Unveiling hidden semantic structures of corpora using Topic Models

Simón Roca and Jerónimo Arenas

May 14, 2018

Universidad Carlos III de Madrid, Dpto. TSC

Presented by Simón Roca (sroca@ing.uc3m.es) in Machine Learning Group, UC3M.

Outline

Outline

1. Introduction
2. Background
3. Paragraph LDA
4. Experiments
5. Conclusions

Structured corpus vs. Unstructured corpus

Some examples: project forms, job offers, course guides, patents, articles...

Núm. 1. Jueves 3 de Noviembre de 1853. Año I.

FARO DE VIGO.

PERIODICO MERCANTIL, AGRICOLA E INDUSTRIAL.

Esté periódico se publica los jueves y domingos de cada semana —Precio de suscricion 6 rs. al mes en esta ciudad, y fuera 20 rs. por trimestre franco de porte: Ultramar 12 rs. al mes, tambien franco.—No se recibe ninguna correspondencia que no venga franca.

Vigo 3 de Noviembre de 1853.

Al presentarnos en el estadio de la imprenta, no son otras nuestras aspiraciones que las de coadyuvar, en cuanto lo permitan nuestras débiles fuerzas, al fomento de los intereses de Galicia, que deplorables circunstancias por una parte, y preocupaciones funestas por otra, contribuyeron á que esta vasta provincia continúe sin participar de los beneficios que otras disfrutan ya. Muévenos tambien el deseo de sofocar esas rivalidades que no han desaparecido del todo entre provincias hermanas, y el de contribuir al mismo tiempo á que las Castillas, uniéndose mas estrechamente á este antiguo reino, puedan dar á su agricultura é industria mayor impulso, como Galicia mas vida á sus intereses mercantiles y marítimos con el ausilio del interior de la Peninsula.

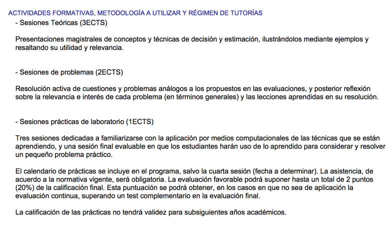
Si los enlaces de familia, el sentimiento relijioso, el elemento monárquico y el principio de unidad administrativa, pudieron reunir bajo un mismo cetro diversos paises, conocidos antes con los nombres de condados, señorios, principados y reinos para constituir una gran unidad política, no han sido capaces, por ahora, de borrar completamente las diferencias que en aquel entonces habia. Mucho, es cierto, se hizo desde el reinado de D. Alonso el Sábio para uniformar social y económicamente la nacion; pero no poco tiene que hacer la administracion y no menos la prensa, para que haya un todo armónico, sea una misma la lejislacion civil y económica, y desaparezcan esas diversidades de dialectos, de pesos, medidas y monedas, y con todo ello, falsas nados por nuestro silencio y las aseveraciones de algun escritor.

Ninguna provincia como la de Galicia es víctima de esas prevenciones desfavorables, de esos equivocados juicios; y por eso comun es el error en creerá este territorio, conquistado desde el siglo VI á los suevos por Leovigildo y erijido en reino por Fernando I en el siglo XI, como un pais triste, pobre y estéril, y á sus naturales como la gente de mas corto entendimiento. Las consecuencias que se siguieron de este fatal error nadié los desconoce, para estendernos á manifestarlas; baste decir, que esta gran porcion de la corona de Castilla no mereció por anteriores administraciones la atencion de que es muy digna, considerándola siempre como la Irlanda española. Contra ese reprensible abandono, contra esos calumniosos epitetos protestaron mas de una vez los hijos del pais; y nosotros que nos gloriamos de haber nacido en él, cumple á nuestra conciencia y á nuestro amor propio ofendido, demostrar que se equivocan grandemente los que toman la laboriosidad gallega por esclavitud, la economía por mezquindad y la prudencia por estupidez. Sin embargo, gracias á la ilustracion de la época no necesitamos esforzar los argumentos, ni aducir demasiadas pruebas en defensa del pais, porque otra es la opinion que se vá formando ya del carácter de sus habitantes, de sus disposiciones naturales y de la fertilidad del suelo.

Los que se dejan llevar de vulgaridades y se detienen poco á examinar un hecho para averiguar la causa que lo produce, los que miran con desden á los que se emplean en trabajos útiles plazo, para suplir muchas veces la falta de brazos que hay en otros puntos, los que, finalmente, juzgan que un territorio que produce lo bastante para sostener su poblacion, no la permite que emigre á otros en demanda de trabajo son víctimas, los unos de su lijereza, y los otros de un falso principio. La emigracion gallega á varias de nuestras provincias y al Portugal, no es hija de la esterilidad del suelo como generalmente se presume, sinó mas bien de los hábitos de laboriosidad y economia que tanto distinguen al labrador gallego.

