myCorpus

Exploring a collection of JP>EN translations by a single translator

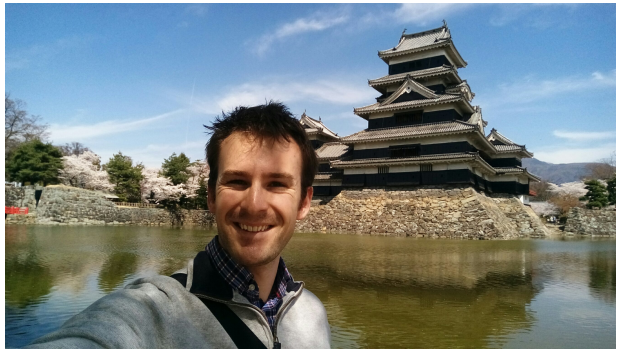
Mark Streer • METIS (DS/ML) • 2021-11-12

Client

Technical translator

Japanese>English

~500 research articles from 2013-2021



How can NLP improve my translation workflow?

How can I apply NLP tools best to add value to language services for Japanese professionals?

Motivation

Use case #1: Translator matcher

User: Contracting agency

Compare source text* with past translations at agency

Prioritize translators most experienced in topic(s)

Use case #2: Genre matcher

User: Translators

Compare source text* with biomedical corpus

Reference terms/language of papers on same topic(s)

Motivation, Ulterior

Machine translation output acceptable:
- To check ‘gist’ of text
- To inspect rare words
- NOT PUBLISHABLE

Significant overlap with NLP
- Tokenization
- Language model training

Neural translation is state-of-the-art
- Preprocessing, tokenization, NER
- Deep learning (‘zero-shot learning’)

Corpus

English translations from 2020-2021 (n=110)

- 90% research articles, 10% abstracts
- Consistent format (IMRAD), voice
- Diverse subdomains in medicine, healthcare

```
Apoptotic effects of a thioether analog of vitamin K3 in a human leukemia cell line
Cancer rehabilitation care as provided by designated cancer care hospitals in Japan
Effectiveness of employment support program in cooperation with psychiatric day care and Hello Work
```

Tasks

Clustering: K-means

Topic modeling: SVD NMF LDA

Corpus

English translations from 2020-2021 (n=110)

- 90% research articles, 10% abstracts
- Consistent format (IMRAD), voice
- Diverse subdomains in medicine, healthcare

