Exploring conceptual dimensions in encyclopedia articles

A study in token encodings and dimension reduction based on concept relation graphs

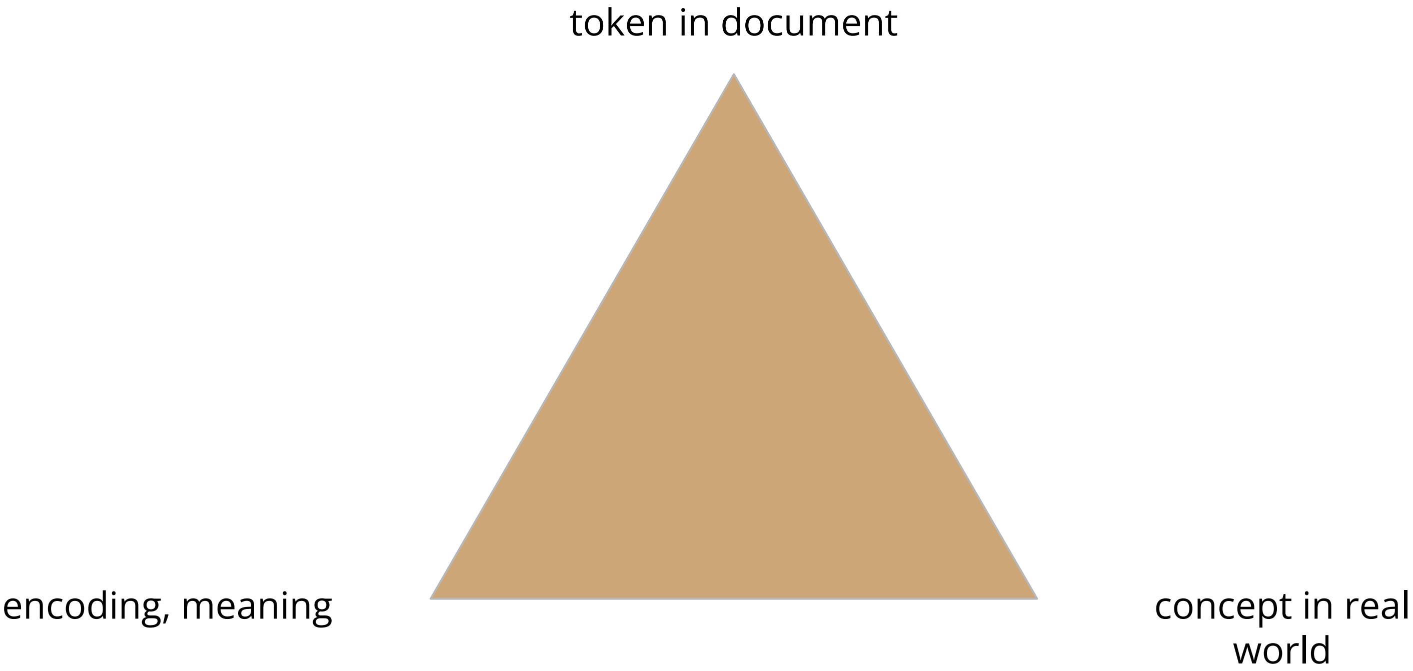
Seppo Nyrkkö
HELSLANG University of Helsinki

October 2022

Structure

1. Knowledge databases and textual articles
2. Conceptual spaces as a framework
3. Extracting dimensions from short encyclopedia articles as an experiment

The semiotic triad



token in document

encoding, meaning

concept in real world

1. Structured knowledge bases for machine reasoning

Machine-readable structured **knowledge bases** have been expected to be the hidden "**barrel of treasure**" for instance in concept recognition or term extraction tasks in NLP.

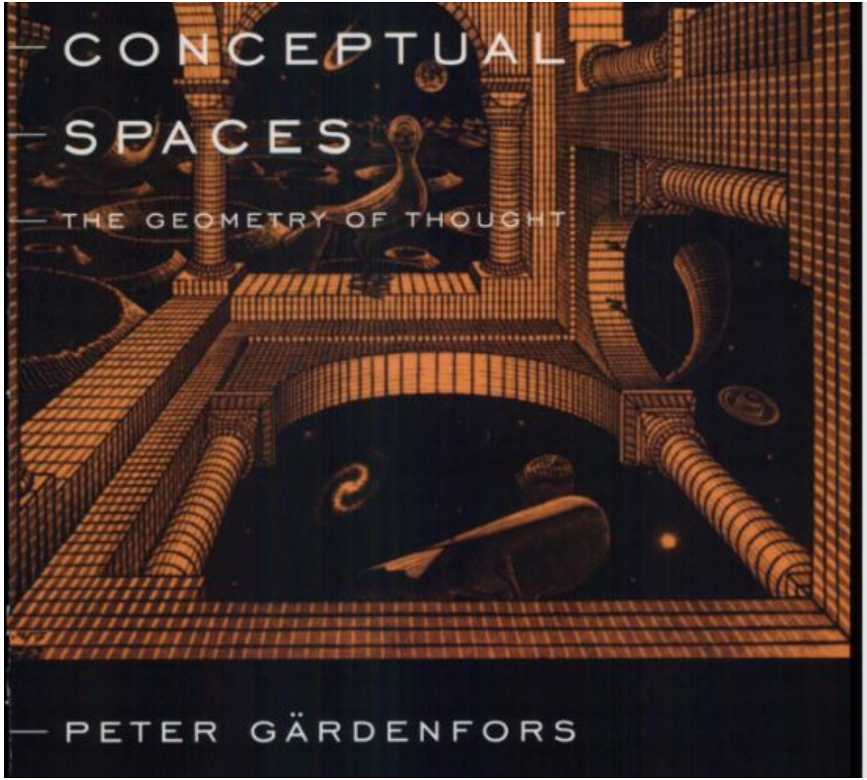
Practical real-life related information has been widely stored in formats such as **web ontology language** or inference rule-based AI languages.

