Trabajo Final

Seminario
Algoritmos supervisados y
convenciones de anotación para tareas
de procesamiento del lenguaje natural

Departamento de Letras
FFyL - UBA
Junio de 2025

Tal como figura en el programa, para aprobar el seminario, quienes hayan regularizado deben presentar un trabajo final. Este trabajo final debe abocarse a una (y solo una) de las siguientes opciones:

- **Opción 1**: Una monografía de entre 8 y 12 paginas (entre 4000 y 6000 palabras).
- **Opción 2**: Un estado del arte breve de entre 6 y 10 páginas (entre 3000 y 5000 palabras) de modelos entrenados mediante algoritmos supervisados para algún tipo particular de tarea de clasificación.
- **Opción 3**: Un experimento completo de anotar datos lingüísticos para una tarea específica y entrenar con esos datos un algoritmo supervisado. Esta opción es obligatoriamente en grupo de hasta cuatro personas.
- **Opción 4**: La anotación de una noticia completa que siga los lineamientos del corpus de Ancora.

La fecha límite para la entrega de trabajos es la que prevea la reglamentación de la facultad. De todos modos, en la medida de lo posible, se agradecerá que no demoren la entrega. No serán recibidos trabajos una vez que venza el plazo que la facultad disponga. Sientanse libres de contactarnos ante cualquier duda a nuestros mails:

fernandocarranza86@gmail.com, schiaffinofernando@gmail.com

Debajo especificamos más detalles de la modalidad de cada una de las opciones.

1 Opción 1: Una monografía

Puede optarse por escribir una monografía en la que se defienda una postura original acerca de algún problema que involucre alguno de los temas vistos durante la cursada. La monografía debe tener entre ocho y doce páginas (entre 4000 y 6000 palabras). A modo ilustrativo, estos pueden ser algunos de los posibles temas para un trabajo monográfico acorde a los temas vistos en el seminario, si bien esta lista es, desde ya, no exhaustiva:

1. Las convenciones de análisis sintáctico de Conll-U.

2. La interpretación semántica de las gramáticas de dependencias.

3. Rol de los distintos tipos de tokenizadores (Byte-Pair Encoding, Morfessor, Finite State Transducer) y su rendimiento en distintas lenguas.

4. Convenciones de anotación para clases de palabra.

5. Utilidad o no de los algoritmos supervisados en la era de los grandes modelos de lenguaje.

Recomendamos que, en cualquier caso, nos consulten por el tema elegido antes de comenzar la escritura.

2 Opción 2: Estado del arte

El estado del arte tiene que revisar alguna tarea particular de clasificación que se haya hecho con algoritmos supervisados (a modo de ejemplo, etiquetamiento automático de clase de palabra, detección de entidades de determinado tipo, lematización, análisis morfológico, análisis sintáctico, detección de enfermedades en textos médicos, detección de empresas en el boletín oficial, etc.). Se recomienda revisar los materiales de alguna hackatón.

3 Opción 3: Anotación y entrenamiento

La tercera opción para el trabajo es anotar y entrenar un algoritmo para alguna tarea específica de clasificación (ejemplo, para análisis de sentimiento, para extracción de entidades no contempladas en Spacy, etc.). Este trabajo tiene que ser grupal, con un mínimo de tres y un máximo de cuatro miembros por grupo. La tarea debe consistir en anotar primeramente los resultados. No es posible, para estar tarea, usar datos anotados por terceros. Sí pueden usarse las convenciones de otro dataset existente, en caso de que quieran después probar contribuir con sus datos a ese dataset. Una vez realizada la anotación, debe usarse al menos un algoritmo (e.g., regresión logística, bayesiano ingenuo, máquina de soporte vectorial) para entrenar un modelo y evaluar su rendimiento. No es necesario que el algoritmo alcance un valor comparable al estado del arte. Al momento de entregar el trabajo, tienen que entregar los datos y el código, acompañado de un breve texto (que puede estar integrado con el código en un google colab) en el que expliquen el trabajo realizado, las convenciones de anotación utilizadas, el objetivo perseguido y los resultados alcanzados.

4 Opción 4: Anotación de una noticia

Anotar una noticia completa en español usando las convenciones de CONLL-U tal como las toma el corpus de Ancora. Esto tiene que estar anotado en un documento con el siguiente nombre:

- es_APELLIDO_ud_train.conllu

En donde dice "APELLIDO", completar con el apellido todo en minúsculas de quien hace la anotación. La extensión tiene que ser conllu (tal como en ocurre con el corpus de ancora `https://github.com/UniversalDependencies/UD_Spanish-AnCora`. Cada dato debe estar separado por un tab, tal como sucede en el corpus de Ancora. Se recomienda crear el archivo con libre office y seleccionar "tabulación" como separador de columna. Se recomienda que sea una noticia breve.