Machine Learning and Content Analytics

Argumentation Mining Project

Argumentation Mining

- ## What is argumentation mining?

  – The process by which the machine learns to recognize argumentative statements

  – Interpreted by the machine with an acceptable degree of accuracy

  – Applied to several sectors (e.g., NLP, social and philosophical science)

Project – Business Goals

- ## Our project

  - Scientific articles related to medical science sector

  - Annotation process to form datasets

  - Classification via ML techniques

  - Clustering via ML techniques

- ## Business Goals

  - Support sense-making (Data driven decisions)

  - Argument retrieval (Support future work for writing scientific articles)

Methodology - Data Collection

- **Annotation Process**
  - Abstracts' distribution to teams
  - Arguments (claim, evidence)
  - Structure (background, aim, results, etc.)
  - Citations (positive, negative, etc.)

- **Team's pairwise agreement**
  - Corrections on annotation process

- **Forming final datasets through all teams' agreement**

Methodology – Data Overview

- Final datasets

  - Arguments

Out[1]:

	document	sentences	labels
749	doi: 10.1161/circep.117.005858	[Rotors Detected by Phase Analysis of Filtered…	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, …
875	doi: 10.1590/s1678-9946201961019	[Zika virus infection among symptomatic patien…	[NEITHER, NEITHER, NEITHER, NEITHER, EVIDENCE,…
165	doi: 10.1016/j.cub.2020.04.070	[Vast Differences in Strain-Level Diversity in…	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, …
319	doi: 10.1021/acschemneuro.7b00314	[Novel Trimodal MALDI Imaging Mass Spectrometr…	[NEITHER, NEITHER, NEITHER, NEITHER, EVIDENCE,…
253	doi: 10.1016/j.redox.2018.101066	[Lipoproteins as targets and markers of lipoxi…	[NEITHER, NEITHER, NEITHER, NEITHER, NEITHER, …

  - Structure

Out[1]:

	document	sentences	labels
628	doi: 10.1093/nar/gkx1109	[GPCRdb in 2018: adding GPCR structure models …	[NEITHER, BACKGROUND, OBJECTIVE, METHOD, METHO…
626	doi: 10.1093/nar/gkw989	[proGenomes: a resource for consistent functio…	[NEITHER, BACKGROUND, BACKGROUND, BACKGROUND, …
469	doi: 10.1038/s41592-019-0630-5	[Tailoring cryo-electron microscopy grids by p…	[NEITHER, OBJECTIVE, BACKGROUND, BACKGROUND, R…
406	doi: 10.1038/nrmicro.2017.75	[Prediction of antibiotic resistance: time for…	[NEITHER, BACKGROUND, BACKGROUND, OBJECTIVE, C…
555	doi: 10.1080/16000870.2019.1699387	[Global variability in radiative-convective eq…	[NEITHER, BACKGROUND, BACKGROUND, OBJECTIVE, M…

Methodology – Data Processing

- **Text cleansing** (Nltk library)
  - Punctuation removal, lowercase transformation
  - Stopwords, Common words removal

- **Split dataset** (train, validation, test)

- **Text transformation for classification process**
  - Fasttext's algorithm specific input format

Methodology – Classification

- **Intuitive Baseline model**
  - Argument labeling using lexicons
  - Extra labeling based on empirical rules
    - Sentences between Evidence - Claim   ---------->   Evidence
    - Sentence exists after first Claim   ---------->   Claim
- **Fasttext algorithm**
  - Open-source library by the Facebook AI Research lab
  - Advanced technique for text classification
  - Specific input format : '__label__' prefix + label + text
  - Model training on train dataset
  - Hyperparameter tuning on validation dataset:
    - **Epochs** (number of times the model sees an input)
    - **Learning rate** (how much the model changes after each sentence is processed)
    - **Loss functions** (computation time speed up)
    - **Word n-gramms** (different sequences of n items)

Methodology – Clustering

- **Graph method** (NetworkX library)
    - Sentence **embeddings** extraction
    - Embeddings' **normalization**
    - **Similarity matrix** creation (1 – distance)
    - **Normalization** of similarity matrix
    - Similarity **threshold** for strong relationships
    - Graph creation
    - Setting similarity scores as **edges**' weights
    - **Communities**' identification for different k-cliques
    - **Threshold tunning** for stronger communities' identification

Experiments / Results - Classification

— **Fasttext's model for arguments:**

- Loss function = hierarchical Softmax
- Learning Rate = 0.7
- Epochs=18
- Word N-Grams = 2

