

Clasificación de documentos científicos usando modelado de tópicos en Python

Luis A. López Espinosa, Siri Jodha S Khalsa

Research Associate, NSIDC
Boulder, Colorado. USA.

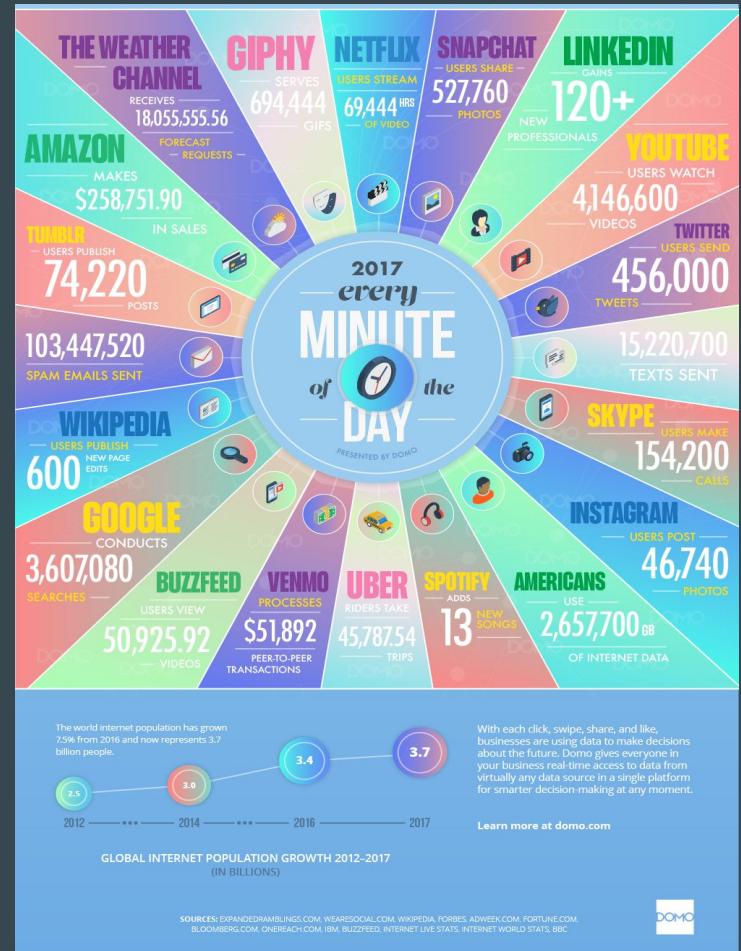
luis.lopez@nsidc.org
[@betolink](https://twitter.com/betolink)

SciPy LATAM 2019, Bogotá Colombia



Introducción

- Vivimos en la era de la información, somos la antesala del “Homo Machina” y generamos mucha más información de la que podemos procesar.
- La ciencia experimental y las colaboraciones científicas a gran escala no son la excepción.
- El contexto de este proyecto es EarthCube una iniciativa de la NSF para avanzar las geo-ciencias en el frente informático.
- BCube @ EarthCube
<https://www.earthcube.org/group/polar-data-insights-search-analytics-deep-scientific-web>



Introducción

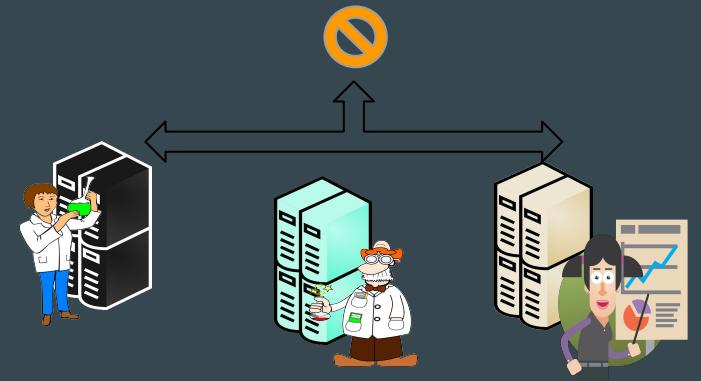
- Buscadores y redes sociales son nuestra ventana a la información.
- Científicos alrededor del mundo comparten su investigación en journals, conferencias (como esta) y sorpresivamente de vez en cuando hasta en redes sociales.
- Buscadores y redes sociales tienen objetivos comerciales y no indexan gran parte de la información que ellos no consideren relevante.
- La información científica es mucho más específica que una noticia o una conversación en lenguaje natural.

Google

Greenland: Aerosol Datasets



Greenland: Mass Spectrum Analysis in Geothermal Areas



Problemas

Encontrar información relevante.

