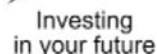


COUNTING WORDS AND DETECTING THEMES WITH PYTHON: CORPUS LINGUISTICS AND TOPIC MODELLING APPROACHES

Yin Yin Lu, Oxford Internet Institute
University of Tartu Digital Methods Summer School
22-23 August 2018

University of Tartu Digital Methods Summer School

22-23 August 2018



- Favourite corpus linguistics tool/concept

Favourite corpus linguistics tool/concept

Digital Economies



Digital Knowledge and Culture



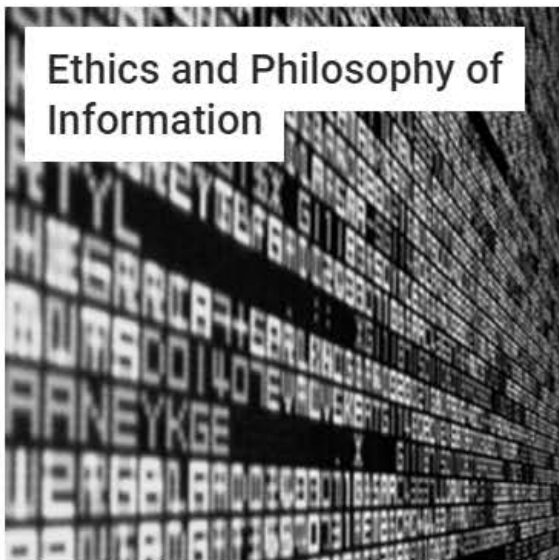
Digital Politics and Government



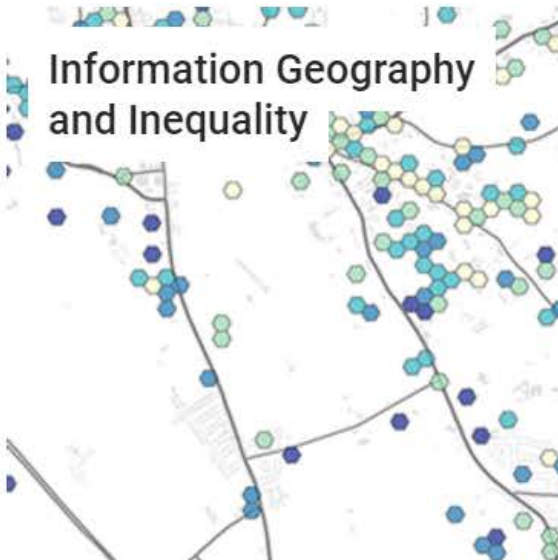
Education, Digital Life and Wellbeing



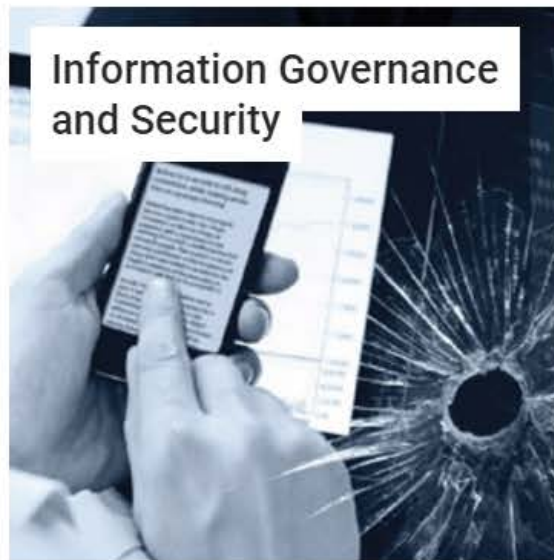
Ethics and Philosophy of Information



Information Geography and Inequality



Information Governance and Security



Social Data Science



THE SCHEDULE

WEDNESDAY

- 10.45-12.45: Introductions & troubleshooting / setup
- 14.00-15.30: Review of Python / Corpus Linguistics with NLTK
- 16.00-17.30: Corpus Linguistics with NLTK continued

