Topic Models

Document Understanding, session 6

Northeastern University
College of Computer and Information Science

CS6200: Information Retrieval

Vocabulary Mismatch

Vocabulary mismatch is a fundamental challenge in IR: people use different words to talk about the same thing.

The document representations in vector space and language models are meant to represent the "topic" of a document, but have trouble recognizing additional terms on the same topic. We've already looked at term co-occurrence to try to address this.

Probabilistic topic models approach the problem by learning how likely each term is to appear when a certain topic is discussed.

Topic Modeling

The main idea of topic modeling is to learn a relatively small number of topics which explain the text in a document collection.

Each topic z has a probability distribution over the entire vocabulary, $\mathrm{P}(w|z)$.

The topics are latent (unobserved) random variables in a generative language model.

$z = 1$	$z = 2$	$z = 3$	$z = 4$
disease	water	mind	story
bacteria	fish	world	stories
diseases	sea	dream	tell
germs	swim	dreams	character
fever	swimming	thought	characters

Most probable words from four topics

Topic Modeling

Each word in each document has an associated (latent) topic, so the document has a probability distribution over topics $\mathrm{P}(z|d)$.

A generalized^{3} fundamental^{146} theorem^{267} of **natural**^{280} **selection**^{280} is derived^{233} for **populations**^{280} incorporating^{149} both **genetic**^{280} and **cultural**^{280} transmission^{25}. The phenotype^{3} is determined^{17} by an arbitrary^{3} number^{287} of multiallelic^{3} loci^{3} with two^{271}-factor^{60} **epistasis**^{280} and an arbitrary^{149} linkage^{3} map^{3}, as well as by **cultural**^{280} transmission^{25} from the **parents**^{280}. **Generations**^{280} are discrete^{69} but partially^{275} overlapping^{146}, and **mating**^{280} may be **nonrandom**^{280} at either the **genotypic**^{280} or the **phenotypic**^{280} level^{199} (or both). I show^{25} that **cultural**^{280} transmission^{25} has several^{173} important^{173} implications^{17} for the **evolution**^{280} of **population**^{280} **fitness**^{280}, most notably^{230} that there is a time^{72} lag^{72} in the response^{213} to **selection**^{280} such that the future^{287} **evolution**^{280} depends^{105} on the past **selection**^{280} **history**^{280} of the **population**^{280}.

Topic assignments here are numbers, and the contrast level indicates the word's probability of coming from the most common topic in the document.

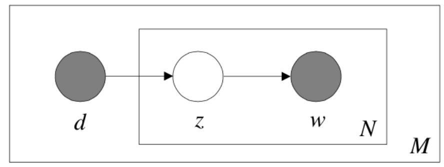
PLSA

Probabilistic Latent Semantic Analysis (PLSA) is a very widely-used topic model. It is a generative model, based on the following process.

1. Select a document d from the collection with probability $\mathrm{P}(d)$.

2. Select a latent topic z with probability $\mathrm{P}(z|d)$.

3. Generate a word w with probability $\mathrm{P}(w|z)$.



M – number of documents
N – document length
d – document, selected with $\mathrm{P}(d)$
z – topic, selected with $\mathrm{P}(z|d)$
w – word, selected with $\mathrm{P}(w|z)$

PLSA Likelihood

To train PLSA for a document collection, we need the likelihood of that data. The log likelihood function is:

$$L = \sum_{n=1}^{N} \sum_{m=1}^{M} tf_{w_m, d_n} \log P(d_n, w_m)$$

$$P(d, w) = P(d)P(w|d) = P(d) \sum_{z} P(w|z)P(z|d)$$

We choose parameters for the distributions $\mathrm{P}(d)$, $\mathrm{P}(w|z)$, and $\mathrm{P}(z|d)$ that maximize the expected log likelihood $\mathbb{E}_{\mathrm{P}(d,w)}[L]$ of our training data using a process known as Expectation Maximization (EM).

Training PLSA with EM

To train PLSA, we keep track of four distributions. We initialize to uniform distributions. Then we alternate between updating $P(z|d,w)$ from the others, and then updating the others from $P(z|d,w)$ and the data.

Each iteration increases the expected log likelihood of the data. These steps are repeated until the distributions converge.

Expectation (E) step:

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}$$

Maximization (M) step:

$$P(z) \propto \sum_d tf_{w,d} P(z|d,w)$$

$$P(d|z) \propto \sum_w tf_{w,d} P(z|d,w)$$

$$P(w|z) \propto \sum_d \sum_w tf_{w,d} P(z|d,w)$$

Wrapping Up

Topic modeling seeks to represent topics as probability distributions over the vocabulary, and thus to address vocabulary mismatch.

The resulting topics can be used as features for IR, for document clustering, or for many other purposes.

PLSA is a commonly-used topic model that's easy to train and gives reasonable performance.

Next, we'll see a selection of variations on this basic topic modeling framework.

Applications of Topic Models

Document Understanding, session 7

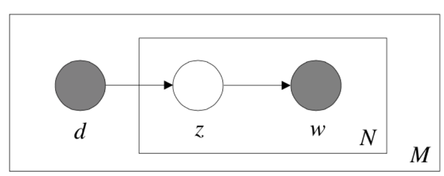
Northeastern University
College of Computer and Information Science

CS6200: Information Retrieval

Extending Topic Models

PLSA is the most basic probabilistic topic model, and the idea has been usefully extended in many ways.