Examples: KIF (Knowledge interchange format), SUMO (suggested upper level ontology)

Understanding text articles with the semantic web?

Unfortunately, utilization or development of **logical knowledge base** (in a larger scale) has proven to be complex and less straightforward than expected due to their "set theoretic" structure, and strict programming-language driven modeling style.

2. Conceptual spaces as a theoretical framework



Gardenfors, Peter. *Conceptual spaces: The geometry of thought*. MIT press, 2004.

Gärdenfors, Peter. "How to make the semantic web more semantic." *Formal ontology in information systems*. 2004.

2. Conceptual spaces as a theoretical framework

The **Conceptual spaces** is a theoretical framework proposed by *Peter Gärdenfors*, built for pragmatic engineering purposes in AI tasks, such as **pattern recognition** – or even human-like **inference** tasks.

The proposed **quality dimensions** in this model represent specific perception domains (like describing complex concepts such as *sweet green apples*), but may also express functional properties between concepts (like *running implies fast movement*).

A key feature along the dimensions is a metric of **similarity** or **distance**

Conceptual spaces as a theoretical framework...

- a proposed **framework for AI and** engineering purposes in AI tasks
- Pragmatic and scalable approach (vs subclass-superclass designs)
- **Suitable for pattern recognition** or even human-like inference tasks involving human perception (visual, tactile, audible, ...).
- The proposed dimensions in this model represent specific **perception domains**
- Able to describing **complex concepts** such as 'sweet red apples')
- Also **functional dependencies** and **expected variation** between quality dimensions
- "Comes with **similarity metric** built-in"

Programming with Conceptual Spaces?

A concept is a "well behaving" convex region in a quality domain, consisting of dimensions:

- Taste (sweetness, sourness ..)
- Color (R,G,B)
- Size (W,L,D)

Similarity is Distance is reversed

Prototypes are central concepts

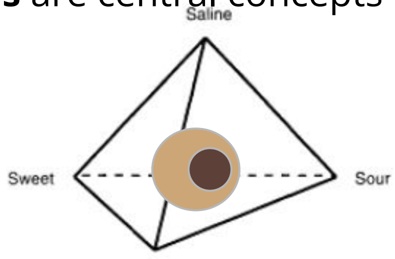


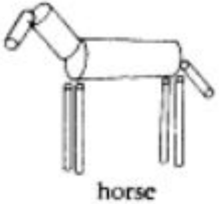
Figure 1.8
Henning's taste tetrahedron.

Allows **Complex concepts** for categories to be defined by prototypes, e.g.

Apple: fruit shape, color, taste

Horse: animal, size, 4 legged, ...

Running: movement, speed ...



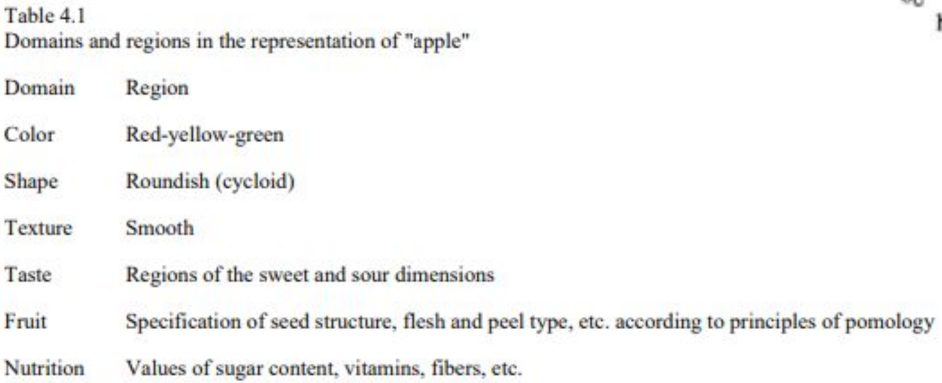


Table 4.1
Domains and regions in the representation of "apple"

Domain	Region
Color	Red-yellow-green
Shape	Roundish (cycloid)
Texture	Smooth
Taste	Regions of the sweet and sour dimensions
Fruit	Specification of seed structure, flesh and peel type, etc. according to principles of pomology
Nutrition	Values of sugar content, vitamins, fibers, etc.