Nadie ignora que cuenta Galicia con multitud de rias y rios abundantes en esquisitos pescados, que tiene profundos valles, dilatadas vegas, frondosas selvas y dehesas, y estensos prados para sostener una poblacion mas numerosa; pero todos se olvidan ó desconocen que nuestros paisanos no se contentan con los ricos frutos que el suelo produce, que no pueden vivir en la ociosidad, ni tranquilamente descansar de sus faenas consumiendo los productos de la tierra, regada y fertilizada con el sudor de su frente. Si estas circunstancias se tuviesen presentes cuando el agricultor gallego deja su hogar para trabajar en algunas de nuestras provincias, seguros estamos que otro seria el juicio formado del pais. Tan arraigadas están en él la sobriedad, el apego al trabajo y la prudente economia, que sus habitantes van en busca de ocupacion en donde quiera que la encuentran, sin por eso abandonar el cuidado de sus terrenos.

Es costumbre inveterada en nuestra gente del campo, concluidas que son las principales labores que exije la



ACTIVIDADES FORMATIVAS, METODOLOGÍA A UTILIZAR Y RÉGIMEN DE TUTORÍAS

- Sesiones Teóricas (3ECTS)

Presentaciones magistrales de conceptos y técnicas de decisión y estimación, ilustrándolos mediante ejemplos y resaltando su utilidad y relevancia.

- Sesiones de problemas (2ECTS)

Resolución activa de cuestiones y problemas análogos a los propuestos en las evaluaciones, y posterior reflexión sobre la relevancia e interés de cada problema (en términos generales) y las lecciones aprendidas en su resolución.

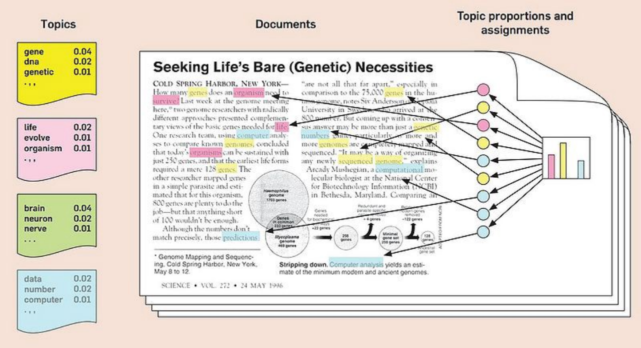
- Sesiones prácticas de laboratorio (1ECTS)

Tres sesiones dedicadas a familiarizarse con la aplicación por medios computacionales de las técnicas que se están aprendiendo, y una sesión final evaluable en que los estudiantes harán uso de lo aprendido para considerar y resolver un pequeño problema práctico.

El calendario de prácticas se incluye en el programa, salvo la cuarta sesión (fecha a determinar). La asistencia, de acuerdo a la normativa vigente, será obligatoria. La evaluación favorable podrá suponer hasta un total de 2 puntos (20%) de la calificación final. Esta puntuación se podrá obtener, en los casos en que no sea de aplicación la evaluación continua, superando un test complementario en la evaluación final.

La calificación de las prácticas no tendrá validez para subsiguientes años académicos.

Reviewing LDA



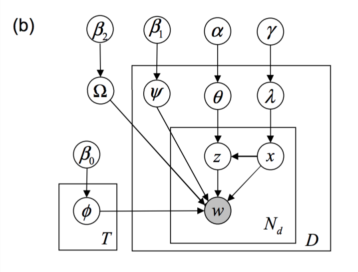
- Inputs: BoW corpus, hyperparameters...
- Outputs: document-topic proportion matrix, topic-word proportion vectors...

Our proposal

- Split the documents into paragraphs.
- New plate/variable: let the paragraphs be semantic or background.
- Use different LDA parameters for each kind.

Outline

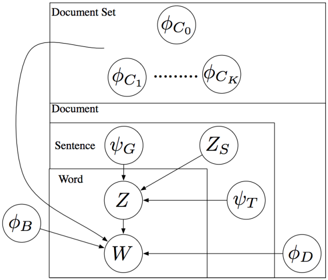
Previous approaches: specific and general topics



Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model

- Three sets of topics: classic topics, corpus specific topic, and document specific topic.

