LeQua@CLEF2022: Learning to Quantify

Andrea Esuli, Alejandro Moreo, Fabrizio Sebastiani

Istituto di Scienza e Tecnologie dell'Informazione - Consiglio Nazionale delle Ricerche
56124, Pisa, Italy

LeQua@CLEF2022

Quantification is the task of predicting the prevalence (i.e., relative frequency) of a property in a sample of elements from a domain.

LeQua 2022 is the first edition of the "Learning to Quantify" lab, hosted within the CLEF 2022 Conference.

The goal of LeQua is to allow the comparative evaluation of methods for "learning to quantify" in textual datasets, i.e., methods for training predictors (called "quantifiers") of the prevalences of the classes of interest in sets of unlabelled documents.

The predictors will be required to issue predictions for several such sets, some of them characterized by prevalence values radically different from the ones of the training set.

Tasks

Two *tasks* are offered:

T1 *The Vector Task*:

- Participant teams are provided with vectorial representations of the documents.
- Mostly for teams not into text mining.

T2 *The Raw-Documents Task*:

- Participant teams are provided with the raw documents.
- Mostly for teams wanting to test end-to-end systems.

For each task, two *subtasks* are offered:

A *The Binary Subtask*:

- 2 classes
- Classes are *sentiment*-related (`Positive` and `Negative`)

B *The Multiclass Subtask*:

- 28 classes
- Classes are *topic*-related (e.g., `Automotive`, `Baby`, `Beauty`, ...)

Dataset

The data are obtained from a crawl of ≥100M Amazon reviews; from these we remove

- all reviews shorter than 200 characters,
- all reviews that have not been recognized as "useful" by any users,
- (for the binary "sentiment-based" task) all reviews with 3 stars.

The 2 training sets L_{B} (binary) and L_{M} (multiclass):

- L_{B} consists of 5,000 documents and L_{M} consists of 20,000 documents
- L_{B} and L_{M} are sampled from the ≥100M-strong dataset Ω via *stratified sampling* on the dimension of interest (resp. sentiment, topic), so as to have "natural" prevalence values for all the class labels.

The 2 development (validation) sets:

- We use 1,000 development samples of 250 documents each for the binary task and 1,000 development samples of 1,000 documents each for the multiclass task.
- The sets of development samples D_{B} and D_{M} are generated from $\Omega \setminus L_{\mathrm{B}}$ and $\Omega \setminus L_{\mathrm{M}}$ via the *Kraemer algorithm* for sampling uniformly from the unit simplex
- The goal of this sampling algorithm is generating samples characterised by a variety of *(equiprobable) class distributions, with class prevalence values* not *from a predefined grid of values.*

The 2 test sets:

- We use 5,000 test samples of 250 documents each for the binary task and 5,000 test samples of 1,000 documents each for the multiclass task.
- The sets of test samples U_{B} and U_{M} are also generated from $\Omega \setminus (L_{\mathrm{B}} \cup D_{\mathrm{B}})$ and $\Omega \setminus (L_{\mathrm{M}} \cup D_{\mathrm{M}})$ via the Kraemer algorithm.

Timeline

Dec 1, 2021	Release of train and dev set	May 13, 2022	Release of results
Apr 22, 2022	**Release of test set**	May 27, 2022	Paper submission deadline (optional)
May 5, 2022	**Run submission deadline**	Sep 5, 2022	**LeQua @ CLEF2022**

Sampling

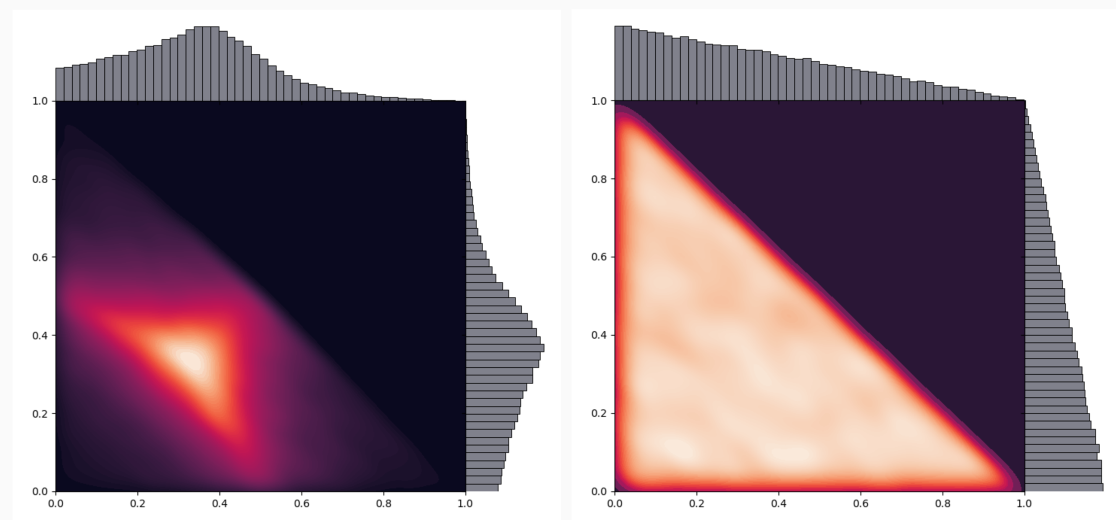
Generating test samples that cover all the possible spectrum of prevalence values is key to test the robustness of quantifiers to the variation of prevalence from training data to test data.

For example, drawing n random values (for a problem with n classes) uniformly at random from the interval [0,1] and then normalizing them so that they sum up to 1 (IID method), produces samples biased towards the centre of the unit $(n-1)$-simplex.

The *Kraemer algorithm* generates samples that uniformly cover the entire spectrum of prevalence values for all classes:

- Given a set of classes $\mathcal{Y}$, generate a vector $A = \langle a_1, ..., a_{(|\mathcal{Y}|-1)} \rangle$ of points sampled uniformly at random from [0,1]
- Sort the a_i's to obtain $B = \langle b_1 \leq ... \leq b_{(|\mathcal{Y}|-1)} \rangle$, and define $b_0 = 0$ and $b_{|\mathcal{Y}|} = 1$
- Obtain a vector $P = \langle p_1, ..., p_{|\mathcal{Y}|} \rangle$ by defining $p_i = b_i - b_{(i-1)}$ for all $i \in \{1, ..., |\mathcal{Y}|\}$
- Use P as the distribution of class prevalence values for generating sample σ

Visualization of distribution of 2-dimensional samples generated using different methods:



IID Kraemer

Evaluation

Relative frequencies of classes are represented by probability distributions.
The true probability distribution p for each set is compared to predicted one $\hat{p}$, by means of *Relative Absolute Error*:

$$\mathrm{RAE}(p, \hat{p}) = \frac{1}{n} \sum_{y \in \mathcal{Y}} \frac{|\hat{p}(y) - p(y)|}{p(y)} \qquad (1)$$

The final score is the mean RAE across all the samples in the test set.

Links

Web: `https://lequa2022.github.io/`
Data: `https://zenodo.org/record/5734465`
Twitter: @LeQua2022

Acknowledgements

This work has been supported by the SoBigdata++ project, funded by the European Commission (Grant 871042) under the H2020 Programme INFRAIA-2019-1, and by the AI4Media project, funded by the European Commission (Grant 951911) under the H2020 Programme ICT-48-2020.