- Billones de documentos en la web
- La mayoría del contenido no es relevante para la ciencia
- Información científica no es siempre indexada.
- Conocimiento científico usa lenguaje especializado.
- ¿Qué es relevante?



Generar conocimiento a partir de información relevante.

- Formatos
- Metadatos
- Información incompleta
- Los problemas de arriba multiplicados x 100000...



Para cada problema...

Generar conocimiento a partir de información relevante.

- Formatos
- Metadatos
- Información incompleta
- Los problemas de arriba multiplicados x 100000...



Extraer conocimiento es un problema difícil.

Florentino Ariza, de los cambios, cuando la navega cuenta de que el de los grandes de la memoria. El cómo con el buques árboles como una expresión o suspiro viajó



sorprendido. El día siguiente, el río Magdalena, uno de los grandes de la memoria. El cómo con el buques árboles como una expresión o suspiro viajó

Resumen. La cuenca del río Magdalena provee 20 millones de m³ de agua, sus bosques almacenan un promedio de 50 toneladas de CO₂ por hectárea y sus poblaciones están ligadas al bosque de la cultura prehispánica. Se plantean propuestas de manejo para cada zona específica considerando los actores sociales involucrados



Python NLP

- Scikit-learn
- Spacy
- Gensim
- GLoVE
- NLTK
-



spaCy



pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Natural Language
Tool Kit (NLTK)
Basic Text Analytics



LDA

- LDA es una forma de modelo probabilístico generativo.
 - asume que los documentos se generaron a partir de un mezcla de temas, cada tema es una distribución sobre palabras perteneciente a ese tema
- Una palabra puede pertenecer a múltiples tópicos.
- Palabras pueden agruparse en n-grams

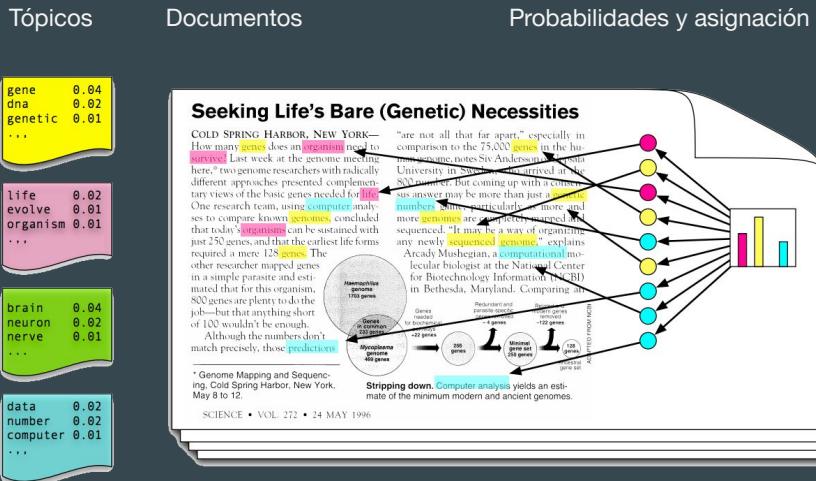


Image credit: M. I. Jordan, T. M. Mitchell

Modelado de topicos con LDA (latent dirichlet allocation)

$$P(Z|W, D) = \frac{\# \text{ of word } W \text{ in topic } Z + \beta_w}{\text{total tokens in } Z + \beta} * (\# \text{ words in } D \text{ that belong to } Z + \alpha)$$

LDA

Florentino Ariza, en efecto, estaba sorprendido de los cambios, y lo estaría más al día siguiente, cuando la navegación se hizo más difícil, y se dio cuenta de que el río padre de La Magdalena, uno de los grandes del mundo, era sólo una ilusión de la memoria. El capitán Samaritano les explicó cómo la deforestación irracional había acabado con el río en cincuenta años: las calderas de los buques habían devorado la selva enmarañada de árboles colosales —que Florentino Ariza sintió como una opresión en su primer viaje.

