Helena Montserrat Gómez Adorno

Investigadora Asociada C – SNI Nivel 1
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Teléfono: +52 5518903203
Dirección: Cda Nahuatlecas 91 1 y 2 Cad de Nahuatlaca,
Colonia Ajusco, C.P. 04300, Delegación Coyoacán, Ciudad de México, México.

helena.adorno@gmail.com, helena.gomez@iimas.unam.mx

Artículos de revista: 19, artículos de conferencia: 33
Citas en Google Scholar: 685, h-index: 16
Google Scholar, LinkedIn, Research Gate, Bitbucket, Github

Mis áreas de investigación son el procesamiento del lenguaje natural, la lingüística computacional y la recuperación de información; específicamente aplicaciones de, búsqueda de respuestas, similitud semántica, atribución de autoría y perfilado de autor. He desarrollado nuevas estructuras de representación de texto para facilitar estas tareas y nuevos métodos para calcular la similitud semántica entre textos. He propuesto una nueva medida de similitud, llamada similitud de coseno suave, que considera la información semántica de las características tales como palabras, n-gramas, etiquetas POS, etc.

Tengo siete años de experiencia trabajando como DBA de Oracle y desarrollador de software en varias empresas del sector público y privado en Paraguay. También tengo mas de tres años de experiencia docente en la Universidad Nacional de Asunción y en la Universidad Nacional Autónoma de México.

Trayectoria Academica

Doctorado, Ciencias de la computación, 2014 – 2018, **Promedio** 10/10 (Mención Honorífica)
Laboratorio de Lenguaje Natural y Procesamiento de Textos,
Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), México
Tesis: Extracción de características de textos basada en Grafos Sintácticos Integrados
Asesores: Dr. Grigori Sidorov, Dr. David Pinto

Maestría, Ciencias de la Computación, 2011 – 2013, **Promedio** 9.41/10 (Mención Honorífica)
Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla (BUAP), México
Tesis: Una metodología para el desarrollo de sistemas de búsqueda de respuestas basadas en pruebas de lectura comprensiva
Asesores: Dr. David Pinto, Dr. Darnes Vilariño

Licenciatura, Análisis de Sistemas Informáticos, 2001 – 2005, **Promedio** 4.10/5
Facultad Politécnica, Universidad Nacional de Asunción, Paraguay
Tesis: Sistema para distribución de Gas en Asunción para la compañía “La Oxigena S.A.”

Experiencia Laboral

Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas – UNAM (México), Agosto 2018-Actualidad.

> **Investigadora asociada C** - Departamento de Ingeniería de Sistemas Computacionales y Automatización.

Instituto de Ingeniería – UNAM (México), Febrero - Julio 2018.

> **Investigadora Post-Doctoral** - Grupo de Ingeniería Lingüística.
> Trabajé en varios proyectos, principalmente en un nuevo modelo para aprendizaje de vectores de palabras y un modelo para detección de lenguaje agresivo en Twitter.

IBM Deutschland Research and Development GmbH (Alemania), Julio - Septiembre 2017.

> **Estancia de lingüística Computacional– Proyecto Watson Analytics for Social Media**

Mejoré los módulos de análisis de texto existentes para español: análisis de sentimientos e identificación de datos demográficos.

Sallustro y cia. – Empresa de importación (Paraguay), 2008-2011

Programador y administrador de sistemas Oracle. Oracle Developer 6i y base de datos Oracle 10g.

Ministerio de Hacienda de Paraguay (*Secretaría de Tributación*), 2005-2007

Programador de sistemas Oracle. Oracle Developer 6i, Java y Visual Basic 6.0. Proyecto de re-ingeniería de sistemas de la Secretaría de Estado de Tributación: realicé la migración de la base de datos de contribuyentes (RUC) y desarrollé el módulo de contribuyentes del nuevo sistema Marangatú.

Grupo Inventiva – Consultora (Paraguay), 2004-2005

Programador de sistemas Oracle. Análisis de sistemas, administración de base de datos Oracle. Implementé y desarrollé sistemas de gestión de ventas para varias empresas.

Experiencia Docente

Universidad Nacional Autónoma de México (México), Agosto 2018 – Actualidad.

Posgrado en Ciencia e Ingeniería de la Computación: Minería de Textos y Minería de Grafos.

Universidad Nacional de Asunción (Paraguay), Agosto 2016, Mayo 2017

Profesora – Curso de posgrado de la "Maestría en tecnologías de la información y la comunicación" Curso de base de datos avanzadas: Base de datos espaciales con SQL server (20 horas).

Universidad Nacional de Asunción (Paraguay), 2009-2011

Profesora – Nivel Licenciatura. Curso de administración y programación de Base de datos.

Premios y Becas

- **Presea "Lázaro Cárdenas 2019"** por desempeño escolar sobresaliente como alumna de doctorado en el área de Ingeniería y Ciencias Fisico Matemáticas del Instituto Politécnico Nacional.
- **1st puesto** (de 6 equipos) en la competencia de agrupamiento de autores del PAN@CLEF 2017.
- **1st puesto** (de 22 sistemas) en la competencia de identificación de perfiles de autor en ruso del PAN@FIRE 2017.
- **Premio al Mejor Artículo** por el trabajo titulado "Author Profiling with doc2vec Neural Network-Based Document Embeddings" del 15th Congreso Internacional México de Inteligencia Artificial (**MICAI**) 2016.
- **Premio al mejor rendimiento académico** de los estudiantes de doctorado del Instituto Politécnico Nacional (IPN) 2016. Existen aproximadamente 50 estudiantes de doctorado en el IPN. El diploma se otorga a los mejores puntajes y publicaciones.
- **Beca de Investigación BEIFI**, IPN, 2014, 2015, 2016
- **Beca** de proyectos de la Red Temática en Tecnologías del Lenguaje (Proyectos CONACYT 260178 y 271622) para el desarrollo de recursos lingüísticos, 2015, 2016
- **Beca** del gobierno Mexicano para estudios de doctorado en programas de alta calidad (PNPC), 2014–2018
- **Beca** del gobierno paraguayo para graduación de maestría en el extranjero, 2013
- **Beca "Premio al Mérito"** para estudiantes internacionales de la Secretaría de Relaciones Exteriores (SRE) del gobierno mexicano para estudios de Maestría, 2011–2013. La SRE otorga dos becas por país por año.

Certificados y Diplomas

- **Applied Data Science with Python Specialization,** Febrero 2018, Coursera.
- **Applied Social Network Analysis in Python,** Febrero 2018, Coursera.
- **Applied Text Mining in Python**, Octubre 2017, Coursera.
- **Applied Machine Learning in Python**, Julio 2017, Coursera.

- **Text Analytics - Level 1**, Julio 2017, IBM.
- **Applied Plotting, Charting & Data Representation in Python**, Julio 2017, Coursera.
- **Introduction to Data Science in Python**, Diciembre 2016, Coursera**.**
- **Certificación de Administrador Linux**, 20 horas, **2014.** Open Intelligence, www.openintelligence.mx, México.
- **Diploma in Java**, 96 horas, **2006–2007.** Universidad Comuneros, Paraguay.

Direcciones de Tesis

Dirección de tesis de maestría terminadas:

- ***Perfilado de autor utilizando técnicas de transferencia de conocimiento*****,** alumno *Aquilino Sotelo* en la disciplina de *Inteligencia Artificial* en el *Posgrado en Ciencia e Ingeniería de la Computación – UNAM*.

Dirección de tesis de maestría en curso:

- ***Aprendizaje automático aplicado a la atribución de autoría*****,** alumno *Alan Emir Araujo Pino* en la disciplina de *Inteligencia Artificial* en el *Posgrado en Ciencia e Ingeniería de la Computación – UNAM*.

- ***Detección de bots en redes sociales usando técnicas de procesamiento de lenguaje natural***, alumno *Daniel Jacob Espinoza González* en la *Maestría en Ciencias de la Computación del Centro de Investigación en Computación – IPN*.

Dirección de tesis de doctorado en curso:

- ***Identificación de la Intención en la gestión de reputación en redes sociales, empleando tecnologías del lenguaje y aprendizaje automático***, alumno *Alex Iván Valencia Valencia* en el Doctorado *en Ciencia e Ingeniería de la Computación – UNAM*.
- **Cross-Genre Multi-label Emotion Analysis of Multi-lingual Text**, alumna *Iqra Ameer* en el Doctorado en *Ciencias de la Computación* del *Centro de Investigación en Computación – IPN.*

Publicaciones

Revistas JCR (11)

1. H. Gómez-Adorno, R. Fuentes-Alba, I. Markov, G. Sidorov, A. Gelbukh. A convolutional neural network approach for gender and language variety identification. **Journal of Intelligent & Fuzzy Systems 36 (5),** pp. 4845–4855, 2019.
   DOI: 10.3233/JIFS-179032
   JCR **impact factor** 2018: 1.426
2. G. Bel-Enguix, H. Gómez-Adorno, J. Reyes-Magaña, G. Sierra. Wan2Vec: Embeddings Learned on Word Association Norms, ***Semantic Web 2019***, pp. 1–16, 2019.
   DOI: 10.3233/SW-190349
   JCR **impact factor** 2018: 2.224
3. J. Reyes-Magaña, G. Bel-Enguix, H. Gómez-Adorno, G. Sierra. A Lexical Search Model based on word association norms, ***Journal of Intelligent & Fuzzy Systems 36(5)***, pp. 4587-459, 2019.
   DOI: 10.3233/JIFS-179010
   JCR **impact factor** 2018: 1.426
4. J.P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, J.J.M. Escobar. Detection of fake news in a new corpus for the Spanish language, ***Journal of Intelligent & Fuzzy Systems 36(5)***, pp. 4869-4876, 2019.
   DOI: 10.3233/JIFS-179034
   JCR **impact factor** 2018: 1.426

5. H. Gómez-Adorno, J.P. Posadas-Durán, G. Sidorov, D. Pinto. Document embeddings learned on various types of *n*-grams for cross-topic authorship attribution, ***Computing*** **2018**, pp. 1–13, 2018. DOI: 10.1007/s00607-018-0587-8

   JCR **impact factor** 2016: 1.589

6. M. A. Sanchez-Perez, A. Gelbukh, G. Sidorov, H. Gómez-Adorno. Plagiarism Detection with Genetic-Based Parameter Tuning, ***International Journal of Pattern Recognition and Artificial Intelligence*** 32(1), **2018**.
   DOI: 10.1142/S0218001418600066

   JCR **impact factor** 2016: 0.994

7. J.P. Posadas-Durán, G.Sidorov, H. Gómez-Adorno, I. Batyrshin, E. Mirasol-Mélendez, G. Posadas-Durán, L. Chanona-Hernández. Algorithm for Extraction of Subtrees of a Sentence Dependency Parse Tree, ***Acta Polytechnica Hungarica*** 14(3), pp. 79–98, **2017**.
   DOI: 10.12700/APH.14.3.2017.3.5

   JCR **impact factor** 2016: 0.745

8. J.P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, I. Batyrshin, D. Pinto, L. Chanona-Hernández. Application of the distributed document representation in the authorship attribution task for small corpora, ***Soft Computing*** 21(3), pp. 1–13, **2016**.
   DOI: 10.1007/s00500-016-2446-x

   JCR **impact factor** 2015: 1.630

9. H. Gómez-Adorno, I. Markov, G. Sidorov, J.P. Posadas-Durán, M.A. Sanchez-Perez, L. Chanona-Hernandez. Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts, ***Computational Intelligence and Neuroscience*** 2016, 13 pp., **2016**. DOI: 10.1155/2016/1638936

   JCR **impact factor** 2015: 0.430

10. H. Gómez-Adorno, G. Sidorov, D. Pinto, D. Vilariño, A. Gelbukh. Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. ***Sensors*** 16(9), **2016**. DOI: 10.3390/s16091374

    JCR **impact factor** 2015: 2.033

11. D. Pinto, H. Gómez-Adorno, D. Vilariño Ayala, V. Kumar Singh. A graph-based multi-level linguistic representation for document understanding. ***Pattern Recognition Letters*** 41, pp. 93–102, **2014**. DOI: 10.1016/j.patrec.2013.12.004

    JCR **impact factor** 2015: 1.586

Revistas en otros índices (8)

1. F. Viveros-Jiménez, MA. Sánchez-Perez, H. Gómez-Adorno, JP. Posadas-Durán, G. Sidorov, A. Gelbukh. Improving the Boilerpipe Algorithm for Boilerplate Removal in News Articles Using HTML Tree Structure. ***Computación y Sistemas*** 22(2), pp. 483–489
   DOI: 10.13053/CyS-22-2-2959

2. H. Gómez-Adorno, G. Rios, J. P. Posadas-Durán, G. Sidorov, G. Sierra. Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. ***Computación y Sistemas*** 22(1), pp. 47–53, 2018
   DOI: 10.13053/CyS-22-1-2882

3. H. Gómez-Adorno, I. Markov, G. Sidorov, J. P. Posadas-Duran, C. Fócil-Arias. Compilación de un lexicón de redes sociales para la identificación de perfiles de autor (Compiling a Lexicon of Social Media for the Author Profiling task). ***Research in Computing Science*** 115, pp. 19–27, 2016.
   PDF

4. G. Sidorov, A. Gelbukh, H. Gómez-Adorno, D. Pinto. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. ***Computación y Sistemas*** 18(3), pp. 491–504, 2014. DOI: 10.13053/CyS-18-3-2043
5. H. Gómez-Adorno, G. Sidorov, D. Pinto, D. Vilariño. Automatic Linguistic Pattern Identification Based on Graph Text Representation. ***Research in Computing Science*** 71, pp. 43–52, 2014.
   PDF
6. H. Reyes Cervantes, N. Loya, Y. Alemán, H. Gómez**.** Predicción de la calidad del aire de la Ciudad de México basado en minería de datos, con soporte para la toma de decisiones (in Spanish). ***Research in Computing Science*** 57, 2012.
7. J. Somodevilla, Y. Alemán, N. Loya, H. Gómez. Almacén de datos espacial para el análisis de desastres naturales en el continente americano (in Spanish). ***Research in Computing Science*** 57, 2012.
8. G. De Ita, H. Gómez, B. Merino. Algorithm to Count the Number of Signed Paths in an Electrical Network via Boolean Formulas. ***Acta Universitaria*** 22(2012), pp. 69-74, 2012
   PDF

Conferencias (33)

1. J. Reyes-Magaña, G. Bel-Enguix, G. Sierra, H. Gómez-Adorno. Designing an Electronic Reverse Dictionary Based on Two Word Association Norms of English Language. *Electronic lexicography in the 21st century (**eLex 2019**): Smart lexicography*, 2019.
2. C. Montaño, G. Sierra, G. Bel-Enguix, H. Gomez-Adorno. A Parallel Corpus Mixtec-Spanish. In *Proceedings of the 2019 Workshop on Widening NLP (**WiNLP**)*, 2019.
3. D. Y. Espinosa, H. Gómez-Adorno, G. Sidorov. Bots and Gender Profiling using Character Bigrams, *Notebook for PAN at **CLEF** 2019*.
4. G. Ortiz, H. Gómez-Adorno, J. Reyes-Magaña, G. Bel-Enguix, G. Sierra. Detection of aggressive tweets in Mexican Spanish using multiple features with parameter optimization. *In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (**IberLEF**), 2019.*
5. L.E.A. Vega, J.C. Reyes-Magaña, H. Gómez-Adorno, G. Bel-Enguix. MineriaUNAM at SemEval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework. Proceedings of the 13th International Workshop on Semantic Evaluation (**SEMEVAL**), 2019.
6. H. Gómez-Adorno, J. Reyes-Magaña, G. Bel-Enguix, G. Sierra. Spanish Word Embeddings Learned on Word Association Norms. Proceedings of the 13th Alberto Mendelzon International Workshop on Foundations of Data Management (**AMW**), 2019.
7. H. Gómez-Adorno, C. M. del-Campo-Rodríguez, G. Sidorov, Y. Alemán, D. Vilariño, D. Pinto. Hierarchical Clustering Analysis: The Best-Performing Approach at PAN 2017 Author Clustering Task. International Conference of the Cross-Language Evaluation Forum for European Languages (**CLEF**), 2018.
8. I. Markov, H. Gómez-Adorno, M. Jasso-Rosales, G. Sidorov. Cic-gil approach to author profiling in Spanish tweets: Location and occupation. Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (**IberEval** 2018).
9. H. Gómez-Adorno, G. Bel-Enguix, G. Sierra, O. Sánchez, D. Quezada. A machine learning approach for detecting aggressive tweets in Spanish. Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (**IberEval** 2018).
10. CM. del Campo-Rodríguez, H. Gómez-Adorno, G. Sidorov, I. Batyrshin. CIC-GIL Approach to Cross-domain Authorship Attribution. *Notebook for PAN at **CLEF** 2018*.
11. M. A. Sanchez-Perez, I. Markov, H. Gómez-Adorno, G. Sidorov. Comparison of Character N-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus. International Conference of the Cross-Language Evaluation Forum for European Languages (***CLEF***), pp. 145-151, 2017.
12. H. Gómez-Adorno, Y. Aleman, D. Vilariño, M.A. Sanchez-Perez, D. Pinto, G. Sidorov. Author Clustering using Hierarchical Clustering Analysis. *Notebook for PAN at **CLEF** 2017*.

13. I. Markov, H. Gómez-Adorno, G. Sidorov. Language-and subtask-dependent feature selection and classifier parameter tuning for author profiling. *Notebook for PAN at **CLEF** 2017*.
14. I. Markov, H. Gómez-Adorno, G. Sidorov, A. Gelbukh. The Winning Approach to Cross-Genre Gender Identification in Russian at RUSProfiling 2017. *Notebook Papers of **FIRE** 2017*.
15. H. Gómez-Adorno, I. Markov, J. Baptista, G. Sidorov, D. Pinto. Discriminating between similar languages using a combination of typed and untyped character n-grams and words. **VarDial** 2017, 2017.
16. I. Markov, H. Gómez-Adorno, J.P. Posadas-Durán, G. Sidorov, and A. Gelbukh. Author Profiling with doc2vec Neural Network-Based Document Embeddings. ***MICAI** 2016*. (**best paper award**)
17. I. Markov, H. Gómez-Adorno, G. Sidorov, A. Gelbukh. Adapting cross-genre author profiling to language and corpus. *Notebook for PAN at **CLEF** 2016*, pp. 947–955.
18. H. Gómez-Adorno, G. Sidorov, D. Vilariño, D. Pinto. CICBUAPnlp at SemEval-2016 Task 4-A: Discovering Twitter Polarity using Enhanced Embeddings. ***SemEval** 2016*, pp. 145–148.
19. J.P. Posadas-Durán, H. Gómez-Adorno, I. Markov, G. Sidorov, I. Batyrshin, A. Gelbukh, O. Pichardo-Lagunas. Syntactic n-grams as features for the author profiling task: Notebook for PAN at **CLEF** 2015.
20. H. Gómez-Adorno, G. Sidorov, D. Pinto, I. Markov. A Graph Based Authorship Identification Approach: Notebook for PAN at **CLEF** 2015.
21. G. Sidorov, H. Gómez-Adorno, I. Markov, D. Pinto, N. Loya. Computing Text Similarity using Tree Edit Distance. *Annual Conf. of North American Fuzzy Information Processing Society (**NAFIPS** 2015)* at 5th World Conference on Soft Computing (**WConSC**), pp. 1–4, 2015.
22. H. Gómez-Adorno, D. Vilariño, D. Pinto, G. Sidorov. CICBUAPnlp: Graph-Based Approach for Answer Selection in Community Question Answering Task. ***SemEval** 2015*, pp. 18–22.
23. H. Gómez-Adorno, D. Pinto, M. Montes, G. Sidorov, R. Alfaro. Content and style features for automatic detection of users' intentions in tweets. ***IBERAMIA** 2014*, *Ibero-American Conference on Artificial Intelligence*, pp. 120–128.
24. H. Gómez-Adorno, G. Sidorov, D. Pinto, A. Gelbukh. Graph-based approach to the question answering task based on entrance exams. *Working Notes of **CLEF** 2014*, pp. 1395–1403.
25. H. Gómez-Adorno, D. Pinto, D. Vilariño. Two Approaches for QA4MRE: Information Retrieval and Graph-based knowledge representation. *Working Notes of **CLEF** 2013***.**
26. D. Vilariño, D. Pinto, H. Gómez-Adorno, Saúl León, Esteban Castillo. Lexical-Syntactic and Graph-Based Features for Authorship Verification, *Notebook for PAN at* **CLEF** *2013***.**
27. H. Gómez-Adorno, D. Pinto, D. Vilariño. A Question Answering System for Reading Comprehension Tests. ***MCPR** 2013*, *Mexican Conference on Pattern Recognition*, pp. 354–363.
28. H. Gómez-Adorno, D. Pinto, D. Vilariño. Semantic Answer Validation in Question Answering Systems for Reading Comprehension Tests. ***AMW** 2013*, *Alberto Mendelzon International Workshop on Foundations of Data Management*, pp. 1–6**.**
29. D. Vilariño, D. Pinto, S. León, Y. Aleman, H. Gómez. BUAP: N-gram based Feature Evaluation for the Cross-Lingual Textual Entailment Task. ***SemEval** 2013 at Joint Conference on Lexical and Computational Semantics (***SEM**)*, pp. 124–127.
30. Nahun Loya, Ivan Olmos, H. Gómez, Y. Alemán, D. Pinto. Forecast of air quality based in Ozone by decision trees and neural networks. ***MICAI** 2012*, pp. 97–106.
31. D. Pinto, D. Vilariño Ayala, H. Jiménez-Salazar, H. Gómez, Y. Alemán, N. Loya. The Soundex Phonetic Algorithm Revisited for SMS Text Representation. **TSD** *2012*, pp. 48–55**.**

32. D. Pinto, D. Vilariño Ayala, H. Gómez, Y. Alemán, Nahun Loya. The Soundex Phonetic Algorithm Revisited for SMS-based Information Retrieval. ***CERI*** *2012, Spanish Conference on Information Retrieval***.**

33. H. Gómez, C. Perez de Celis, N. Loya, Y. Alemán. Modelo de Recuperación utilizando relaciones semánticas entre los objetos de una colección heterogénea y multiclasificada por conceptos (In Spanish). ***CONACI*** 2012.

Proyectos

Proyecto CONACYT 298036: Feria de talleres "Mexicanas del Futuro: Trazando conciencias, pensando en TI. Edición IIMAS UNAM" – Responsable Técnico.

El objetivo del proyecto fue fomentar la vocación científico-tecnológica entre las mujeres estudiantes de nivel medio-superior de la UNAM. Para ello invitamos a las chicas a estudiar alguna carrera de ciencia y tecnología a través de conferencias y talleres en un ambiente ameno e interesante para demostrarles que pensar que las mujeres y la tecnología no se llevan es un prejuicio sin fundamento. Realizamos una feria de talleres en el IIMAS y charlas de divulgación y orientación, impartidas por mujeres profesionales del mundo de la academia y la industria, centradas en la población estudiantil de la UNAM.

Duración: 4 meses desde Agosto hasta Noviembre de 2019.

Financiamiento: 245,000 pesos mexicanos.

Proyecto PAPIIT TA100520: Análisis de autoría de documentos con técnicas de aprendizaje profundo – Responsable Técnico.

El objetivo del proyecto es Desarrollar métodos que permitan extraer características relevantes de un documento para el análisis de autoría, utilizando arquitecturas neuronales profundas que permitan obtener propiedades léxicas, sintácticas y semánticas de los textos.

Duración: 2 años desde Enero 2020 hasta Diciembre 2021.

Financiamiento: 1er. año: 170,790 pesos mexicanos.

Proyecto SECTEI/284/2019: "Diseño de un programa de estudios para la capacitación en programación y habilidades en tecnologías de información y comunicación para la escuela de código dentro de PILARES de la Ciudad de México"- Participante.

Participante en el proyecto como co-responsable del diseño del modulo de "Base de Datos" para la escuela de código.

Duración: 1 año desde Diciembre 2019 hasta Noviembre 2020.

Financiamiento: 1,000,000 pesos mexicanos.

Corpus multi-etiquetado para estudios de autoría en español (2016)

El objetivo del proyecto es la recopilación de noticias en español de sitios de medios digitales y su categorización en tres áreas: variación del español, auto y, género del autor. La recolección se realizó de forma semiautomática con un *crawler* web desarrollado para este fin.

Financiamiento: 30,000 Mxn. Red Temática de Tecnologías del Lenguaje

Descripción del recurso: https://link.springer.com/chapter/10.1007/978-3-319-65813-1_15

Recurso léxico para el procesamiento de datos de redes sociales (2015)

El objetivo del proyecto es la colección de diccionarios de palabras de argot, contracciones, abreviaturas y emoticones comúnmente utilizados en las redes sociales. Cada uno de los diccionarios está diseñado para los idiomas inglés, español, holandés e italiano.

Financiamiento: 30,000 Mxn. Red Temática de Tecnologías del Lenguaje

Descripción del recurso: https://www.hindawi.com/journals/cin/2016/1638936/

Participaciones en Competencias

Author Clustering at PAN - CLEF 2017: 1er. lugar.

El sistema desarrollado agrupa documentos escritos por un mismo autor.

Código fuente disponible: https://github.com/helenpy/clusterPAN2017

Descripción del sistema: http://ceur-ws.org/Vol-1866/paper_108.pdf

RUSProfiling at PAN - FIRE 2017: 1er lugar.

Desarrollo de un modulo del sistema para la competencia de identificación de genero de autores de textos de redes sociales en idioma Ruso basado en aprendizaje automático (*Machine Learning*)
Descripción del sistema: http://ceur-ws.org/Vol-2036/T1-5.pdf

Author Profiling at PAN - CLEF 2016, 2017: 5to lugar.

Desarrollo completo del sistema para la competencia de identificación de edad, genero y lenguaje nativo de autores de textos de redes sociales basado en aprendizaje automático (*Machine Learning*)
Descripción del sistema: http://ceur-ws.org/Vol-1609/16090947.pdf

Sentiment Analysis in Twitter Task 4-A at SemEval 2016.

El sistema desarrollado clasifica mensajes de Twitter de acuerdo a su carga emocional, positivo, negativo o neutro. Para la clasificación, entrenamos un clasificador de Maquinas de Soporte Vectorial (SVM). Desarrollé el sistema en Python.
Código fuente disponible: https://bitbucket.org/helenpy/twittersentimentanalysis/overview

Authorship Identification at PAN - CLEF 2015: 10mo lugar.

El sistema desarrollado determina si un documento ha sido escrito por un autor o no. Calculamos la similitud entre documentos conocidos y desconocidos utilizando características extraídas de Grafos Sintácticos Integrados (GSI).
Descripción del sistema: http://ceur-ws.org/vol-1391/135-CR.pdf

Author Profiling at PAN - CLEF 2015: sin ranking.

El sistema desarrollado utiliza n-gramas sintácticos de varios tipos para predecir género, edad y rasgos de personalidad del autor de un texto dado.
Descripción del sistema: http://ceur-ws.org/vol-1391/136-CR.pdf

Automatic Machine Reading at QA4MRE - CLEF 2013 and 2014: 2do lugar.

El sistema desarrollado lee un documento dado y responde preguntas de opción múltiple al respecto.
Descripción del sistema: http://ceur-ws.org/vol-1180/CLEF2014wn-QA-GomezAdornoEt2014.pdf

Habilidades Técnicas

- **Lenguajes de Programación**: Python, AWK, PL/SQL, SQL (lenguajes principales), JAVA (experiencia básica), C/C++ (cursos de licenciatura), Power Builder 9 (tesis de licenciatura)
- **Procesamiento de Lenguaje Natural**: Modelos de espacio vectorial, ingeniería de características, WordNet; Herramientas: FreeLing, NLTK, Stanford parser, Stanford POS tagger, GATE NLP, Gensim.
- **Machine Learning**: Scikit-learn, Scipy, Gensim, Weka, Word2vec, Doc2vec.
- **Base de Datos**: Oracle, SQL Server, MySql, PostgreSQL, SyBase.
- **Sistemas Operativos**: Windows, Linux.
- **Lingüística**: Descripción del comportamiento semántico-sintáctico de las palabras, análisis estadístico de corpus, gramática y formalismos semánticos, morfología.

Lenguajes

Inglés (fluido), Español (nativo), Guarani (Segunda lengua oficial del Paraguay; lectura)