Technical Specification Double-side Page

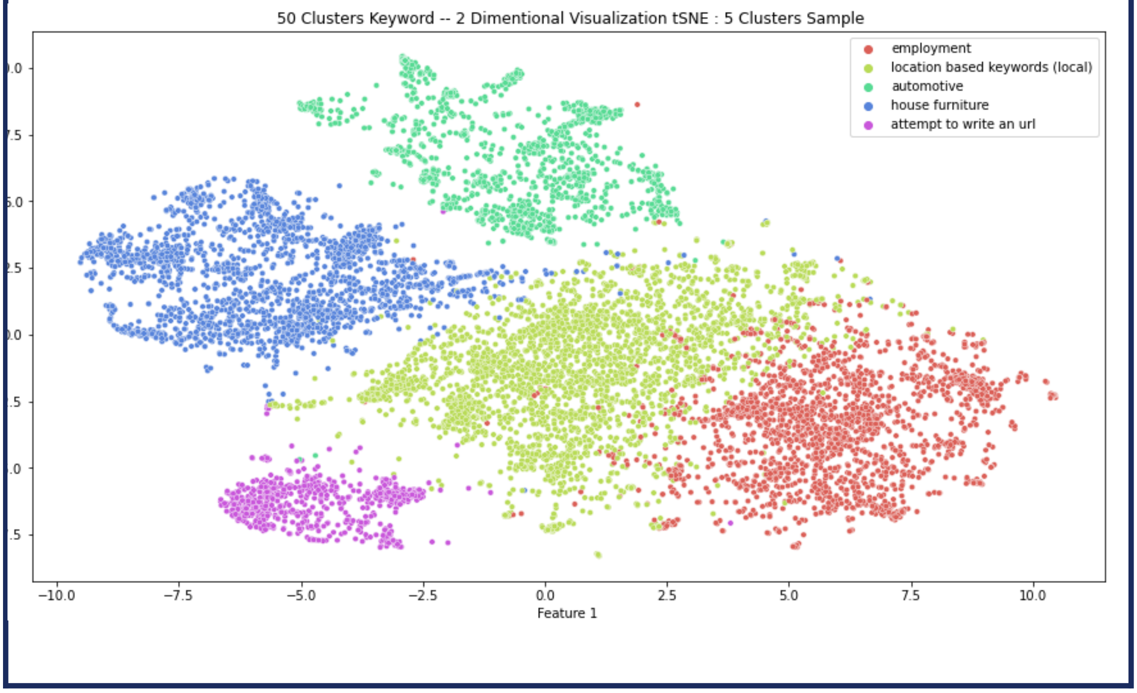
1. **TECHNICAL SCOPE:** Summarize the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

We designed a keyword clustering tool that is able to tackle two different objectives at the same time:

1. A campaign level clustering (adgroup) tool that works on a given set and is able to create a set of adgroups satisfying some given constraints (that are usually queries).
2. A global clustering model that is able to label correctly a keyword into its industry and topic reducing dependence of google's categories that are not always how we'd like them to be / not specific enough in some industries (ex: <<SaaS Marketing Tools>> doesn't exist as a category even though we think it should)

➔ Our solution's processing speed is a real asset: more than 3 Millions Keywords per minute can be classified using only 4 CPUs

➔ Language independence achieved by a corrective mapping translate first to a neutral language (Shown technology)



2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

1. We have a layered sentence topic modelling using deep learning neural networks
   a. Sentence Level (Language independent)
   b. Sentence Level (Language dependent)
   c. Word Level (Language dependent)
   d. Character Level (Language independent)
2. 1.b to 1.d are needed to account for language specificities that cannot be contained in translations, movie names are a good exemples these two are the same movie:
   a. Millénium : Les Hommes qui n'aimaient pas les femmes = <<the guys who didn't like women>>
   b. The Girl with the Dragon Tattoo
3. Our provided data also contains Queries that we can use to build AdGroups ideally sized, it's recommended to have at least 1k queries per adgroup
4. We also have a keyword tier and CPC that we can use to finetune our clustering capabilities, however we'll avoid depending too much on it: one has to first spend money on that given keyword to get those data. Our

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

tool is meant to be used as a first line of contact, before actually spending money. Usually only semantics and data available beforehand

We believe our proposed solution is a perfect fit for the keyword clustering / AdGroups builder described in the challenge

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains

1. **Scalability**:

   For the MVP our non optimized solution was able to cluster 1.7M keywords in 27.5s reaching an average speed of 3.7 Millions keywords classified per minute. This can be scaled by adding more CPUs to the computing machine

2. **Flexibility:**
   The current solution is built using a general model that trains in less than an hour on 1M entries. It allows faster iterations and we are therefore able to change/iterate/improve faster.
   Our model can be used in a multitude of different fields where topics are of relevance, for example clustering website's page topics to select better keywords related to a given topic.
   Social media performance can also be improved by analysing topics that create more engagements.

Data for training (JOT-IM data provided through Reach Incubator) are stored on our secured server that are password protected and can only be accessed from our intranet.

We will not be storing any customer's data on our servers, they'll only transit for the duration of the processing.

All of our servers and endpoints are protected with SSL (https) and cannot be accessed without https from any modern browser (HSTS registry) preventing any mistake from an attentive user.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planned for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment...) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

1. Technological risk: although our MVP clearly indicates that the final solution is feasible there is still a risk that it could take longer than expected to develop the final solution tackling the real need of the challenge.

2. Challenge definition Risk: The challenge describes what JOT-IM (and ourselves) expects will improve marketing results, the actual solution that'll improve marketing results can be slightly different. As we get closer to a final solution we'll reach a point where we'll have to tweak a bit the actual requirements to actually address business outcomes.

3. Business : Selling our products to our customers although we already sell marketing services can require some tweaks

Even if all of these risks manifest themselves the worst that we expect to happen is to lengthen the development time. We do not expect a reasonable risk that could put the project in jeopardy as the MVP is already a viable product.

Quality Control

1. We work using very strict continuous development, deployment and testing in isolated containers
2. Development is made respecting TDD and each step and intermediate results are validated and tested
3. Before each deployment we manually test that the changed behaviours are actually doing what is expected of them

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981