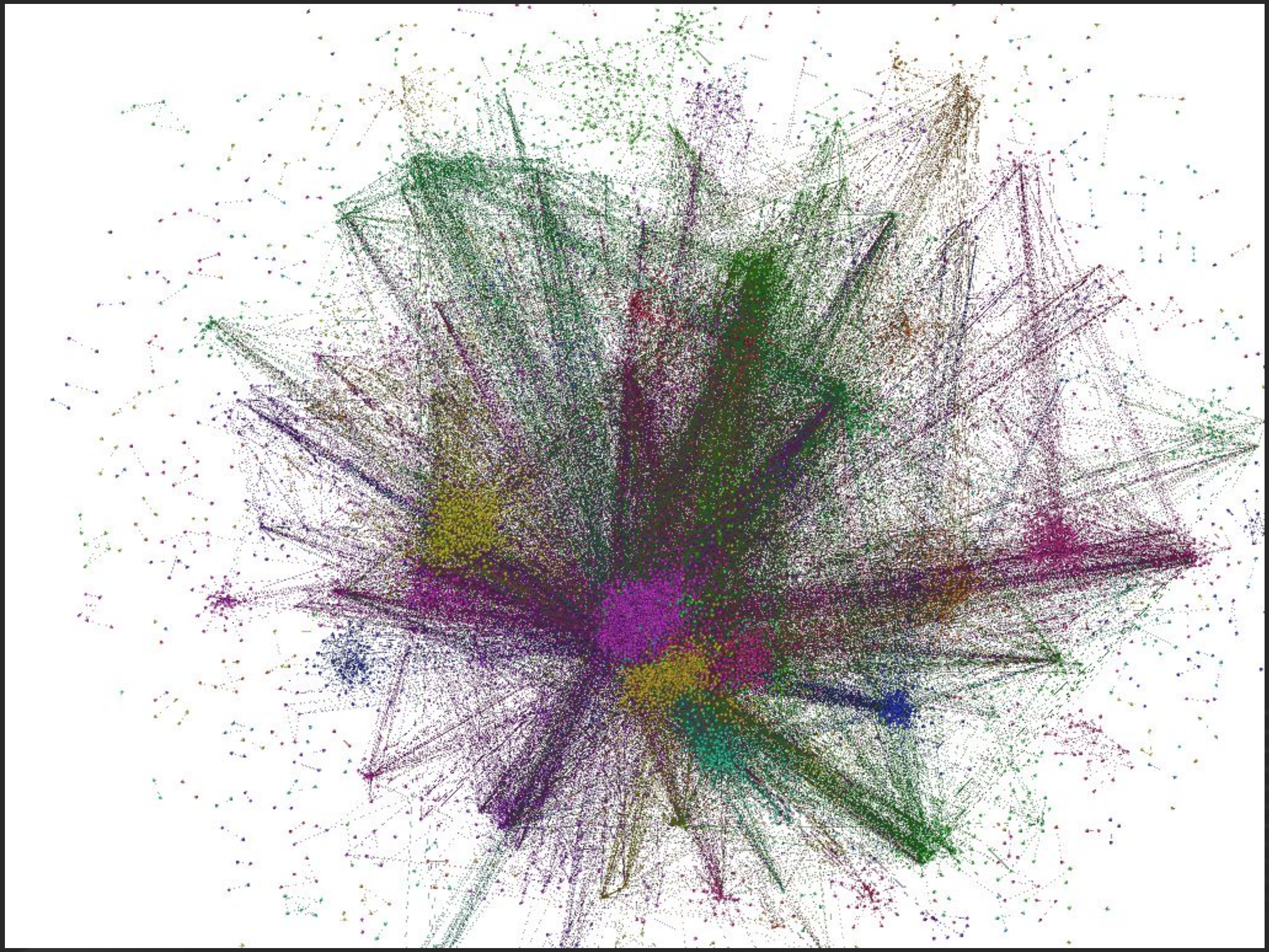
Adding data from previous pan years

- Train on 2016, test on 2017
- Vice versa
- :(

Add Twitter 14k dataset

- Typically ‘male’ and ‘female’ words
- :(

Adding features will help!



Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset

More features

- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers

- Fast Text, Decision Trees, Neural Networks

More Features

Tokenizers

- TweetTokenizer (NLTK)
- Happy Fun Tokenizer (emoticons)
- :(

More Features

Tokenizers

- TweetTokenizer (NLTK)
- Happy Fun Tokenizer (emoticons)
- :(

POS Tags

- :(

More Features

Tokenizers

- TweetTokenizer (NLTK)
- Happy Fun Tokenizer (emoticons)
- :(

POS Tags

- :(

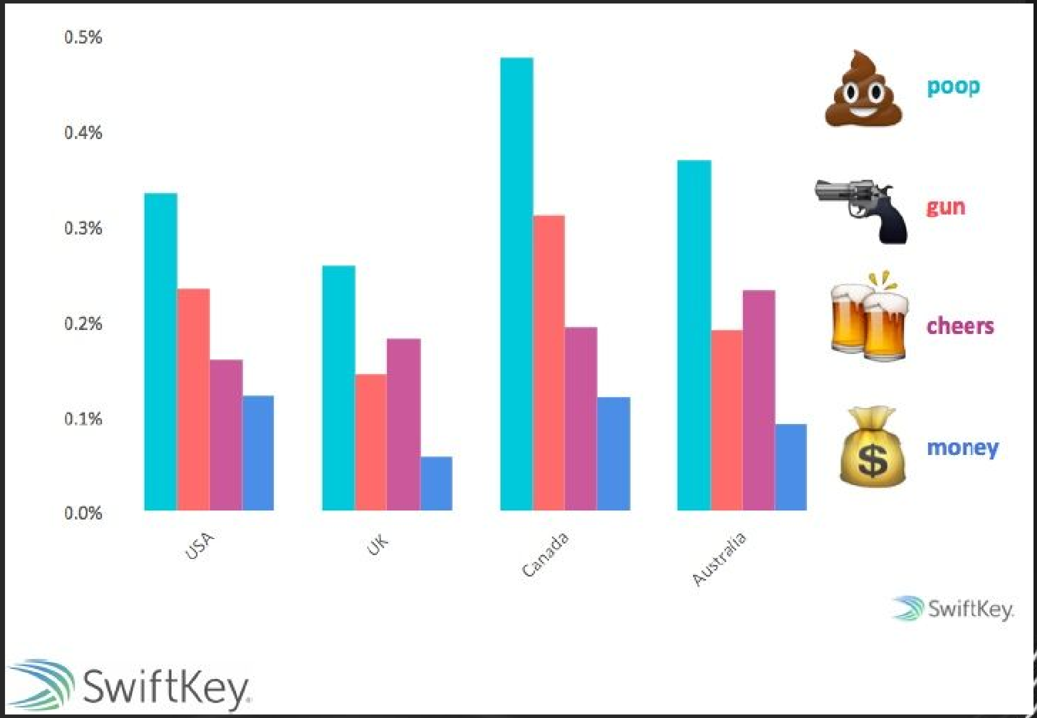
Twitter Handles + Place names (Variety)

- Collect corpus of associations with common towns/ handles
- :(

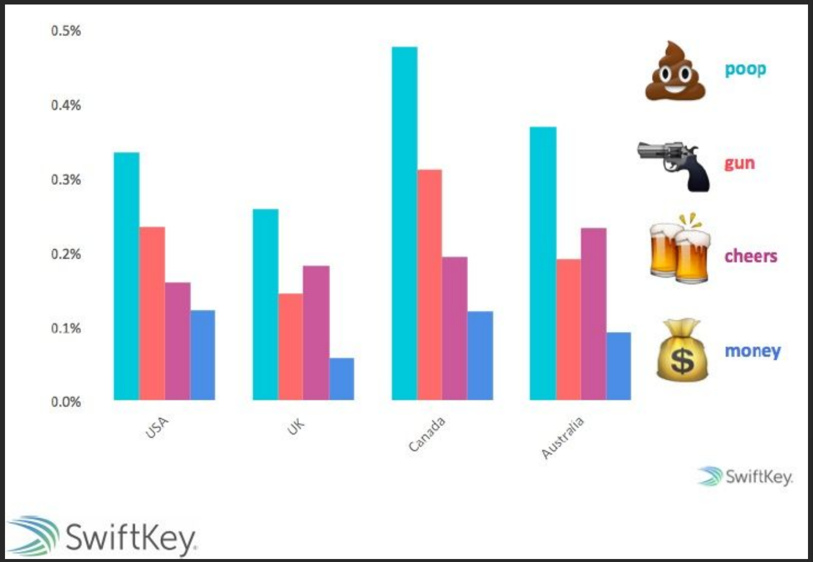
More Features (2)

Emoji

- SwiftKey report
- :(



More Features (2)



Emoji

- SwiftKey report
- :(

GronUP

- Punctuation, word length, capitals, vocabulary, etc, etc
- :(

Gotta use it all

More data

- Previous PAN data
- Twitter14k dataset

More features

- Tokenizers
- POS tags
- Twitter Handles + Place Names
- Emojis

More classifiers

- Neural Networks (!!!!!)

