Generating online text types: A cluster analysis of predicted document embeddings

Veronika Laippala, Aki-Juhani Kyröläinen, Filip Ginter, Jesse Egbert and Douglas Biber

Online text data and novel possibilities

- Major benefit
  - Linguistics
    - Rare expressions, language varieties that could not be easily studied elsewhere (e.g., Biber & Egbert 2018; Schäfer, 2016; Berber-Sardinha, 2018).
  - Natural Language Processing (NLP)
    - Extreme size of data provides improved performance of automatic syntactic analysis and machine translation for example (Tiedemann et al., 2016; Srivastava et al., 2016; Zeman et al., 2017)
- Major drawback
  - No information on the kinds of texts included
  - ”Traditional” language resouces typically divided into sub-corpora representing registers, i.e., language varieties with specific situational characteristics (Biber, 1988; Biber and Conrad, 2009)

This presentation

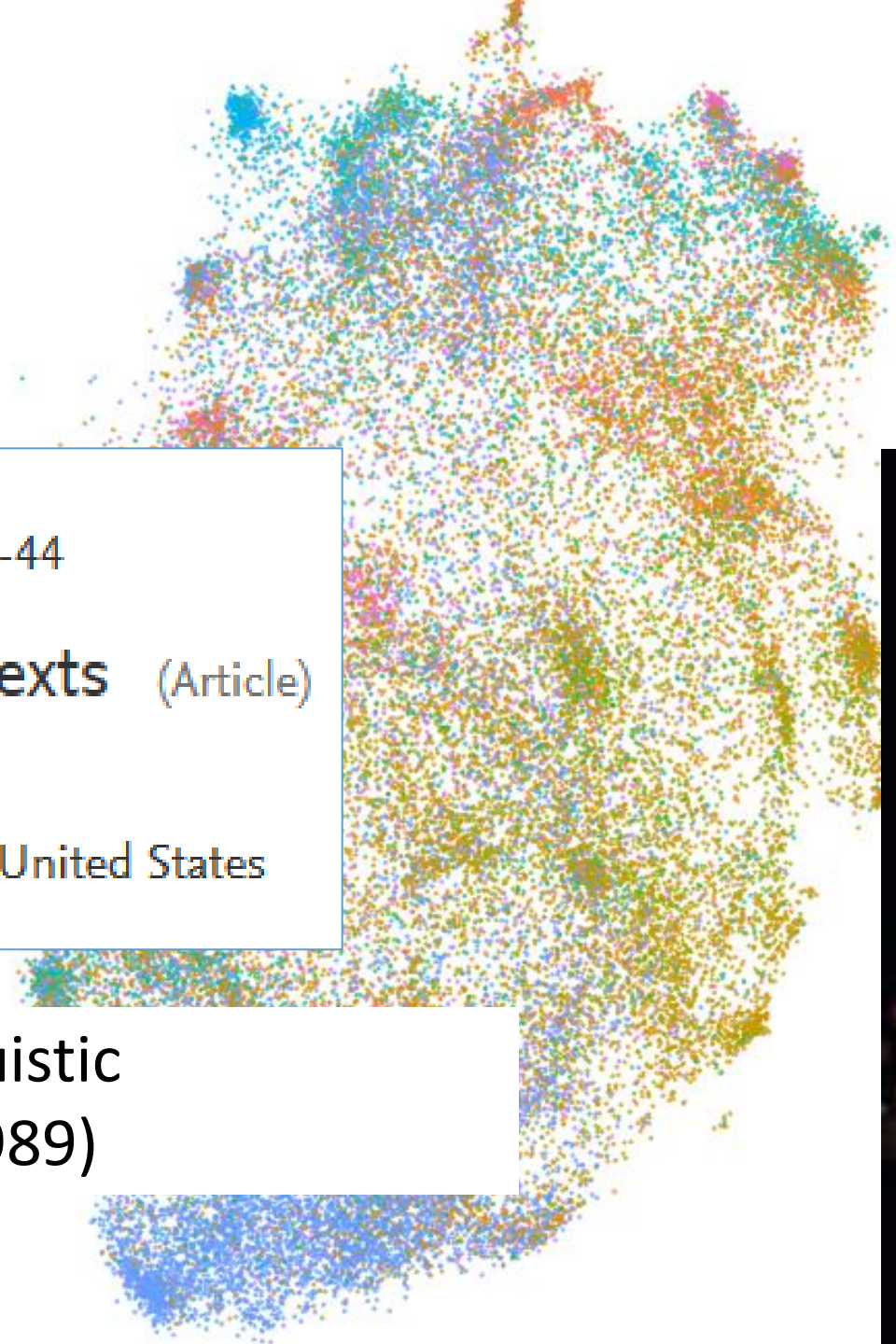
- Finding structure in online text data with clustering
  - Clustering: a data-driven machine learning method to automatically group similar items together

  → (Hopefully) Functionally similar texts form clusters

  → Provide functional information associated with a given document

  → Improve the usability of online text data
- Manually labeled training data **not** required!
- Ideally, a cross-linguistically valid way to group similar online texts together