Similarity is Distance is reversed

- City block metric
- Separable dimensions

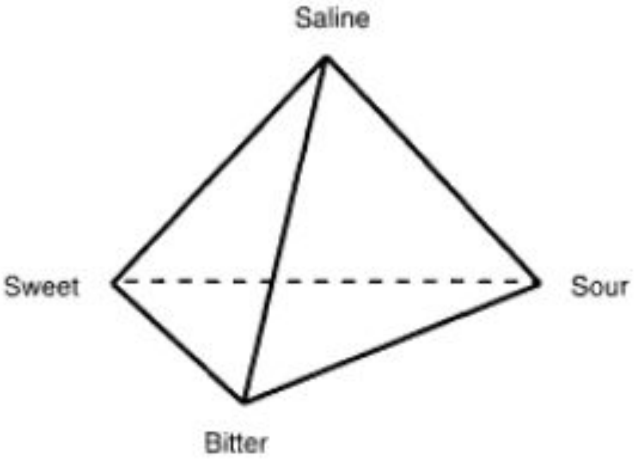
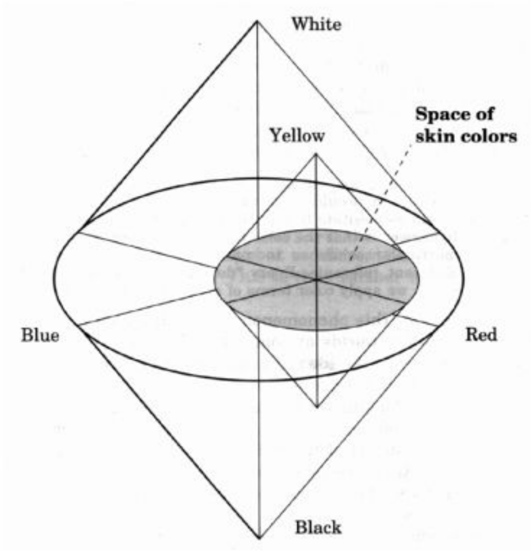


Figure 1.8
Henning's taste tetrahedron.

- Euclidean
- Integral dimension



3. Quality dimension extraction experiment

I am approaching the task of **concept/knowledge extraction** using the BBC Wildlife ontology data set as a test bed (rdmpage at github), and an associated collection of Wikipedia articles on the species in the test set.

Instead of using machine-learnt vectors to encode word meanings, I am using **information-theoretic** approaches for finding axes, which might provide useful for specifying a **semantic dimension** connected to a property in the knowledge set.

The pipeline

Lots of processing

Input consists of CONLL-U dependency parsed text articles. The goal is to find frequent patterns in terms syntactic behavior.

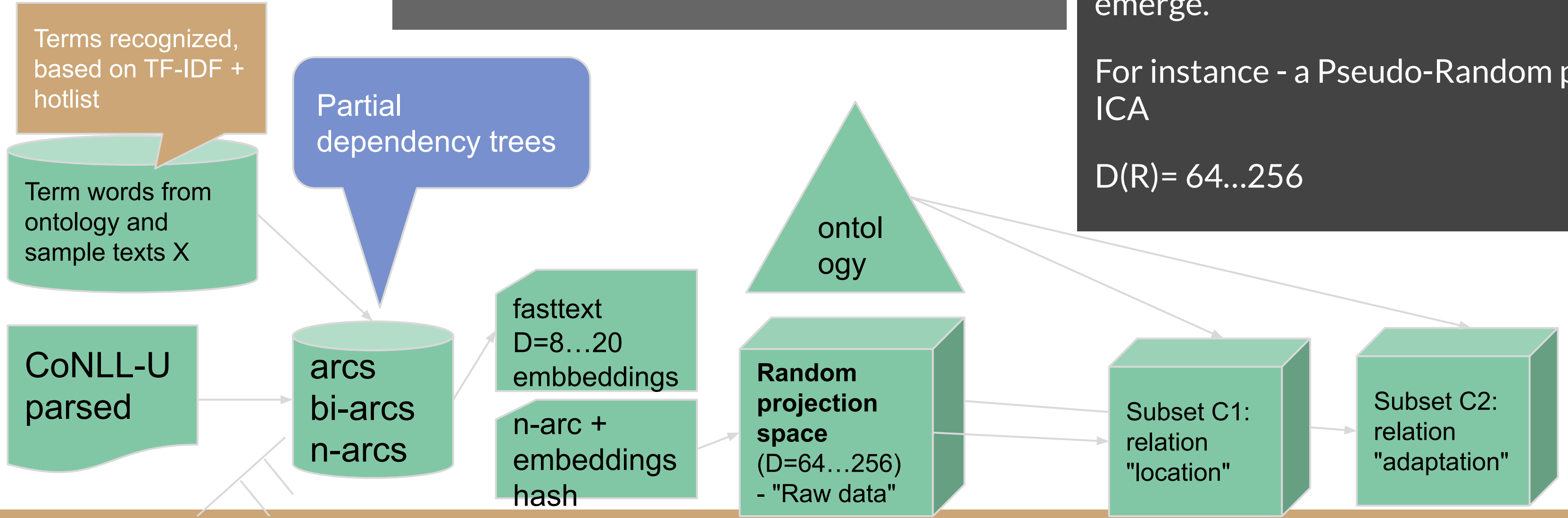
The analysis problem is divided by ontological predicates.

(wildlife ontology: adaptation, location, nutrition)

Results in vector data points in which clusters for sought relations are expected to emerge.

