UNIVERSITÉ CÔTE D'AZUR

UNIVERSITÉ CÔTE D'AZUR
CENTER OF MODELING
SIMULATION AND
INTERACTIONS

UNIVERSITÉ CÔTE D'AZUR
MEDICAL DATA
LABORATORY

ARGUMENTATIVE ANALYSIS OF CLINICAL TRIALS FOR EVIDENCE-BASED MEDICINE

marco milanesio
AI4HEALTH - 7-8/1/2021

Inria
inventeurs du monde numérique

Epione
e-patient / e-medicine

Outline

- NLP
- Argument Mining
- Classification
- Hands-on

NLP

- Natural Language Processing
- Starts with:
  - Tokenise
  - Normalise
  - Remove stop words
  - Count vectorise
  - Tf-idf representation

Bag-of-Words

- Tokenise
  - Split text into tokens
- Normalise
  - Transform a token into its root form
  - e.g., "sharing", "shared" → "share"
- Remove stop words
  - "and", "or", "but", "while", …
- Count vectorise
  - Tokens to matrix of counts

TF-IDF

- Term frequency Inverse Document Frequency
- More importance to less frequent words
- Idf is the "weight"
- BoW "improved"

A word on normalisation

- Stemming refers to a heuristic process that chops off the ends of words
- Lemmatization refers to doing things properly with the use of a vocabulary
- [Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze: Introduction to Information Retrieval, 2008]
- Stemming uses predefined rules to transform the word into a stem whereas lemmatisation uses context and lexical library to derive lemma. Stem doesn’t always have to be a valid word whereas lemma will always be a valid word because lemma is a dictionary form of a word.

Argument Mining

- Extracting natural language arguments and their relations from text, with the final goal of providing machine-processable structured data for computational models of argument

- Classic pipeline:
  - **Identification of arguments**
  - Definition of boundaries
  - Definition of intra-argument relations

Argument Mining

- Identification of arguments
- Classification task
  - Argumentative vs non-argumentative
  - Premise vs Claim
  - Claim vs Major Claim

Hands-on

- https://github.com/marcomilanesio/ai4health.git

- Python3 (3.7+)
  - jupyter
  - colab

- If jupyter (use pip):
  - nltk
    - nltk.download("book")
  - glove
    - wget http://nlp.stanford.edu/data/glove.6B.zip
  - sklearn
  - keras
  - tensorflow

Hands-on

- Dataset:
  - https://gitlab.com/tomaye/abstrct.git

- What we will do:
  - Setup/Installation
  - Preprocessing
  - Classification Argumentative vs Non Argumentative
  - Classification Premise vs Claim (with Glove)
  - PubMed

- Difficulty: entry-level (sometimes intermediate)

Live Coding

-