Home Office

Yet More Topic Modelling

Constructing a topic modelling/clustering, validation & summarization pipeline

Martin Wood – Data Scientist, Home Office
(+ ONS, + ESSnet + Brookes…)

OXFORD BROOKES UNIVERSITY

EUROPEAN STATISTICAL SYSTEM

Office for National Statistics

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

2) Introduction

Home Office

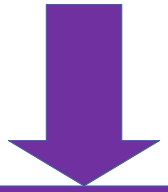
Latent Dirichelet Allocation (BOW)

HDBSCAN (Doc2Vec)

OPTIMAL clusters

Coherence

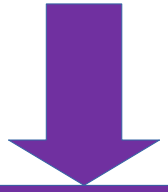
Perplexity



VALIDATE clusters

Centroid separation?

Coherence measure?



Extractive (representative sentences)

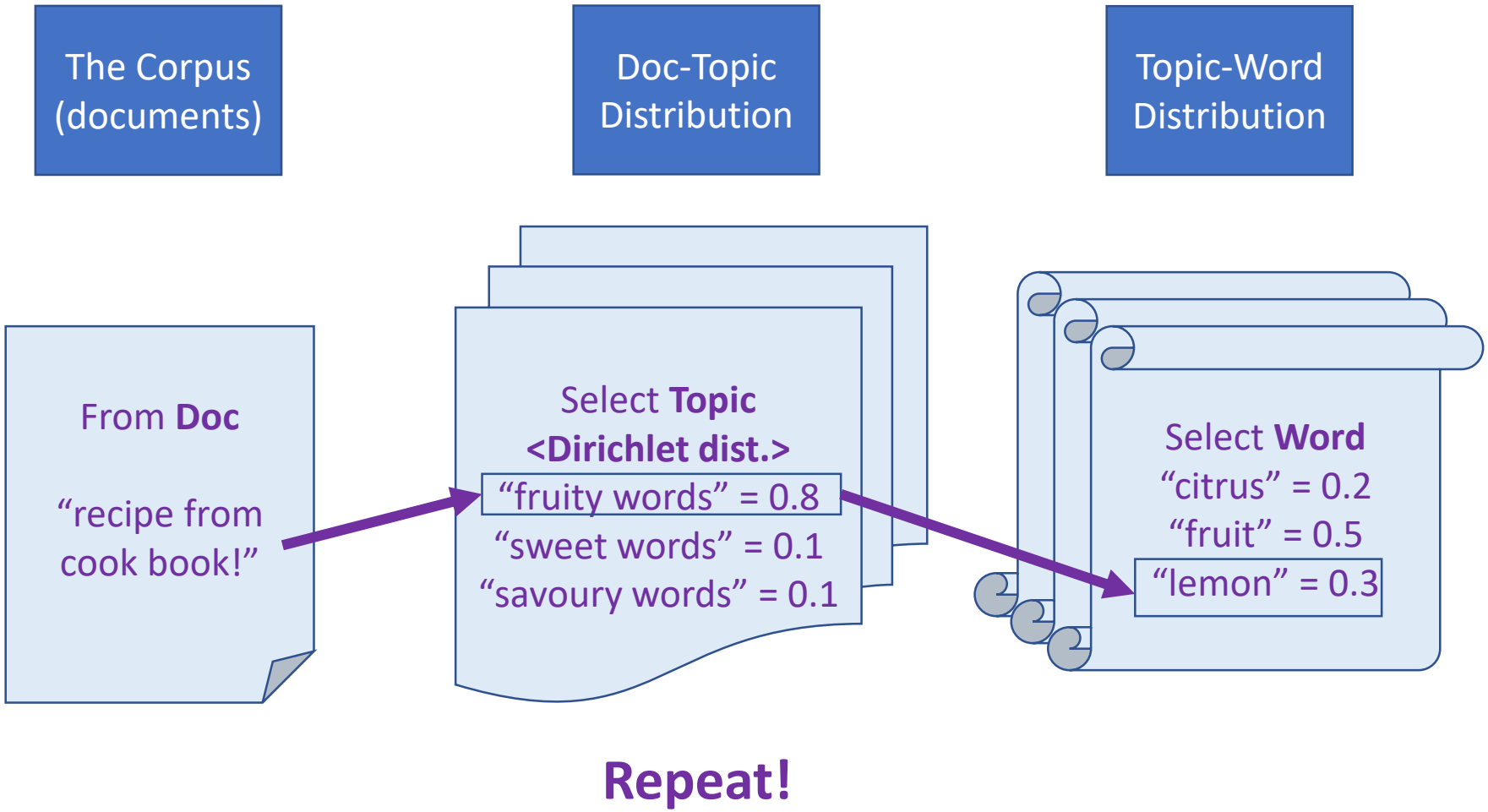
Augmented (named entity recognition?)