Clustering
*text types**

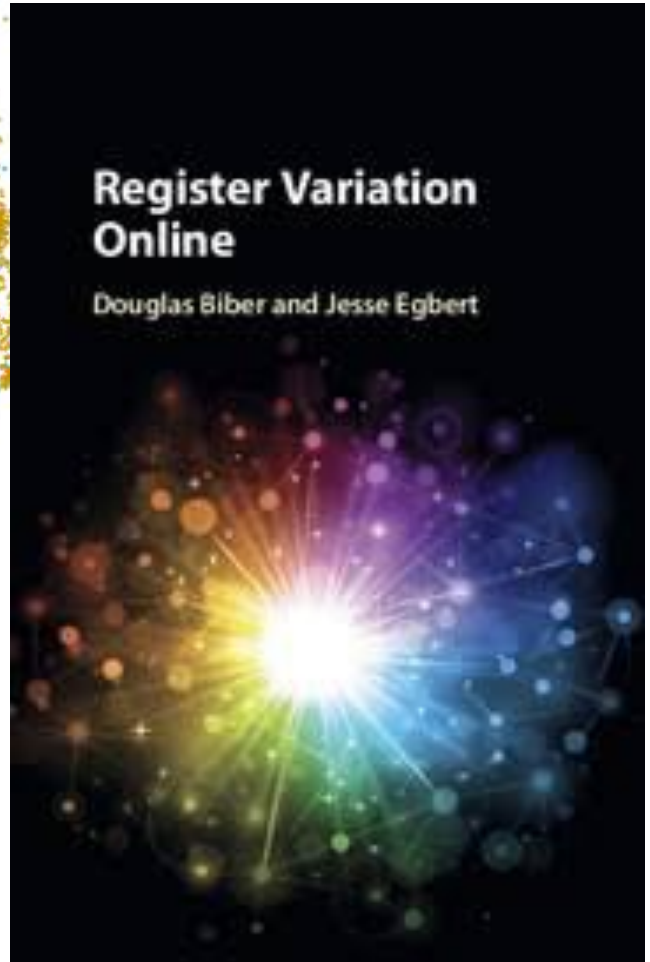


Linguistics Open Access
Volume 27, Issue 1, 1989, Pages 3-44

A typology of English texts (Article)

Biber, D.

University of Southern California, United States



* Texts with similar linguistic characteristics (Biber, 1989)

Sparse data

$m > 100{,}000$

	able	about	actually	after	… word *m*
Text 1	0	2	0	1	
Text 2	0	4	0	0	
Text 3	0	0	0	0	
Text… N					

$N > 10{,}000$

Specific objectives

- Go beyond traditional sparse document representation
  - SVM-based classification of online registers with Biber tags in Laippala et al., submitted

1) Do predicted document embeddings capture linguistic meaning?
2) To what extent do they reflect differences between registers?
3) Does the clustering of these embeddings produce functionally interpretable groupings?

Idea behind word2vec

- “You shall know a word by the company it keeps”