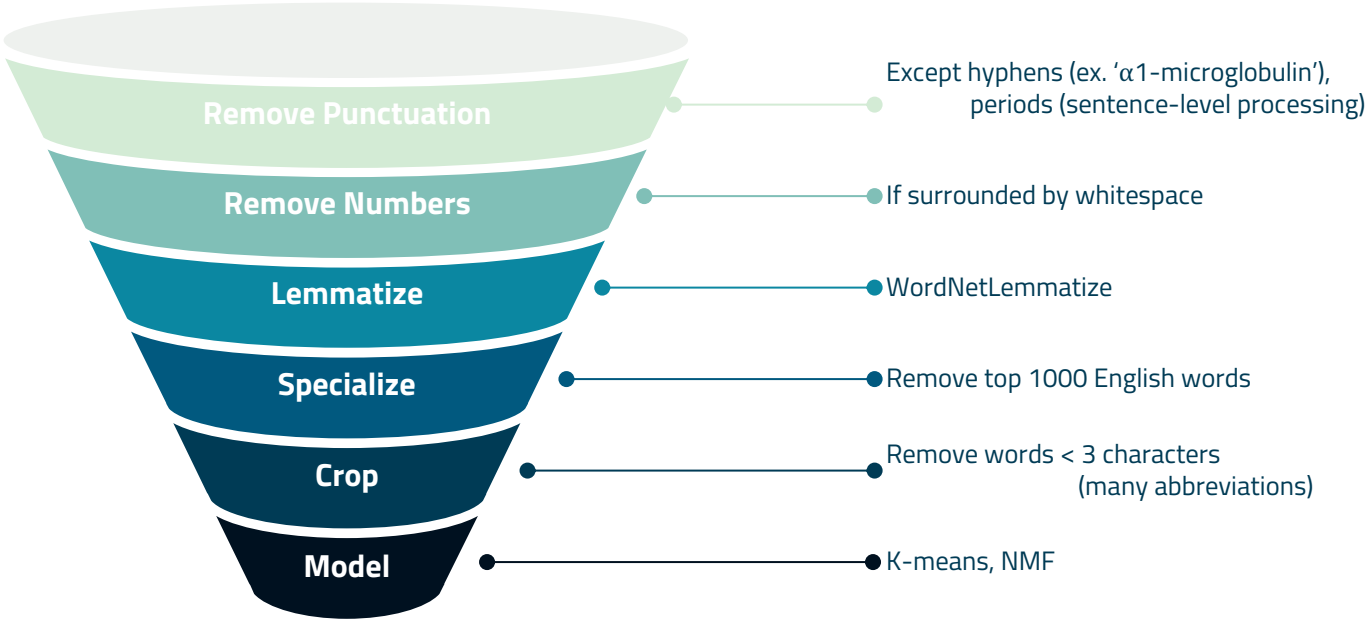
```
Apoptotic effects of a thioether analog of vitamin K3 in a human leukemia cell line
Cancer rehabilitation care as provided by designated cancer care hospitals in Japan
Effectiveness of employment support program in cooperation with psychiatric day care and Hello Work
```

