Taxonomy Generation Using Topic Modeling and Semantic Relationships

Elisabeth Daley

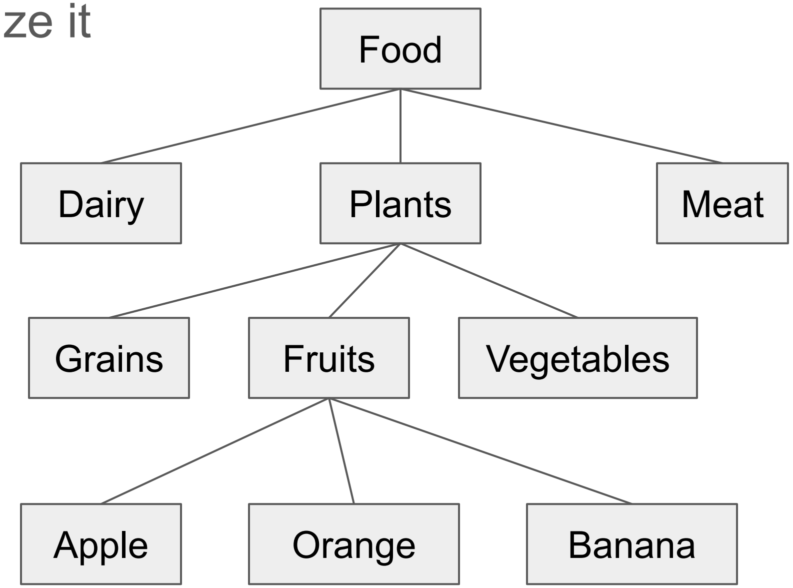
Introduction

Taxonomy: a classification structure of the entities or concepts related to a domain

A framework to understand the structure of information in a domain is useful:

- To give human’s a high level view of the concepts and how they relate
- As tools for classification, search, query expansion

As the volume of text documents proliferates, there is increasing value in being able to automatically process and organize it



Methods of Taxonomy Generation

Pattern based: Hearst patterns (“x is a y”) (Hearst, 1992)

Clustering Based:

- Distributional hypothesis: terms which share linguistic context are similar (Harris, 1968)
- Infer a relationship based on document co-occurrence or syntactic co-occurrence (Schmidt-Thieme, et al 1999)

Knowledge source: integrate information from existing databases or ontologies (Steyvers et al. 2011)

- Probase
- Wikipedia
- Cambridge Advanced Learner’s Dictionary
- WordNet

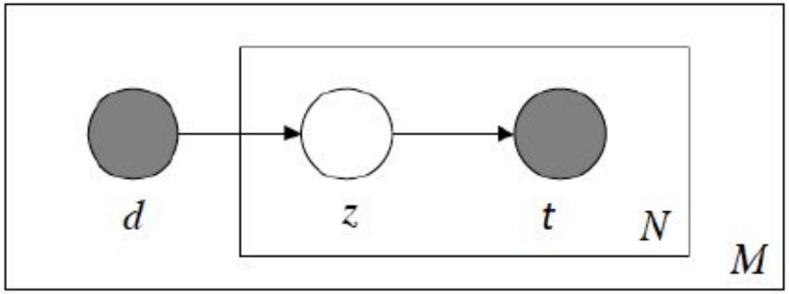
Topic Modeling

Probabilistic Latent Semantic Analysis (LSA) - hidden topics (z) probabilistically generate terms (t) in documents (d) (Hofmann, 1999)

$$P(d, t) = P(d)\, P(t \mid d)$$

$$P(t \mid d) = \sum_z P(t \mid z)\, P(z \mid d)$$

Assumes each document is a member of a single topic

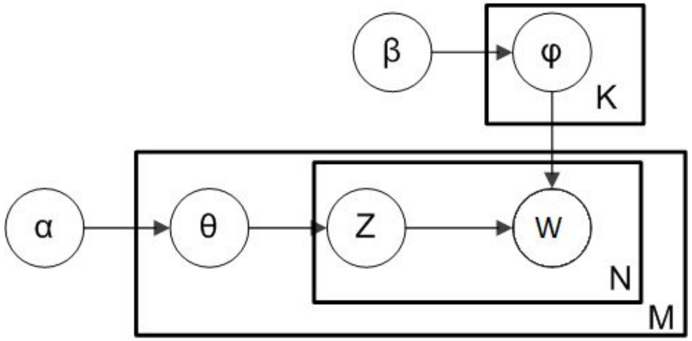


Topic Modeling - LDA

Latent Dirichlet Allocation (LDA) - documents are modeled as mixtures of *K* latent topics (Ng et al, 2003)

Dirichlet = conjugate prior distribution of categorical distribution and multinomial distributions, parametrized by:

- α: prior weights of a topic in a document
- β: prior weight of a term in a topic