Previous approaches: specific and general topics

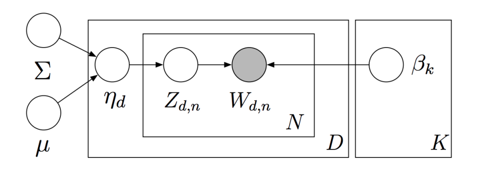


(b) HIERSUM Graphical Model

Exploring Content Models for Multi-Document Summarization

- Used for summarizing documents.
- They assume constant topic across a single sentence.

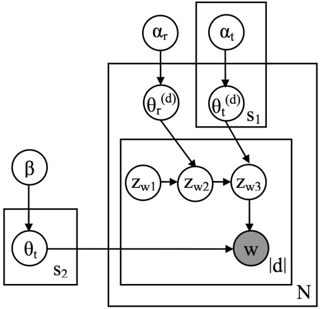
Previous approaches: Hierarchies among topics



Correlated Topic Models

- Logistic Normal distribution as a prior: correlation among topics can be obtained.
- Non Conjugacy, but permits clustering topics.

Previous approaches: Hierarchies among topics

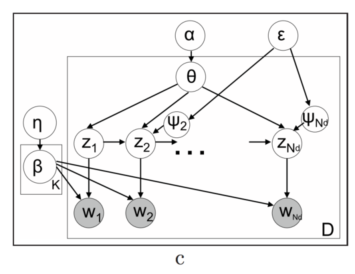


(b) Four-Level PAM

Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations

- It allows learning topics, sub-topics, super-topics...
- Due to its tree structure (DAG), it permits obtaining correlations.

Previous approaches: Word dependency



Hidden Topic Markov Models

- Each word in a sentence comes from the same topic.
- Transition between sentences modeled with a Markov Chain.

Previous approaches: Word dependency

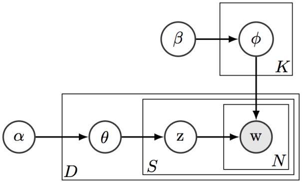


Figure 1: The *senLDA* model. The words w of a sentence share the same topic z.

On a topic model for sen-
tences

- Each word in a sentence comes from the same topic.
- Modeled as an extra plate in model.

Is our model different?

- We propose a model combining some of the previous approaches.
- However, that's not the main contribution.
- Why not selecting in which parts useful topics may be learned, and then looking there for best quality topics?

Outline

Graphical Model

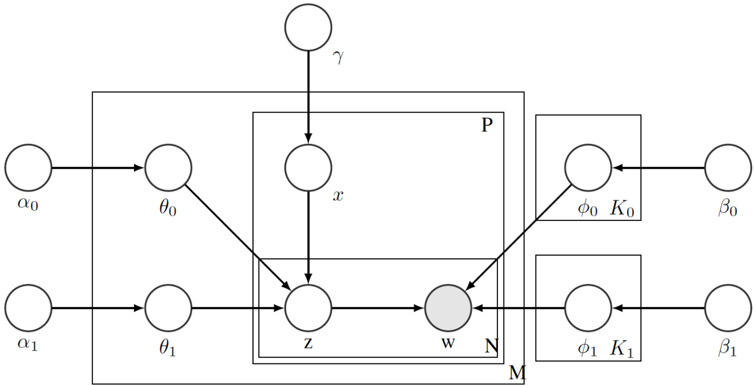


Figure 2: Paragraph LDA graphical model

Generative Model

1. $\phi_1 \sim Dir_{V_1}(\beta_1)$
2. $\phi_0 \sim Dir_{V_0}(\beta_0)$
3. For each document,
   (a) $\theta_1 \sim Dir_{K_1}(\alpha_1 << 1)$
   (b) $\theta_0 \sim Dir_{K_0}(\alpha_0 \geq 1)$
   (c) For each paragraph,
      - $\psi_x \sim Dir_2(\gamma)$
      - $x \sim Ber(\psi_x)$
        - if $x = 1$, for each word:
          * $x_w \sim Ber(m)$
          * if $x_w = 1$:
            · $z \sim Mult(\theta_1)$
            · $w \sim Mult(\phi_{1,z})$
          * if $x_w = 0$:
            · $z \sim Mult(\theta_0)$
            · $w \sim Mult(\phi_{0,z})$
        - if $x = 0$, for each word:
          * $z \sim Mult(\theta_0)$
          * $w \sim Mult(\phi_{0,z})$

