# Identifying topics for content creation

Fiona Chow Pui Sann

September 2016

Dissertation submitted in partial fulfilment for the degree of
Master of Science in Big Data

Computing Science and Mathematics
University of Stirling

# Abstract

This MSc project describes the prototype Model ("the Model") developed for Wood Mackenzie LTD, a global energy consultancy and content provider which provides topic insights for content creation. The Project was designed based on the Company's defined business problem, to bridge the information gap between available content and extent information requirements of the Users are not met, without their direct involvements.

This project was segregated into three (3) main components where the Company's captured portal search interaction data was utilised to:-

- train a decision tree model for predicting success of search sessions by partitioning search sequence based on search goals using a combination of timeout threshold and query reformulation as the key point for determining search goal;
- detect and correct misspellings within search terms submitted via a context-sensitive spelling algorithm - in the designing of spelling checker, the process was separated into query pre-processing, isolated spell check and context based spell check;
- group and rank unique terms accordingly, giving high priority to highly queried terms which have few existing related content and suffers from high user dissatisfaction.

The prediction model and spelling algorithm were evaluated using traditional metrics. The prediction model had achieved an overall accuracy of 0.84 and true negative rate (fail search sessions) of 0.87; and the spelling algorithm had achieved 0.85 error correction rate and 0.99 error detection rate.

The term ranking aspect and the final deliverable of the Model were jointly evaluated using intrinsic measures. MUsic was adapted and applied to assess the usuability of final deliverable in a qualitative manner whilst Subsequent Event is used to evaluate the extent final deliverable is able to benefit and support the Research team in discovering and responding to rising interest based on how long it takes the team to discover and create the content of interest.

The final outcome is that intended users of the Model had expressed positive feedbacks on the relevance and usability of the deliverable, furthermore, the Model had shown potential of improving the content creation responsiveness towards users information demands by 2 months.