Tasks

Clustering: K-means

Topic modeling: ~~SVD~~ **NMF** ~~LDA~~

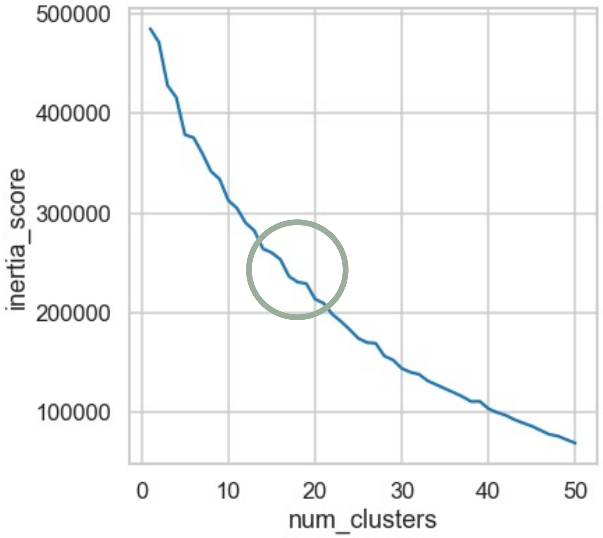
PREPROCESSING



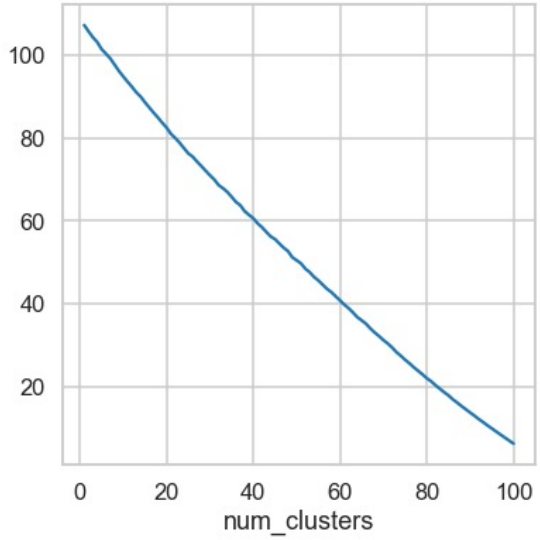
num_topics = 16

K-MEANS CLUSTERING

Count



TF-IDF



Hyperparameter Decisions

Don’t want words unique to single doc

min_df = 2

Don’t want really common words

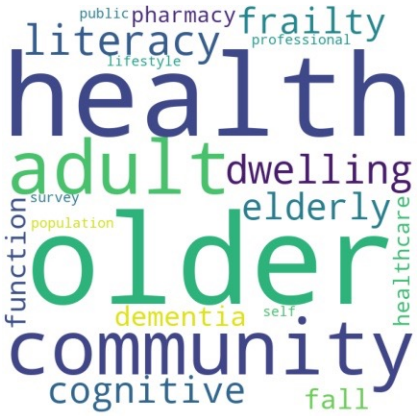
max_df = 0.5

Reward rare words (likely domain-specific)

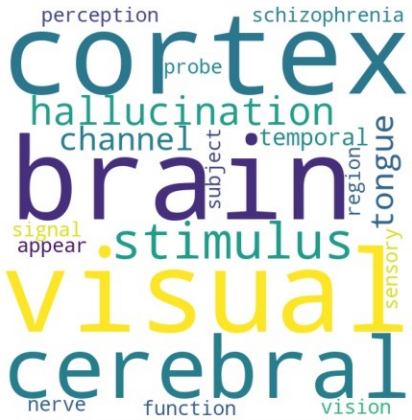
TfidfVectorizer > CountVectorizer

NMF Tuning

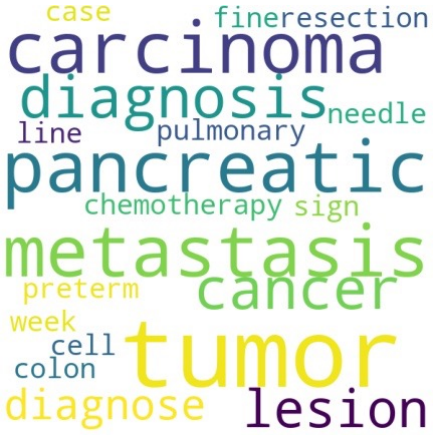
	CountVectorizer	TfidfVectorizer min_df=2, max_df=0.5, max_features=1000
Topic 1	adult, older, physical, environment, factor, population, health, elderly, live, association	older, **health, community**, adult, literacy, dwelling, cognitive, elderly, frailty, dementia
Topic 2	healthcare, professional, older, study, community, review, experience, approach, relationship, participant	**surgery**, case, procedure, cancer, operation, laparoscopic, surgical, complication, gastric, underwent
Topic 3	screen, health, rate, cancer, population, score, higher, factor, knowledge, subject	**covid**, infection, pandemic, emergency, pulmonary, chest, suspect, nurse, sars, cov
Topic 4	cancer, provide, professional, survey, staff, program, set, regard, provider, physical	**cell, culture**, cartilage, tissue, expression, protein, organ, **induce,** differentiation, apoptosis



Community Health



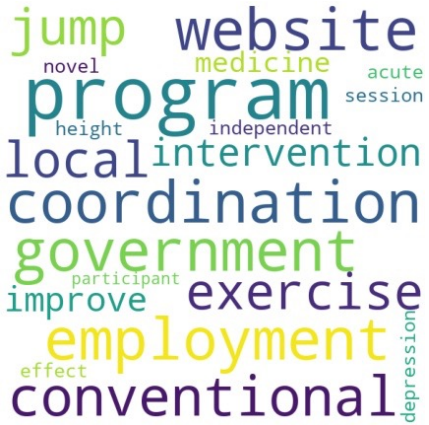
Neuroscience



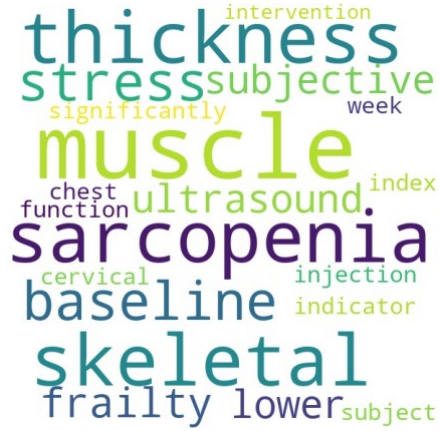
Cancer

COVID-19

Surgery



SocWelfare



Physiology

StemCell

How Many Topics?

