

Topic Analysis of RDA Activities

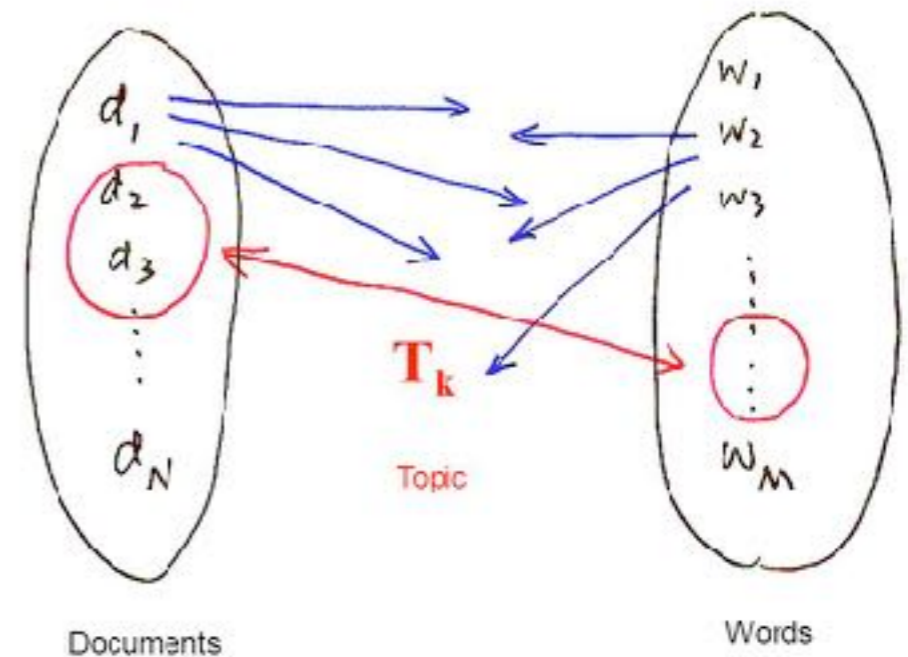
A preliminary report

RDA TAB and IGWG Chairs Meeting, Gaithersburg, MD January 2017

Motivation

- ❖ Supplement the “Six Words” exercise
- ❖ Learn what RDA groups are doing by analyzing the documents they produce
- ❖ Create a corpus by crawling the RDA web pages
- ❖ Apply well-established techniques of Probabilistic Latent Semantic Analysis

Latent Semantic Analysis (LSA)



Methodology

- ❖ Crawl rd-alliance.org using Apache Nutch, put docs in Solr (N=6,102)
- ❖ Select all documents containing “case-statement” (N=196)
- ❖ Remove stop words, do lemmatization
- ❖ Use *Gensim* toolkit to do Latent Dirichlet Allocation (LDA)



Stopwords



Basic Concepts

- ❖ Topic Modeling identifies *topics* and their *distributions* across the *documents* in a *corpus*.
- ❖ Generative probabilistic models use statistical methods to discover hidden (i.e., "latent") themes (topics) in documents.
- ❖ "Generative" means we assume the source documents were generated from a mixture of topics, with each topic being a distribution over words belonging to that topic.
- ❖ A word can belong to multiple topics.
- ❖ Neither the order of the documents in the corpus, nor the order of the words in each document are taken as significant.

LDA results w/ 15 topics, 8 words

Topic								
0	plan	platform	dmp	fabric	requirement	course	management	dmps
1	provenance	disciplinary	discipline	professor	reproducibility	illinois	collaboration	need
2	collection	privacy	pid	metadata	quality	digital	set	related
3	link	literature	infrastructure	publishing	article	publisher	scholix	hub
4	brokering	mediation	component	registry	resource	standard	provide	need
5	national	workflow	publishing	tool	publication	preservation	policy	activity
6	indigenous	sov	international	network	sovereignty	security	citation	trust
7	cost	database	field	recovery	farmer	centre	network	management
8	storage	qos	provider	datalc	vocabulary	multiple	document	adoption
9	legal	interoperability	law	domain	codata	international	national	ccm
10	type	creating	past	registry	standard	national	record	joining
11	agriculture	interoperability	agricultural	semantic	marine	semantics	infrastructure	vre
12	brokering	governance	model	infrastructure	middleware	interoperability	approach	support
13	metadata	standard	sample	practice	fishery	vocabulary	digital	humanity
14	rice	interoperability	ontology	policy	wheat	standard	organization	common

Topic Words



Sample Topics

- ❖ Case Statement from “Agrisemantics” WG has high probability ($P=0.999$) that it contains Topic 11:
 - ❖ *agriculture, interoperability, semantic, marine, semantics, infrastructure, vre*
 - ❖ Also “Virtual Research Environments” IG ($P=0.997$) and “Marine Data Management” WG ($P=0.672$)
- ❖ Statement for “On-Farm Data Sharing” WG contains Topic 7:
 - ❖ *cost, database, field, recovery, farmer, centre, network, management*
- ❖ Earlier LDA analysis had “Array Database” WG in Topic with:
 - ❖ *array, database, big, domain, document, report, plan*
 - ❖ but in this analysis has $P=0.995$ that contains Topic 7

LDA results can be viewed at:

<https://goo.gl/V6Yraj>

What's Next?

- ❖ Refine list of stop words
- ❖ Remove duplicate documents (last version only?)
- ❖ Further tune algorithms
- ❖ Aggregate topic probabilities by group
- ❖ Include other documents (e.g. outputs)
- ❖ Assign dates to document, show Topic trends
- ❖ Other?

Thanks for your attention!
