



ANÁLISIS DE TEXTO AUTOMATIZADO EN CIENCIAS POLÍTICAS Y SOCIALES

David Losada Carril (Universidad de Santiago de Compostela, Spain)
30th September – 2th October

Teacher e-mail:
david.losada@usc.es

Language: English
Schedule: 15:00-19:00

Computer requirement

For the practical part of the course, the participants are encouraged to bring a laptop with access to Internet.

Class Schedule

Module 1: Introduction to Python for Data Manipulation and Text Processing

- **Text Data Manipulation Fundamentals:**
 - Working with native Python data structures for text (strings, lists, dictionaries).
 - Introduction to the Python data science ecosystem: loading, inspecting, and cleaning datasets using *Pandas*.
- **Essential Text Preprocessing:**
 - Tokenization techniques and sentence segmentation.
 - Text normalization: case folding, punctuation removal, and stop-word handling.
 - Morphological reduction: Stemming vs. Lemmatization using *NLTK* and *spaCy*.
 - Advanced usage of Regular Expressions (RegEx) for unstructured text cleaning and pattern matching.

Module 2: Introduction to Machine Learning for Text Analysis

- **Text Representation (Feature Engineering):**
 - Vector Space Models: Bag-of-Words (BoW) and term frequency counts.
 - Statistical weighting: Normalization and document modeling with TF-IDF (Term Frequency - Inverse Document Frequency).
 - Intuitive introduction to dense word embeddings vs. sparse vector representations.
- **NLP Machine Learning Workflow:**
 - Data splitting strategies (train, validation, test) tailored for imbalanced text corpora.
 - Evaluation metrics specific to NLP: Precision, Recall, F1-Score, and Confusion Matrices.

Module 3: Automated Text Classification

- **Classic Supervised Algorithms:**
 - Naive Bayes for rapid filtering and baseline categorization.

- Support Vector Machines (SVM) and Logistic Regression optimized for high-dimensional text vectors.
- **Classifier Optimization:**
 - Dimensionality reduction strategies applied to sparse text data.
 - Hyperparameter tuning using cross-validation

Module 4: Thematic Classification

- **Multi-class and Multi-label Categorization:**
 - One-vs-Rest (OvR) and One-vs-One (OvO) strategies for handling multiple mutually exclusive categories.
 - Multi-label classification: simultaneously assigning multiple topics or tags to a single document.
- **Practical Case Studies:**
 - Automated news classification by sections (e.g., Politics, Sports, Economy).
 - Automated indexing of scientific literature using predefined taxonomies or ontologies.

Module 5: Sentiment Analysis and Opinion Mining

- **Methodological Approaches:**
 - Lexicon and rule-based methods vs. supervised machine learning approaches.
 - Quantitative polarity analysis (positive, negative, neutral) and subjectivity detection.
- **Advanced Opinion Mining:**
 - Aspect-Based Sentiment Analysis: identifying specific targets and features within reviews.
 - Linguistic challenges: detecting and handling negation, irony, and sarcasm.

Module 6: Automated Text Summarization

- **Extractive Summarization:**
 - Graph-based and centrality algorithms (TextRank, LexRank).
 - Sentence selection and ranking using relevance and redundancy metrics.
- **Abstractive Summarization and Evaluation:**
 - Introduction to Sequence-to-Sequence (Seq2Seq) neural models for abstractive text generation.
 - Automated quality evaluation: ROUGE metrics (ROUGE-N, ROUGE-L) and their limitations.

Module 7: Topic Modeling in Document Corpora

- **Mathematical and Probabilistic Modeling:**
 - Latent Dirichlet Allocation (LDA): core concepts, latent variables, and Dirichlet distributions.
 - Non-Negative Matrix Factorization (NMF) applied to unsupervised theme discovery.
- **Interpretation and Visualization:**
 - Evaluating topic models using coherence scores and perplexity metrics.
 - Interactive topic map visualization using tools like *pyLDavis*.

Module 8: Other Machine Learning Applications for Text

- **Content-Based Recommender Systems:**
 - Measuring textual similarity (Cosine Similarity) to recommend related articles, books, or documents.
- **Content Moderation and Trust & Safety:**
 - Automated detection of offensive language, hate speech, and cyberbullying.
 - Spam filtering and fake profile identification based on textual patterns.
- **Text Attribute Identification:**
 - Automated language identification.
 - Basic introduction to Stylometry and authorship attribution.

Module 9: Exploiting LLMs for NLP Tasks (Prompt Engineering)

- **Foundations of LLM Interaction:**
 - Basic architecture of Large Language Models and interfacing with inference APIs.
 - The role of generation hyperparameters (temperature, Top-P, frequency/presence penalties).
- **Prompt Engineering Techniques:**
 - Instruction design patterns: Zero-shot, Few-shot learning, and role-play prompting (Persona Prompting).
 - Advanced reasoning techniques: Chain-of-Thought (CoT) and task decomposition.
- **Structured Outputs and Automation:**
 - Techniques for enforcing structured programmatic formats (JSON, CSV) from LLM outputs.
 - Case studies: replicating traditional NLP tasks (classification, summarization, entity extraction) via guided prompting.