num_topics=16

- CommHealth
- Surgery
- COVID-19
- StemCell
- Physiology
- Cancer
- Neuroscience
- SocWelfare
- Rehabilitation
- ObGyn
- NursingEd
- Endovascular
- Stroke
- ClinTrial
- PubPolicy
- Genomics

num_topics=8

- Nursing
- Surgery
- COVID-19
- MolBioTherap
- FuncHealth
- Cancer
- Neuroscience
- Rehabilitation

How Many Topics?

num_topics=16

- CommHealth
- Surgery
- COVID-19
- StemCell
- Physiology
- Cancer
- Neuroscience
- SocWelfare
- Rehabilitation
- ObGyn
- NursingEd
- Endovascular
- Stroke
- ClinTrial
- PubPolicy
- Genomics

num_topics=8

- Nursing
- Surgery
- COVID-19
- MolBioTherap
- FuncHealth
- Cancer
- Neuroscience
- Rehabilitation

How Many Topics?

num_topics=16

- CommHealth
- Surgery
- COVID-19
- StemCell
- Physiology
- Cancer
- Neuroscience
- SocWelfare
- Rehabilitation
- ObGyn
- NursingEd
- Endovascular
- Stroke
- ClinTrial
- PubPolicy
- Genomics

num_topics=8

- Nursing
- Surgery
- COVID-19
- MolBioTherap
- FuncHealth
- Cancer
- Neuroscience
- Rehabilitation

How Many Topics?

num_topics=16

- CommHealth
- Surgery
- COVID-19
- StemCell
- Physiology
- Cancer
- Neuroscience
- SocWelfare
- Rehabilitation
- ObGyn
- NursingEd
- Endovascular
- Stroke
- ClinTrial
- PubPolicy
- Genomics

num_topics=8

- Nursing
- Surgery
- COVID-19
- MolBioTherap
- FuncHealth
- Cancer
- Neuroscience
- Rehabilitation

How Many Topics?

num_topics=16

- CommHealth
- Surgery
- COVID-19
- StemCell
- Physiology
- Cancer
- Neuroscience
- SocWelfare
- Rehabilitation
- ObGyn
- NursingEd
- Endovascular
- Stroke
- ClinTrial
- PubPolicy
- Genomics

num_topics=8

- Nursing
- Surgery
- COVID-19
- MolBioTherap
- FuncHealth
- Cancer
- Neuroscience
- Rehabilitation

TOPIC-CLUSTER LOADINGS

```
Top Three Topics in Cluster C_6 (mean of docs)
('10.ObGyn', 0.3830142787789891)
('02.Surgery', 0.07472117118840474)
('13.Stroke', 0.02816125709660942)


Top Three Docs in Cluster (i.e. nearest to centroid):
Doc 38: Cluster 6
Successful IVF pregnancy and delivery in an infertile patient with true hermaphroditism
Doc 104: Cluster 6
An exploration of factors guiding expectant mothers' decisions to keep or give up their child for
icted pregnancies:
Doc 56: Cluster 6
Midwife education programs to foster maternal identity in women who become pregnant with twins via
ctive technology
```

TOPIC-CLUSTER LOADINGS

```
Top Three Topics in Cluster C_3 (mean of docs)
('13.Stroke', 0.774323278068024)
('10.ObGyn', 0.014845037241769018)
('12.Endovascular', 0.008097451545758697)


Top Three Docs in Cluster (i.e. nearest to centroid):
Doc 92: Cluster 3
Development and feasibility of an intervention model for family surrogate decision-makers of acute stroke
Doc 14: Cluster 3
Everyday experiences of wives of elderly stroke victims after home discharge
Doc 72: Cluster 10
Protection Stroke Code for COVID-19 based on Task Calc. Stroke: Ensuring the continued provision of stroke
```