SUMMARIZE clusters

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

3) Optimising Latent Dirichlet Allocation

Home Office

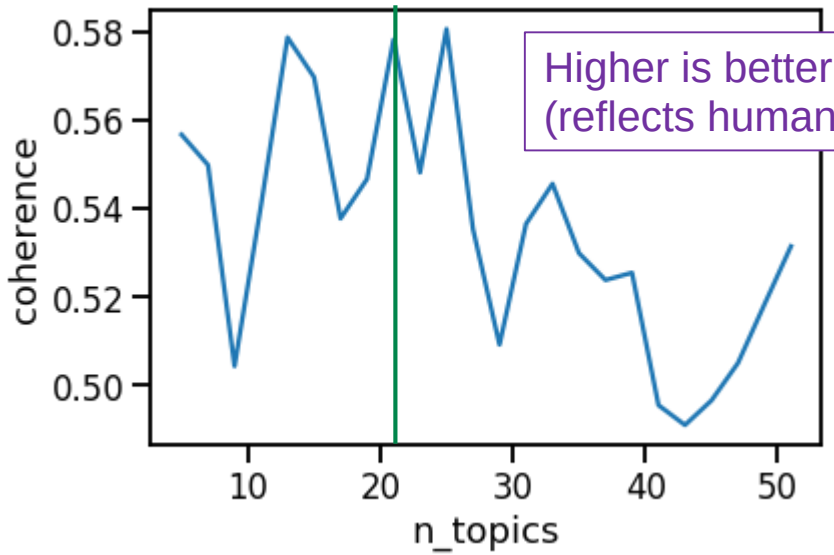


Repeat!

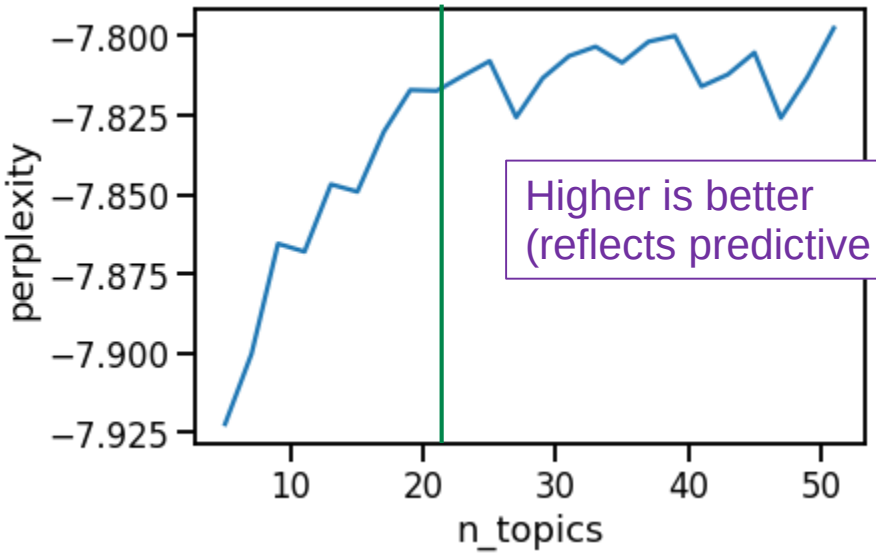
Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

