COMP9417 Machine Learning and Data Mining T1, 2020 - Group Project: Multiclass Text Classification

# ML Wizards - Chani Charak, Connor Clark, Raymon Qian, Tal Weiner, Yannick Schnider

# Introduction

## Main Issues to be Addressed

* Unbalanced distribution of number of Articles per Topic

**Ways to Address:**

* Over-sampling of under represented classes
  + SMOTE (Synthetic Minority Over-Sampling Technique)
  + Use from imblearn.over\_sampling import SMOTE
  + <https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-data-set-29f6a177c1a>
* Use TF-IDF (Term Frequency - Inverse Document Frequency)
  + Words that are common for a specific document, but rare among all other documents are given higher weighting
* For each class:
  + Look at the terms with highest frequency and lowest IDF’s
  + Then combine with all data: look at terms that had highest frequency and now highest IDF’s
    - These terms should be common for a specific topic and rare among other topics, making it have high weight for predicting a specific topic.
* Turn our Classification Problem into 2 different Classification problems:
  + Problem 1 - Binary Classification between RELEVANT and IRRELEVANT articles
  + Problem 2 - Multiclass Classification between the 10 RELEVANT articles.
* Now the 2 different Problems have a much more balanced distribution of number of articles per class.

# Methodology

* Accuracy is not a good metric if classes aren’t balanced.
  + Model might be really bad at classifying one topic, but if that topic has a small amount of articles, then the accuracy of our model can still be high, not telling the true capability of our model.
* Confusion matrix is better metric when classes aren’t balanced.
  + You can tell straight away if the metric is lower for lower represented topics.
  + Precision = Out of all predictions that predicted ‘Class A’, how many were actually ‘Class A’.
  + Recall = For all of ‘Class A’, how many were correctly classified as ‘Class A’.
* To decide what metric is better for us we ask ourselves 2 different questions.
  + Out of all the recommendations we make, how many are relevant recommendations (PRECISION)
  + For each specific class, how many relevant recommendations are we making (RECALL)
  + Looking at both these questions, we see that in essence, we want both.
    - We want all our recommendations to be relevant, so users are more likely to read something we recommend
    - And for any rare topics, we need to make sure whatever we do recommend is relevant, not to discourage users from reading.