More Classifiers

FastText

- It’s fast!
- :(

More Classifiers

FastText

- It’s fast!
- :(

scikit-learn MLP

- Not so fast
- :(

More Classifiers

FastText

- It’s fast!
- :(

scikit-learn MLP

- Not so fast
- :(

Keras

- Had fun with generative models
- :(

MORE MORE MORE

Data

Features

Classifiers

MORE MORE MORE

Data

Features

Classifiers

MORE MORE MORE

Data

Features

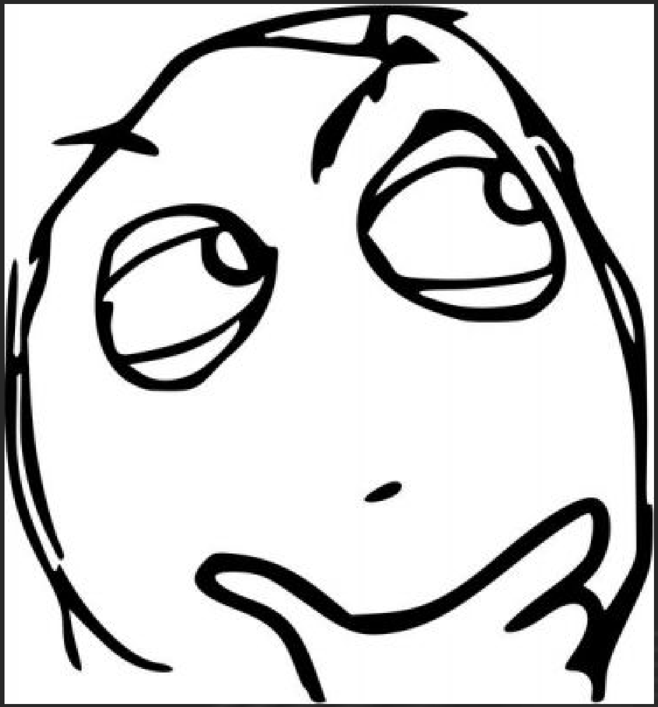
Classifiers

MORE MORE MORE

Data

Features

Classifiers



Grid search for results

64 cores, 1TB RAM, 1 day

Grid search for results

64 cores, 1TB RAM, 1 day

Tune parameters per language / task ?

- Not necessary this time

Grid search for results

64 cores, 1TB RAM, 1 day

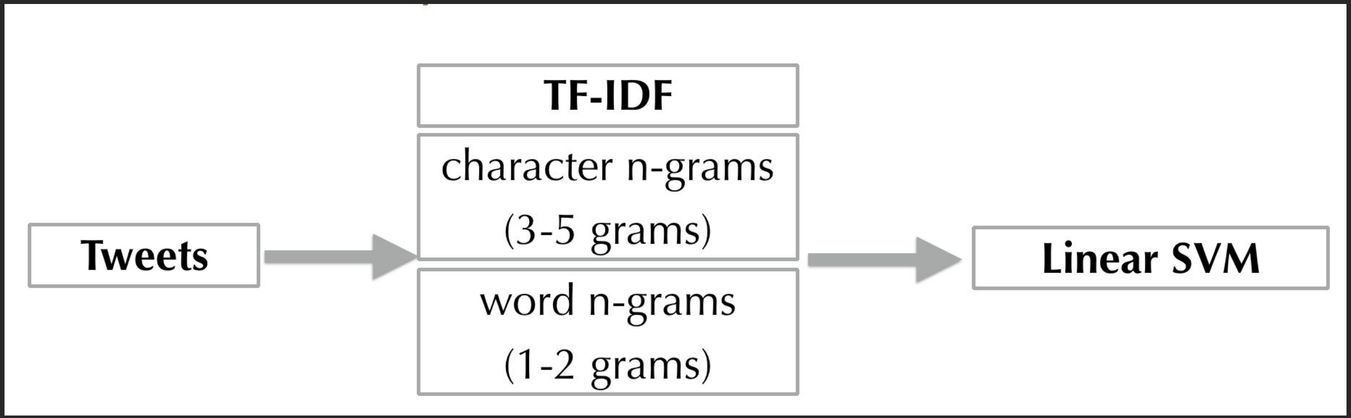
Tune parameters per language / task ?