3) Optimising Latent Dirichlet Allocation

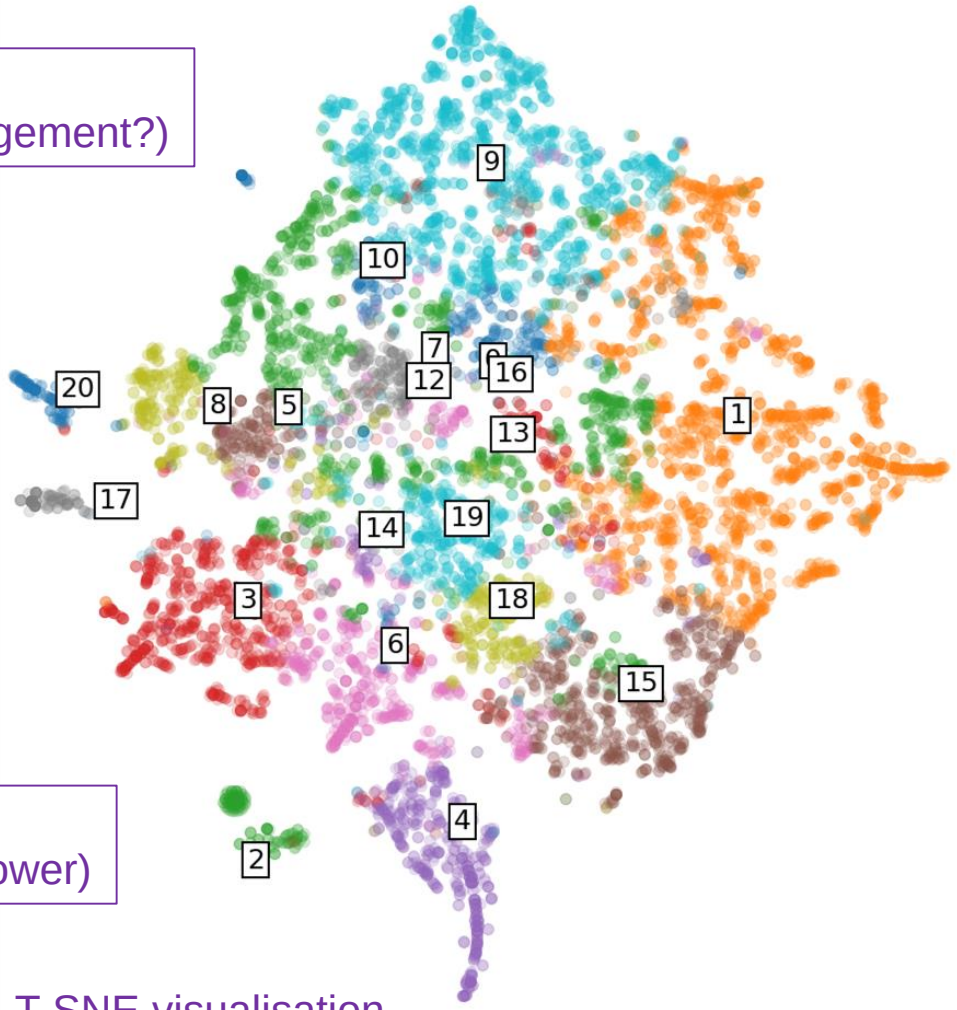
Home Office



Higher is better
(reflects human judgement?)



Higher is better
(reflects predictive power)



T-SNE visualisation

No “true” topics, decide how many clusters are useful to you, then locally optimise?

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

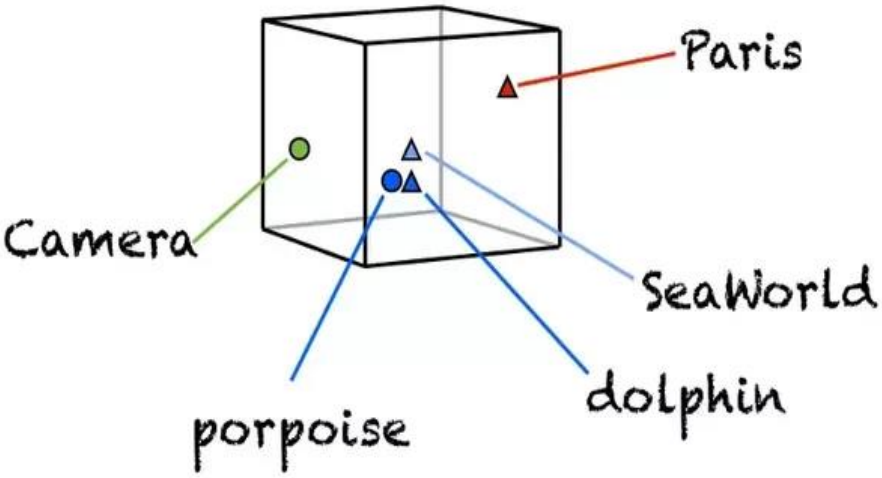
3) Optimising HDBSCAN

Home Office

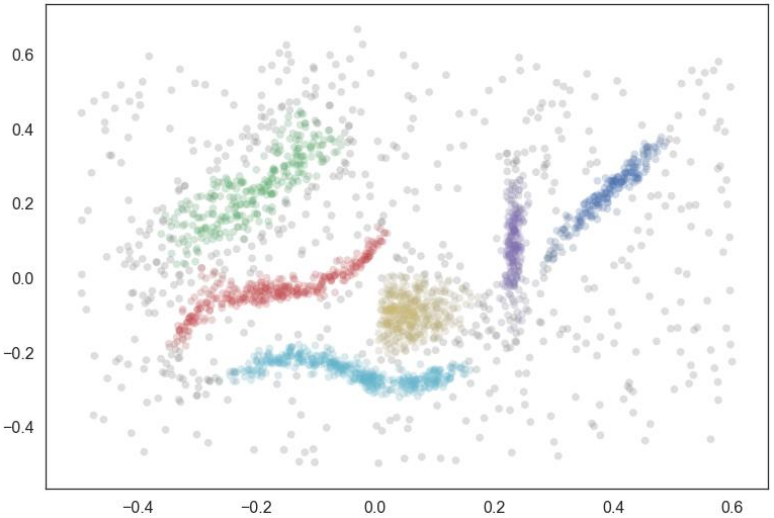
<Hasn't produced pretty graphs yet>

Doc2Vec and then find high-dimensional clusters

Good for very short records, not enough "innate" data to link – use pre-trained models