Inference

$$p(\vec{z}, \vec{x}, \vec{w}) = p(\vec{w}|\vec{z}, \vec{\beta})p(\vec{z}|\vec{x}, \vec{\alpha})p(\vec{x}|\gamma) \tag{1}$$

$$p(\vec{x}|\gamma) = \frac{\Delta(\vec{n}'_d + \gamma)}{\Delta(\gamma)} \tag{2}$$

$$p(z_i = k|\vec{z_{\neg i}}, \vec{w}) \propto (n_{m,\neg i}^{(t)} + \alpha_k)\frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \tag{3}$$

$$p(x_p = s|\vec{w}, \vec{x}_{\neg p}, \vec{z}, \alpha, \beta, \gamma) \propto \frac{\prod_{w \in p}(n_{s,\neg p}^{(w)}+\gamma)...(n_{s,\neg p}^{(w)}+\gamma+(n_{s,p}^{(w)}-1))}{(\sum_{w \in V}(n_{s,\neg p}^{(w)}+\gamma))...(\sum_{w \in V} n_{s,\neg p}^{(w)}+\gamma+(n_{s,p}^{(w)}-1))} \tag{4}$$

Outline

1. Introduction
2. Background
3. Paragraph LDA
4. Experiments
5. Conclusions

Evaluation metrics: Histogram intersection distances

For matching real topics and predicted ones in synthetic dataset.
Checking how much they differ.

$$exp(-\alpha * \sum n_{j=1} min(I_j, M_j)) \tag{5}$$

Evaluation metrics: Topic coherence

- Perplexity (held-out probability) is not the best choice for topic quality/coherence.
- Some measurements based on PMI (word co-ocurrences), WordEmbeddings (word vectors), high correlation with human judgement.
- Learned on reference corpus, strange behaviour with specific vocabulary.
- David Mimno stands that PMI can be obtained in the same corpus.

Experiments on synthetic dataset: some numbers

When the generative model is met, is our method worth?

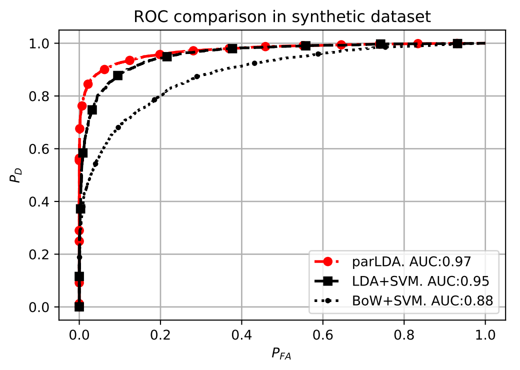
Attribute	Docs(test)	Paragraphs	Total Words	$K_0(K_1)$	V	$\alpha_0(\alpha_1)$	$\beta_0(\beta_1)$
Value	3000(500)	?	?	10 (30)	5000	2 (0.1)	0.1(0.1)

Table 1: Synthetic dataset generated using our generative model

Noisy: variable proportion of background words in semantic paragraphs.

Experiments on synthetic dataset: ROCs

It seems so!



Experiments on real datasets

After several runs in Patents, SCIELO and NIPS:

- Good intuition about learned topics...
- ...But referenced corpus coherence doesn't claim so.
- In some corpus there are too many semantic paragraphs.

Experiments on real datasets: Patents

Some background topics:
['temperature', 'gas', 'liquid', 'water', 'heat', 'heating', 'pressure', 'oil', 'fluid', 'chamber']
['compound', 'reaction', 'mixture', 'mmol', 'substituted', 'solution', 'atom', 'formula', 'acid', 'stirred']
['page', 'subject', 'action', 'sequence', 'active', 'report', 'activity', 'factor', 'total', 'property']

Some semantic topics: ['channel', 'transmission', 'antenna', 'receiver', 'symbol', 'transmit', 'transmitted', 'transmitter', 'resource', 'carrier']
['server', 'client', 'call', 'network', 'web', 'request', 'status', 'telephone', 'manager', 'local']
['frequency', 'phase', 'filter', 'digital', 'noise', 'pulse', 'band', 'amplitude', 'clock', 'gain']

Proper experiments:

- Feasible: running parLDA in these datasets and compare some topics to those obtained from classic LDA.
- Just in time: small validation process on hyperparameters based on topic coherence.
- ...? Deadline on Friday, 8pm.

Outline

1. Introduction
2. Background
3. Paragraph LDA
4. Experiments
5. Conclusions

Conclusions

- Our inference performs well enough when the generative model is met.
- In that scenario, identifies better semantic paragraph than other methods, even without labels.
- On real datasets, it finds different paragraphs, learning reasonable topics.

Future Work

- Deep analysis of the quality of these new topics.
- Analysis of topic coherence measurements.
- Getting richer models.
- ...