- Not necessary this time

Scikit-learn defaults are well chosen

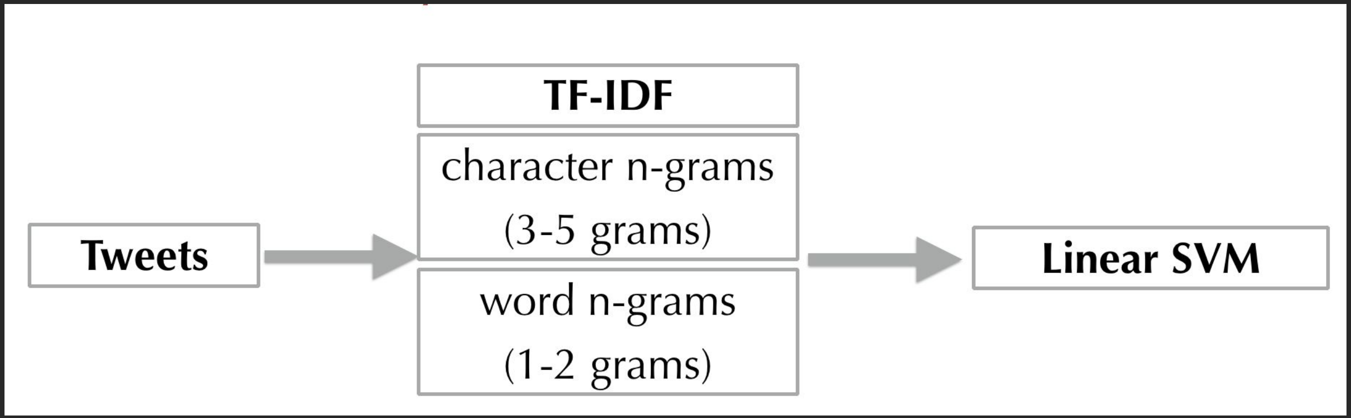
- `min_df=2, sublinear_tf=True`

Start



Tweets
TF-IDF
character n-grams
(3-5 grams)
word n-grams
(1-2 grams)
Linear SVM

End



Tweets
TF-IDF
character n-grams
(3-5 grams)
word n-grams
(1-2 grams)
Linear SVM

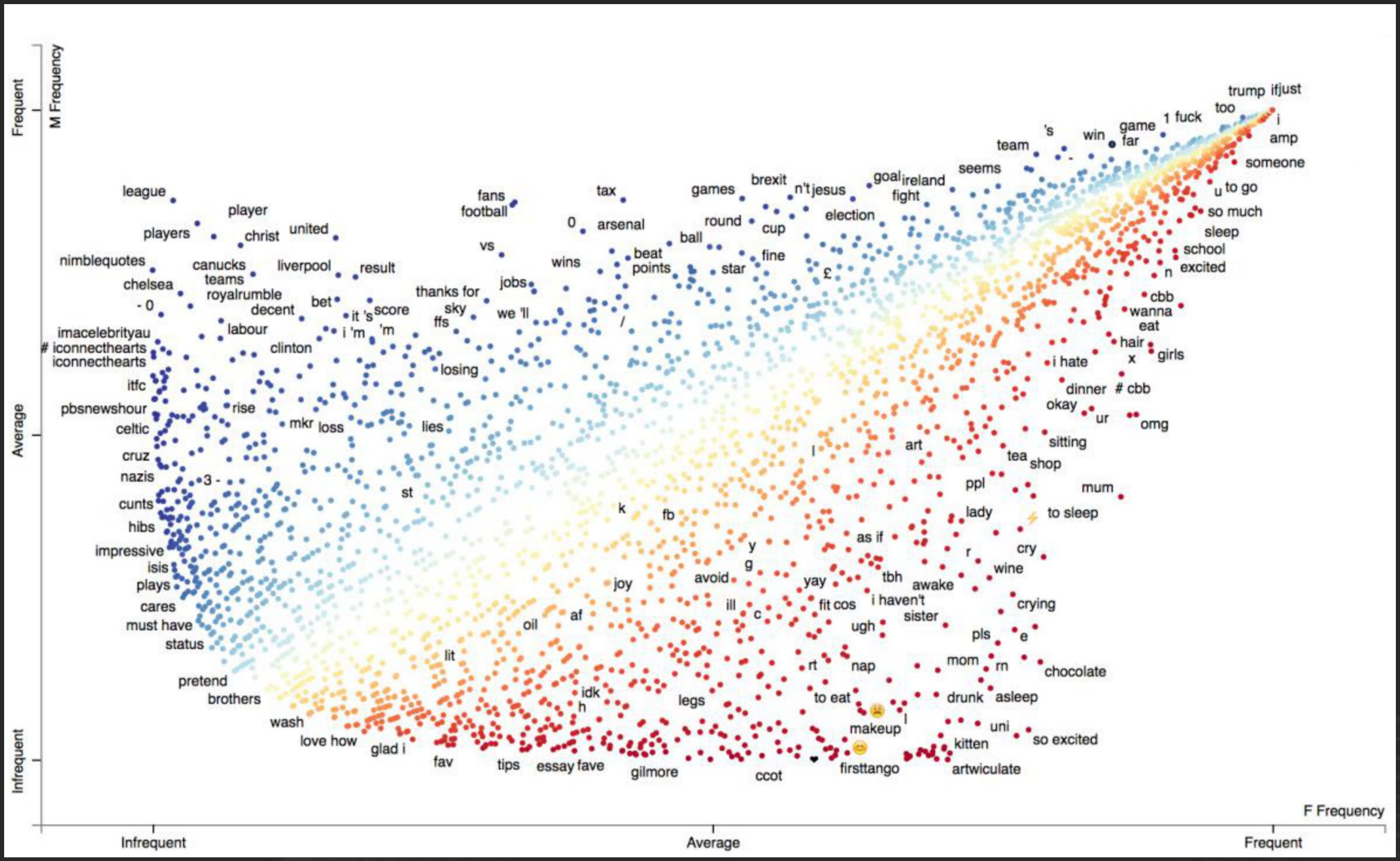
Results

Task	System	Arabic	English	Portuguese	Spanish	Average	+ 2nd
Variety	N-GrAM	**0.8313**	0.8988	0.9813	0.9621	0.9184	0.0013
	LDR	0.8250	**0.8996**	**0.9875**	**0.9625**	**0.9187**	
Gender	N-GrAM	**0.8006**	**0.8233**	**0.8450**	**0.8321**	**0.8253**	0.0029
	LDR	0.7044	0.7220	0.7863	0.7171	0.7325	
Joint	N-GrAM	**0.6831**	**0.7429**	**0.8288**	**0.8036**	**0.7646**	0.0101
	LDR	0.5888	0.6357	0.7763	0.6943	0.6738	