https://www.kaggle.com/sbongo/do-pretrained-embeddings-give-you-the-extra-edge



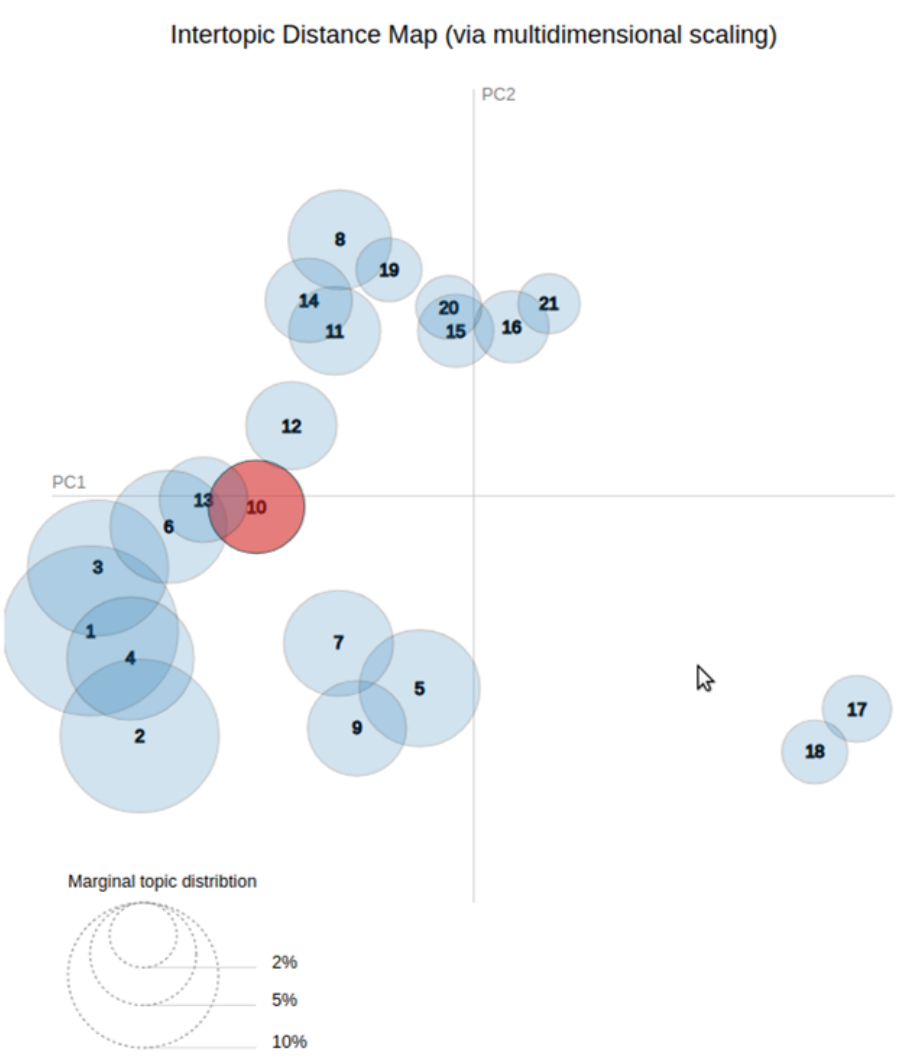
https://hdbscan.readthedocs.io/en/latest/index.html

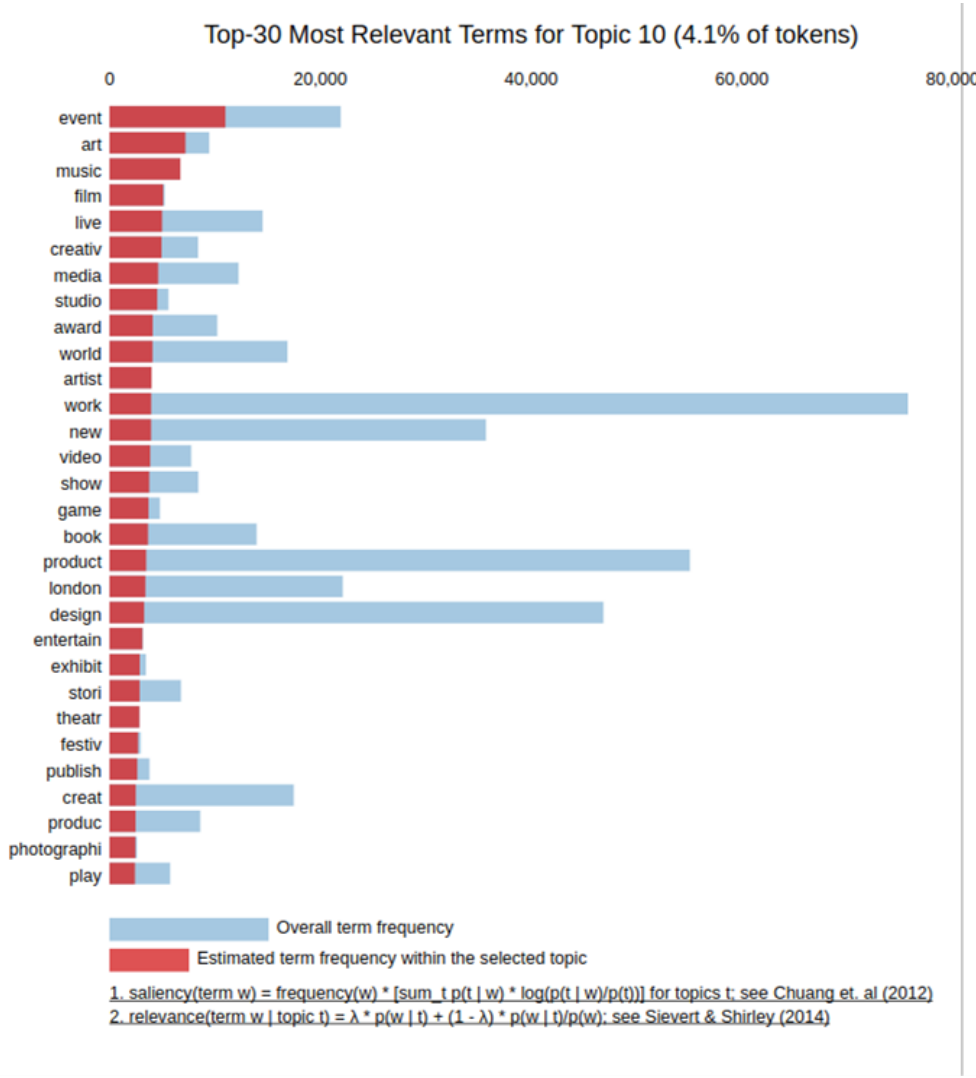
Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