For instance - a Pseudo-Random projection + ICA

D(R)= 64…256



Terms recognized, based on TF-IDF + hotlist

Term words from ontology and sample texts X

CoNLL-U parsed

Partial dependency trees

arcs
bi-arcs
n-arcs

fasttext
D=8…20
embbeddings

n-arc + embeddings hash

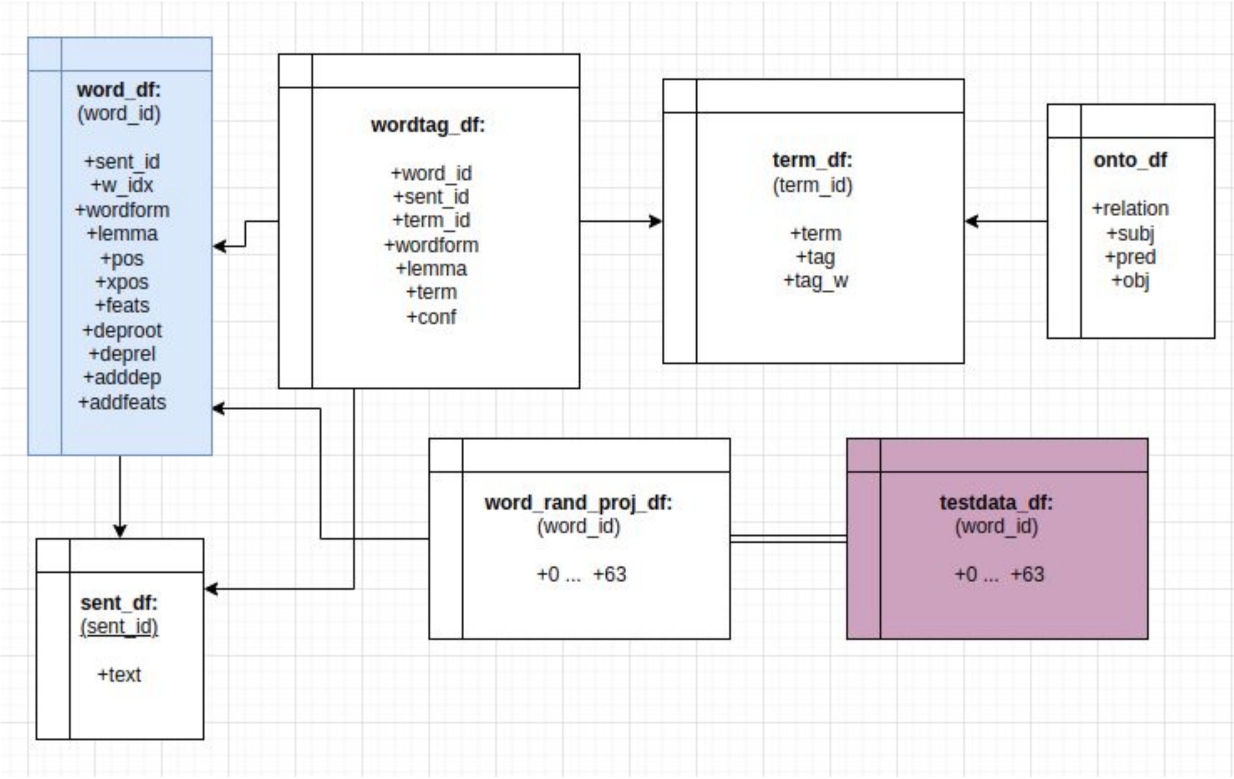
ontol
ogy

Random projection space
(D=64…256)
- "Raw data"

Subset C1: relation "location"

Subset C2: relation "adaptation"

Database model



word_df:
(word_id)

+sent_id
+w_idx
+wordform
+lemma
+pos
+xpos
+feats
+deproot
+deprel
+adddep
+addfeats

wordtag_df:

+word_id
+sent_id
+term_id
+wordform
+lemma
+term
+conf

term_df:
(term_id)

+term
+tag
+tag_w

onto_df

+relation
+subj
+pred
+obj

sent_df:
(sent_id)

+text

word_rand_proj_df:
(word_id)

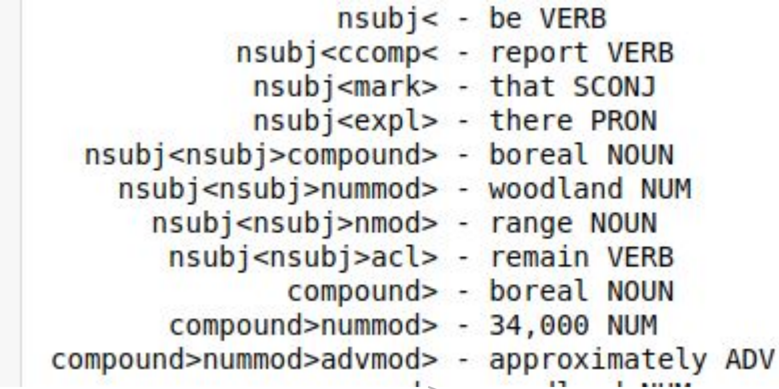
+0 ... +63

testdata_df:
(word_id)

+0 ... +63

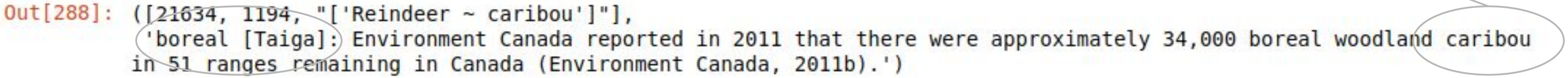
Hypothetical information projection model (1/3)