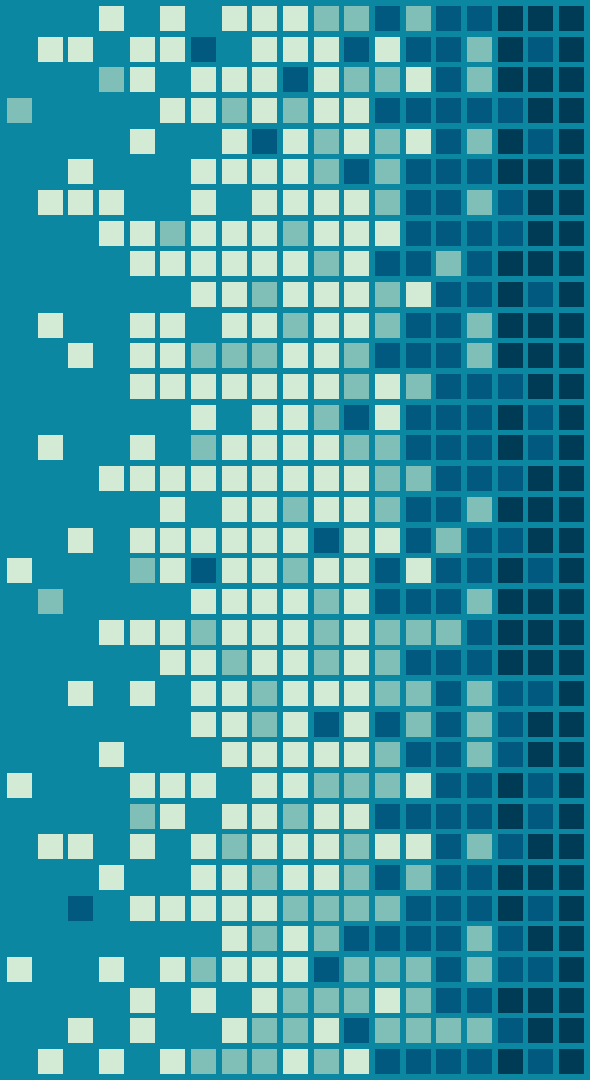
TOPIC-CLUSTER LOADINGS

```
Top Three Topics in Cluster C_10 (mean of docs)
('03.COVID', 0.3731914206943709)
('13.Stroke', 0.060268688207677966)
('02.Surgery', 0.02998768228506464)


Top Three Docs in Cluster (i.e. nearest to centroid):
Doc 74: Cluster 10
Extracorporeal CPR should not be performed on confirmed or suspected COVID-19 patients
Doc 101: Cluster 10
Protection from tuberculosis for infants in the COVID-19 era.
Doc 9: Cluster 10
A case of bilateral pulmonary embolism occurring under self-isolation during the COVID-19 pandemic
```

Next steps:

- *Same pipeline for Japanese docs*
  - *Do same topics emerge?*
  - *Do same top words?*
- *Post-MT topic modeling for paper matching*
  - *Syntax not important in BOW models*
  - *Requires preprocessing of large corpora*
- *Document similarity in neural translation*
  - *Topic weights as word features?*



THANKS!

Any questions?