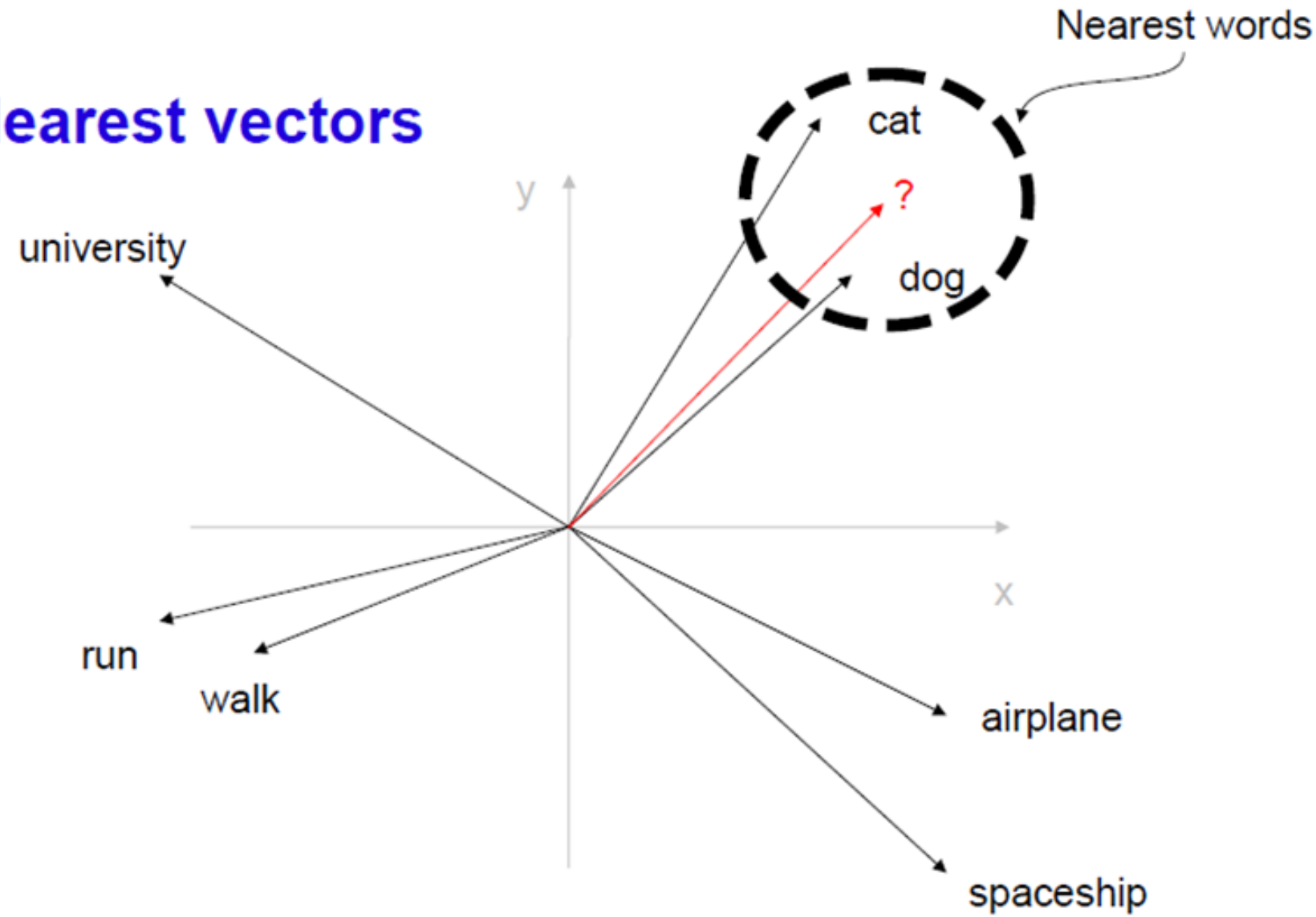
J.R. Firth, 1957

- “We found a cute, hairy wampimuk sleeping behind the tree.”

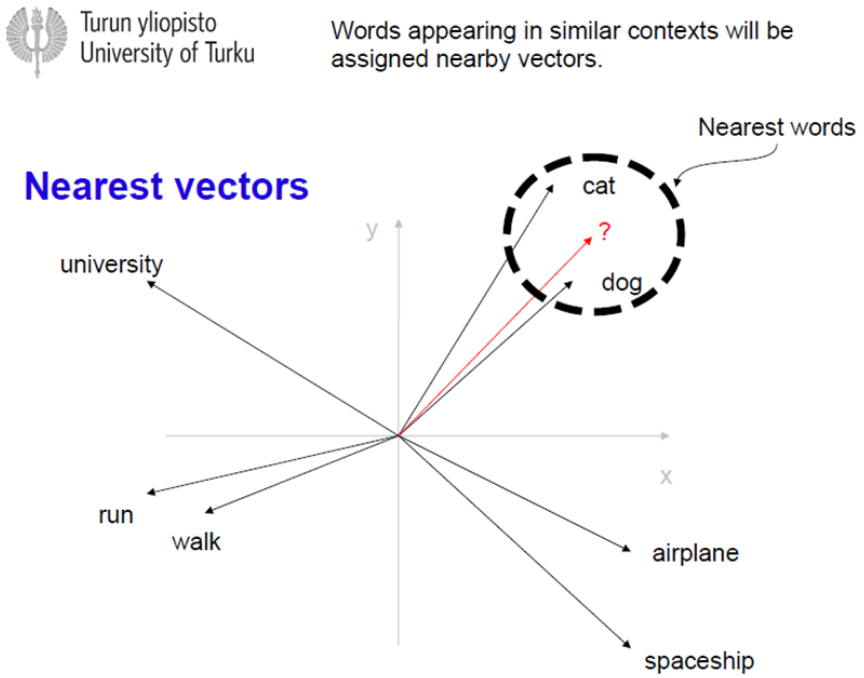
Turun yliopisto
University of Turku

Words appearing in similar contexts will be assigned nearby vectors.

Nearest vectors



Clustering registers



- Original word2vec
  - Context: co-occurrence of words
  - Output: predicted word embedding
  - man _ hamburgers

- From word-level representation to document-level
  - Context: word co-occurrence enriched with syntax
  - Output: predicted document embedding

Corpus of online registers of English (CORE)
(Egbert, Biber and Davies 2015)

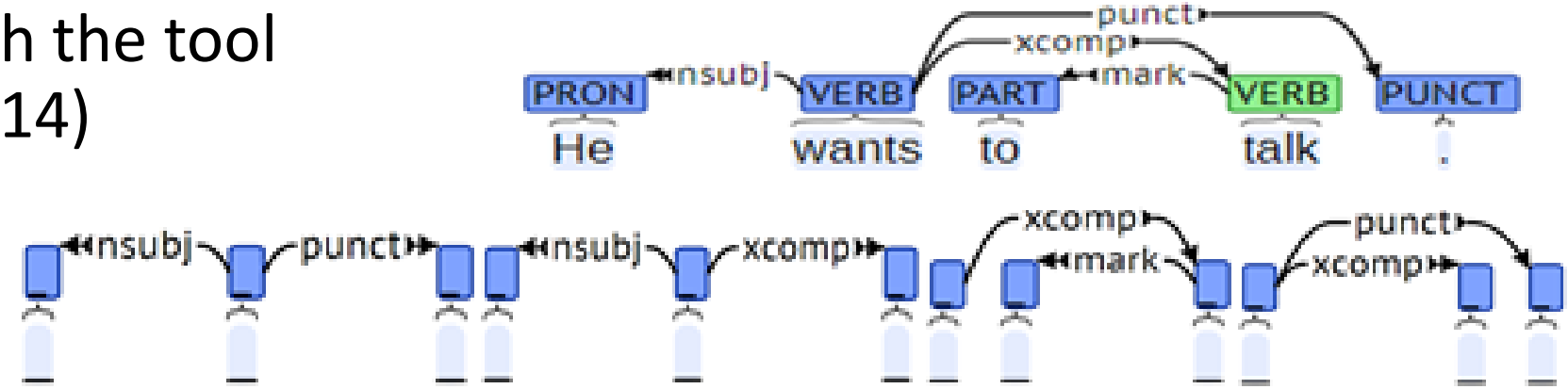
- Unrestricted sample of the searchable English web
- 48,571 documents, > 50 million words
- Manually coded for registers
- 8 main registers functional labels
  - *narrative, informational description, lyrical, informational persuasion, spoken, opinion, how-to, interactive discussion*
- Divided into 33 sub-registers
  - Such as *news, personal blogs, encyclopedia articles, how-to*
- Highly imbalanced!

