Experiments with GORC Dataset using Topic Models

Swapnil Hingmire

GORC: A large contextual citation graph of academic papers[1]

- GORC : the Semantic Scholar Graph of References in Context
- A large contextual citation graph of 81.1M academic publication
  - Each citation edge in the graph includes the context in which a paper cites another paper
- Contains parsed full text for 8.1M open access papers
- Sources include PubMed, PubMed Central, arXiv and ACL Anthology
- The dataset also includes *word2vec* based vector representation of papers based on their citation contexts.

[1] Lo et al., "GORC: A large contextual citation graph of academic papers", 2019

- Experiments with ACL Anthology abstracts (# 27691) in GORC dataset using Latent Dirichlet Allocation (LDA)[2]

[2] Blei et al., "Latent Dirichlet Allocation", 2003

Latent Dirichlet Allocation (LDA)

- Topic: a probability distribution over words
- Document: a probability distribution over topics

Latent Dirichlet Allocation (LDA)

- Topic: a probability distribution over words
- Document: a probability distribution over topics

Gist...

As the topics are interpretable, and we know how a document exhibits each topic, we can infer the **gist** of the document (Griffiths et al., 2007)[a].

[a] Griffiths et al., "Topics in Semantic Representation", 2007

LDA : Macro-analysis of a Document Corpus

LDA : Macro-analysis of a Document Corpus

A subset of topics on NIPS corpus:

Topic-1	**Topic-2**	**Topic-3**	**Topic-4**
speech	grammar	pixel	gene
frame	sentence	colour	dna
band	language	segment	species
speaker	acquisition	scene	cell
audio	syntax	texture	tissue
utterance	corpus	light	genome
voice	fsm	surface	biological
acoustic	syntactic	affinity	genomic
channel	child	intensity	cancer
pitch	symbol	histogram	molecular

LDA : Macro-analysis of a Document Corpus

A subset of topics on NIPS corpus:

Topic-1	Topic-2	Topic-3	Topic-4
speech	grammar	pixel	gene
frame	sentence	colour	dna
band	language	segment	species
speaker	acquisition	scene	cell
audio	syntax	texture	tissue
utterance	corpus	light	genome
voice	fsm	surface	biological
acoustic	syntactic	affinity	genomic
channel	child	intensity	cancer
pitch	symbol	histogram	molecular

LDA : Macro-analysis of a Document Corpus

A subset of topics on NIPS corpus:

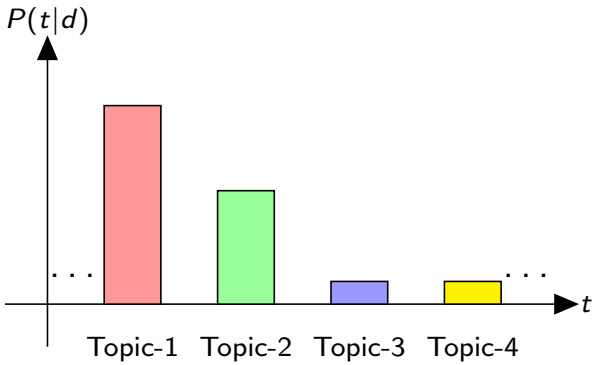
Topic-1	**Topic-2**	**Topic-3**	**Topic-4**
speech	grammar	pixel	gene
frame	sentence	colour	dna
band	language	segment	species
speaker	acquisition	scene	cell
audio	syntax	texture	tissue
utterance	corpus	light	genome
voice	fsm	surface	biological
acoustic	syntactic	affinity	genomic
channel	child	intensity	cancer
pitch	symbol	histogram	molecular

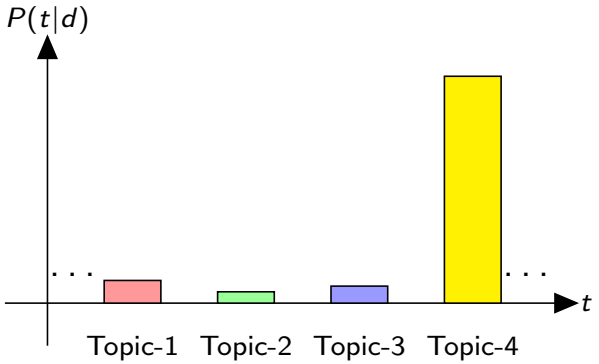
speech processing	**natural language**	**image processing**	**genetics**