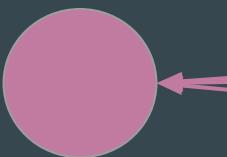
{ El amor en tiempos del cólera }

{ Navegación }

{ Ecología }

LDA

{ Ecología }



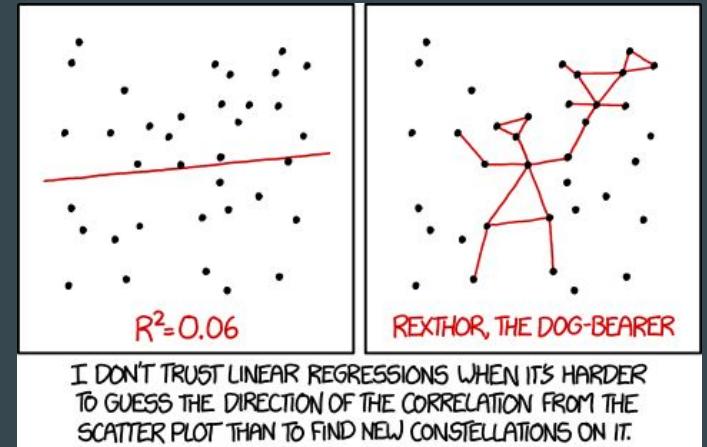
{ ¿Impacto ambiental? }



Resumen. La cuenca del río Magdalena provee 20 millones de m³ de agua, sus bosques almacenan un promedio de 50 toneladas de carbono por hectárea y sus pobladores han estado ligados al bosque desde la época prehispánica. Se plantean propuestas de manejo para cada zona específica considerando los actores sociales involucrados

Limitaciones del modelado de tópicos

- Recursos computacionales
- Priors (condiciones a priori)
- Modelos probabilísticos no tienen conocimiento del lenguaje
- Inestabilidad en los tópicos dependiendo de los parámetros usados
- Sensible al preprocesamiento de nuestro corpus.
- Malos resultados en documentos homogéneos



Ventajas del modelado de tópicos

- Recursos computacionales
 - No tan intensivo como redes neuronales
- Modelos probabilísticos no tienen conocimiento del lenguaje
 - Podemos usarlo sin tener conocimiento del lenguaje del corpus
- Muchas librerías implementadas en Python

Particularidades del modelado de tópicos en documentos científicos

- Lenguaje especializado
- Abreviaciones y puntuación
- Uso de múltiples lenguajes (Latín)
- Frases cortas
- Pérdida de información al pre-procesar
- Ruído

QoS Routing Algorithms based on Multi-Objective Optimization for Mesh Networks

M. Camelo, Member, IEEE, C. Omaña and H. Castro

Abstract— In this paper we present a new alternative for routing with quality of service (QoS) problem solution in Wireless Mesh Networks (WMN). This problem has the special characteristic that more than one objective functions are competing between them. A multi-objective model is proposed for this problem and includes QoS parameters such as bandwidth, packet loss rates, delay and power consumption. The classical approach consists in optimizing a single objective (the QoS parameter), however does not take into account the conflicting nature of this parameters leading to suboptimal solutions. In this work is proposed the use of multi-objective evolutionary algorithms (MOEA), particularly NSGA II which allow finding an optimal solution taking into account all the objectives as QoS parameters.

Keywords— Genetic Algorithms, meta-heuristics, multi-objective optimization, wireless mesh networks.

I. INTRODUCCIÓN

LAS REDES inalámbricas Mesh (WMN - Wireless Mesh Networks) son una nueva tecnología que promete jugar un importante rol en el futuro de la siguiente generación de redes móviles. Este tipo de redes tiene como características el diseño de red, la recuperación de errores, la eficiencia en la organización y recuperación de fallos, lo cual permite un rápido despliegue, fácil mantenimiento, bajo costo, alta escalabilidad, servicios confiables así como a aumentar la capacidad de red, conectividad y capacidad de recuperación [1]. La arquitectura de un WMN está compuesta por Enrutadores y clientes Mesh inalámbricos. Los Clientes Mesh inalámbricos (WMCs Wireless Mesh Clients) son dispositivos que necesitan transmitir y recibir datos. Los Enrutadores Mesh inalámbricos (WBRs - Wireless Mesh Router) trabajan como puntos de acceso y/o interfaces. Los puntos de acceso trabajan como parte de la red inalámbrica Multi-Salto y son utilizados como backbone de la red, permitiendo el acceso a la red a los WMCs. Las interfaces son los puntos de interconexión entre diferentes redes y son llamados normalmente Gateways. Tradicionalmente los WMNs se implementaron:

La Calidad de Servicio (QoS - Quality of Service) es un valor cuantitativo o cualitativo que define un contexto de desempeño entre el proveedor y el cliente [3]. Cuando la QoS es garantizada por la red, la red es capaz de satisfacer un conjunto predeterminado de restricciones sobre el rendimiento del servicio a través de una comunicación de extremo a

M. Camelo, Universidad de Granada, miguel.camelo@ugr.es
C. Omaña, Universidad de los Andes, c.omanal@egresados.uniandes.edu.co
H. Castro, Universidad de los Andes, hcastro@uniandes.edu.co

extremo en términos de retraso, ancho de banda disponible, la tasa de pérdida de paquetes, etc. Aplicar QoS implica encontrar un conjunto de valores que garanticen el mejor rendimiento, lo cual conlleva a la necesidad de optimizar simultáneamente las funciones que representan los objetivos, por ejemplo, "minimizar el retraso" y "minimizar la tasa de pérdida de paquetes".