ARGUMENT CLASSIFICATION		precision	recall	f1-score
Intuitive baseline model	claim	0.35	0.36	0.35
	evidence	0.43	0.45	0.44
Fasttext model	claim	0.52	0.32	0.4
	evidence	0.62	0.54	0.58
Fasttext model tuned	claim	0.49	0.41	0.45
	evidence	0.56	0.61	0.58

— **Fasttext's model for structure:**

- Loss function = 'one vs all'
- Learning Rate = 0.1
- Epochs=18

STRUCTURE CLASSIFICATION		precision	recall	f1-score
Fasttext model	result	0.46	0.76	0.57
	background	0.4	0.48	0.44
	objective	0.43	0.61	0.5
	conclusion	0.49	0.2	0.28
	method	1	0	0.01
Fasttext model tuned	result	0.59	0.64	0.61
	background	0.59	0.6	0.6
	objective	0.52	0.55	0.54
	conclusion	0.47	0.35	0.4
	method	0.44	0.55	0.49

Experiments / Results - Clustering

➢ Starting **threshold = 0.6**

➢ **Communities with k=3 cliques**

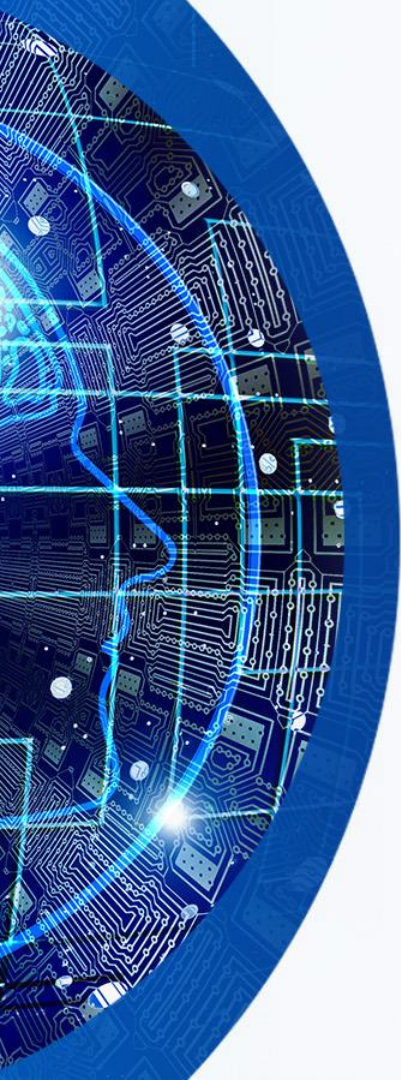| Number of observations	Similarity range
23315511	0 – 0.1
2044619	0.1 – 0.2
9443238	0.2 – 0.3
9397561	0.3 – 0.4
2049584	0.4 – 0.5
100369	0.5 – 0.6
2912	0.6 – 0.7
163	0.7 – 0.8
18	0.8 – 0.9
4843	0.9 – 1

Experiments / Results - Clustering

➢ Similarity > 0.6

➢ Communities of k = 7 cliques
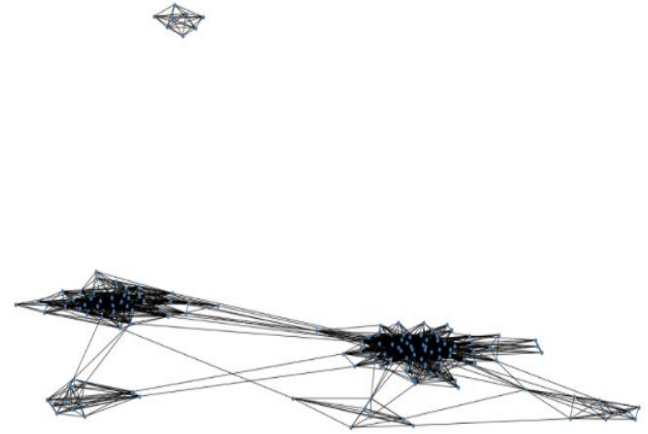
Removing edges

➢ Similarity > 0.7

➢ Communities of k = 3 cliques

Error Analysis - Comments

- **Good clustering :**
  - ➢ IDs: [2132 , 2031]
  - ➢ Topic: 'elderly patients treatments'

- **Poor clustering:**
  - ➢ IDs: [796 , 2002]
  - ➢ 'Electronic supplementary material The online version of this article contains supplementary material, which is available to authorized users.'
  - ➢ Problematic annotation process

- **Graph using only "Claim arguments"**
  - ➢ Weaker clustering

- **Comments**:
  - ➢ Better text pre-processing
  - ➢ Different ML model for embeddings extraction

Team members

Kyriakakis Nikolaos

Papadomarkakis Georgios

Thank you for your attention!