My hypothesis states that certain patterns in the **syntactical structure** in natural language reflect similar statements of information which are stored in a knowledge base, or a semantic web ontology.



```
                  nsubj<  - be VERB
           nsubj<ccomp<  - report VERB
            nsubj<mark>  - that SCONJ
            nsubj<expl>  - there PRON
  nsubj<nsubj>compound>  - boreal NOUN
    nsubj<nsubj>nummod>  - woodland NUM
      nsubj<nsubj>nmod>  - range NOUN
       nsubj<nsubj>acl>  - remain VERB
              compound>  - boreal NOUN
       compound>nummod>  - 34,000 NUM
compound>nummod>advmod>  - approximately ADV
```

RDF ontology statement (S–P→O):
Reindeer – livesIn→Taiga

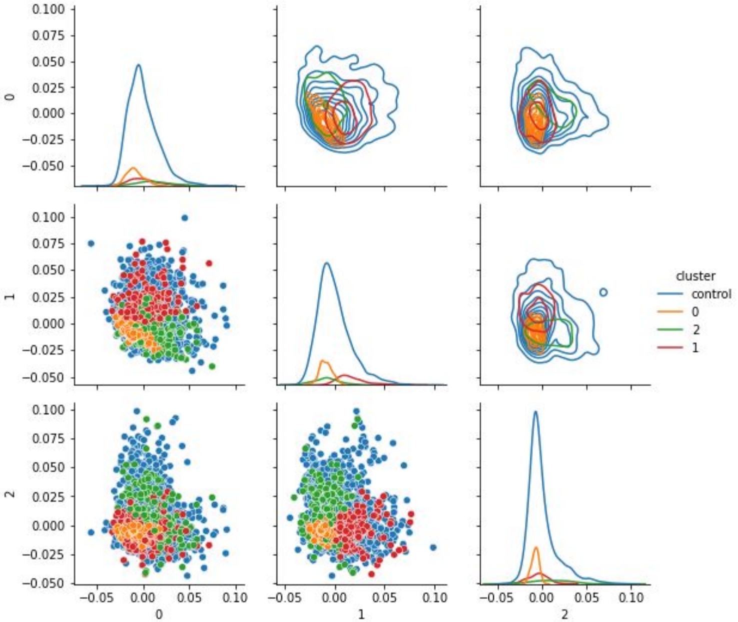


```
Out[288]: ([21634, 1194, "['Reindeer ~ caribou']"],
           'boreal [Taiga]: Environment Canada reported in 2011 that there were approximately 34,000 boreal woodland caribou
          in 51 ranges remaining in Canada (Environment Canada, 2011b).')
```

Hypothetical information projection model (2/3)

I assume these "statements of information" will be shown as **emergent structures** in a **dimension reduced projection**. This requires a feature analysis of the syntactic dependencies.

In the study, I will show how **measures of mutual information** can be applied to extract features from a dependency syntax analysis.



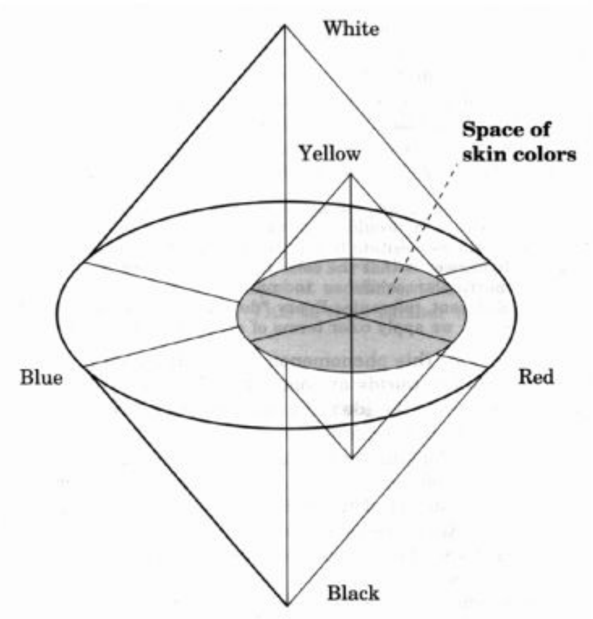
Hypothetical information projection model (3/3)

Expected conceptual quality dimensions are found if

- meaningful clusters are found
- boundaries reflect the statements and concepts in the ontology

If such can be extracted from a language resource data set, they might turn out helpful in NLP applications

- Pattern recognition in syntactic tree fragments
- Term harvesting patterns
- Further understanding in information retrieval (IR) and corpus search
- Especially interesting for categorizing newly found terms and concepts



Data set

"Parameters" for my experiment

Data sources:

- BBC Wildlife Ontology (BBCWO)
- Related Wikipedia articles (Wild animals) with matching titles
- UD dependency parsed sentences from both

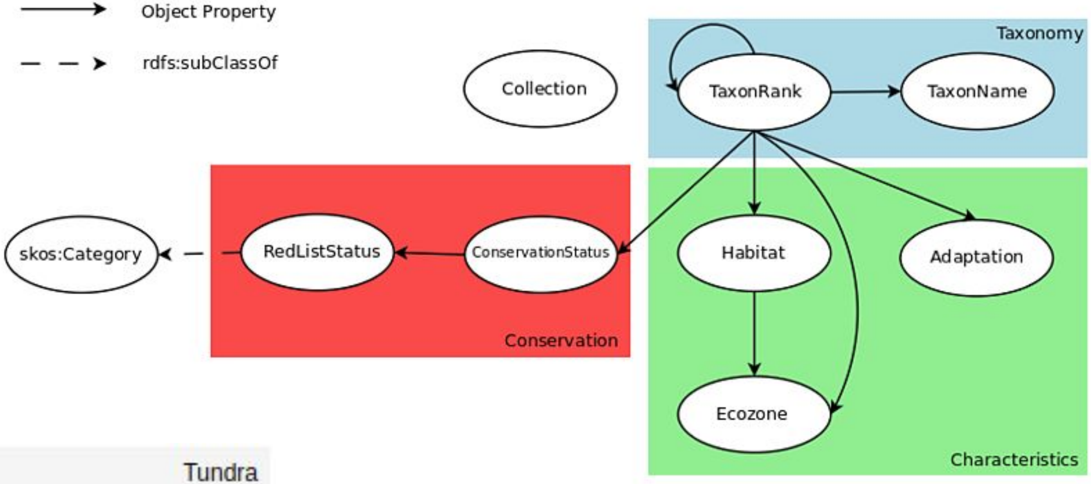
Goal function for capturing an "axis set" describing an ontology feature

Approaches to cluster formation: ICA for dimension reduction, GMM for cluster segmentation

The analyzed **Data points** are an unique term occurrences (token) in a sentence.

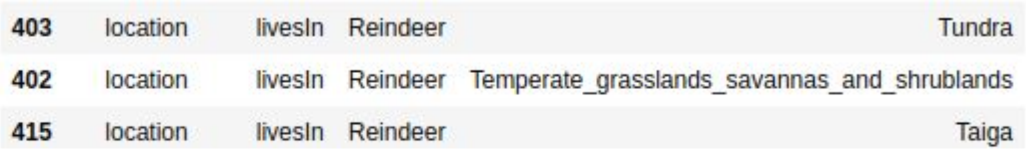
BBC Wildlife ontology as a practical knowledge base

BBC Wildlife ontology explained



conceptual relations (RDF predicate)
growsIn
livesIn
adaptation

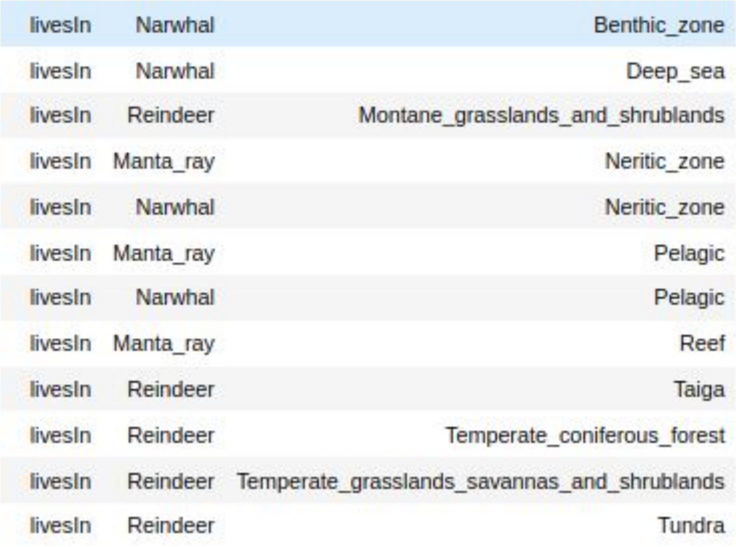
example



403	location	livesIn	Reindeer	Tundra
402	location	livesIn	Reindeer	Temperate_grasslands_savannas_and_shrublands
415	location	livesIn	Reindeer	Taiga

(BBC WO data available at github.com/rdmpage)

Some data for development



livesIn	Narwhal	Benthic_zone
livesIn	Narwhal	Deep_sea
livesIn	Reindeer	Montane_grasslands_and_shrublands
livesIn	Manta_ray	Neritic_zone
livesIn	Narwhal	Neritic_zone
livesIn	Manta_ray	Pelagic
livesIn	Narwhal	Pelagic
livesIn	Manta_ray	Reef
livesIn	Reindeer	Taiga
livesIn	Reindeer	Temperate_coniferous_forest
livesIn	Reindeer	Temperate_grasslands_savannas_and_shrublands
livesIn	Reindeer	Tundra

Summarized (before demo)

I am approaching a task of concept-knowledge "text pattern rule extraction" using the BBC Wildlife ontology data set as a test bed, and an associated collection of Wikipedia articles on the species in the test set.

Instead of using machine-learnt vectors to encode words (word2vec out-of-box), I am using dimension reduction based on mutual information, for finding an axis set of "**meaningful variance**". I am also using entropy based measures for finding out the typical structures in clusters found in the data points.

These meaningful clusters and variances could be analogous to the **conceptual spaces** and **quality dimensions**. These quality dimensions might turn out useful when describing a property or statement in the ontology, in contrast with related concepts.