THURSDAY

- 10.45-12.45: Topic modelling overview
- 14.00-15.30: Vectorization & topic modelling with scikit-learn
- 16.00-17.30: Topic modelling visualisation / closing remarks

Trusted

Python [default]

GITHUB REPOSITORY

```
print('First 100 tokens in cars corpus:', tokenized_cars[:100])
print('First 100 tokens in space corpus:', tokenized_space[:100])
print('First 100 tokens in guns corpus:', tokenized_guns[:100])
```

First 100 tokens in cars corpus: ['from', 'article', 'ucsu', 'colorado', 'edu', 'lorenzo', 'rintintin', 'colorado', 'edu', 'eric', 'lorenzo', 'let', 'put', 'like', 'this', 'the', 'only', 'similarity', 'between', 'the', 'three', 'models', 'the', 'liter', 'engine', 'displacement', 'actually', 'the', 'the', 'coupe', 'and', 'the', 'the', 'funky', 'looking', 'new', 'sedan', 'share', 'the', 'same', 'liter', 'inline', 'six', 'and', 'the', 'popular', 'small', 'sedan', 'uses', 'shared', 'with', 'the', 'camry', 'the', 'luxury', 'sports', 'coupe', 'the', 'the', 'new', 'luxury', 'sedan', 'and', 'the', 'the', 'base', 'executive', 'sedan', 'all', 'three', 'look', 'completely', 'different', 'aamir', 'qazi', 'perhaps', 'because', 'witnesses', 'who', 'have', 'left', 'the', 'compound', 'have', 'all', 'testified', 'that', 'the', 'bat', 'shot', 'first', 'they', 'they', 'did', 'not', 'identify', 'these', 'before', 'telling', 'conversations']

First 100 tokens in space corpus: ['any', 'lunar', 'mission', 'need', 'full', 'regular', 'orbit', 'correct', 'and', 'when', 'its', 'fuel', 'run', 'out', 'will', 'crash', 'within', 'months', 'the', 'orbital', 'relationships', 'change', 'd', 'noticeably', 'during', 'lunar', 'missions', 'lasting', 'only', 'few', 'days', 'possible', 'that', 'there', 'are', 'stable', 'orbits', 'here', 'and', 'there', 'the', 'moon', 'gravitational', 'field', 'poorly', 'mapped', 'but', 'know', 'none', 'perturbations', 'from', 'sun', 'and', 'earth', 'are', 'relatively', 'minor', 'issues', 'low', 'altitudes', 'the', 'big', 'problem', 'that', 'the', 'moon', 'own', 'gravitational', 'field', 'quite', 'lumpy', 'due', 'the', 'irregular', 'distribution', 'mass', 'within', 'the', 'moon', 'glad', 'see', 'griffin', 'spending', 'his', 'time', 'engineering', 'rather', 'than', 'ritual', 'purification', 'the', 'language', 'pity', 'got', 'stuck', 'with', 'the', 'turkey', 'rather']

First 100 tokens in guns corpus: ['that', 'revisionist', 'account', 'what', 'happened', 'gritz', 'was', 'well', 'aware', 'duke', 'presence', 'the', 'ticket', 'given', 'that', 'gritz', 'not', 'all', 'shy', 'about', 'associating', 'and', 'promoting', 'other', 'white', 'supremacists', 'such', 'the', 'christian', 'identity', 'movement', 'willis', 'carto', 'whatever', 'reasons', 'gritz', 'had', 'leave', 'the', 'ticket', 'had', 'nothing', 'with', 'duke', 'presence', 'believe', 'chip', 'berlet', 'has', 'populist', 'party', 'newsletter', 'from', 'the', 'time', 'with', 'photo', 'gritz', 'happily', 'shaking', 'hands', 'with', 'duke', 'having', 'lived', 'played', 'and', 'worked', 'and', 'near', 'the', 'navajo', 'reservation', 'for', 'number', 'years', 'can', 'confirm', 'this', 'ancient', 'pattern', 'found', 'petroglyphs', 'dated', 'years', 'old', 'also', 'the', 'indians', 'never', 's