- Its probability estimates have been regularized to improve output quality, most notably by Latent Dirichlet Allocation (LDA).
- The document collection has been grouped in various ways (e.g. by language or publication date) to give topics more flexibility.
- Additional data can be included, such as sentiment labels, to condition the vocabulary distribution on new factors.



M – number of documents
N – document length
d – document, selected with $\mathrm{P}(d)$
z – topic, selected with $\mathrm{P}(z|d)$
w – word, selected with $\mathrm{P}(w|z)$

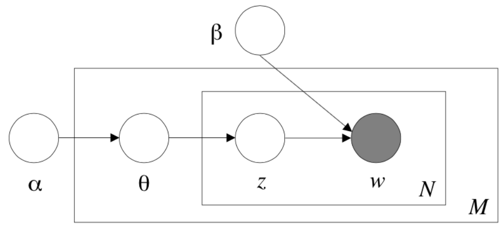
Latent Dirichlet Allocation

Latent Dirichlet Allocation regularizes PLSA by using Dirichlet priors for its Multinomial topic distributions. Most topic models extend LDA, not PLSA.

The distributions α and β are Bayesian posteriors, whose priors work like smoothing parameters to limit how extreme the document and vocabulary distributions can become.

The data likelihood is given by:

$$P(\mathcal{D}|a,\beta) = \prod_{d=1}^{M} \int p(\vartheta_d|a) \left(\prod_{n=1}^{N_d} \sum_{z} p(z|\vartheta_d) p(w_n|z,\beta) \right) d\vartheta$$



M – number of documents
N – document length
α – Multinomial dist. over documents
β – Multinomial dists. over words
θ – document, selected with $P(d|\alpha)$
z – topic, selected with $P(z|\theta)$
w – word, selected with $P(w|z, \beta)$

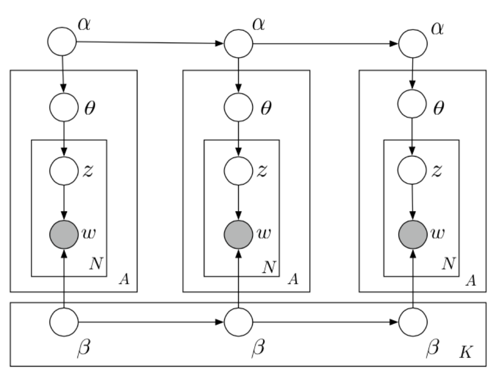
David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation.

Dynamic Topic Models

Language usage changes over time, due to vocabulary drift and communities' changing interests. Dynamic Topic Models capture that change by learning how topics drift as time goes on.

Documents are grouped into time steps, according to their publication dates.

The distributions over vocabulary and documents, α and β, are constrained to drift only gradually from the distributions in the preceding time step.

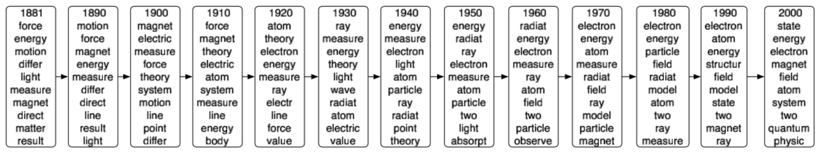


Three time steps of the model. α and β drift slightly in each time step.

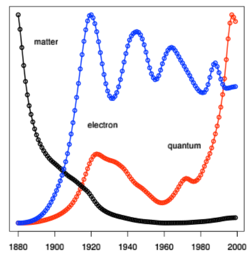
David M. Blei and John D. Lafferty. 2006. Dynamic topic models.

Topics over Time

The resulting topics show how language usage changes within each topic.



"Atomic Physics"



1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

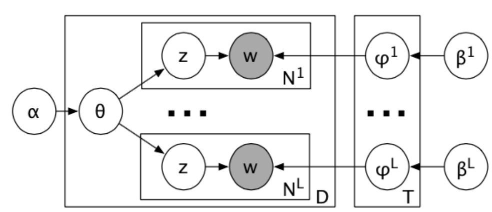
David M. Blei and John D. Lafferty. 2006. Dynamic topic models.

Polylingual Topic Models

Can we learn how topics are expressed by speakers of different languages?

Polylingual Topic Models accomplish this by training on a collection of document tuples: each tuple has a representative document from each language.

Tuples may be translations, or just Wikipedia pages in each language – even though they don't cover the same subtopics.



θ is a tuple of related documents, one in each language.
φ is a language-specific vocabulary distribution.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models.

Polylingual Topic Models

Two topics from EU Parliament Proceedings (direct translations)

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

DA	børn familie udnyttelse børns børnene seksuel
DE	kinder kindern familie ausbeutung familien eltern
EL	παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής
EN	**children family child sexual families exploitation**
ES	niños familia hijos sexual infantil menores
FI	lasten lapsia lapset perheen lapsen lapsiin
FR	enfants famille enfant parents exploitation familles
IT	bambini famiglia figli minori sessuale sfruttamento
NL	kinderen kind gezin seksuele ouders familie
PT	crianças família filhos sexual criança infantil
SV	barn barnen familjen sexuellt familj utnyttjande

Two topics from Wikipedia (related pages)

CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

Wrapping Up

There are many ways to group documents or include additional data to extend topic modeling. The resulting topics are useful for data exploration and categorization.

Topic models are not sufficient alone to yield good IR ranking performance, but they are a useful set of supplementary features for document understanding.

Next, we'll look at how to cluster documents together using any set of features.