Solving for

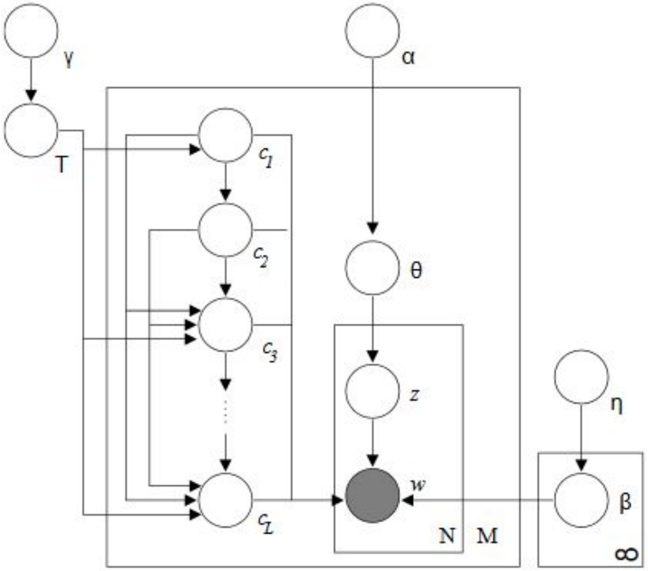
- φ: distribution of terms in each topic
- θ: distribution of topics in each document

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

Topic Modeling

Hierarchical LDA (hLDA) - each document is a mixture topics and their parent topics, up to the top of a tree (Blei et al, 2004)

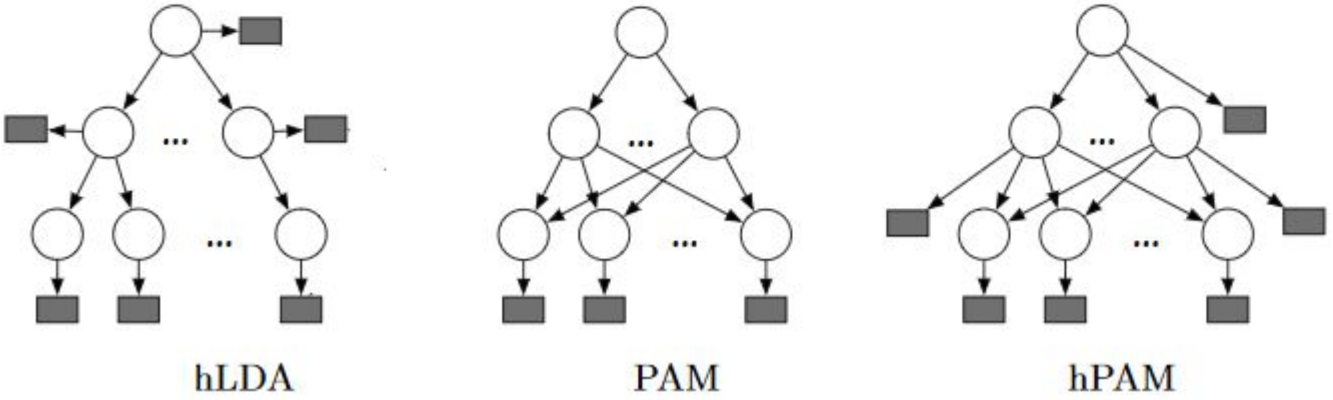
Still assumes that topics are independent from ‘cousins’



Topic Modeling - PAM and hPAM

Pachinko Allocation (PAM): Models correlations between topics as a directed acyclic graph (DAG), where the interior nodes represent topics, and the leaves terms. (Li & McCallum, 2006)

Hierarchical Pachinko Allocation (hPAM): allows for arbitrary levels of topics in the DAG (Mimno et al, 2007)



Taxonomy Generation with Topic Modeling

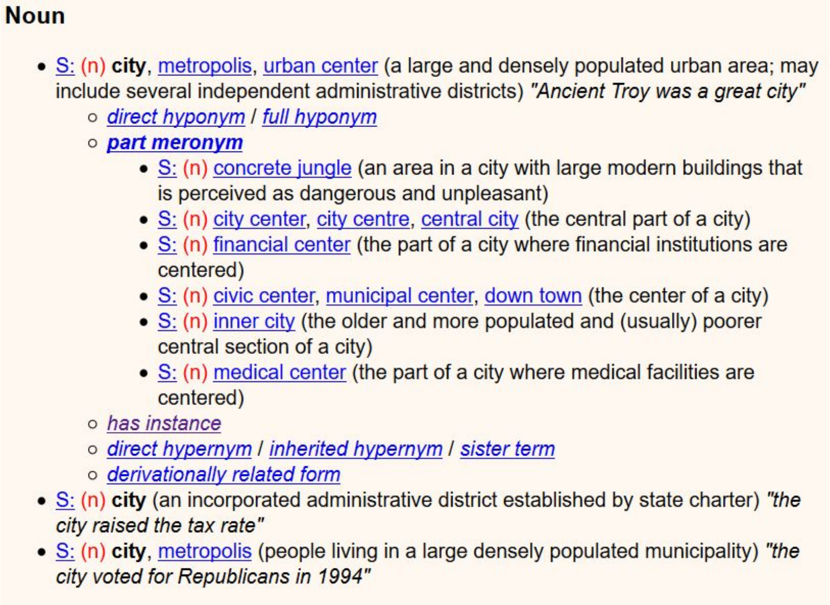
There have been several studies in the past decade using topic modeling for taxonomy generation:

- Steyvers et al, 2011: developed their own ‘concept-topic’ model, related to hPAM, where both latent topics and concepts drawn from a background dictionary were modeled in a DAG structure
- Bakalov et al, 2012: augmented an existing taxonomy with multi-labeled documents. They compared hLDA and hPAM, and found better performance with hPAM.
- Tang et al, 2017, developed a variant of LDA, and integrated information from the Probase ontology to develop a concept hierarchy

WordNet

WordNet - Princeton University’s collection of 117000 synsets and relations

- Synset = an unordered group of synonyms describing a concept



Noun

- S: (n) **city**, metropolis, urban center (a large and densely populated urban area; may include several independent administrative districts) *"Ancient Troy was a great city"*
  - *direct hyponym* / *full hyponym*
  - ***part meronym***
    - S: (n) concrete jungle (an area in a city with large modern buildings that is perceived as dangerous and unpleasant)
    - S: (n) city center, city centre, central city (the central part of a city)
    - S: (n) financial center (the part of a city where financial institutions are centered)
    - S: (n) civic center, municipal center, down town (the center of a city)
    - S: (n) inner city (the older and more populated and (usually) poorer central section of a city)
    - S: (n) medical center (the part of a city where medical facilities are centered)
  - *has instance*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*
- S: (n) **city** (an incorporated administrative district established by state charter) *"the city raised the tax rate"*
- S: (n) **city**, metropolis (people living in a large densely populated municipality) *"the city voted for Republicans in 1994"*

- Hypernym = “is a” relationship (ex. ‘city’ is a hypernym of ‘Boston’)
- Meronym = “part of” relationship (ex. ‘branch’ is a meronym of ‘tree’)

Hypothesis

By expanding documents with each entity’s hypernym chain, topic modeling will result in ‘topics’ which approximate a taxonomy of hierarchical concepts.

- Each document is reduced to nouns and named entities
- Each entity is found in WordNet, and inserted it’s next three parent hypernyms
- hPAM run on the resulting document corpus

Datasets

- **Lonely Planet** (PASCAL Ontology Learning Challenge, 2010): (1801) travel descriptions and a complete taxonomy definition.

- **Twenty Newsgroups** (Mitchell, 1999): 2000 Usenet articles from 1999 in 20 newsgroups.

- **NASA Lessons Learned** (NASA, 2018): 1574 problem descriptions and recommendations for avoiding them on NASA programs

Approach

- Python’s Natural Language Toolkit (NLTK) was used for the parsing and extraction of named entities, and the interface to WordNet
- McCallum’s MALLET was used for the topic modeling:
  Java implementation of LDA, hLDA, PAM, hPAM, and other topic modeling algorithms
- Initial tests determined the number of topics at which the log-likelihood of each dataset stopped improving
- The results of hPAM on each original corpus were compared with the original corpus reduced to nouns, and then to the nouns + hypernyms

	a		**b**		**c**	
	Super Topics	Sub Topics	Super Topics	Sub Topics	Super Topics	Sub Topics
Lonely Planet	10	5	20	5	50	10
NASA	20	10	40	10	50	10
20 Newsgroups	10	5	50	5	50	10

Evaluation

Appropriate metrics for evaluation depends on the intended use of the taxonomy

Clustering metrics: distortion, cohesion, separation, silhouette coefficient,

- For number of clusters: Dunn index, Pseudo-F Index
- For comparing sets of clusters: Rand Index

Evaluating topic models:

- Log-likelihood: how well does the trained model predict the distributions in a held-out test set?

$$L(\mathrm{d})=\log p(d|\varphi,\alpha)=\sum_{d}\log p(\mathrm{d}_{test}|\varphi,\alpha)$$

$$\mathrm{Perplexity}(\mathrm{d}_{\mathrm{test}})=\exp\{-L(\mathrm{d}) / \#\mathrm{terms}\}$$

Evaluation

Taxonomies are often evaluated either by hand, or by comparison to a ‘gold-standard’ taxonomy

- Neither of these approaches is satisfactory:
    - There may be multiple subjectively valid ways of organizing information
    - It may change over time, as more information is added
    - Varying levels of detail may be desirable

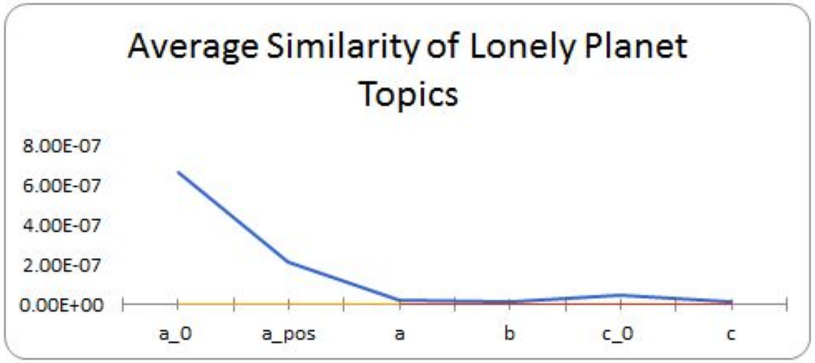
- Semantic distance: how many synsets separate two words in WordNet?