tinyurl.com/TartuNLP

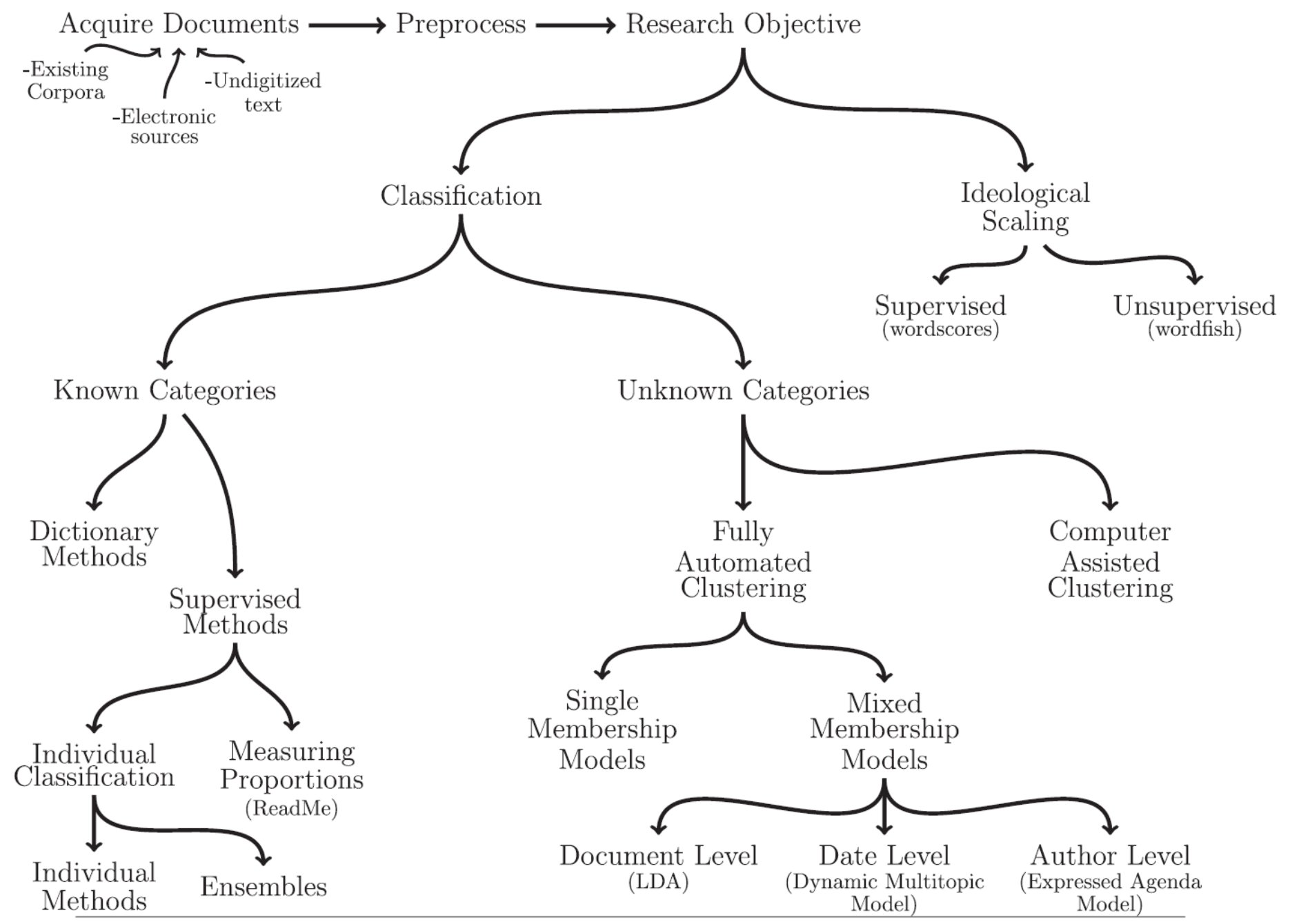


Fig. 1 An overview of text as data methods.