Current study

- Sub-register level
- The most frequent registers with high level of inter-annotator agreement
- → 25,038 documents
- → 23 (sub-)registers
- Registers were not balanced (e.g., News cover for almost 30% of the data)
- Realistic distribution

Document embeddings for online English

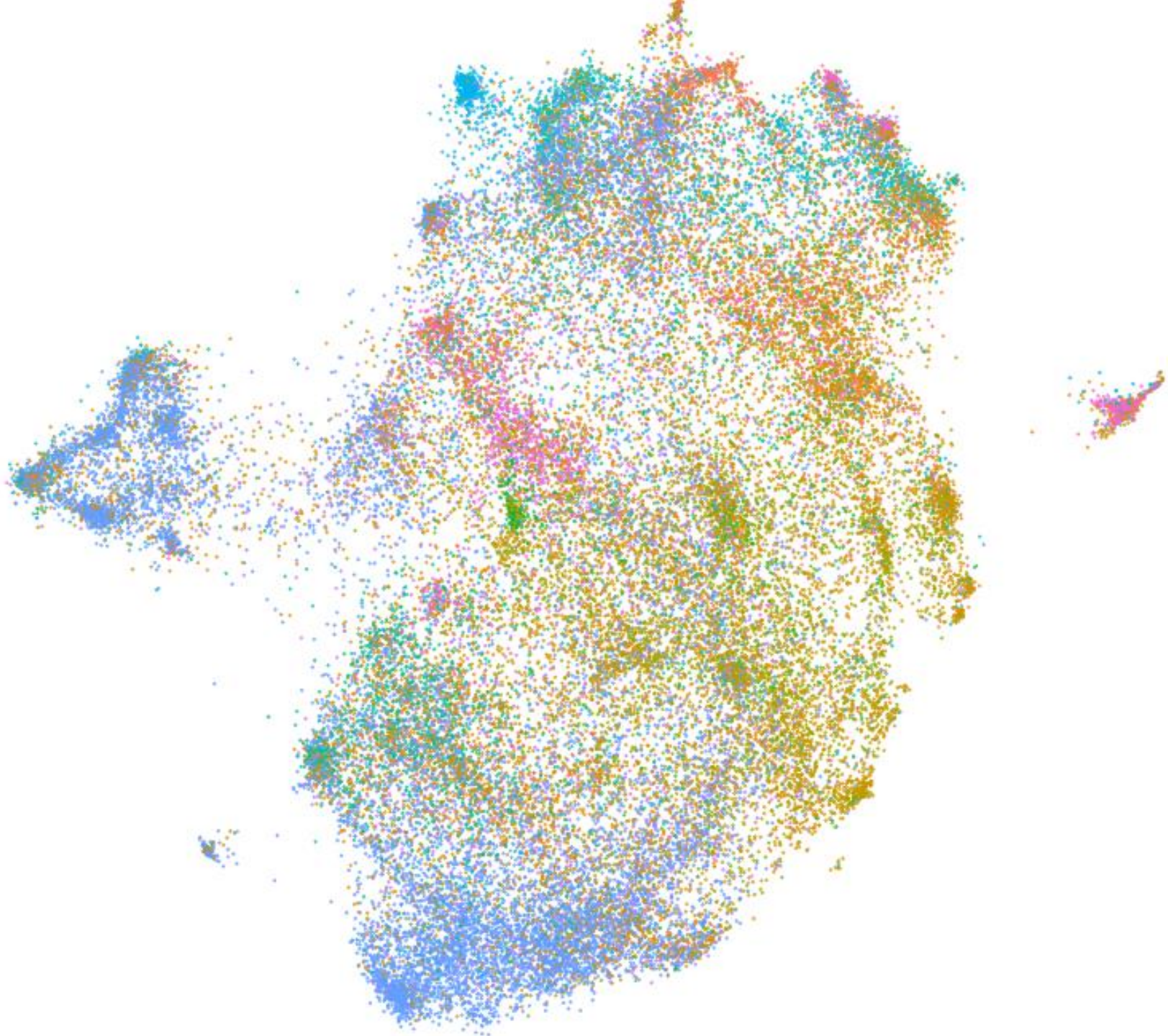
- Model: Word2vecf (Levy and Goldberg 2014)
- Training data CORE + documents from Global Web-Based Corpus of English → 120,000 documents
- Contexts: lemmata + *unlexicalized syntactic biarcs*
- Syntactic parser by Dozat et al. (2017) following Universal Dependency Schema, available for 50+ languages
- Biarcs generated with the tool by Kanerva et al. (2014)



Document embeddings for online English

- Dense representation of a document

	1	2	3	4	… 300
Text 1	0.233608	0.542426	0.376280	0.313996	0.186521
Text 2	0.132363	0.141933	-0.197238	-0.012108	0.046967
Text N					



Evaluation

Part 1: Evaluation of document embeddings

- Qualitative inspection of documents with nearby embeddings
- Text classification experiments to quantify the goodness of the document embeddings