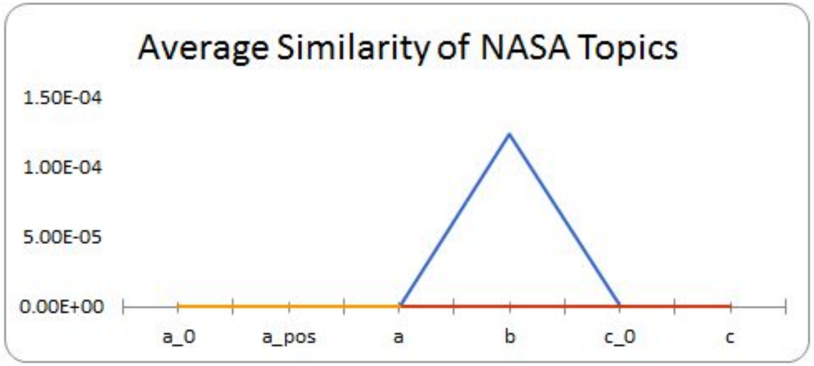
  Product of the pairwise similarity of the top *n* words used in each topic

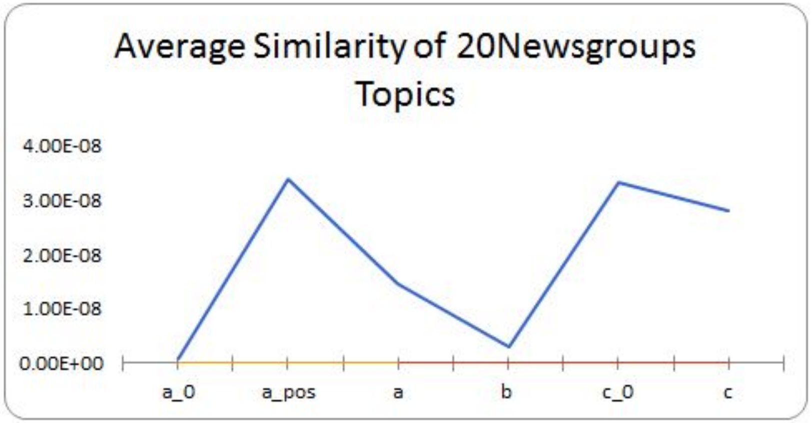
Results

Similarity (as measured by the distance between the top 5 words in each topic) did not improve in the hypernym expanded corpora

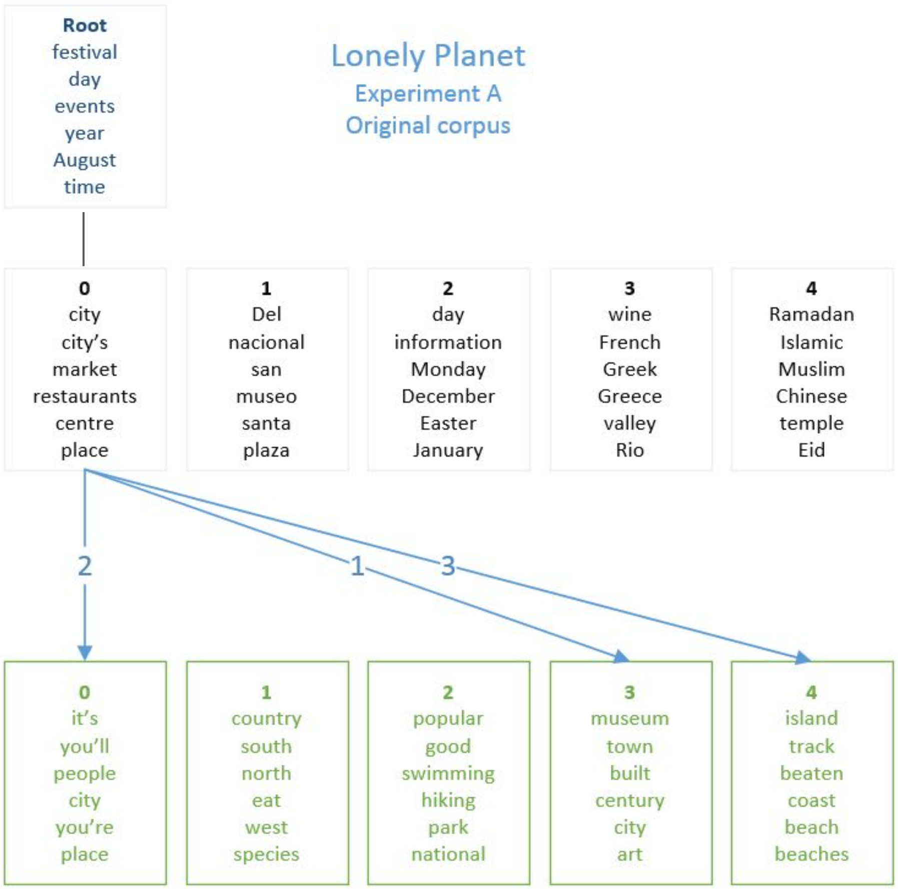
Subjectively, the top words for each topic seem to relate more to individual concepts than in the original, unmodified corpora, however the hierarchical relationships are not evident