du', 'eri
'liter',
'share',
camry',
'sedan',
'left',
identif

nd', 'whe
, 'change
'stabl
e', 'pert
problem',
mass', 'w
'purific

', 'duk
ing', 'ot
sons', 'g
s', 'popu
'duke',
'can',
ever', 's

TOPIC MODELLING APPLICATIONS

Topic modelling is an extremely popular NLP technique with applications across many domains. It is broadly connected to **text summarisation**. Some specific industrial and academic examples:

1. Categorisation of limitless volumes of legal documents and news stories by lawyers and journalists.
2. Restriction of JSTOR search results to specific categories.
3. Agenda setting of U.S. Congressional statements.
4. Author gender in 19th-century literature.
5. Trends in academic fields based on PhD abstracts.

TOPIC MODELLING ASSUMPTIONS

All algorithms share the same core assumptions:

1. Documents are composed of mixtures of topics.

2. Topics are composed of mixtures of words.

3. Topics can be *inferred* from word-document co-occurrences.

On a broader level, topic models are grounded upon the idea that *meanings of documents are governed by latent variables* (topics). The goal is to uncover them, and there are various approaches for doing so.



TOPIC MODELLING ALGORITHMS

```
print('First 100 tokens in cars corpus:', tokenized_cars[:100])
print('First 100 tokens in space corpus:', tokenized_space[:100])
print('First 100 tokens in guns corpus:', tokenized_guns[:100])
```

There are two basic types of topic models:

1. **Matrix decomposition**, as represented by *Latent Semantic Analysis* (LSA, also known as truncated Singular Value Decomposition).
2. **Probabilistic inference**
 - *Probabilistic LSA* (pLSA)—rarely used on its own. Document probabilities are fixed.
 - *Latent Dirichlet Allocation* (LDA)—most popular and generalizable (‘distribution over distributions’). Bayesian pLSA.

The approaches have the same input and similar output, but different maths.

BAG OF WORDS (VECTORIZATION)

```
print('First 100 tokens in cars corpus:', tokenized_cars[:100])
print('First 100 tokens in space corpus:', tokenized_space[:100])
print('First 100 tokens in guns corpus:', tokenized_guns[:100])
```

"This is how you get ants."

First 100 tokens in cars corpus: ['from', 'article', 'ucsu', 'colorado', 'edu', 'lorenzo', 'rintintin', 'colorado', 'edu', 'eric', 'lorenzo', 'let', 'put', 'like', 'this', 'the', 'only', 'similarly', 'between', 'the', 'three', 'models', 'the', 'liter', 'engine', 'displacement', 'actually', 'the', 'the', 'coupe', 'and', 'the', 'the', 'funky', 'looking', 'new', 'sedan', 'share', 'the', 'same', 'liter', 'inline', 'six', 'and', 'the', 'popular', 'small', 'sedan', 'uses', 'shared', 'with', 'the', 'camry', 'the', 'luxury', 'sports', 'coupe', 'the', 'the', 'new', 'luxury', 'sedan', 'and', 'the', 'the', 'base', 'executive', 'sedan', 'all', 'three', 'look', 'completely', 'different', 'admir', 'qazi', 'perhaps', 'because', 'witnesses', 'who', 'have', 'left', 'the', 'compound', 'have', 'all', 'testified', 'that', 'the', 'batf', 'shot', 'first', 'they', 'they', 'did', 'not', 'identify', 'themselves', 'before', 'tossing', 'concussion']