DATA INSIGHTS

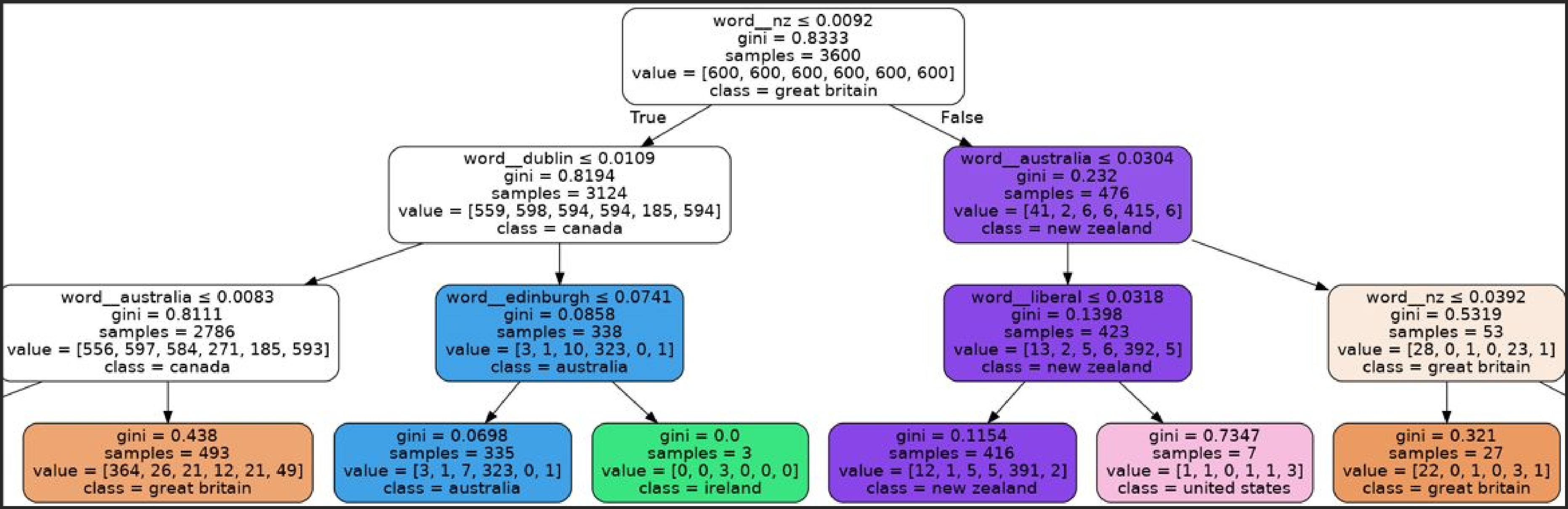
Stereotypes ahead!

English gender visualisation



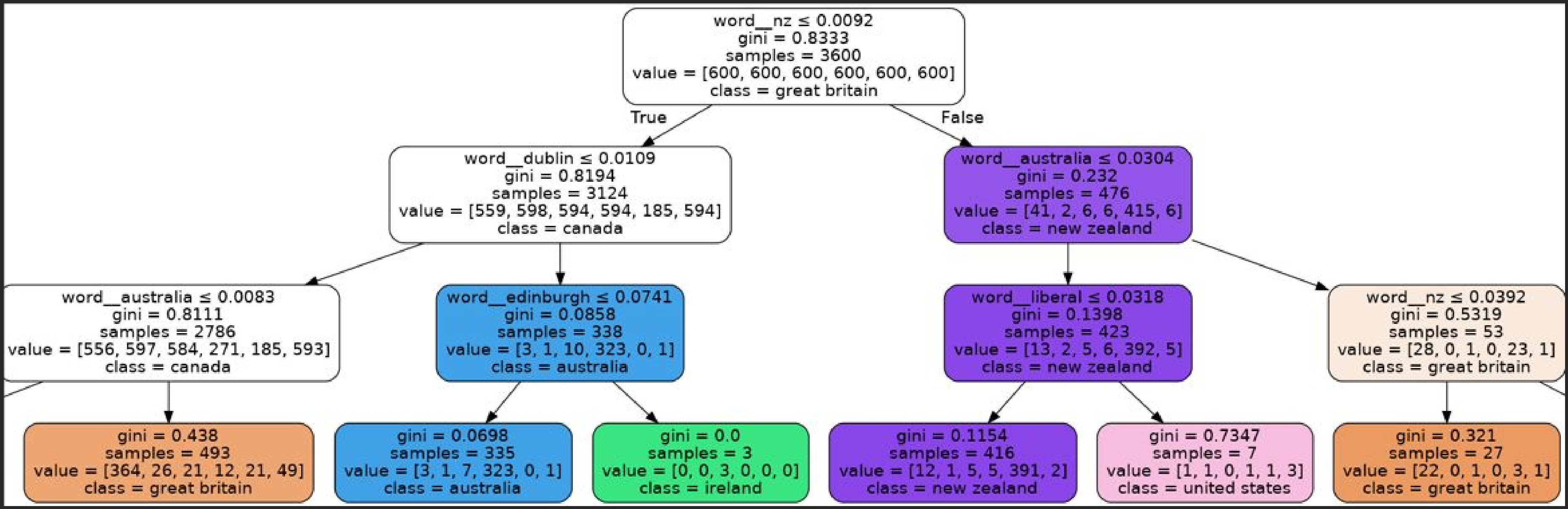
Made with https://github.com/JasonKessler/scattertext

English variety visualisation



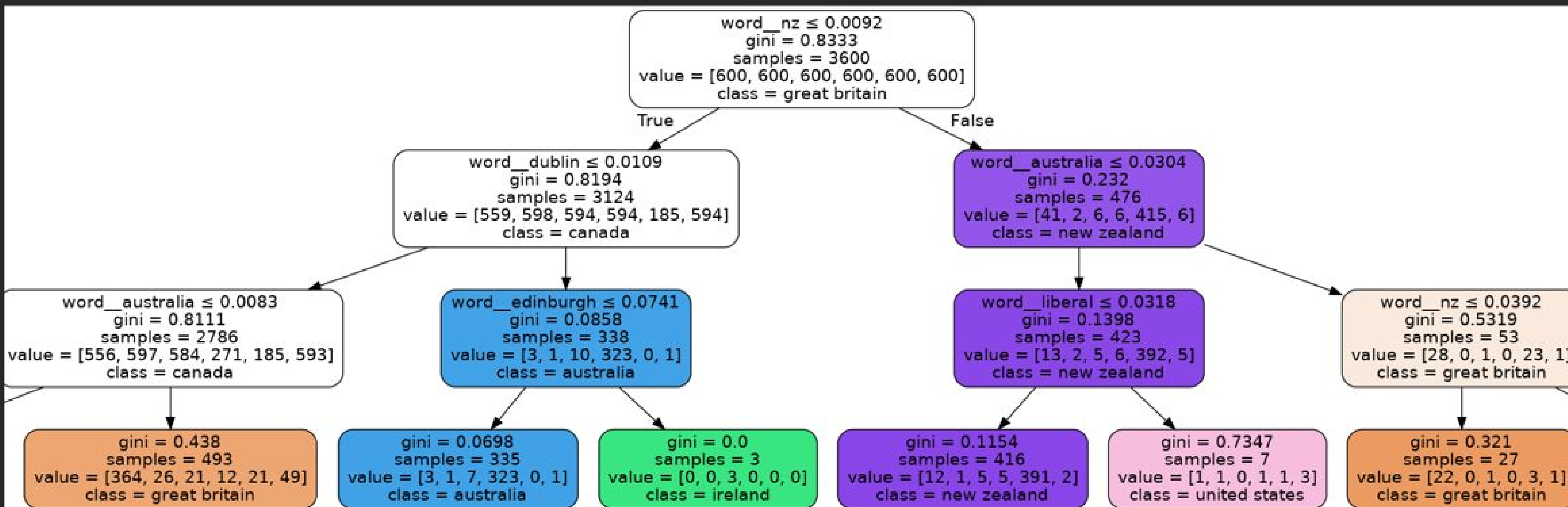
Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

English variety visualisation



Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

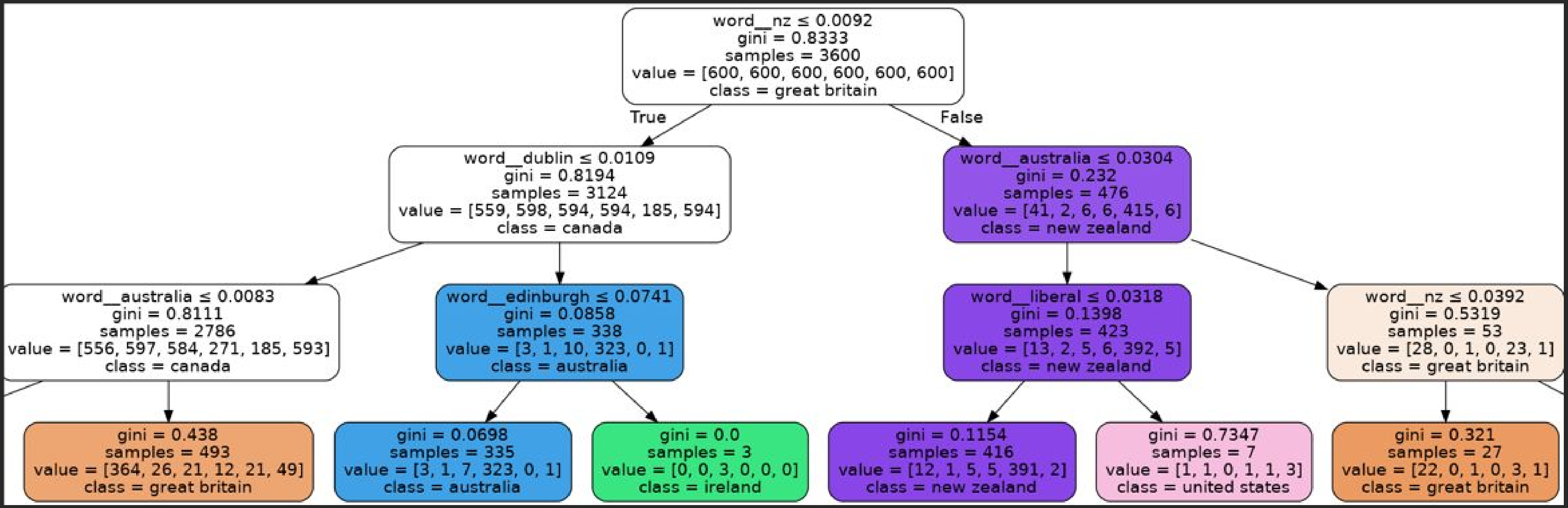
English variety visualisation



Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”

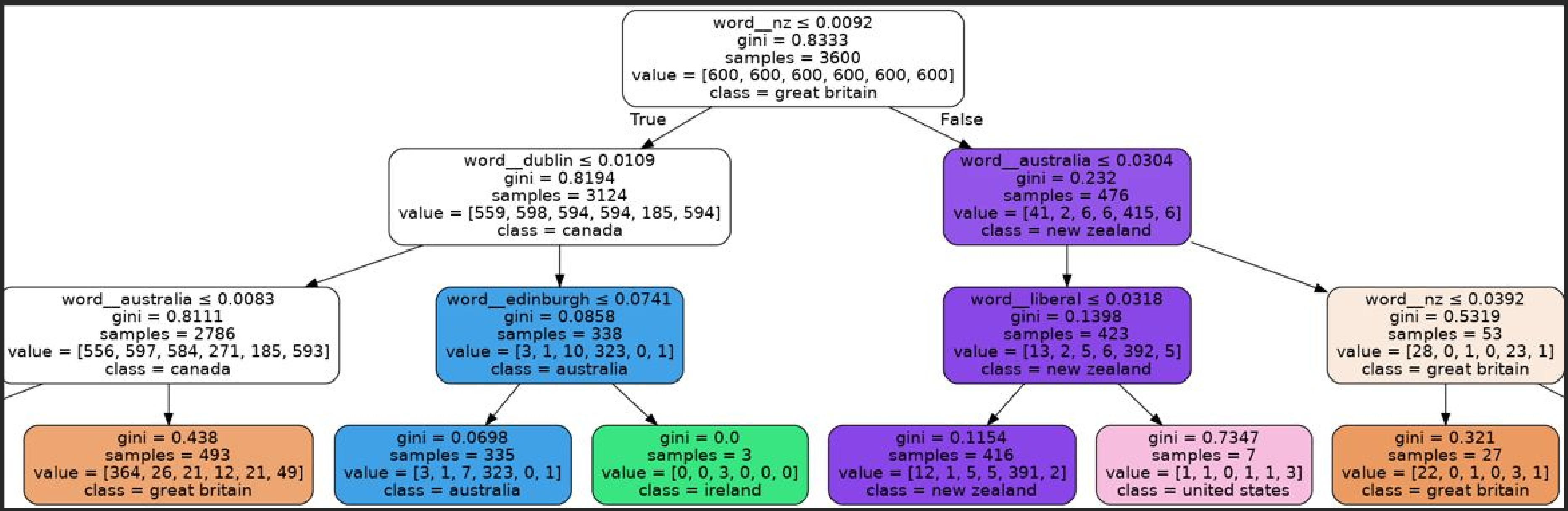
English variety visualisation



Colour/color? lift/elevator? Toilet/Loo/WC/Dunny? ✗

“Australia”, “Dublin”