4) Validation – Valid LDA topics

Home Office

From a corpus of business's descriptions of themselves on the internet:

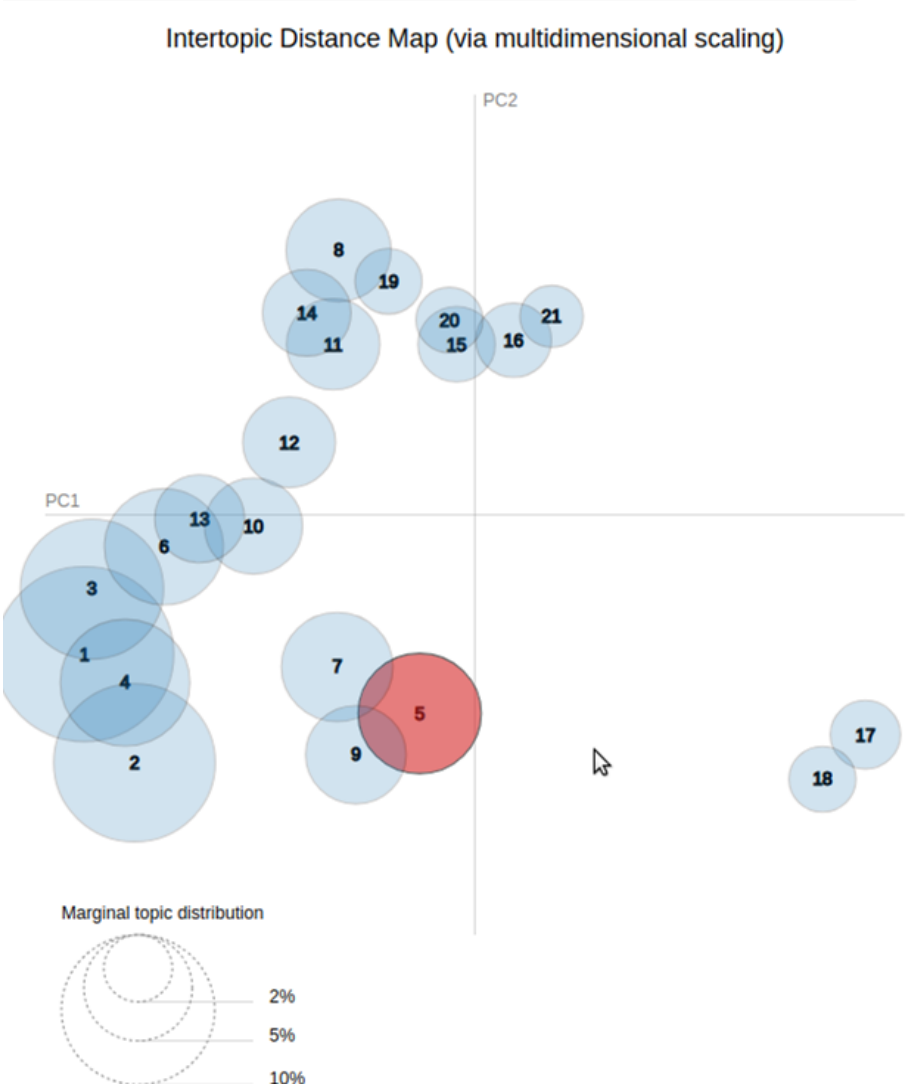


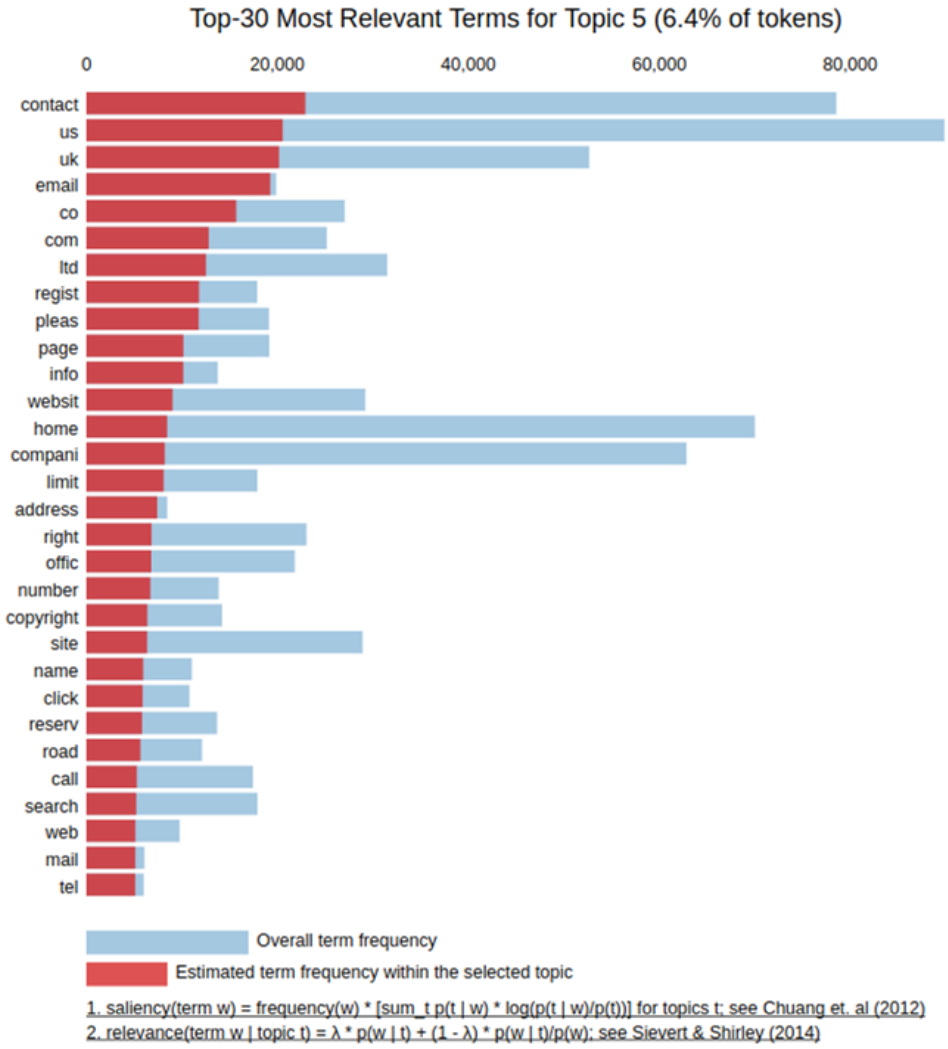


Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

4) Validation – Invalid LDA topics

Home Office



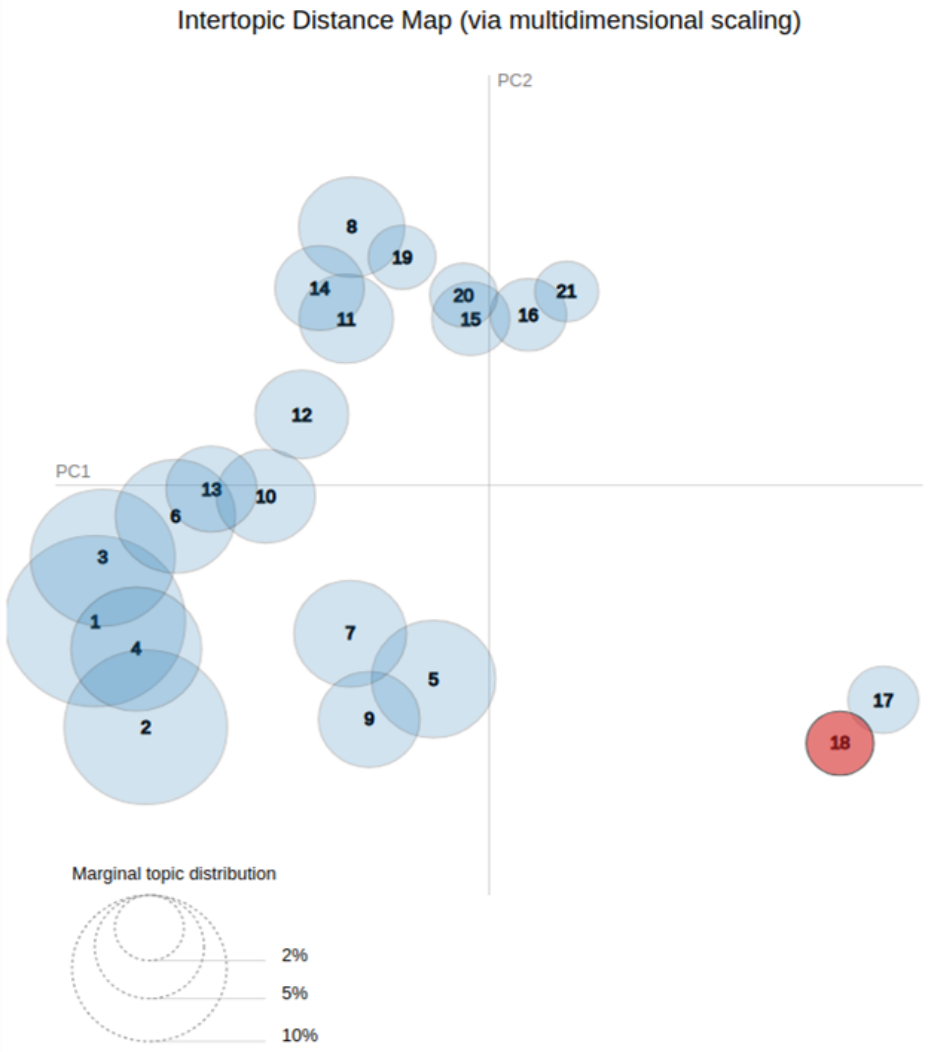


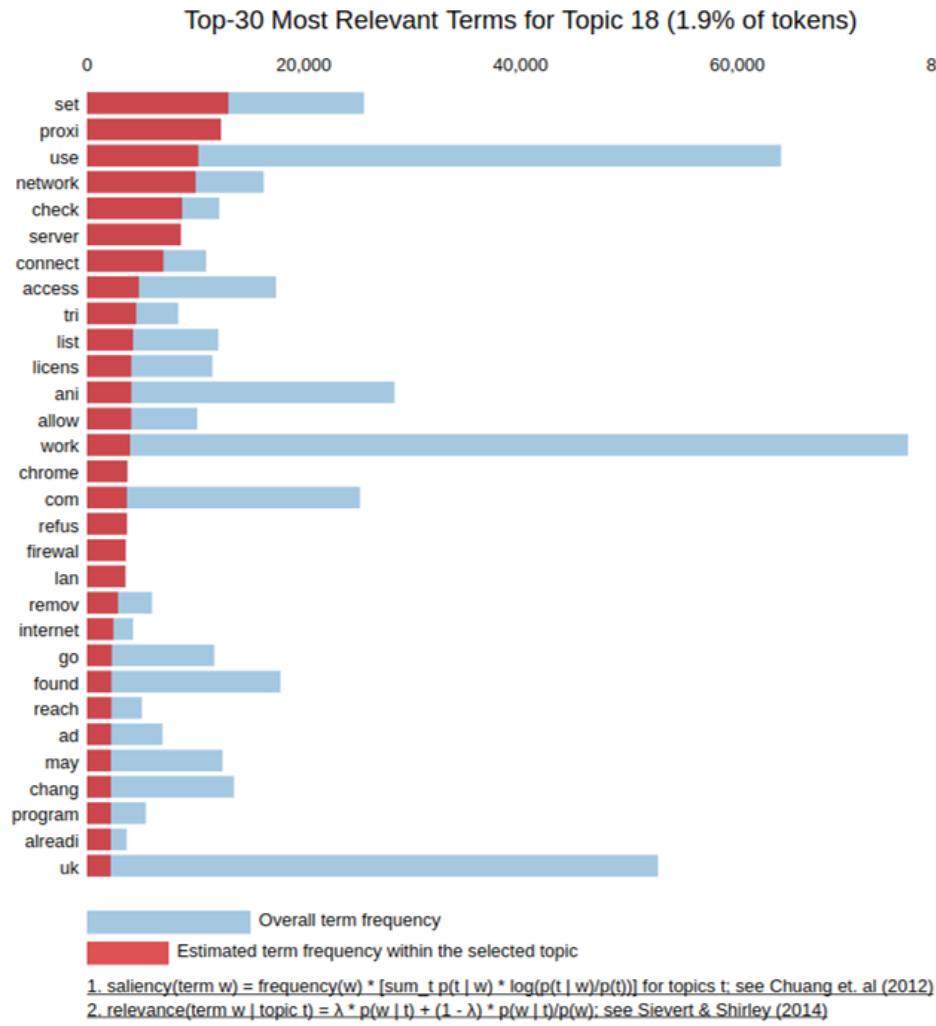
Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

4) Validation – Invalid DATA

Home Office

Invalid topics WELL SEPARATED in topic/vocabulary space, discriminate automatically?

