Inspira Crea Transforma

H. Z. Stephen Robertson, The Probabilistic Relevance Framework:

UNIVERSIDAD EAFIT®

Proyecto Integrador

Procesamiento de texto

Equipo 4

Juan Diego Estrada Pérez

Ing. Sistemas

Andrés Franco Zapata

Ing. Sistemas

Liceth Mosquera Galvis

Ing. Eléctrica

Alejandro Palacio Vásquez

Ing. Matemático

Johan Steward Rios

Ing. Matemático

Contexto

Procesamiento de Texto

Minería o Analítica de texto

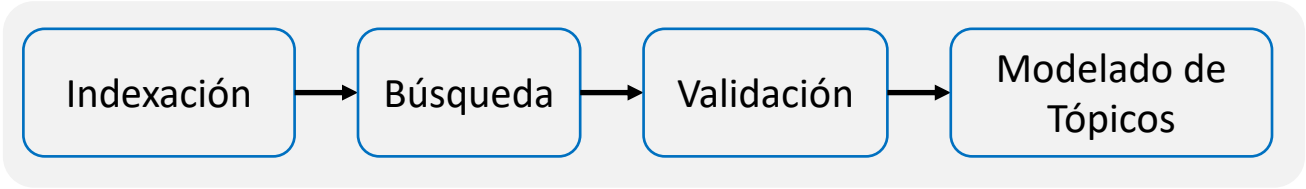
Respuesta a Tendencias y Avances Tecnológicos

Primeros estudios 1.950[1]

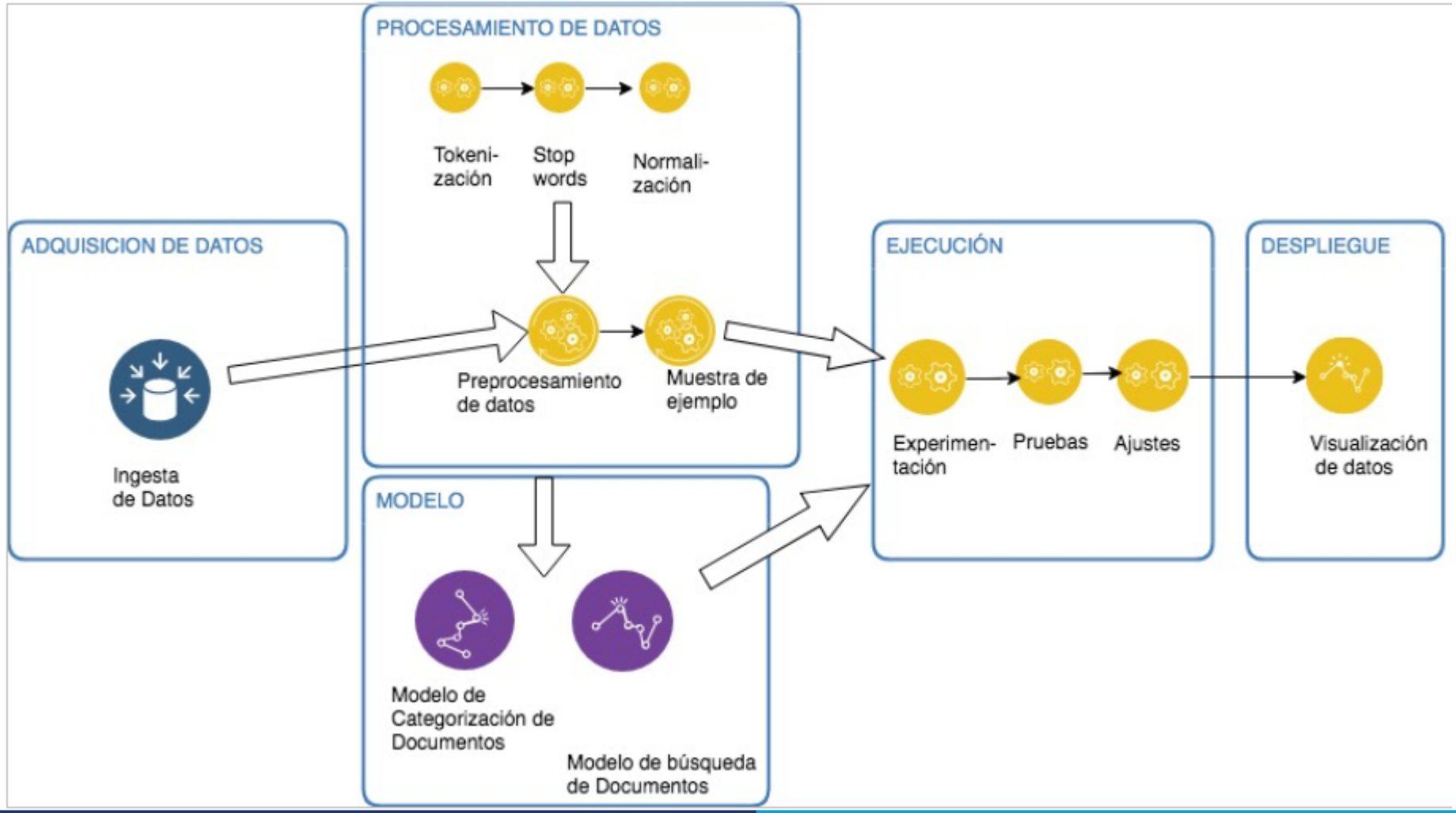
Analizar datos no estructurados

Objetivo General

1. Recuperación de la información
2. Procesamiento de Lenguajes Naturales (NLP)
3. Extracción de la información
4. Minería de datos

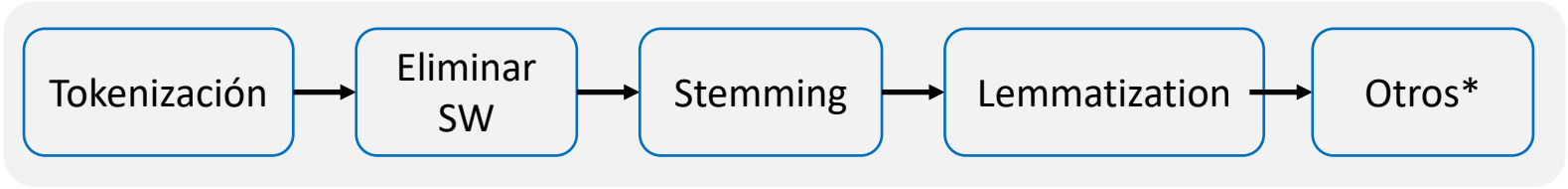


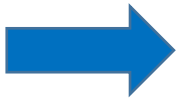
Arquitectura

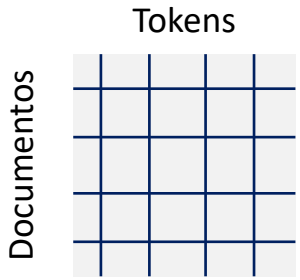


Proceso de Indexación

- Formatos .txt , .pdf y .dc
- Mismo dominio de interés
- 980 documentos
- 13 millones de palabras
- Promedio 12.900 palabras x documento







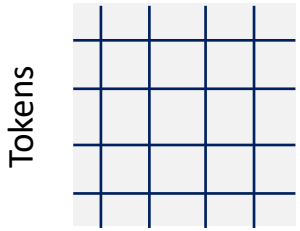
Bag of Words []

- 980 documentos
- 73,643 tokens

Proceso de Búsqueda

Índice Invertido

<Doc, TFIDF, TF, Length>



Estructura de datos más utilizada en los sistemas de búsqueda y recuperación de información

Query

Para **calificar** las búsquedas usamos la función *Okapi BM25 [3]*

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) * \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})}$$

$$k_1 = 1.2,\ b = 0.75$$

Existen varios tipos de modelos para **calificar** y **recuperar** información

Proceso de Validación

Valoración Experto

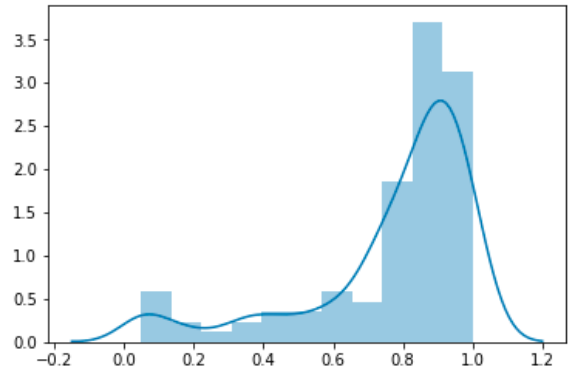
Para hacer las comparaciones seleccionamos a **MetaPy** como el experto [4]

Validación

Para realizar la validación analizamos la matriz de **confusión** [5]

Nuestro Buscador \ MetaPy	SI	NO
SI	*TP*	*FP*
No	*FN*	*TN*

- Precisión = $\frac{TP}{TP+FP} \approx 1$
- Recall = $\frac{TP}{TP+FN}$



Recall

Proceso de Validación

Valoración Experto

	Indri *IR*	Lucene *IR*	MALLET *ML/NLP*	LIBLINEAR *ML*	SVM^{MULT} *ML*	scikit *ML/NLP*	CoreNLP *ML/NLP*	MeTA *all*
Feature generation	✓	✓	✓			✓	✓	✓
Search	✓	✓						✓
Classification			✓	✓	✓	✓	✓	✓
Regression			✓	✓	✓	✓	✓	✓
POS tagging			✓				✓	✓
Parsing							✓	✓
Topic models			✓			✓		✓
n-gram LM								✓
Word embeddings			✓				✓	✓
Graph algorithms								✓
Multithreading		✓	✓			✓	✓	✓

Ejemplo

Palabras	Precision	Recall
biology	1	1
activity	1	0.8
machine learning	1	0.8
machine	1	1
math	1	0.7
magazine	1	0.9
mahalanobis distance	1	1
kruskal algorithm	1	0.8
mathematician	1	0.9
norm	1	0.8

P. Z. Stephen Robertson, The Probabilistic Relevance Framework:

Modelado de Tópicos

Meta Data

MetaData Paser: Desarrollada para navegar sobre el xml utilizando xmlToDict. Convierte XML a diccionario.

Categorización

- Latent Dirichlet Allocation (LDA)[6]
- Clasificar cada documento según su conjunto de palabras
- 10 categorías o Tópicos

topic	topic_desc
0	[map, data, point, curv, loss, biolog, cross, algorithm, surfac, cell]
1	[quantum, game, strategi, logic, classic, agent, mathcal, version, automata, lambda]
2	[power, delta, comput, size, parallel, total, independ, et, al, test]
3	[graph, tree, color, edg, time, algorithm, problem, np, planar, vertic]
4	[channel, bind, sourc, big, node, degre, relay, input, case, polynomi]
5	[matric, group, random, block, error, rule, complex, entropi, distribut, correl]
6	[protocol, receiv, attack, secur, messag, design, matrix, wireless, order, fix]
7	[social, estim, detect, data, learn, network, decis, softwar, commun, test]
8	[code, user, control, system, scheme, energi, video, inform, decod, interact]
9	[method, model, imag, optim, network, nois, search, sampl, cost, learn]

Referencias

1. A. Turing, “Mind a quarterly review of psychology and philosophy,” 1950.
2. Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (ICML '06). ACM, New York, NY, USA, 977-984. DOI: https://doi.org/10.1145/1143844.1143967
3. H. Z. Stephen Robertson, The Probabilistic Relevance Framework: BM25 and Beyond. Editorial Board, 2009.
4. Massung, S., Geigle, C., & Zhai, C. (2016). MeTA: A Unified Toolkit for Text Retrieval and Analysis, 91–96. https://doi.org/10.18653/v1/p16-4016
5. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
6. David M Blei, Andrew Y Ng, Michael I JordanJournal of machine Learning research 3 (Jan), 993-1022, 2003