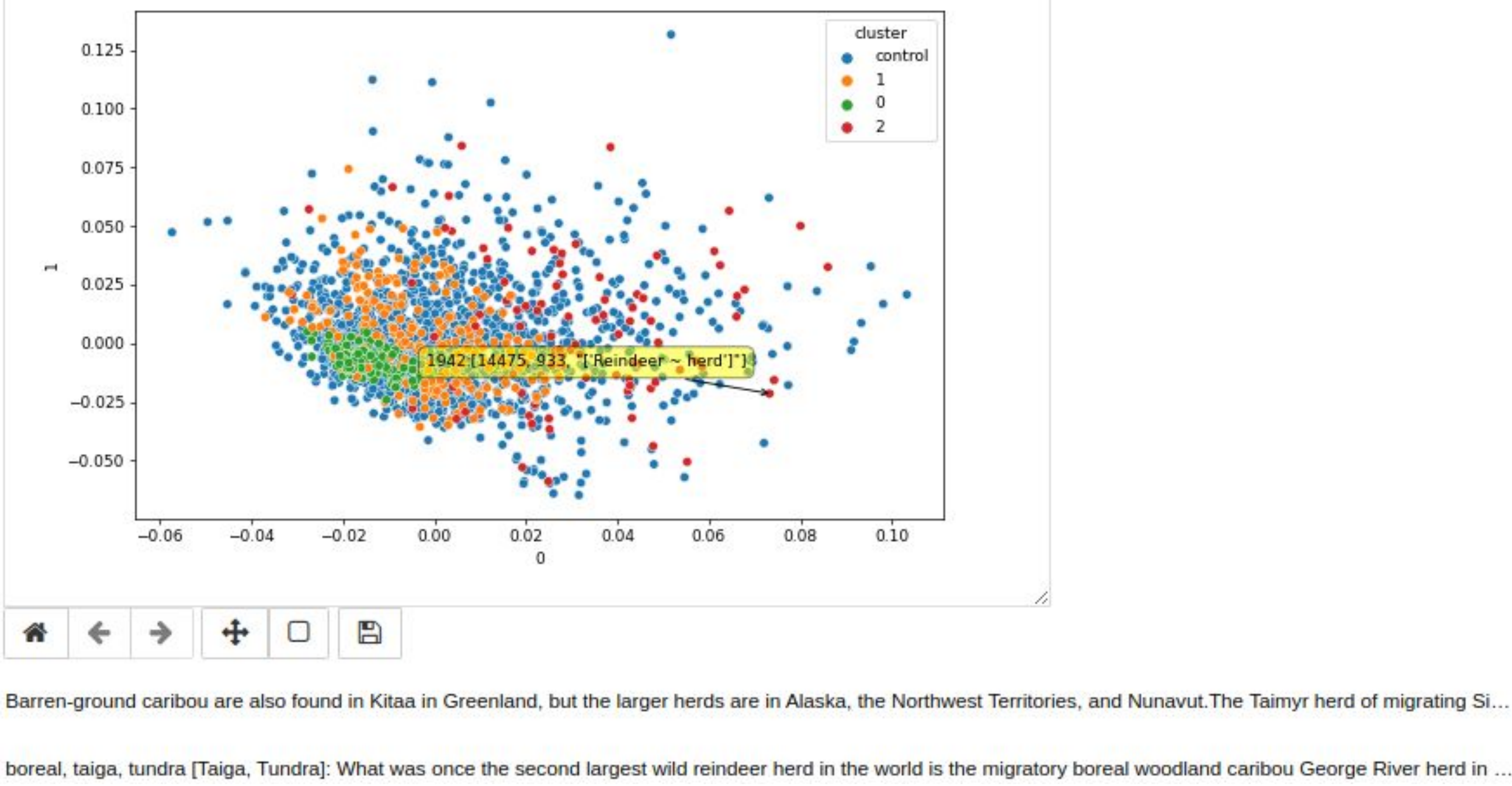
Aim for evaluation phase: I really would like to validate the found clusters, whether they "really" reflect the given "gold standard" in the ontology, or even something different.

Conclusions

- ICA seems to be a wonderful tool for dimension reduction - with a goal function "built in" to find divergent regions in output space - "camel back shape" optimization.
- ICA transformation can be trained on a smaller, subset of data (certain sample types, statements, etc), which can lead to nice-looking data distributions over all data points.
- GMM (Gaussian mixture model) is fast and sufficient for clustering the outcomes this far.
- Lots of ideas from the conceptual spaces model and the implicit similarity metrics.
- Emergent hypothesis in a nutshell - if "it" is frequent in the input, "it" should build up in clusters in the output as well.
- Q: How this can be reflected back in the given source ontology? How is this evaluated?
- A: Work in progress - lots of findings to be written into the manuscript at this moment.

demo

Pointwise exploration in clusters - in notebook demo



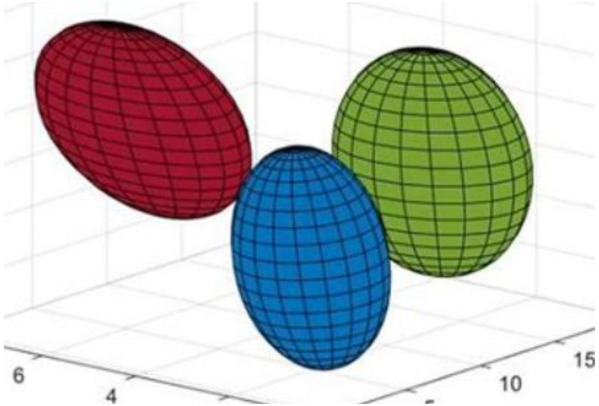
Barren-ground caribou are also found in Kitaa in Greenland, but the larger herds are in Alaska, the Northwest Territories, and Nunavut.The Taimyr herd of migrating Si...