Part 2: Evaluation of cluster solution

- Optimal number of clusters
- Text classification experiments to quantify the goodness of the cluster solution
- Evaluation of internal structuring of the cluster solution

Qualitative evaluation

Jelavic calls for Old Firm to come to the **Premier League**

New Everton signing Nikica Jelavic fears Rangers and Celtic will become second-class clubs if they remain in Scottish football. The Croatian striker moved to Goodison Park on **transfer** deadline day, with Rangers manager Ally McCoist powerless to prevent the **6m** move. "Rangers and Celtic can't compete with English clubs **financially,** so it would be very important to them if they could join the Premier League one day,"

Jelavic calls for Old Firm to come to the **Premier League**

New Everton signing Nikica Jelavic fears Rangers and Celtic will become second-class clubs if they remain in Scottish football. The Croatian striker moved to Goodison Park on transfer deadline day, with Rangers manager Ally McCoist powerless to prevent the **6m** move. "Rangers and Celtic can't compete with English clubs **financially,** so it would be very important to them if they could join the Premier League one day,"

The Premier League took a big step towards introducing a break-even rule following a meeting in London. There was no formal agreement between the 20 chairmen and chief executives over how to introduce **costs controls**. But clubs agreed to focus on a model similar to the **Financia**l Fair Play **regulations** introduced by Uefa, which require teams to avoid making losses.

Document embeddings predicting the registers

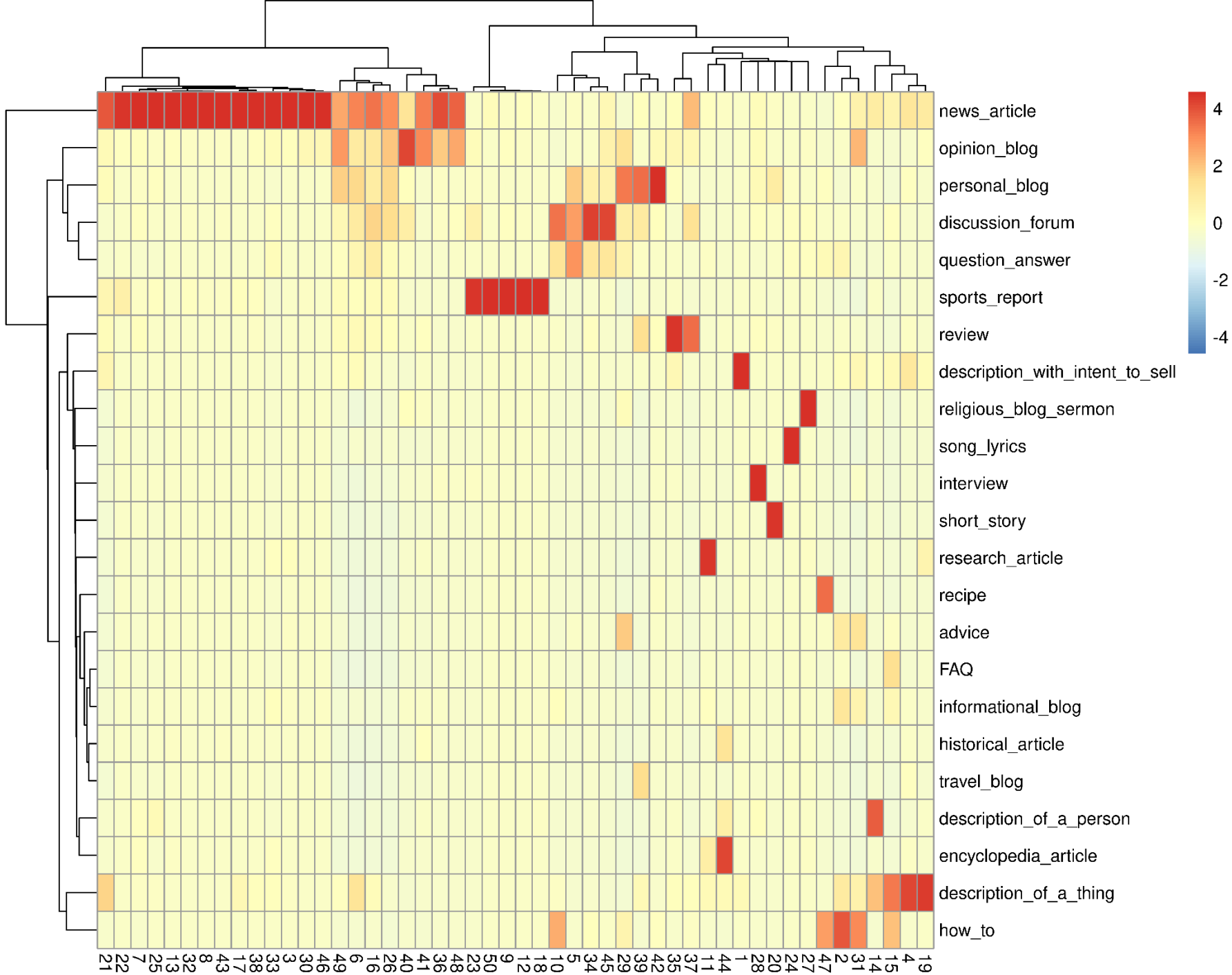
- Text classification with a polynomial support vector machine
  - svm.SVC with a polynomial kernel in Scikit learn, 80%/20% train / test sets
- Two experiments
  - Full dataset
  - Naïve dataset with four registers with equal class sizes (*Travel blog, news, discussion forum, sport report)*