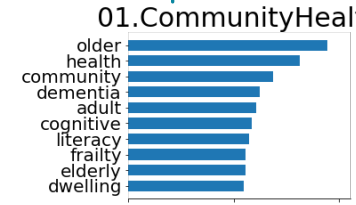
Github: streerm

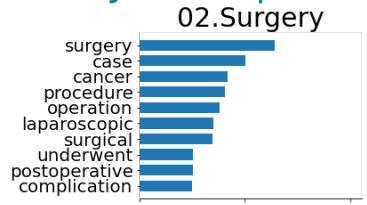
mark.streer@gmail.com

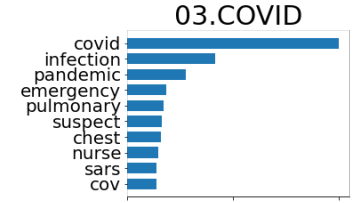
appendix

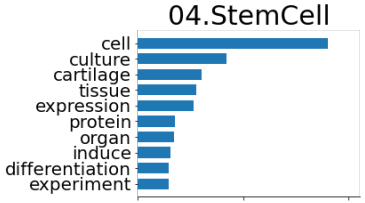
Bring the attention of your audience over a key concept using icons or illustrations

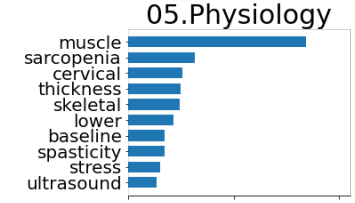
Top Words by Topic (n=16

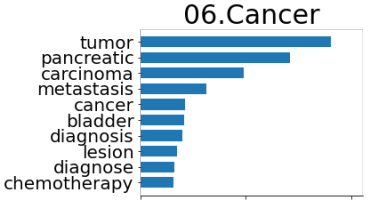


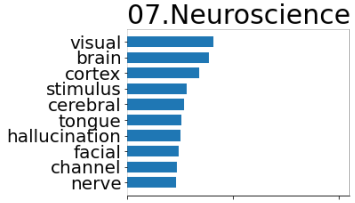