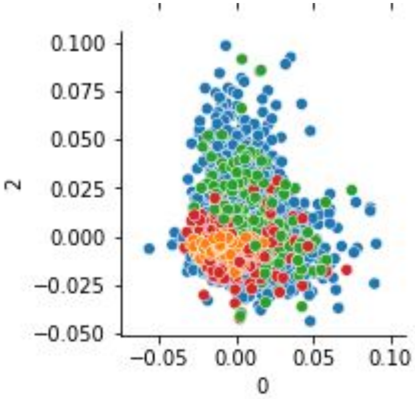
boreal, taiga, tundra [Taiga, Tundra]: What was once the second largest wild reindeer herd in the world is the migratory boreal woodland caribou George River herd in ...

Gaussian Mixture - refresher

Find dense "blob components" of data points - centers, deviations per axis and covariances between axis.

Example on right-hand side: Gaussian Mixture components (GMM) - calculated from sentence data with **scikit-learn**

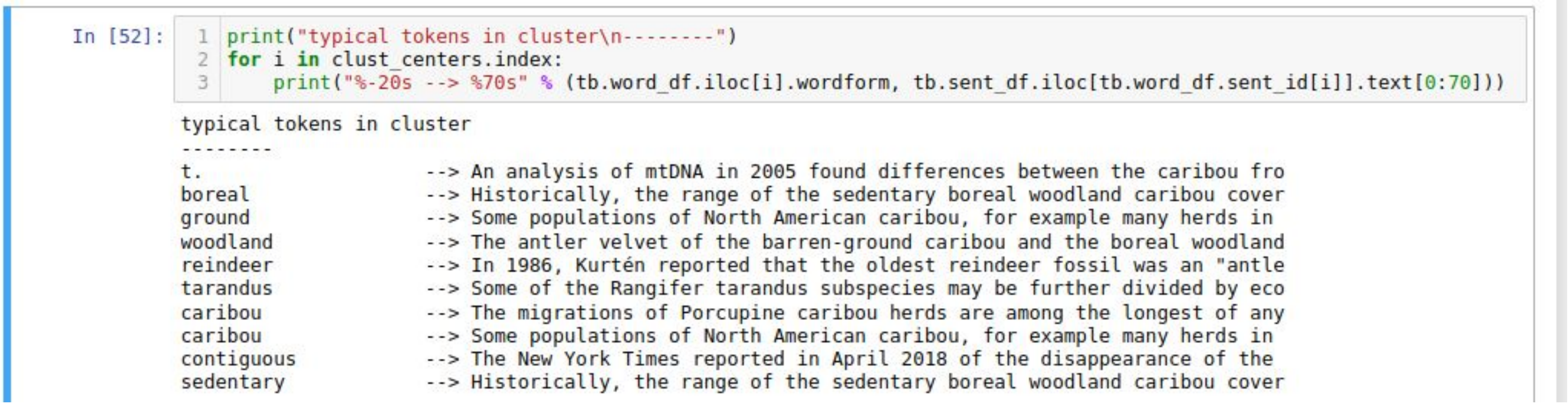




NB! Multiple alternatives for GMM: e.g. SOM, t-SNE, Agglomerative clustering, LVQ ...

Distributional analysis of tokens in clusters (demo)

K-L divergence between single-occurrence syntactic arc distribution and cluster-picked group → rates "typicalness" of the analyzed token's usage within the cluster. This helps finding out the "most typical" subtrees in the cluster - ie. what has been clustered.



In [52]:

```
print("typical tokens in cluster\n--------")
for i in clust_centers.index:
    print("%-20s --> %70s" % (tb.word_df.iloc[i].wordform, tb.sent_df.iloc[tb.word_df.sent_id[i]].text[0:70]))
```

```
typical tokens in cluster
--------
t.                   --> An analysis of mtDNA in 2005 found differences between the caribou fro
boreal               --> Historically, the range of the sedentary boreal woodland caribou cover
ground               --> Some populations of North American caribou, for example many herds in
woodland             --> The antler velvet of the barren-ground caribou and the boreal woodland
reindeer             --> In 1986, Kurtén reported that the oldest reindeer fossil was an "antle
tarandus             --> Some of the Rangifer tarandus subspecies may be further divided by eco
caribou              --> The migrations of Porcupine caribou herds are among the longest of any
caribou              --> Some populations of North American caribou, for example many herds in
contiguous           --> The New York Times reported in April 2018 of the disappearance of the
sedentary            --> Historically, the range of the sedentary boreal woodland caribou cover
```

Information metrics, relative entropy

Kullback-Leibler divergence

zero = full match

describes a distance, how divergently a sample distribution stands out from a "global" distribution

$$d = \sum_k p_k \log_2 \left(\frac{p_k}{q_k} \right)$$

Dependency syntax - refresher

```
# sent_id = 5

# text = Aardvarks are incredible diggers, so well equipped with powerful spoon shaped claws they can dig
a hole faster than several men with shovels.

1       Aardvarks      Aardvark    NOUN    NNS     Number=Plur     4       nsubj   _       _

2       are            be          AUX     VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        4       cop
_       _

3       incredible     incredible  ADJ     JJ      Degree=Pos      4       amod    _       _

4       diggers        digger      NOUN    NNS     Number=Plur     0       root    _       SpaceAfter=No

5       ,              ,           PUNCT   ,       _       4       punct   _       _
```