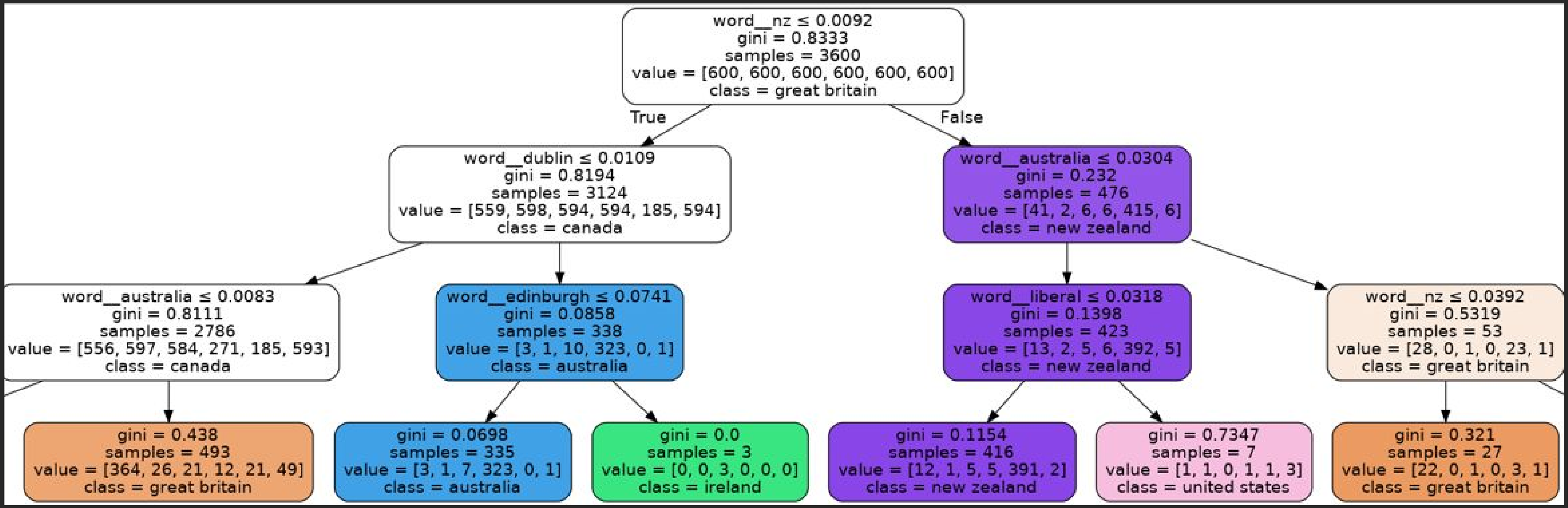
English variety visualisation



Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”, “Dublin”, “NZ”

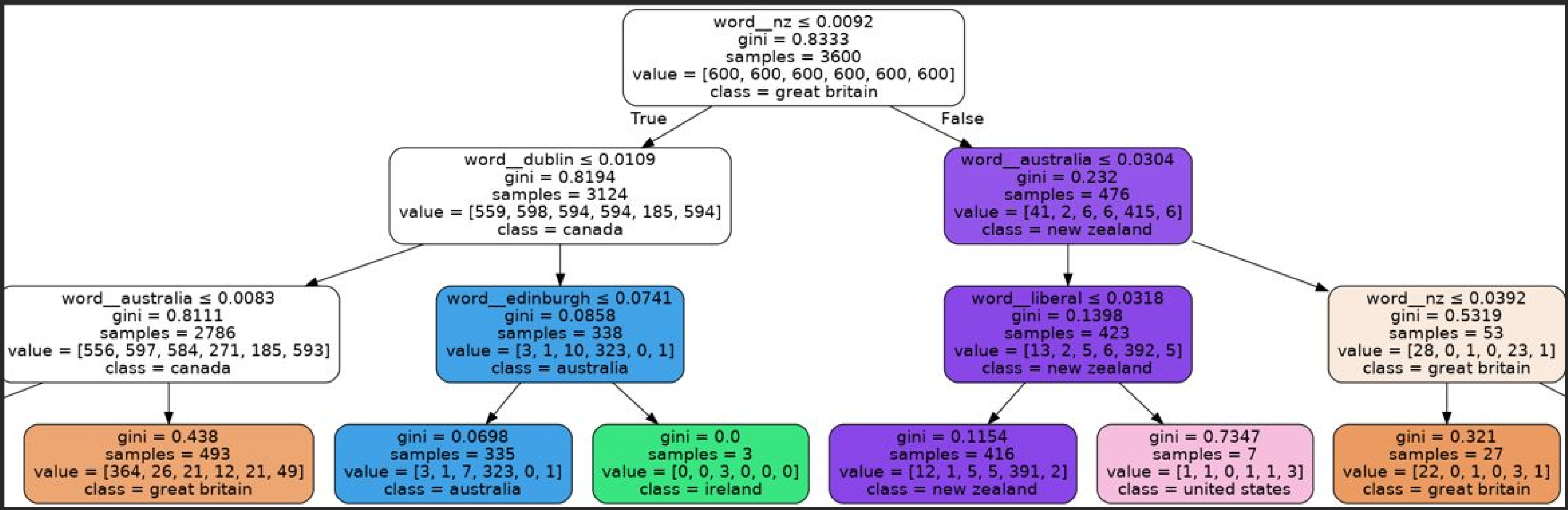
English variety visualisation



Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”, “Dublin”, “NZ”, “Edinburgh”

English variety visualisation



Colour/color? lift/elevator? Toilet/Loo/WC/Dunny?

“Australia”, “Dublin”, “NZ”, “Edinburgh”, “Liberal”

CONCLUSION

Conclusion

N-grams + SVM is (still?) a powerful combo

Conclusion

N-grams + SVM is (still?) a powerful combo

Adding data and features doesn’t always help (and can harm)

Conclusion

N-grams + SVM is (still?) a powerful combo

Adding data and features doesn’t always help (and can harm)

Neural Networks are tricky

Conclusion

N-grams + SVM is (still?) a powerful combo

Adding data and features doesn’t always help (and can harm)

Neural Networks are tricky

Assumptions are wrong

FIN

Questions?
Suggestions?
Answers?
Money?

*With apologies to James Connan