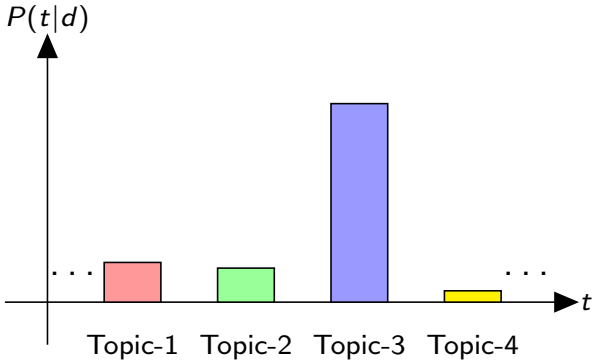
LDA : Macro-analysis of a Document Corpus

A subset of topics on NIPS corpus:

Topic-1	Topic-2	Topic-3	Topic-4
speech	grammar	pixel	gene
frame	sentence	colour	dna
band	language	segment	species
speaker	acquisition	scene	cell
audio	syntax	texture	tissue
utterance	corpus	light	genome
voice	fsm	surface	biological
acoustic	syntactic	affinity	genomic
channel	child	intensity	cancer
pitch	symbol	histogram	molecular
speech processing	**natural language**	**image processing**	**genetics**





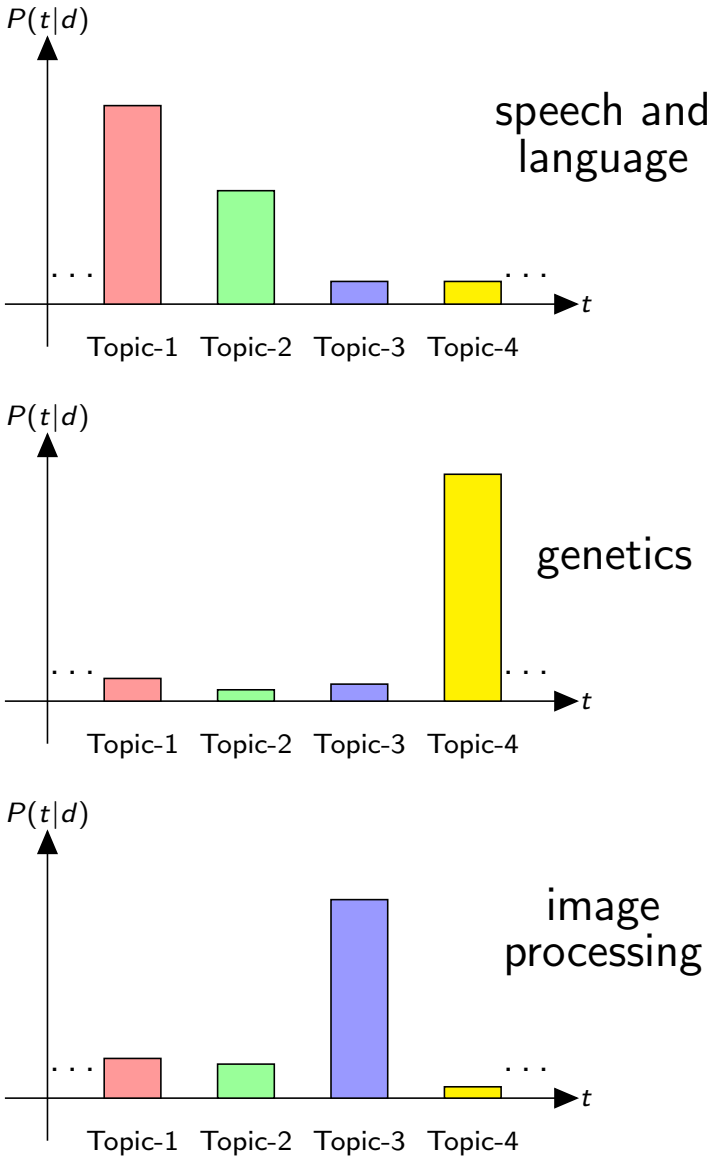


LDA : Macro-analysis of a Document Corpus

A subset of topics on NIPS corpus:

Topic-1	Topic-2	Topic-3	Topic-4
speech	grammar	pixel	gene
frame	sentence	colour	dna
band	language	segment	species
speaker	acquisition	scene	cell
audio	syntax	texture	tissue
utterance	corpus	light	genome
voice	fsm	surface	biological
acoustic	syntactic	affinity	genomic
channel	child	intensity	cancer
pitch	symbol	histogram	molecular

speech processing	natural language	image processing	genetics



Annotation of LDA Topics using LectureBank Taxonomy-I

Most probable words of the topic	Taxonomy ID
semantic role labeling syntactic predicate **frame srl framenet** argument **predicate-argument propbank** thematic shallow assignment labels **frame-semantic** proposition verbnet nombank nominal	Semantic Role Labeling (367)
tagging pos tags tagger shallow unknown tag **chunking** tokens part token chunk memory-based **tagset** sequences transformation-based predicted chunker **pos-tagging part-of-speech brill**	Part of Speech (142)
grammars context-free tree parsing lexicalized derivation formalism **probabilistic** transduction adjoining bottom-up strings linear trees extension **pcfg** stochastic **cfg** top-down	Context-Free Grammars (315)
sense disambiguation wordnet senses wsd lexical ambiguous sample senseval disambiguate **polysemy synsets** inventory all-words knowledge-based coarse-grained semeval glosses ambiguity sense-tagged	Word Sense Disambiguation (39)
entity named entities recognition names name **ner** linking person mentions proper personal location types nested base nes **named-entity** wikipedia recognize newswire gazetteers link geographic f-measure gazetteer muc persons lists organizations ace	Information Extraction and NER (232)

Annotation of LDA Topics using LectureBank Taxonomy-II

Most probable words of the topic	Topic Label
extraction events **temporal ordering tense** causal complex **tempeval** narrative **times date** triggers key types aspectual **timeline intervals duration temporally timeml**	Temporal reasoning
multimodal visual image spatial descriptions instructions **video** grounded **objects** grounding human **vision** actions world **modalities multi-modal** robot **captions scene** ground situated environment **multimedia**	Multimodal
resolution coreference anaphora pronoun resolving **zero ellipsis mentions** coordination **chains anaphoric antecedent mention bridging** salience antecedents **muc** definite reference refer **pronominal** ontonotes centering null	Co-reference resolution
deep noisy noise **bias** uncertainty **gender** real-world auxiliary **biases** scenarios sampling suffer inherent tackle poor bert low real incorporating augment popular leads pre-trained **biased** suffers fine-tuning	Bias in NLP
evidence arguments stage **claim** preference selectional **argumentation clues** determine **argumentative** restrictions essential **mining** determining **propositions** inferring separate **supporting** specificity **argue** positions	Argumentation mining

Bias in NLP

- N19-1061: Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them
  - Word embeddings are widely used in NLP for a vast range of tasks. It was shown that word embeddings derived from text corpora reflect gender biases in society...
- D18-1302: Reducing Gender Bias in Abusive Language Detection
  - Abusive language detection models tend to have a problem of being biased toward identity words of a certain group of people because of imbalanced training datasets. For example, "You are a good woman" was considered "sexist" when trained on an existing dataset....
- P19-1160: Gender-preserving Debiasing for Pre-trained Word Embeddings
  - Word embeddings learnt from massive text collections have demonstrated significant levels of discriminative biases such as gender, racial or ethnic biases..
- N15-1084: Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test
  - The Bechdel test is a sequence of three questions designed to assess the presence of women in movies. Many believe that because women are seldom represented in film as strong leaders and thinkers...

Approach for Classification of AAN Documents

- Learn a LDA model on NLP related papers of GORC corpus
- Manually tag each LDA topic to one or more entries in the taxonomy

Approach for Classification of AAN Documents

- Learn a LDA model on NLP related papers of GORC corpus
- Manually tag each LDA topic to one or more entries in the taxonomy
- For each document in AAN corpus, infer distribution over the topics using the LDA model
- Classify an AAN document based on tags corresponding its most probable topics

Key Aspects

- Lesser annotation efforts:
  #topics is likely to be small as compared to #documents

- Annotation of topics is likely to be sensitive to interpretation of topics and domain knowledge of annotators

Key Challenges

- Labelling topics may not be enough, need an iterative process where annotators annotate a few documents, words, phrases and topics.
- Some topics may be noisy, may not represent an entity from the taxonomy
- LectureBank Taxonomy is hierarchical. How to capture the hierarchy using a topic model?
- Visualize the corpus to facilitate taxonomy construction and text classification.

- Thank you!