Data	Response	Precision	Recall
Full data (N=25,038)	Registers (N=23)	71%*	64%*
Naïve data (N=472)	Registers (N=4*)	96%	96%

* SVM-based results using lexical information + Biber tags Prec 74%, Rec 75% in Laippala et al. (submitted)

Part 2: Clustering

- Generative mixture-model approach based on the von Mises-Fisher (MVMF) distribution (Banerjee et al., 2005)
  - Distribution arises naturally for data distributed on the unit (hyper)sphere, such as our embeddings with normalized lengths
  - Performs well with high-dimensional data
- Evaluation of the cluster solution
  - BIC (Bayesian information criterion) to evaluate the optimal number of clusters
  - Kappa to analyze the concentration of the cluster (similar to standard deviation in Gaussian distributions), how concentrated the documents vectors are around the mean vector
  - Alpha to estimate the probability of a given cluster

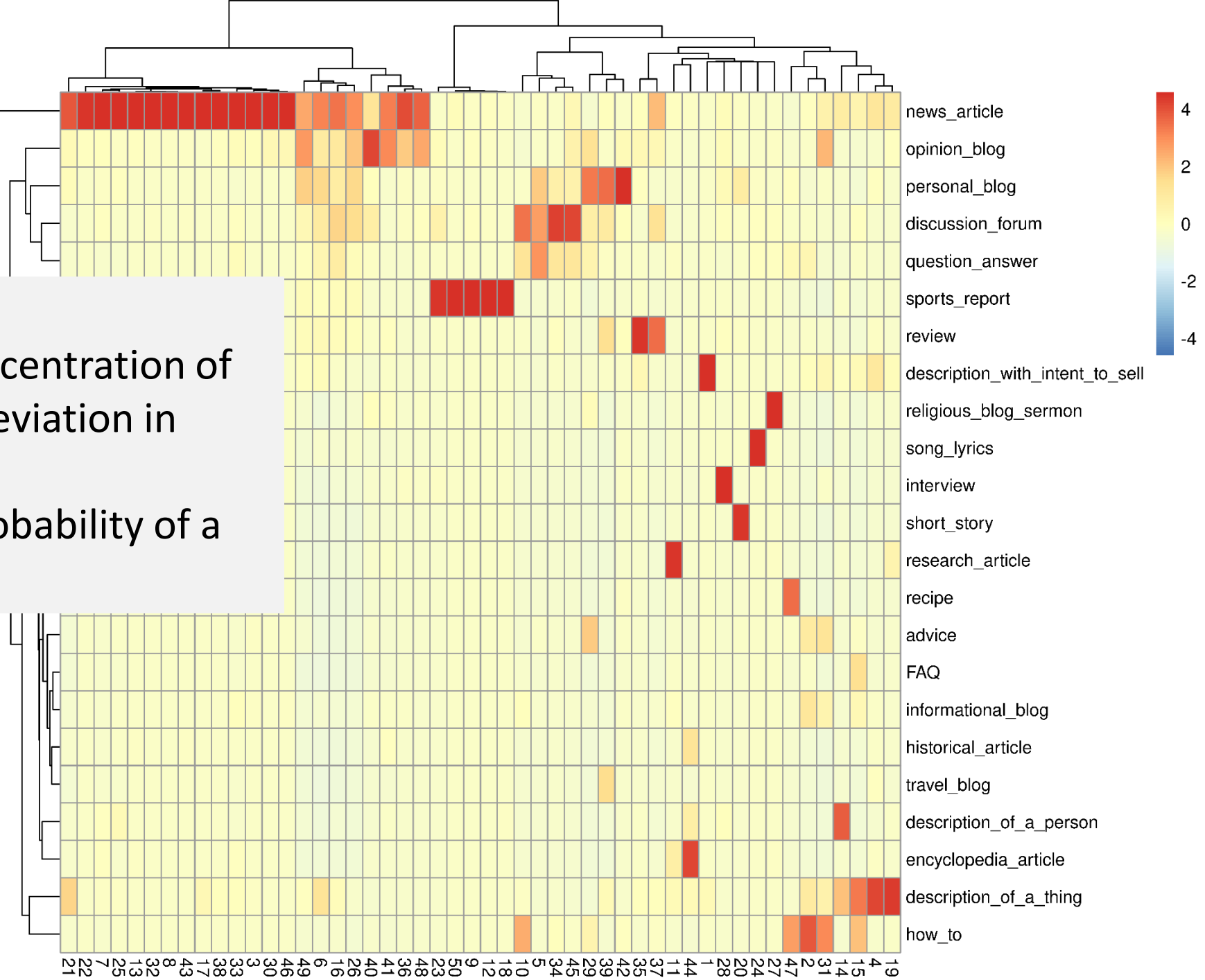

news_article
opinion_blog
personal_blog
discussion_forum
question_answer
sports_report
review
description_with_intent_to_sell
religious_blog_sermon
song_lyrics
interview
short_story
research_article
recipe
advice
FAQ
informational_blog
historical_article
travel_blog
description_of_a_person
encyclopedia_article
description_of_a_thing
how_to
4
2
0
-2
-4
21 22 7 25 13 32 8 43 17 38 33 3 30 46 49 6 16 26 40 41 36 48 23 50 9 12 18 10 5 34 45 29 39 42 35 37 11 44 1 28 20 24 27 47 2 31 14 15 4 19

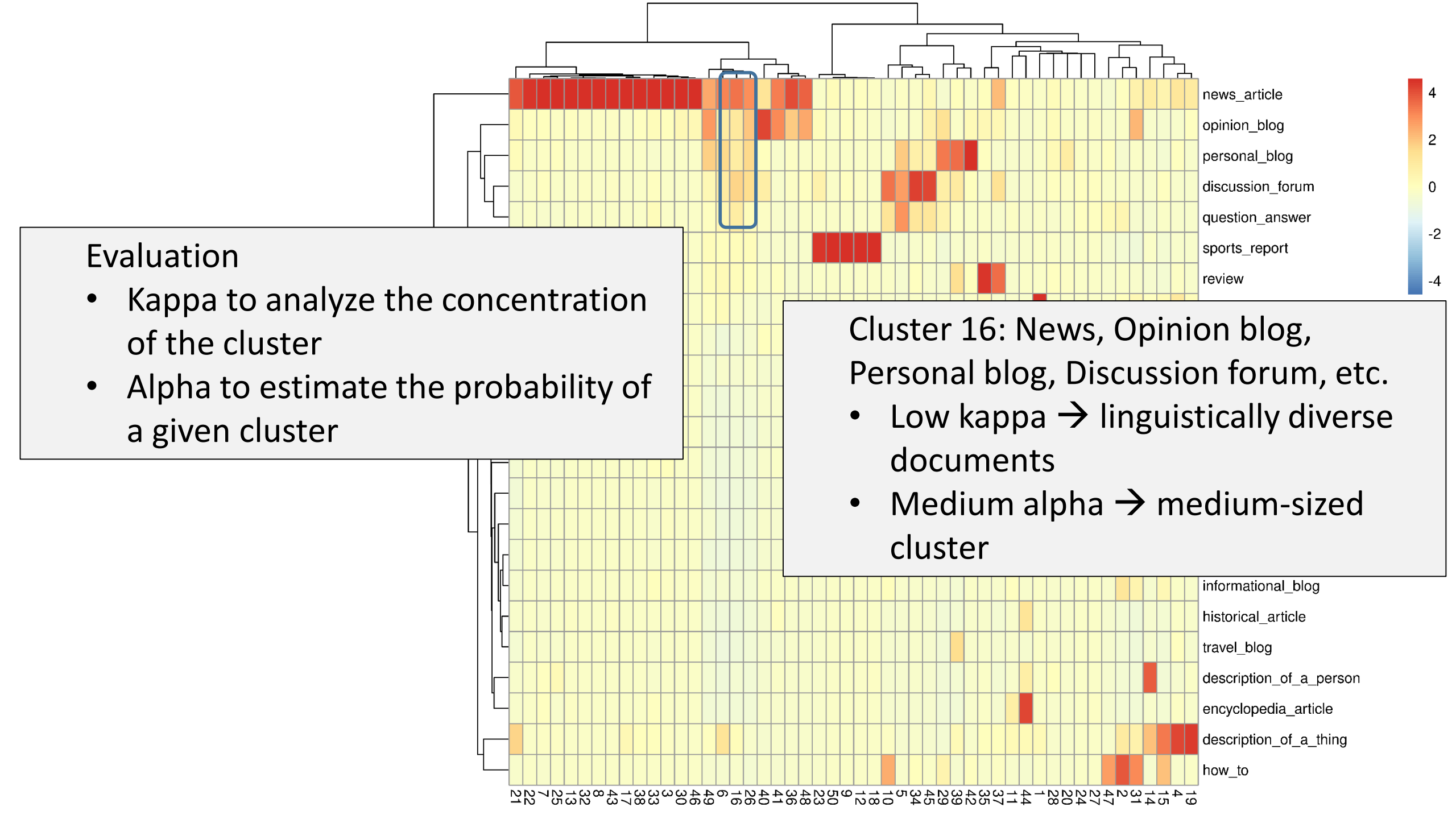
The goodness of the cluster solution

Data	Response	Precision	Recall
Full data (N=25,038)	Clusters (N=50)	84%	81%
Naïve data (N=472)	Clusters (N=4)	99%	99%
Random data (N=472)	Clusters (N=4)	34%	34%

Evaluation

- Kappa to analyze the concentration of the cluster (~ standard deviation in Gaussian distributions)
- Alpha to estimate the probability of a given cluster