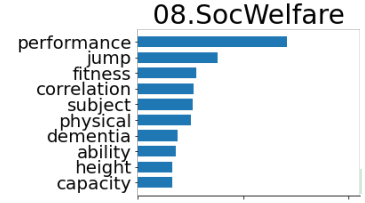


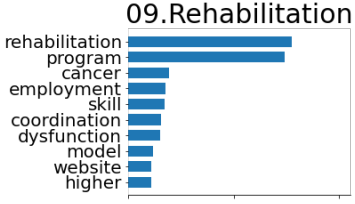


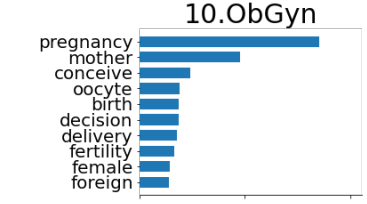


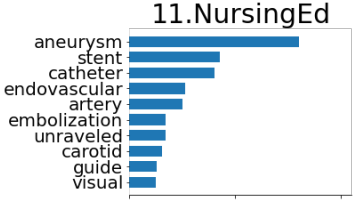


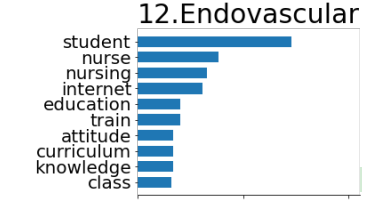


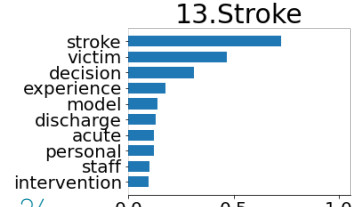


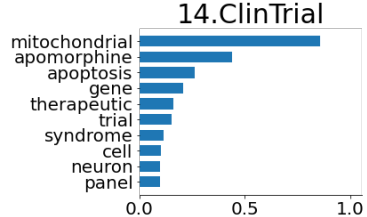


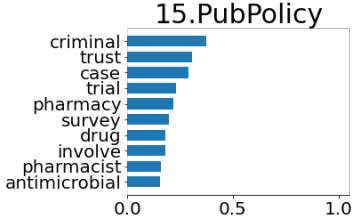


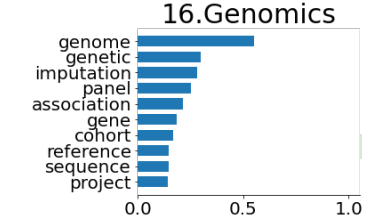




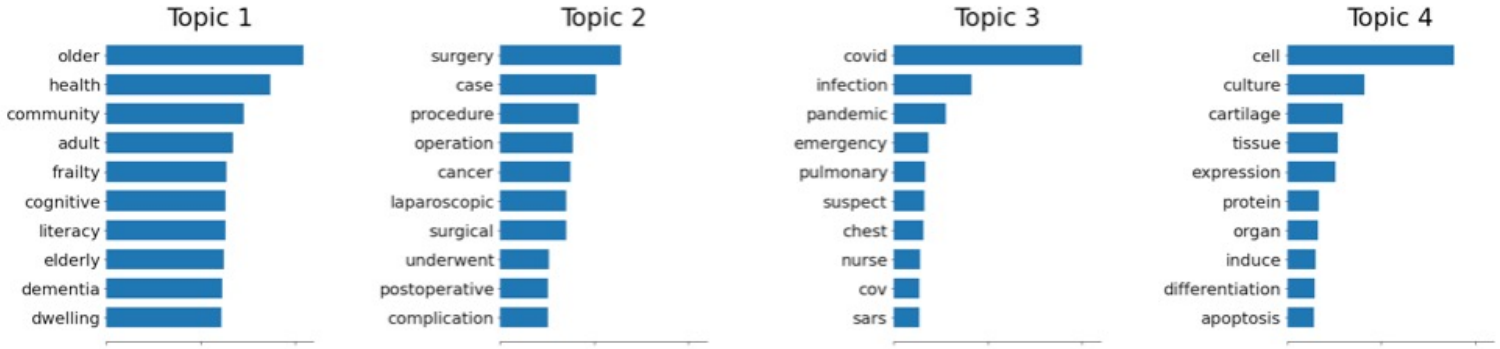