First 100 tokens in space corpus: ['any', 'lunar', 'satellite', 'and', 'when', 'its', 'fuel', 'runs', 'out', 'will', 'crash', 'within', 'months', 'the', 'orbits', 'the', 'apollo', 'motherships', 'changed', 'noticeably', 'during', 'lunar', 'missions', 'lasting', 'only', 'few', 'days', 'possible', 'that', 'there', 'are', 'stable', 'orbits', 'here', 'and', 'there', 'the', 'moon', 'gravitational', 'field', 'poorly', 'mapped', 'but', 'know', 'none', 'perturbations', 'and', 'renewable', 'resources', 'low', 'altitudes', 'which', 'big', 'problem', 'that', 'the', 'moon', 'own', 'gravitational', 'field', 'quite', 'lumpy', 'due', 'the', 'irregular', 'distribution', 'mass', 'within', 'the', 'moon', 'glad', 'see', 'griffin', 'spending', 'his', 'time', 'engineering', 'rather', 'than', 'ritual', 'purification', 'the', 'language', 'pity', 'got', 'stuck', 'with', 'the', 'turkey', 'rather']

First 100 tokens in guns corpus: ['that', 'revisionist', 'account', 'that', 'was', 'well', 'aware', 'duke', 'presence', 'the', 'ticket', 'given', 'that', 'gritz', 'not', 'all', 'shy', 'about', 'associating', 'and', 'promoting', 'other', 'white', 'supremacists', 'such', 'the', 'christian', 'identity', 'movement', 'willis', 'carto', 'whatever', 'reasons', 'gritz', 'had', 'leave', 'the', 'ticket', 'had', 'nothing', 'with', 'duke', 'presence', 'believe', 'chip', 'berlet', 'has', 'populist', 'party', 'newsletter', 'time', 'with', 'photo', 'you', 'happily', 'making', 'hands', 'with', 'duke', 'having', 'lived', 'played', 'and', 'worked', 'and', 'beach', 'the', 'vaio', 'reservation', 'for', 'number', 'years', 'can', 'confirm', 'this', 'ancient', 'pattern', 'found', 'petroglyphs', 'dated', 'years', 'old', 'also', 'the', 'ancient', 'history']

tokenizer

['this', 'is', 'how', 'you', 'get', 'ants']

Build a vocabulary over all document

['aardvak', 'amsterdam', 'ants', '...', 'you', 'your', 'zyxst']

Sparse matrix encoding

aardvak ants get you zyxst

[0, ..., 0, 1, 0, ..., 0, 1, 0, ..., 0, 1, 0, ..., 0]

File

Edit

+

Corpus
(document-term frequency matrix)

	Word 1	Word 2	Word 3	...	Word k
Doc 1	c_{11}	c_{12}	c_{13}	...	c_{1k}
Doc 2	c_{21}	c_{22}	c_{23}	...	c_{2k}
Doc 3	c_{31}	c_{32}	c_{33}	...	c_{3k}
...
Doc n	c_{n1}	c_{n2}	c_{n3}	...	c_{nk}

WORDS

Vocabulary

SVD

Inference

	Topic 1	Topic 2	Topic 3
Doc 1	0.2	0.4	0.4
Doc 2	0.9	0.05	0.05
Doc 3	0.1	0.1	0.8
...
Doc n	0.8	0.1	0.1

$p(\text{topics})$

LSA

LDA

$p(\text{words} | \text{topic})$

LDA

	Topic 1	Topic 2	Topic 3
Word 1	0.001	0.002	0.005
Word 2	0.009	0.019	0.009
Word 3	0.002	0.001	0.004
...
Word n	0.000	0.001	0.002

LSA

William Zheng, 2017

Python [default]

Presentation

'edu', 'eri', 'liter', 'share', 'camry', 'sedan', 'left', 'identif', 'and', 'whe', 's', 'change', 'stabl', 'one', 'pert', 'problem', 'mass', 'w', 'purific', 're', 'duk', 'oting', 'ot', 'easons', 'g', 'has', 'popu', 'duke', 'can', 'never', 's