Lonely Planet - original vs hypernym expanded



Lonely Planet
Experiment A
Original corpus

Root
festival
day
events
year
August
time

0
city
city's
market
restaurants
centre
place

1
Del
nacional
san
museo
santa
plaza

2
day
information
Monday
December
Easter
January

3
wine
French
Greek
Greece
valley
Rio

4
Ramadan
Islamic
Muslim
Chinese
temple
Eid

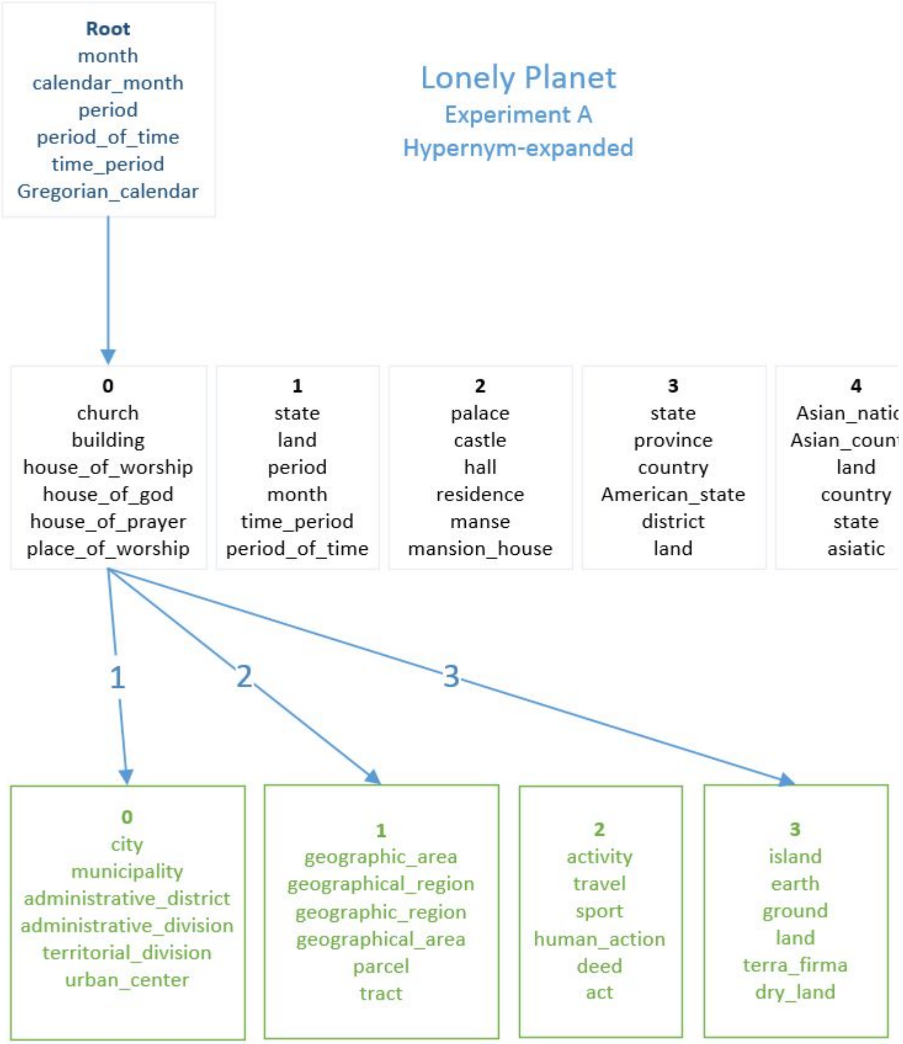
0
it's
you'll
people
city
you're
place

1
country
south
north
eat
west
species

2
popular
good
swimming
hiking
park
national

3
museum
town
built
century
city
art

4
island
track
beaten
coast
beach
beaches



Lonely Planet
Experiment A
Hypernym-expanded

Root
month
calendar_month
period
period_of_time
time_period
Gregorian_calendar

0
church
building
house_of_worship
house_of_god
house_of_prayer
place_of_worship

1
state
land
period
month
time_period
period_of_time

2
palace
castle
hall
residence
manse
mansion_house

3
state
province
country
American_state
district
land

4
Asian_natic
Asian_coun
land
country
state
asiatic

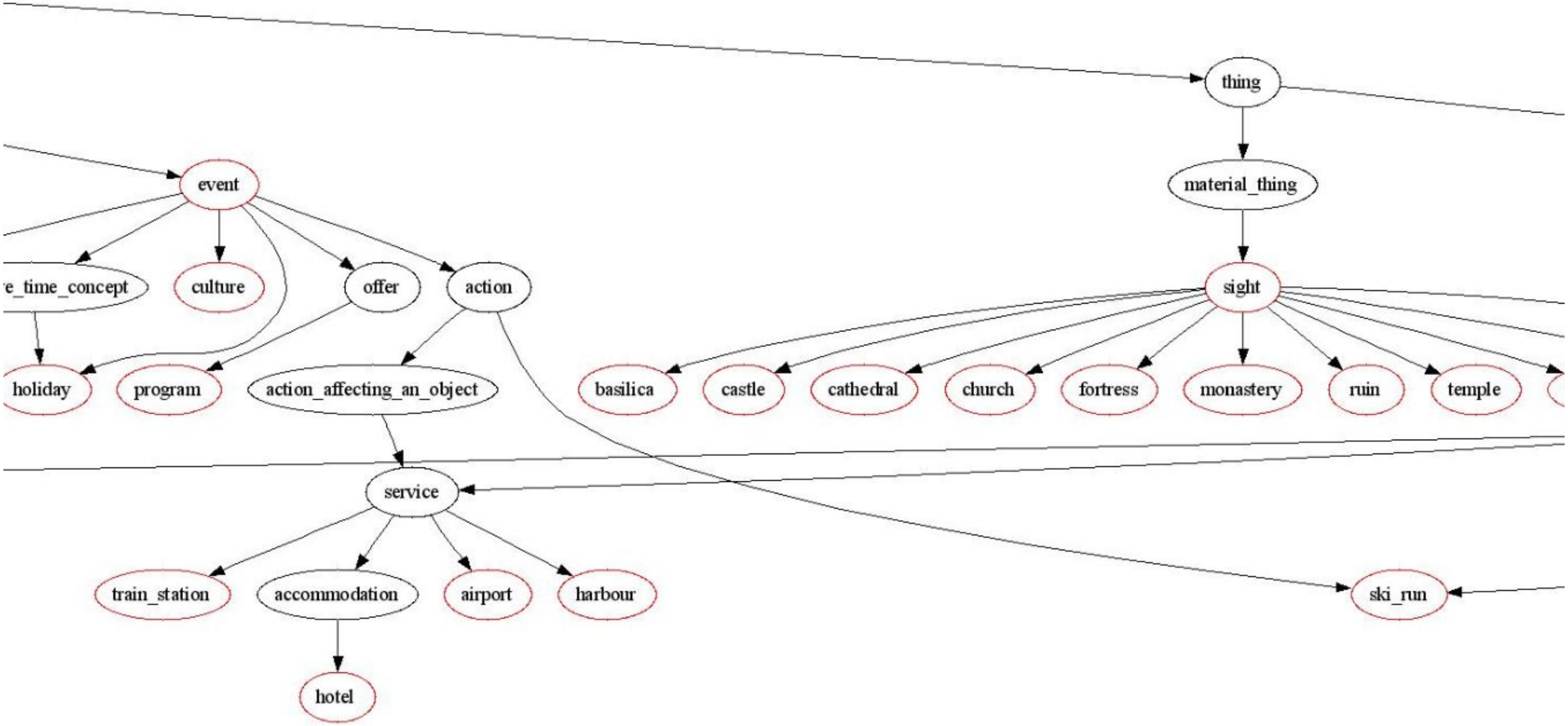
0
city
municipality
administrative_district
administrative_division
territorial_division
urban_center

1
geographic_area
geographical_region
geographic_region
geographical_area
parcel
tract

2
activity
travel
sport
human_action
deed
act

3
island
earth
ground
land
terra_firma
dry_land

Lonely Planet “gold standard”



Future Work

Potential methodology improvements:

- hLDA (with many levels), rather than hPAM
- Model with more topics
- Include more levels of hypernyms
- Use other semantic relationships (metonyms)
- Include all original terms from the documents

More nuanced methods of evaluation:

- Coherence
- Summarize concept with a single term

Thank you!

Z. Harris. Mathematical Structures of Language. Wiley, 1968.

M. Hearst. Automatic acquisition of hyponyms from large text corpora. Proceedings of the14th International Conference on Computational Linguistics, pages 539-545, 1992

L. Schmidt-Thieme, P. Cimiano, A. Pivk and S. Staab. Learning taxonomic relations from heterogeneous sources of evidence. Springer Verlag, 1999.

M. Steyvers, P. Smyth, and C. Chemuduganta. Combining background knowledge and learned topics. Topics in cognitive science, 3(1):18-47, 2011.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the Twenty-Second Annual International SIGIR Conference, pages 50–57.

A. Ng, D. Blei and M. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

D. Blei, T. M. Jordan, T. Griffiths and J. J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. NIPS, 2004.

W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. ICML, 2006.

D. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In Proceedings of the 24th international conference on Machine learning, pages 633-640. ACM, 2007.

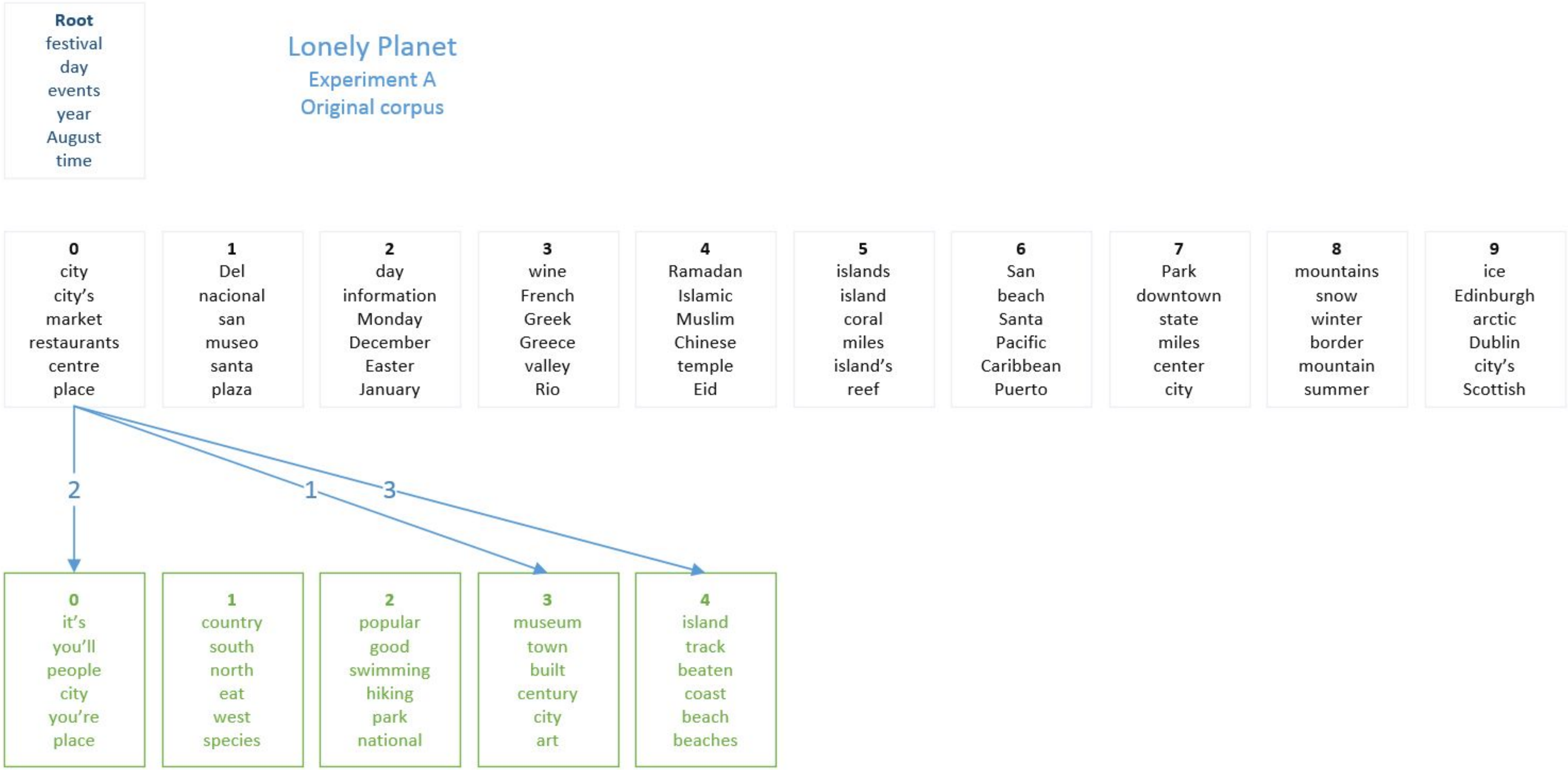
A. Bakalov, A. McCallum, H. Wallach, and D. Mimno. Topic models for taxonomies. In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, pages 237-240. ACM, 2012.

Thank you!

Y. Tang, X. Mao, H. Huang, X. Shi, and G. Wen. Conceptualization topic modeling. Multimedia Tools and Applications, 77(3):3455-3471, 2018.

D. Mimno, Hanna M Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In Proceedings of the conference on empirical methods in natural language processing, pages 262-272. Association for Computational Linguistics, 2011.