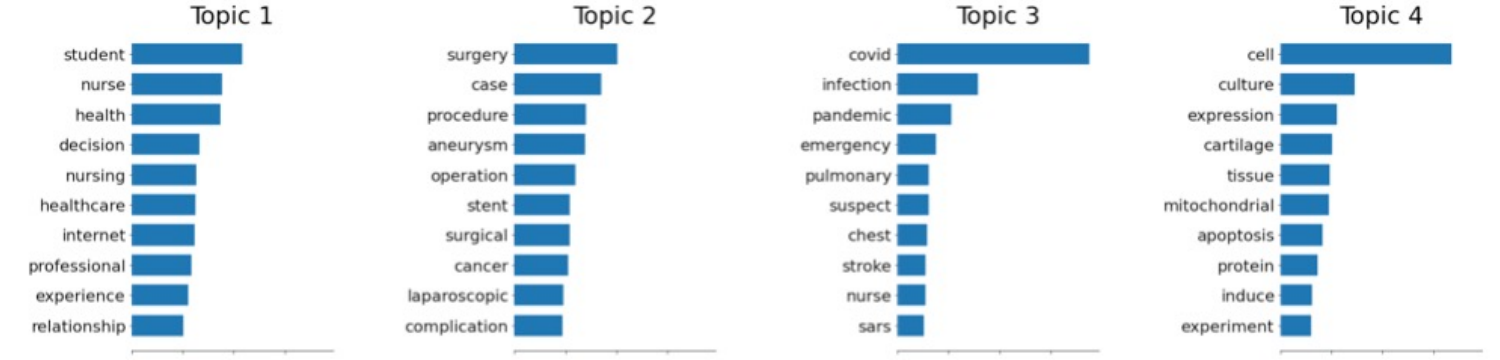




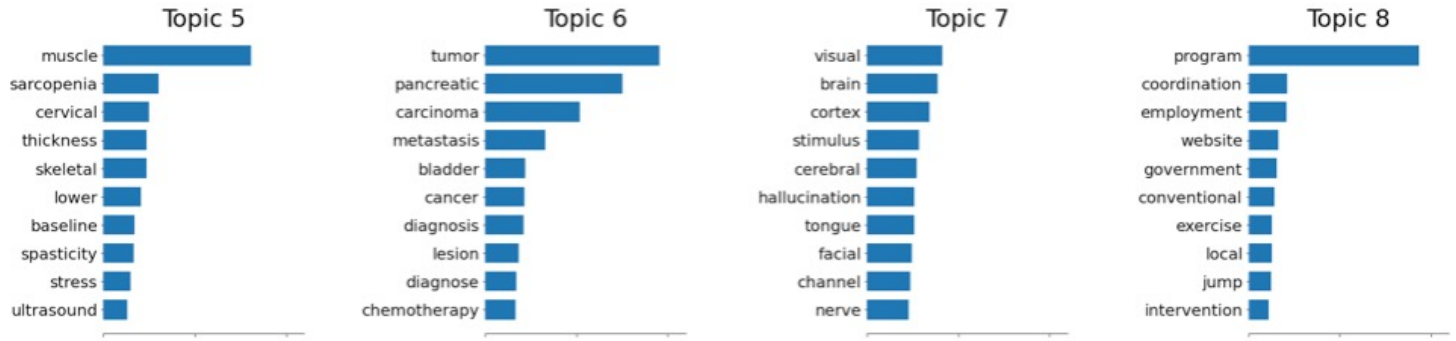


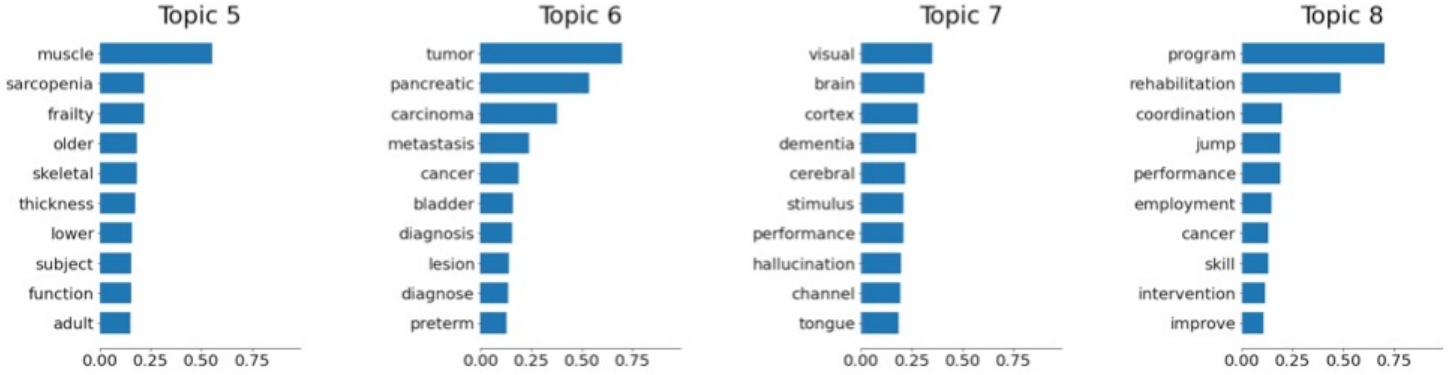
Top Words by Topic (n=16 v. n=8) p.1





Top Words by Topic (n=16 v. n=8) p.2





Motivation

Languages express the same ideas using different words and syntax.

EN: I am hungry. PRON-V(to be)-ADJ

DE: Ich habe Hunger. PRON-V(to have)-N

JP: O-naka ga suita. POL-N-PART-V(past)

お腹が空いた

(lit. (my) stomach is empty)