LATENT DIRICHLET ALLOCATION

LDA is a *generative model*: it specifies a procedure through which documents are written (generated). Its assumption about how to write a text is quite naïve:

1. Choose the number of words N in your document
2. Choose a topic mixture θ for your document (e.g., 60% about topic 1 and 40% about topic 2)
3. While the number of generated words is smaller than N , generate a word w_i by:
 - Choosing a topic according to the chosen topic mixture θ
 - Choosing a word according to the topic's vocabulary distribution



LATENT DIRICHLET ALLOCATION

```
print('First 100 tokens in cars corpus:', tokenized_cars[:100])  
print('First 100 tokens in space corpus:', tokenized_space[:100])  
print('First 100 tokens in guns corpus:', tokenized_guns[:100])
```

According to this procedure, we might write (generate) a document as follows:

1. We decide that it will be 4 words long.
2. 25% will be about topic 1 (religion) and 75% about topic 2 (cars).
3. By choosing from the vocabularies of topic 1 and topic 2, we generate four words: *Ferrari* (T2), *engine* (T2), *Jesus* (T1), *drive* (T2). This is our document.

This **bag of words (BoW)** approach is linguistically absurd, but allows us to *infer* topic mixtures and their associated words.

Real-World Example

Finally, I applied LDA to a set of Sarah Palin's emails a little while ago (see [here](#) for the blog post, or [here](#) for an app that allows you to browse through the emails by the LDA-learned categories), so let's give a brief recap. Here are some of the topics that the algorithm learned:

- **Trig/Family/Inspiration:** family, web, mail, god, son, from, congratulations, children, life, child, down, trig, baby, birth, love, you, syndrome, very, special, bless, old, husband, years, thank, best, ...
- **Wildlife/BP Corrosion:** game, fish, moose, wildlife, hunting, bears, polar, bear, subsistence, management, area, board, hunt, wolves, control, department, year, use, wolf, habitat, hunters, caribou, program, denby, fishing, ...
- **Energy/Fuel/Oil/Mining:** energy, fuel, costs, oil, alaskans, prices, cost, nome, now, high, being, home, public, power, mine, crisis, price, resource, need, community, fairbanks, rebate, use, mining, villages, ...
- **Gas:** gas, oil, pipeline, agia, project, natural, north, producers, companies, tax, company, energy, development, slope, production, resources, line, gasline, transcanada, said, billion, plan, administration, million, industry, ...
- **Education/Waste:** school, waste, education, students, schools, million, read, email, market, policy, student, year, high, news, states, program, first, report, business, management, bulletin, information, reports, 2008, quarter, ...
- **Presidential Campaign/Elections:** mail, web, from, thank, you, box, mccain, sarah, very, good, great, john, hope, president, sincerely, wasilla, work, keep, make, add, family, republican, support, doing, p.o, ...

'having', 'lived', 'played', 'and', 'worked', 'and', 'near', 'the', 'navajo', 'reservation', 'for', 'number', 'years', 'can',
'confirm', 'this', 'ancient', 'pattern', 'found', 'petroglyphs', 'dated', 'years', 'old', 'also', 'the', 'Edwin Chen, 2011', 's

DEEP LEARNING & LDA

One of the most recent extensions of LDA is **lda2vec**, which is also an extension of **word2vec**. It is a deep learning model that jointly learns word, document, and topic vectors (embeddings).

- **word2vec** uses a skipgram neural network model to generate **word vectors**. These vectors are used to predict **context words**.
- **lda2vec** uses a **context vector** to make predictions: the **word vector + the document vector** (topic weight vector + topic matrix).

More information: www.github.com/cemoody/lda2vec

THANK YOU & STAY IN TOUCH! 😊

```
print('First 100 tokens in cars corpus:', tokenized_cars[:100])  
print('First 100 tokens in space corpus:', tokenized_space[:100])  
print('First 100 tokens in guns corpus:', tokenized_guns[:100])
```

First 100 tokens in cars corpus: ['from', 'article', 'ucsu', 'colorado', 'edu', 'lorenzo', 'rintintin', 'colorado', 'edu', 'eric', 'lorenzo', 'let', 'put', 'like', 'this', 'the', 'only', 'similarity', 'between', 'the', 'three', 'models', 'the', 'liter', 'engine', 'displacement', 'actually', 'the', 'comp', 'an', 'it', 'the', 'funky', 'looking', 'new', 'sedan', 'share', 'the', 'same', 'lita', 'lilia', 's', 'the', 'popular', 'the', 'sedan', 'with', 'the', 'camry', 'the', 'luxury', 'sports', 'coupe', 'the', 'the', 'new', 'luxury', 'sedan', 'and', 'the', 'the', 'base', 'executive', 'sedan', 'all', 'three', 'look', 'completely', 'different', 'aamir', 'qazi', 'perhaps', 'because', 'witnesses', 'who', 'have', 'left', 'the', 'compound', 'have', 'all', 'testified', 'that', 'the', 'batf', 'shot', 'first', 'they', 'they', 'did', 'not', 'identify', 'themselves', 'before', 'tossing', 'the', 'guns', 'only']

First 100 tokens in space corpus: ['and', 'the', 'satellite', 'needs', 'fuel', 'regular', 'orbit', 'corrections', 'and', 'when', 'its', 'fuel', 'runs', 'out', 'will', 'crash', 'within', 'months', 'the', 'orbits', 'the', 'apollo', 'motherships', 'changed', 'noticeably', 'during', 'lunar', 'missions', 'lasting', 'only', 'few', 'days', 'possible', 'that', 'there', 'are', 'stable', 'orbits', 'here', 'and', 'there', 'the', 'moon', 'gravitational', 'field', 'poorly', 'mapped', 'but', 'know', 'none', 'perturbations', 'from', 'sun', 'and', 'earth', 'are', 'relatively', 'minor', 'issues', 'low', 'altitudes', 'the', 'big', 'problem', 'that', 'the', 'moon', 'now', 'gravitational', 'field', 'quite', 'lumpy', 'the', 'distribution', 'mass', 'within', 'the', 'moon', 'is', 'not', 'even', 'spending', 'this', 'time', 'engineering', 'this', 'path', 'that', 'the', 'initial', 'purification', 'the', 'language', 'pity', 'got', 'stuck', 'with', 'the', 'turkey', 'rather']

First 100 tokens in guns corpus: ['that', 'revisionist', 'account', 'what', 'happened', 'gritz', 'was', 'well', 'aware', 'duke', 'presence', 'the', 'ticket', 'given', 'that', 'gritz', 'not', 'all', 'shy', 'about', 'associating', 'and', 'promoting', 'other', 'white', 'supremacists', 'such', 'the', 'christian', 'identity', 'movement', 'willis', 'carto', 'whatever', 'reasons', 'gritz', 'had', 'leave', 'the', 'ticket', 'had', 'nothing', 'with', 'duke', 'presence', 'believe', 'chip', 'berlet', 'has', 'populist', 'party', 'newsletter', 'from', 'the', 'time', 'with', 'photo', 'gritz', 'happily', 'shaking', 'hands', 'with', 'duke', 'having', 'lived', 'played', 'and', 'worked', 'and', 'near', 'the', 'navajo', 'reservation', 'for', 'number', 'years', 'can', 'confirm', 'this', 'ancient', 'pattern', 'found', 'petroglyphs', 'dated', 'years', 'old', 'also', 'the', 'indians', 'never', 's']

yin.lu@oii.ox.ac.uk

@yinneth

[linkedin.com/in/periwynkle](https://www.linkedin.com/in/periwynkle)