Lonely Planet - original corpus



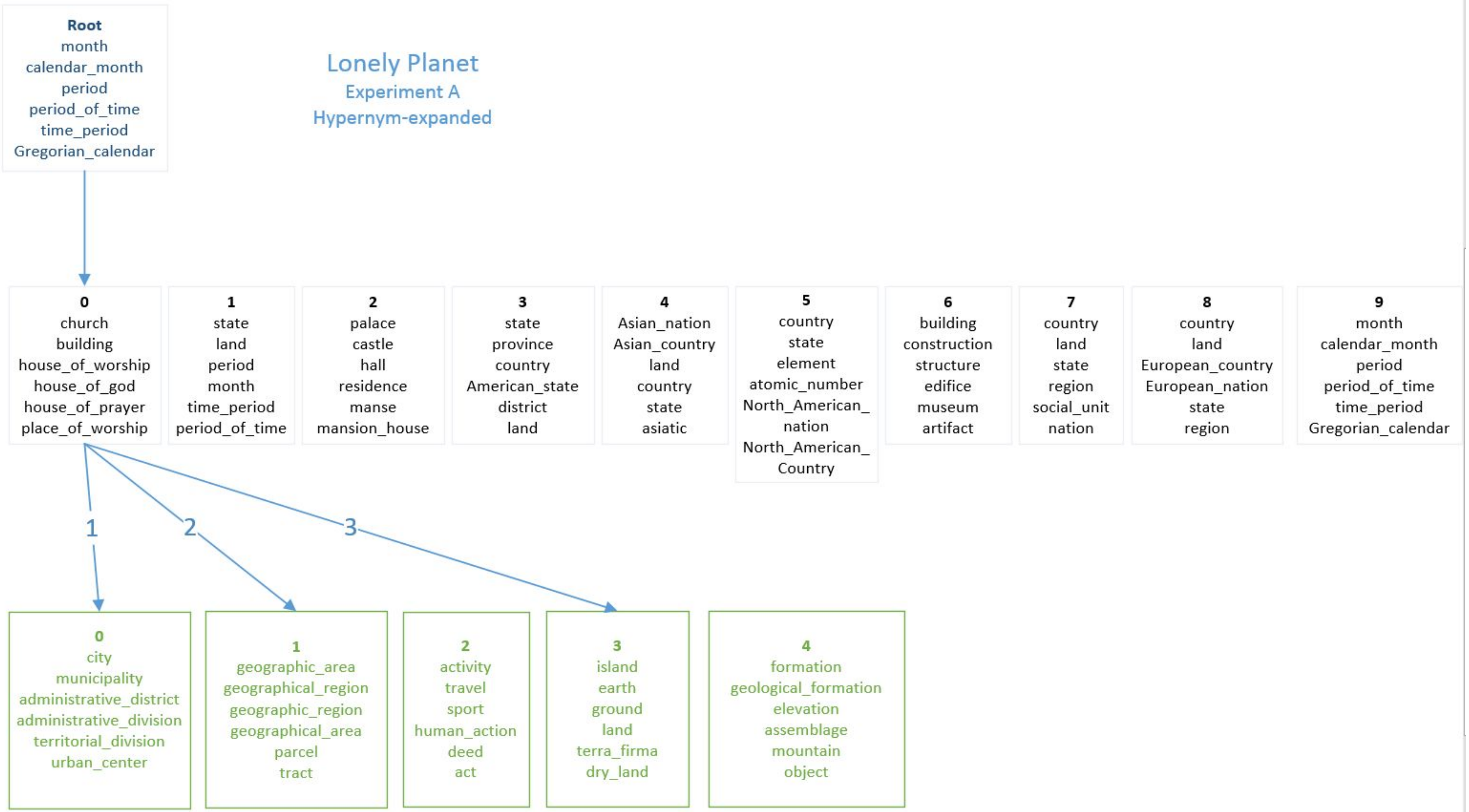
Root
festival
day
events
year
August
time

Lonely Planet
Experiment A
Original corpus

0	1	2	3	4	5	6	7	8	9
city	Del	day	wine	Ramadan	islands	San	Park	mountains	ice
city's	nacional	information	French	Islamic	island	beach	downtown	snow	Edinburgh
market	san	Monday	Greek	Muslim	coral	Santa	state	winter	arctic
restaurants	museo	December	Greece	Chinese	miles	Pacific	miles	border	Dublin
centre	santa	Easter	valley	temple	island's	Caribbean	center	mountain	city's
place	plaza	January	Rio	Eid	reef	Puerto	city	summer	Scottish

0	1	2	3	4
it's	country	popular	museum	island
you'll	south	good	town	track
people	north	swimming	built	beaten
city	eat	hiking	century	coast
you're	west	park	city	beach
place	species	national	art	beaches

Lonely Planet - hypernym expanded



Root
month
calendar_month
period
period_of_time
time_period
Gregorian_calendar

Lonely Planet
Experiment A
Hypernym-expanded

0
church
building
house_of_worship
house_of_god
house_of_prayer
place_of_worship

1
state
land
period
month
time_period
period_of_time

2
palace
castle
hall
residence
manse
mansion_house

3
state
province
country
American_state
district
land

4
Asian_nation
Asian_country
land
country
state
asiatic

5
country
state
element
atomic_number
North_American_
nation
North_American_
Country

6
building
construction
structure
edifice
museum
artifact

7
country
land
state
region
social_unit
nation

8
country
land
European_country
European_nation
state
region

9
month
calendar_month
period
period_of_time
time_period
Gregorian_calendar

0
city
municipality
administrative_district
administrative_division
territorial_division
urban_center

1
geographic_area
geographical_region
geographic_region
geographical_area
parcel
tract

2
activity
travel
sport
human_action
deed
act

3
island
earth
ground
land
terra_firma
dry_land

4
formation
geological_formation
elevation
assemblage
mountain
object

Data

Dataset	Method	Average Similarity	Log-Likelihood
LP	a_0	6.59E-07	-9.20959
LP	a_pos	2.13E-07	-8.6913
LP	a	2.05E-08	-8.45075
LP	b	1.27E-08	-8.44522
LP	c_0	4.60E-08	-9.04806
LP	c	1.37E-08	-8.32801
NASA	a_0	1.22E-07	-8.51425
NASA	a_pos	1.01E-08	-7.76723
NASA	a	9.55E-09	-8.12076
NASA	b	1.24E-04	-8.01261
NASA	c_0	1.18E-07	-8.25621
NASA	c	1.49E-08	-7.96715
News	a_0	1.06E-09	-8.89428
News	a_pos	3.39E-08	-9.55352
News	a	1.47E-08	-8.87058
News	b	3.06E-09	-8.83002
News	c_0	3.31E-08	-8.55286
News	c	2.80E-08	-8.64083