Práctica 01 UD03 PLN.

Modelos de Inteligencia Artificial
Álvaro Martínez Lineros.

APARTADO 01. Analizador de Sintaxis.

Corpus Español.

Existen varios organismos que cuentan con corpus del español. Destacando:

- CREA, CORPES, CDH y CORDE: Creado por la RAE, es posible extraer información para analizar las palabras y su significado para entrenar un modelo de procesamiento del lenguaje natural. Incluye tanto textos modernos (1974-2004), como históricos desde el inicio del idioma y un diccionario histórico de la lengua española.
- Corpus del Español: Incluye palabras y textos de diversos puntos históricos. Pudiendo analizar palabras del s.XII al s.XXI. Cuenta con miles de millones de palabras, tanto extraídas de la web como de libros históricos y artículos académicos.

Corpus Inglés.

Los principales corpus en inglés son:

- Gutenberg: Colección de más de 75 mil ebooks, una gran parte en inglés. Una desventaja es que no incluye anotaciones semánticas y sintácticas.
- BNC (British National Corpus): Es una colección de 100 millones de palabras tanto escritas como habladas. Principalmente usado para el análisis de gramática y lingüística computacional.
- Penn Treebank: Es un corpus lingüístico con la peculiaridad de que cada palabra y frase ha sido parseada (anotada con su estructura sintáctica).

Herramientas de análisis de texto y procesamiento del lenguaje natural.

- Preprocesamiento del texto: regex y beautifulsoup.
- Separación del texto en palabras: spacy, NLTK y Stanza.
- Análisis sintáctico y semántico: Stanford CoreNLP y Benepar.
- Análisis de sentimientos: Google Cloud Natural language API e IBM Watson Tone Analyzer.
- Extracción de entidades: spacy NER y Polyglot. Para representación contextual BERT y GPT.