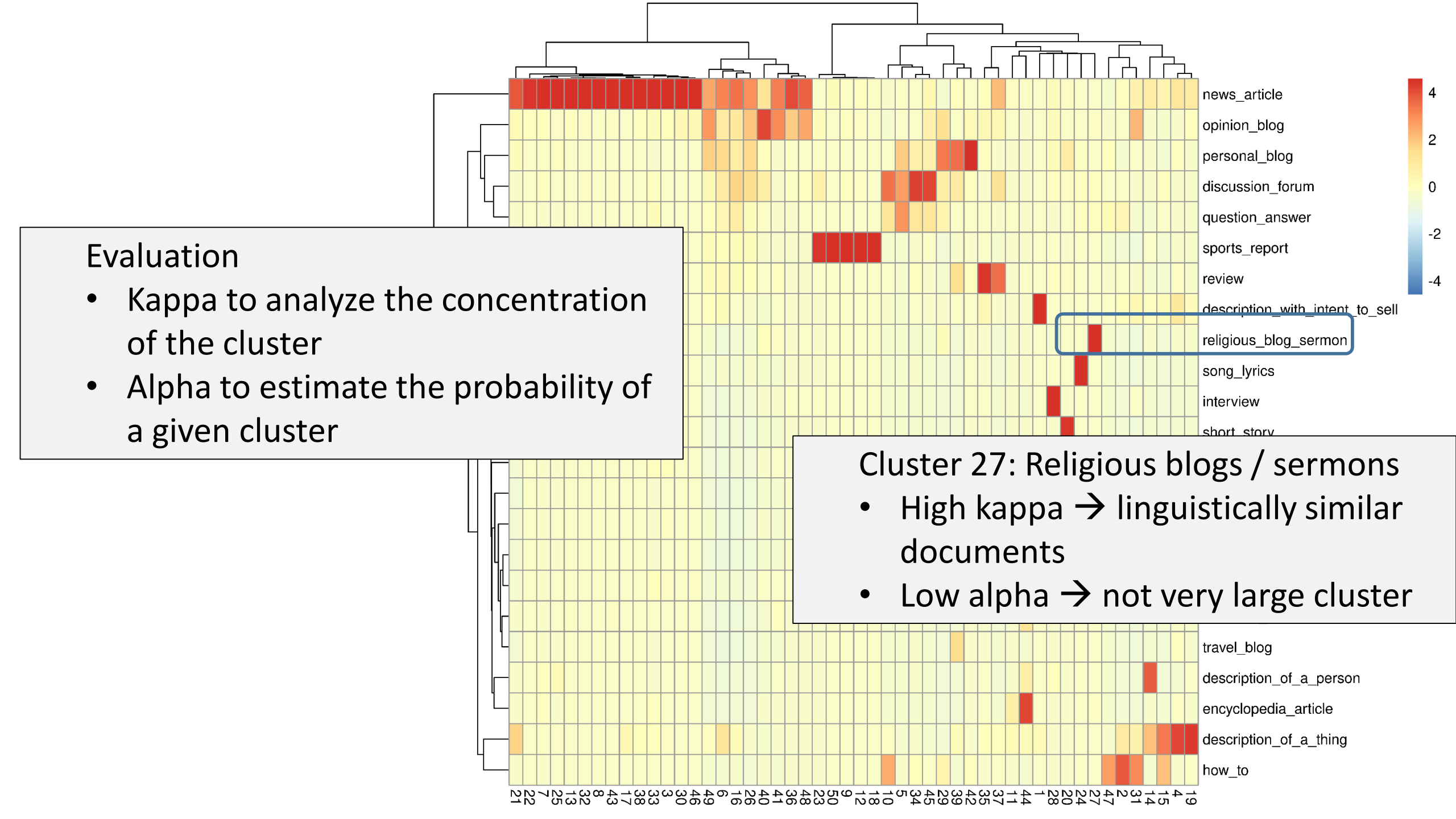


Evaluation

- Kappa to analyze the concentration of the cluster
- Alpha to estimate the probability of a given cluster

Cluster 16: News, Opinion blog, Personal blog, Discussion forum, etc.

- Low kappa → linguistically diverse documents
- Medium alpha → medium-sized cluster



Evaluation

- Kappa to analyze the concentration of the cluster
- Alpha to estimate the probability of a given cluster

Cluster 27: Religious blogs / sermons

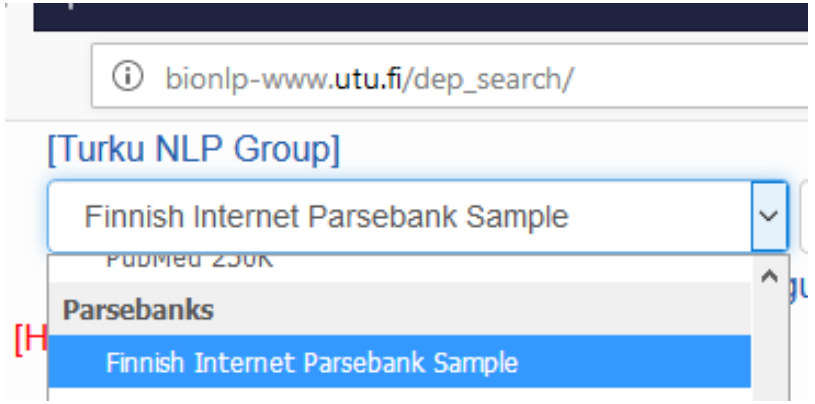
- High kappa → linguistically similar documents
- Low alpha → not very large cluster

To conclude

1) Predicted document embeddings based on combination of words and syntax
    1) Capture linguistic meaningful functions of online text data
    2) Can be used for supervised and unsupervised tasks
2) Mixture model clustering with von Mises-Fisher distribution
    1) Excellent fit to the data
    2) Functionally interpretable clusters
    3) Unique perspective on the internal structuring of a given cluster

Future perspectives

- Extending the use of predicted document embeddings to other languages (Finnish, Swedish, French + others in UD)
- Explaining the linguistic dimensions of document embeddings
- Compare cluster solution to supervised (multilingual) experiments
  - Laippala et al. 2019: *Toward Multilingual Identification of Online Registers* In Proceedings of NoDaLida in September 2019.
- Establishing register-specific sub-corpora to Universal Parsebanks, available through Dep search



Thanks!

FULBRIGHT FINLAND FOUNDATION

EMIL AALTOSEN SÄÄTIÖ