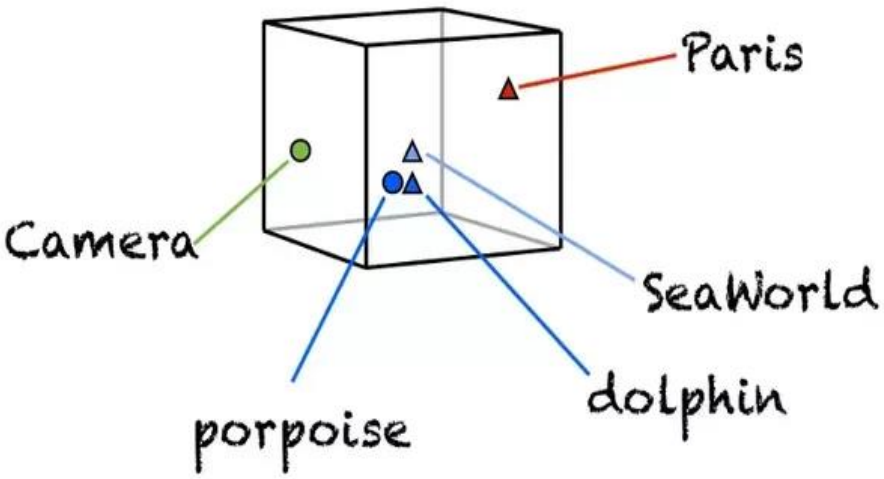




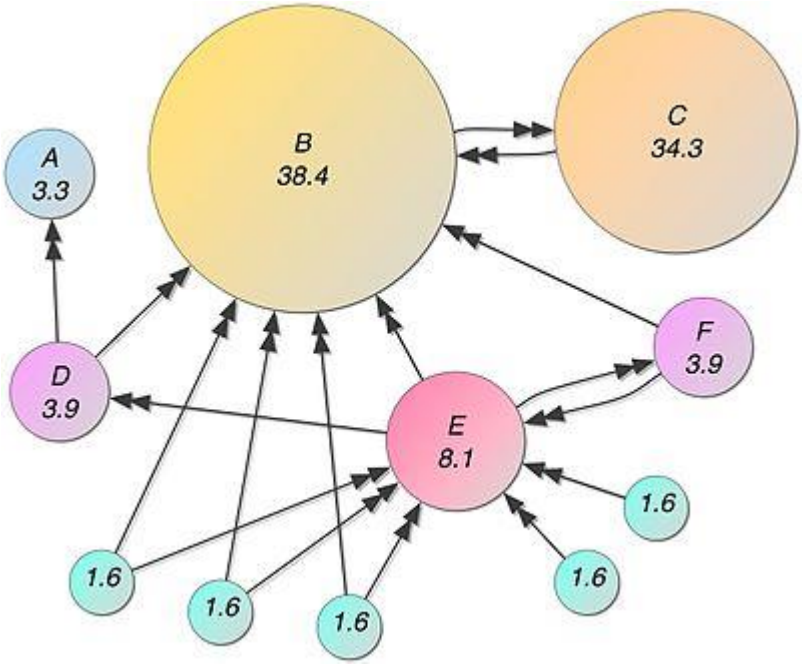
Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

5) Summarizing Clusters

Home Office



https://www.kaggle.com/sbongo/do-pretrained-embeddings-give-you-the-extra-edge



https://en.wikipedia.org/wiki/PageRank

TextRank – variant of Google PageRank Algorithm

Create representation of sentences (summed Word2Vec, Doc2Vec, TF-IDF, edit distance, combo?)

Create giant similarity matrix (cosine distance)

Extract X sentences MOST similar to all other sentences

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

5) Summarizing Clusters

Home Office

Chosen text representation = implicit assumptions

Summed Word2Vec (GloVe, from Stanford NLP)	'Many researchers predict that such narrow AI work in different individual domains will eventually be incorporated into a machine with artificial general intelligence (AGI), combining most of the narrow skills mentioned in this article and at some point even exceeding human ability in most or all these areas.' 'Some of the learners described below, including Bayesian networks, decision trees, and nearest-neighbor, could theoretically, if given infinite data, time, and memory, learn to approximate any function, including whatever combination of mathematical functions would best describe the entire world.'
Doc2Vec (Lau & Baldwin, 2016)	'They solve most of their problems using fast, intuitive judgements.' 'Emergent behavior such as this is used by evolutionary algorithms and swarm intelligence.' 'Many learning algorithms use search algorithms based on optimization.'

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

5) Summarizing Clusters

Home Office

Some Humans!

'In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals. Computer science defines AI research as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.[1] More in detail, Kaplan and Haenlein define AI as “a system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation”.[2] Colloquially, the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".[3] The scope of AI is disputed...’

Alternative - Abstractive Summarization?

Nallapati et al – 2016 – *“Abstractive Text Summarization Using Sequence-to-Sequence RNN’s and Beyond”*

Really hard + “state of the art” results don’t look great (so far)

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

5) Summarizing Clusters

Home Office

text	POS	count
US	GPE	51
Donald Trump	PERSON	29
U.S.	GPE	25
Brexit	ORG	17
British	NORP	15
French	NORP	15
Russian	NORP	15
Syria	GPE	14
China	GPE	12

Named entity recognition RSS feeds and SpaCy…

- Prepackaged NN's label Parts of Speech (POS)
- Pick out common entities
- Distinguish between people, countries, orgs…
- Pick out adjectives? Sentiment?

Martin.Wood5@homeoffice.gov.uk, https://github.com/ozwaldcavendish

5) Conclusions

Home Office

Conclusions

- A lot of work to be done
- Clustering can only be locally optimised
- Concept of “true” topics waiting for us is misleading!
- True abstractive summarization out of reach for now
- Aim to improve extractive to point where humans don’t have to open the source documents

Future Work

- HDBSCAN, more robust for sparse datasets?
- Use Doc2Vec features, similarities to “landmark” statements, statistical properties of vectors of descriptive sentences…

The End

Home Office

Questions?

My email: Martin.Wood5@homeoffice.gov.uk
My git thingy: https://github.com/ozwaldcavendish

My map of wine tasting notes: