PROGRAMA DE ASIGNATURA

1. NOMBRE DE LA ASIGNATURA

Análisis Computacional de Datos Lingüísticos

2. NOMBRE DE LA ASIGNATURA EN INGLÉS

Computational Analysis of Linguistic Data

3. TIPO DE CRÉDITOS DE LA ASIGNATURA

SCT/ UD/ OTROS/

4. NÚMERO DE CRÉDITOS

10 créditos

5. HORAS DE TRABAJO PRESENCIAL DEL CURSO

3 horas

6. HORAS DE TRABAJO NO PRESENCIAL DEL CURSO

6 horas

7. OBJETIVO GENERAL DE LA ASIGNATURA

Revisar junto con los estudiantes técnicas básicas de programación en Python, enfocadas al análisis de datos lingüísticos.

8. OBJETIVOS ESPECÍFICOS DE LA ASIGNATURA

1. Manejar los fundamentos de programación en Python.

2. Aplicar nociones básicas de programación en problemas de análisis de datos lingüísticos.

3. Resolver problemas computacionales relacionados a lenguas de Sudamérica.

9. SABERES / CONTENIDOS

1. Introducción a la programación para lingüistas: ¿Qué problemas son abordables con un computador? ¿Qué nuevas perspectivas se abren?
1.1. Disciplinas relacionadas: Lingüística Computacional, Procesamiento del Lenguaje Natural y Análisis de Datos.
1.2. Tipos de datos lingüísticos.
1.3. Bases de datos lingüísticas: SAILS, WALS, corpus paralelos, datos sobre obsolescencia. Datos y lenguas de Sudamérica.

2. Nociones de Python
2.1. Jupyter Notebooks y plataformas para Python.
2.2. Elementos básicos: variables, listas, diccionarios, funciones.
2.3. Estructuras de control: if, else, for. Nociones básicas de lógica para computación.
2.4. Manejo de datos.

3. Manipulación de bases de datos lingüísticas.
3.1. Herramientas básicas de análisis: estadística descriptiva, distancia de Hamming, frecuencia de palabras, entropía.
3.2. Fundamentos de Análisis Automático de Datos.
3.3. Relaciones entre problemas lingüísticos y análisis computacionales.

4. Visualización de información lingüística.
4.1. Construcción de mapas.
4.2. Gráficos y presentación de resultados.

10. METODOLOGÍA

El seminario se estructura en torno a clases en formato taller, en que los conceptos
lingüísticos son estudiados mediante Python. Las clases están orientadas a la práctica guiada de implementación computacional de problemas lingüísticos. Además, el seminario supone el desarrollo de un proyecto de análisis de datos lingüísticos de lenguas de Sudamérica.

11. METODOLOGÍAS DE EVALUACIÓN

- Pre-entregas trabajo final (50%)
- Trabajo final (50%)

12. REQUISITOS DE APROBACIÓN

ASISTENCIA *(indique %)*: La reglamentaria

NOTA DE APROBACIÓN MÍNIMA *(Escala de 1.0 a 7.0)*: 4.0

REQUISITOS PARA PRESENTACIÓN A EXAMEN: No se rinde examen.

13. PALABRAS CLAVE

Programación en Python, Datos Lingüísticos, Lingüística Computacional, Lenguas de Sudamérica.

14. BIBLIOGRAFÍA OBLIGATORIA

Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly Media, Inc., 1st edition, 2009.

Michael Hammond. Python for Linguists. Cambridge University Press, 2020.

G.K. Zipf. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press, 1949.

Muysken, Pieter, Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn & George Saad. 2016. South American Indigenous Language Structures (SAILS) Online. Jena: Max Planck Institute for the Science of Human History. (Available at https://sails.clld.org)

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2021-06-09.)

15. BIBLIOGRAFÍA COMPLEMENTARIA

Bentz, C., Alikaniotis, D., Cysouw, M., Ferrer-i-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy* 19, no. 6: 275. https://doi.org/10.3390/e19060275

Christodouloupoulos, C., Steedman, M. (2015). A massively parallel corpus: the Bible in 100 languages. *Lang Resources & Evaluation* **49,** 375–395. https://doi.org/10.1007/s10579-014-9287-y

Bentz, C., Ruzsics, T., Koplenig, A. and Samardžić, T. (2016). A comparison between morphological complexity measures: typological data vs. language corpora. In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japan.

Corral, A; Boleda, G; Ferrer-i-Cancho, R. (2015). Zipf's law for word frequencies: word forms versus lemmas in long texts. *PLoS ONE*, 10 , pp. e0129031.

16. RECURSOS WEB

www.u-cursos.cl

https://github.com/christos-c/bible-corpus

http://www.nltk.org/

https://sails.clld.org/

https://wals.info/

https://glottolog.org/

https://www.unicode.org/udhr/index.html

Profesor Responsable:

Javier Vera Zúñiga, 15.331.351-2,
javier.vera@pucv.cl