Sin perder generalidad, en este tipo de problemas los objetivos pueden estar en conflicto y a cada objetivo le corresponde a una solución óptima diferente. La Optimización Multi-Objetivo (MO) no calcula una solución única, sino que por el contrario, permite obtener un conjunto de soluciones óptimas (frente de Pareto Óptimo [4]) que representan valores de equilibrio entre los diferentes objetivos.

En el campo de las telecomunicaciones el MO puede ser resuelto mediante dos enfoques. Los *métodos clásicos*, los cuales optimizan una única solución en cada iteración, siguiendo una regla de transición o usando *algoritmos evolutivos*, que imitan los principios evolutivos para conducir su búsqueda hacia una solución óptima. Este trabajo presenta un nuevo enfoque para garantizar determinados niveles de calidad de servicio en una red de WMN usando técnicas de optimización de uso común en algoritmos Evolutivos Multi-Objetivo (MOEA, Multi-Objective Evolutionary Algorithms).

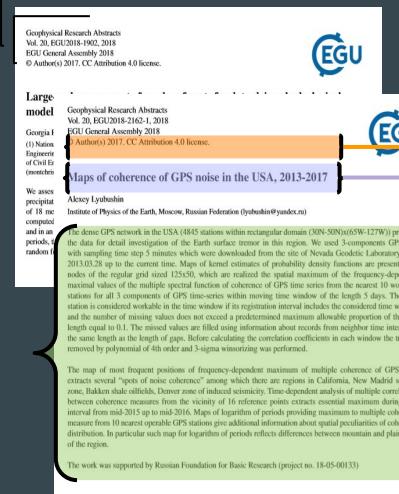
Los MOEA incorporan todas las características de los algoritmos Meta-heurísticos usados en problemas de optimización mono-objetivo pero extendidos a los problemas Multi-Objetivo. Los MOEA también evitan los problemas que se encuentran en los métodos clásicos como:

- La necesidad de la selección correcta de los parámetros iniciales de los algoritmos para ayudar a determinar la solución óptima. Por ejemplo, los algoritmos basados en Simplex presentan el problema de tener que determinar una solución base factible.
- La aproximación punto-a-punto, usada en los métodos clásicos no permiten explotar completamente las ventajas de sistemas en paralelo.

- Ineficiencia e ineficacia al aplicar el algoritmo en multiples tipos de problemas de optimización. Un algoritmo diseñado para un tipo de problema específico tal vez no sea eficiente para resolver eficiente mente otros problemas de optimización.
- Deficiencias de los algoritmos cuando son aplicados en problemas con espacios de búsqueda discretos.

Al analizar las anteriores limitaciones de los métodos clásicos, las cuales son eficientemente superadas por los MOEA, el presente artículo muestra que el problema de

Jupyter notebooks



```
Out[5]:  
Clearly if we were to reduce the main sessions to 3 we end up with less fancy names that the current ones but kind of make sense.  
This modeling is context-independent, we could bring some context via Word2Vec and will discuss that later on. Now let's make an  
experiment on the current corpus, we trained a simple model with 3 topics, let's classify some abstracts and see where they fall into.  
In [ ]:  
# Cell 6: Classifying an abstract using our GENSIM model  
  
# ESSI abstracts taken from https://meetingorganizer.copernicus.org/EGU2018/EGU2018-9778.pdf  
  
document = """  
submarine slope failure is a ubiquitous process and dominant pathway for sediment and organic carbon flux f  
rom  
continental margins to the deep sea. Slope failure occurs over a wide range of temporal and spatial scale  
s, from small (10e4-10e5 m3/event), sub-annual failures on heavily sedimented river deltas to margin-altering  
and  
tsunamigenic (10-100 km3/event) open slope failures occurring on glacial-interglacial timescales.  
Despite their importance to basic (closing the global source-to-sink sediment budget) and applied  
(submarine geohazards) re- search, submarine slope failure frequency and magnitude on most continental mar  
gins  
remains poorly constrained. This is primarily due to difficulty in 1) directly observing events, and 2) rec  
onstructing  
age and size, particularly in the geologic record. The state of knowledge regarding submarine slope failur  
e  
preconditioning and triggering factors is more qualitative than quantitative; a vague hierarchy of factor  
importance  
has been established in most settings but slope failures cannot yet be forecasted or hindcasted from  
a priori knowledge of these factors.  
www  
  
vec = dictionary.doc2bow(clean_document(document))  
predicted_topics = lda_model[vec]  
print(predicted_topics)  
  
Now let's increment the number of topics to 8 and see what we get  
In [8]:  
# Cell 6: LDA Topic Modeling expanding our topics  
  
from collections import defaultdict  
import re  
p = re.compile('.{1,400}')  
topic_list = defaultdict(list)  
# num passes should be adjusted, 5 is just a guesstimate of when convergence will be achieved.  
num_passes = 10  
num_topics = 8  
words_per_topic = 7  
  
lda_model = models.ldamodel.LdaModel(lda_corpus,  
                                         num_topics=num_topics,  
                                         id2word = dictionary,  
                                         passes=num_passes,  
                                         chunksize=17)  
topics = lda_model.print_topics(num_topics=num_topics, num_words=words_per_topic)  
print("Topic List:\n")  
for topic in topics:  
    weighted_terms = topic[1].split(' + ')  
    terms = [t[6:] for t in weighted_terms]  
    for term in terms:  
        topic_list[topics.index(topic)].append(term.replace('"', ''))  
print(topic)
```

The Need for Open Source Software in Machine Learning

Yann LeCun Et. Al.

Open source tools have recently reached a level of maturity which makes them suitable for building large-scale real-world systems.



open_source tool have recent reach level maturity which make suitable build large_scale real_world systems.

Paper: Analysis of geothermal energy as an alternative source for electricity in Colombia

Samuel S. Salazar, Yecid Muñoz & Adalberto Ospino

The conclusion is that geothermal energy is a good alternative to help achieve this objective. By 2025, geothermal sources are expected to generate at least 1400 GWh of electric power per year, equivalent to 1.65% of total electricity estimate demand in Colombia. If the full potential that has been assessed were exploited, generation capacity could reach up to 17,400 GWh/year (equivalent to close to 20% of the country's demand) by 2025.



conclusion geothermal energy good_alternative help achieve objective 2025 geothermal sources expected generate at_least 1400 gwh electric_power per_year equivalent 1.65% total electricity estimate demand colombia full potential asses is exploit generate capacity reach up_to 17400 gwh/year equivalent close_to 20% country demand 2025.

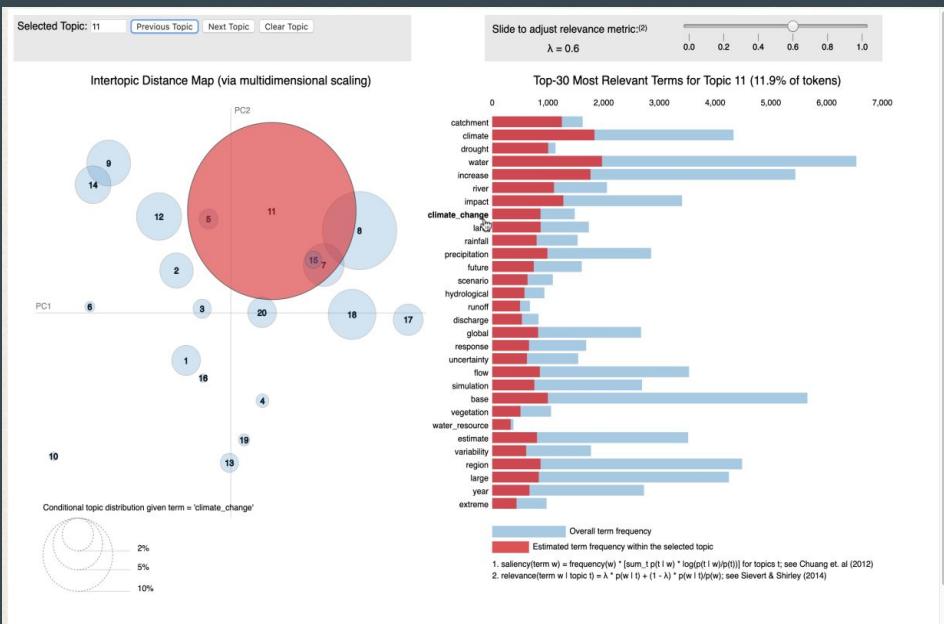
Palabras, tópicos y documentos

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	dominant_topic
Doc0	0	0	0	0	0	0.14	0	0	0	0.84	9
Doc1	0	0.05	0	0	0.05	0.24	0	0.65	0	0	7
Doc2	0	0	0	0	0	0.08	0.2	0	0	0.71	9
Doc3	0	0.55	0	0	0	0.44	0	0	0	0	1
Doc4	0.16	0.29	0	0	0	0.53	0	0	0	0	5
Doc5	0	0	0.05	0	0	0	0	0.12	0	0.83	9
Doc6	0	0	0	0	0	0.88	0.1	0	0	0	5
Doc7	0	0	0	0	0	0.99	0	0	0	0	5
Doc8	0	0	0.08	0.67	0	0	0	0	0.24	0	3
Doc9	0	0	0.74	0	0	0.14	0	0	0.11	0	2
Doc10	0	0	0	0	0.41	0.16	0	0.06	0	0.36	4
Doc11	0	0	0	0	0	0	0	0.97	0	0	7
Doc12	0	0	0	0.44	0	0.04	0	0.27	0	0.24	3
Doc13	0.14	0	0	0	0	0.07	0.57	0.08	0	0.13	6
Doc14	0	0	0	0	0.78	0.22	0	0	0	0	4

Visualizando nuestros modelos

PyLDA

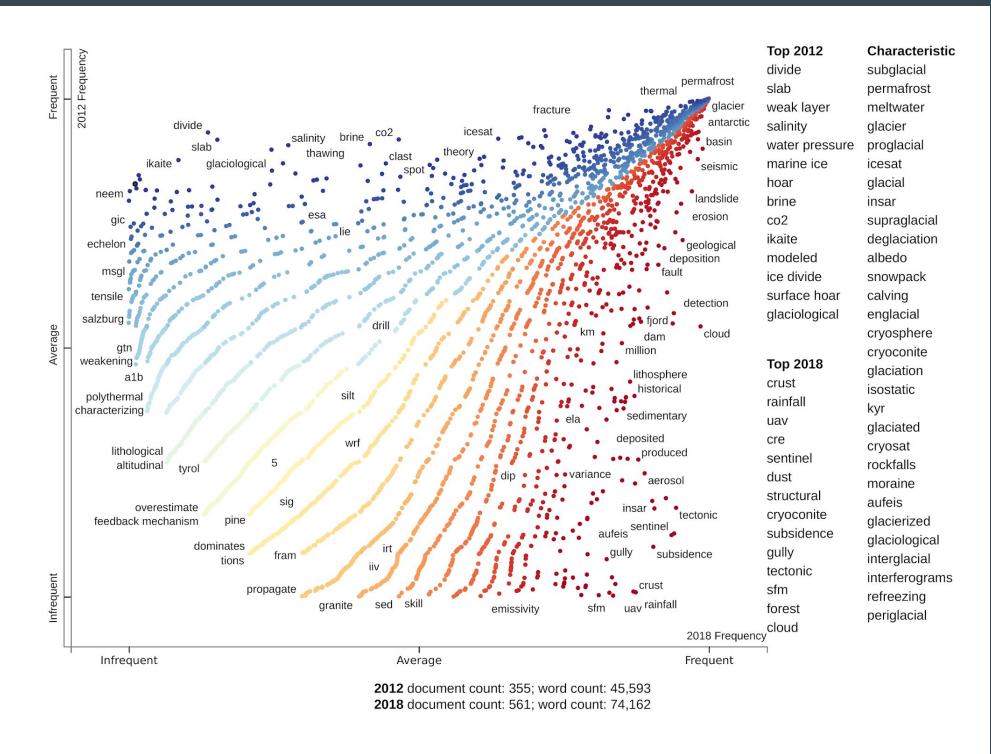
- Visualiza tópicos usando PCA
- Interactivo
- Intuitivo



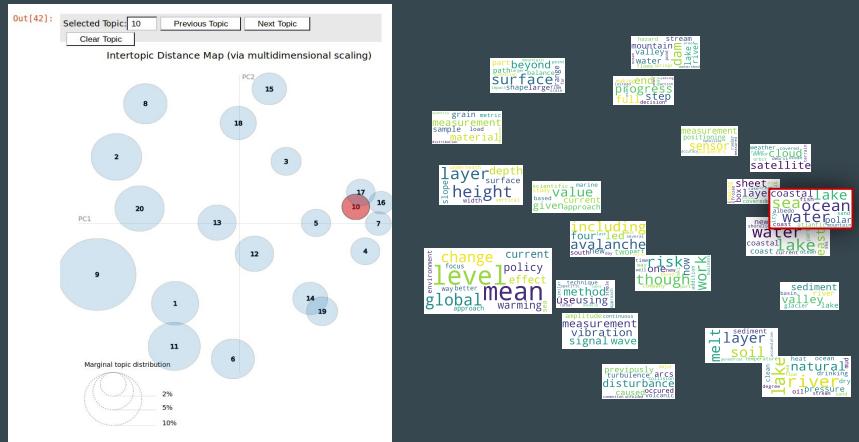
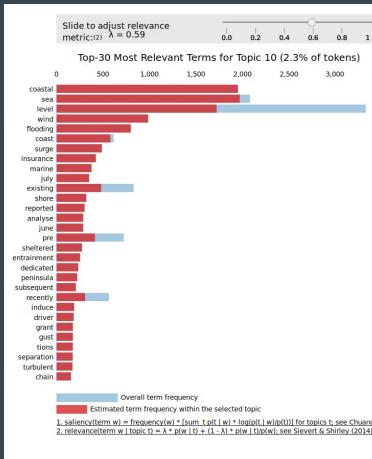
Visualizando nuestros modelos

ScatterText

- Visualiza características en nuestro corpus
- Interactivo
- Intuitivo



Modelado con LDA: Resultados



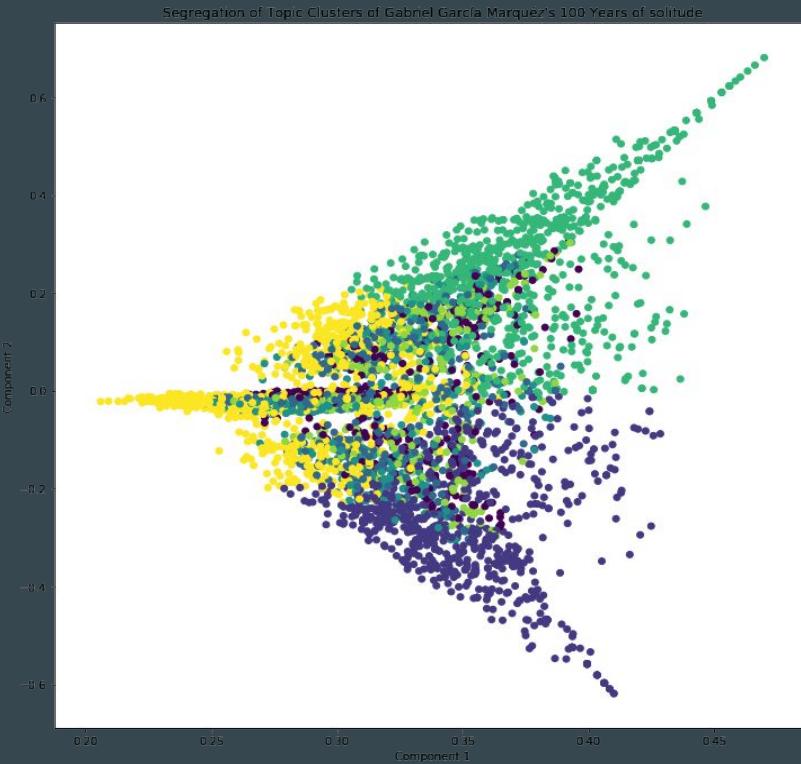


Modelando el realismo mágico

github.com/betolink/scipy-topics



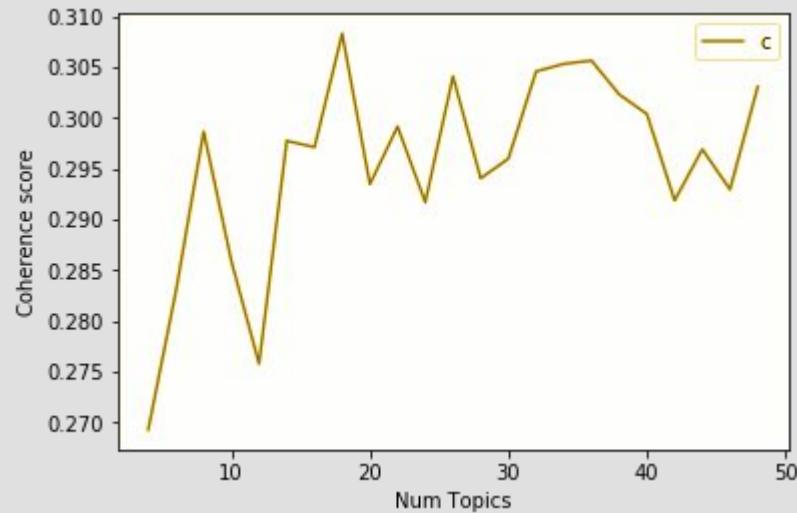
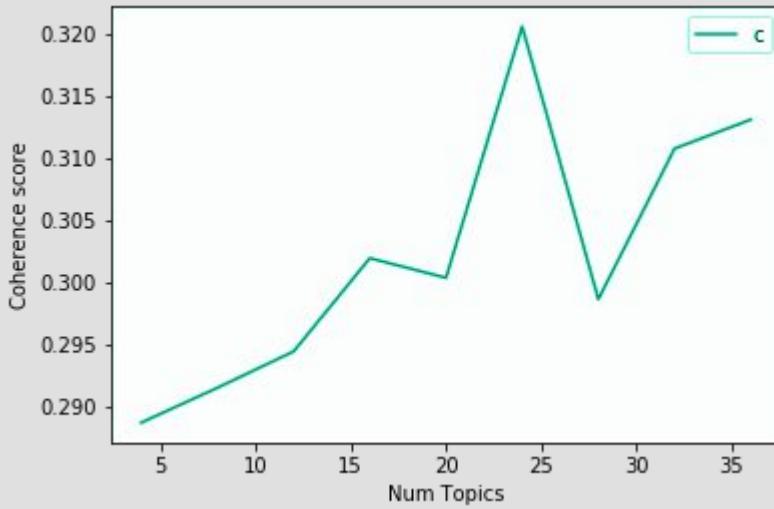
100 Años de Soledad



Perdónanos Gabo...



Evaluación de los modelos: coherencia



Aplicaciones del modelado de tópicos

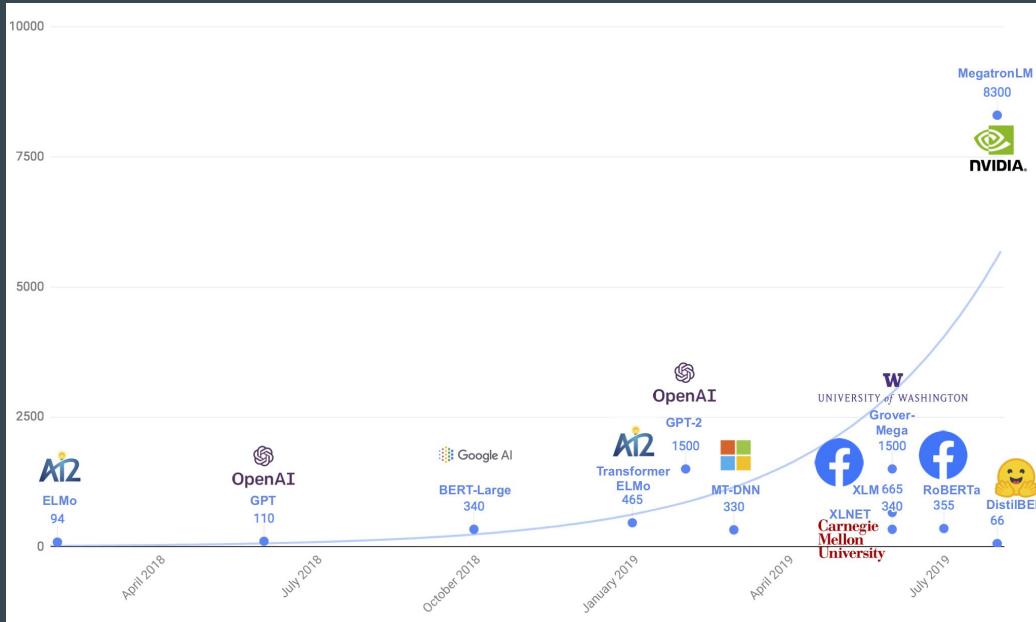
- Clasificación de textos
 - Bibliotecas temáticas
 - Asignación de textos en conferencias y journals (spam!)
 - más!
- Sistemas de recomendaciones
 - Facebook, Twitter (Este artículo te puede interesar)
- Bioinformática
 - Genes como palabras
- Inferir tendencias
 - Investigaciones científicas, uso de tecnologías...

Más allá de los recursos lingüísticos aprendizaje profundo

- Hasta ahora no nos hemos centrado en el significado de las palabras y sus contextos
- Wordnet y el aprendizaje profundo
- El modelado de tópicos
- NLP y aprendizaje profundo



Más allá de los modelos probabilísticos: Wordnet y aprendizaje profundo



Referencias

- Advanced Topic Modeling @ Cornell
 - <https://mimno.infosci.cornell.edu/info6150/>
- PDI topics
 - github.com/USCDataScience/pdi-topics
- Tutorial LDA
 - <https://www.machinelearningplus.com/nlp/topic-modelling-python-sklearn-examples/>
- Modelando realismo mágico
 - <http://github.com/betolink/scipy-latam>

... Google

¡Gracias!